Non-allelic homologous recombination of Alu and LINE-1 elements generates somatic complexity in human genomes

G. Pascarella[1]*, K. Hashimoto[1], A. Busch[1], J Luginbühl[1], C. Parr[1], C. C. Hon[1], W. H. Yip[1], A. Kratz[2], A. Bonetti[1,3,4], F. Agostini[5,6], J. Severin[1], S. Murayama[7], S. Gustincich[8], M. Frith[9,10,11] and P. Carninci[1]*.

[1]RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Japan

[2]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA

[3]Department of Cell and Molecular Biology, Karolinska Institutet, Sweden

[4]Stockholm University, Sweden

[5]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

[6]Science for Life Laboratory, Stockholm, Sweden

[7]Department of Neuropathology, Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology (TMGHIG)

[8]Central RNA Laboratory and Department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia (IIT), Genova, Italy

[9]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

[10]Graduate School of Frontier Sciences, University of Tokyo, Chiba, Japan

[11]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), AIST, Tokyo, Japan

*Correspondence to: carninci@riken.jp; giovanni.pascarella@riken.jp.

Abstract

Millions of Alu and LINE-1 copies in our genomes contribute to evolution and genetic disorders via non-allelic homologous recombination (NAHR), but the somatic extent of these rearrangements has not been systematically investigated. Here we combined high-throughput capture and sequencing of repeat elements with a new bioinformatic pipeline to show that somatic NAHR of Alu and LINE-1 elements is common in human genomes. We describe tissue-specific hallmarks of NAHR, and show that retroelements acting as recombination hotspots are enriched in cancer genes and structural variants. Analysis of recombination in human induced pluripotent stem cells and differentiated neurons revealed a neuron-specific recombination signature suggesting that the emergence of cell type-specific recombination profiles accompanies cell-fate determination. Finally, we found that somatic NAHR profiles are altered in Parkinson's and Alzheimer's disease, indicating a link between retroelements recombination and genomic instability in neurodegeneration. This work shows that somatic recombination of repeat elements contributes massively to genomic diversity, and that extensive recombinogenic activity of retroelements may act as a grey eminence in the transition from health to disease.

Introduction

Alu and Long Interspersed Nuclear Element-1 (LINE-1, abbr. L1) are the two most abundant retrotransposons in the human genome, with ~1.2 and ~1 million annotated copies that together account for almost 30% of the genome (*1*). The main Alu subfamilies, in order of increasing evolutionary age, are AluY, AluS and AluJ (*2*). Primate-specific L1 subfamilies are classified as L1PA1-16, from most recent to oldest; L1PA1 (also known as L1HS) includes the only known autonomously active human retrotransposons (*3*). Key discoveries in recent years have transformed our view of genomic repeats from just parasites to evolutionarily co-opted symbionts with important functions in chromatin and gene regulation (*4–8*). Alu and L1 can also alter the genomic information via retrotransposition and recombination. Retrotransposition originates from a variable but small number of active young L1 and Alu copies per genome and is restrained by several layers of genomic surveillance (*9–12*). Homology-based recombination of repeat elements is not restricted by the same rules governing retrotransposition: any pair of the millions of homologous Alu and L1 elements in the genome can be the substrate for non-allelic homologous recombination (NAHR), a class of rearrangements believed to be a major driving force in genome evolution and a source of pathogenic structural variants (*13–16*). Erroneous pairing during cell division and repair of DNA damage via the DNA double-strand repair (DSBR) pathway are believed to be the major trigger for NAHR in the human genome (*17*). During repair, components of the DSBR complex scan for regions homologous to the damaged locus and fix the breaks via the formation of heteroduplexes and other unstable structures (*18*). The resolution of these temporary configurations can be neutral (non-crossover NAHR), or can lead to dramatic chromosomal rearrangements in the case of crossover between the recombined loci (*19*, *20*). NAHR with crossover can therefore disrupt the genetic information causing aberrant phenotypes; repeat elements have often been found at the breakpoints of NAHR events associated with cancer and

genomic disorders caused by erroneous meiotic pairing (*21–23*). Considering the substantial number of Alu and L1 elements interspersed throughout the genome, the mutational burden imposed by NAHR has been hypothesized to exceed that of other types of structural variations. Although several studies have sought to reveal the mechanisms behind NAHR and its contribution to diseases (*24–27*), a comprehensive investigation of somatic Alu and L1 NAHR in different cells, tissues or biological contexts is unavailable. Here, we combined high-throughput capture and sequencing of retroelements with a new bioinformatic pipeline to comprehensively investigate somatic NAHR of Alu and L1 in the human genome. We describe new features of tissue-specific retroelement recombination in various biological contexts, and show that somatic recombinogenic activity of Alu and L1 elements is an important contributor of genomic structural variants in normal physiological and pathological conditions.

Results

High-efficiency capture and sequencing of Alu and L1 elements from genomic DNA

At the time we started this study, the rate of somatic recombination for repeat elements in the human genome had not been reported. Therefore, considering the low prevalence of somatic structural variants associated with repeat elements (*28–31*), we developed a protocol to maximize the discovery power and enrich for genomic retroelement sequences prior to sequencing ("capture-seq"). We designed tiled DNA capture probes to span the full model sequences of young AluY elements (*32*) and to cover ~250bp of the 5'- and 3'-regions of the youngest L1 element consensus sequence (L1HS) (*33*) (Table S1). This design coupled to random shearing of genomic DNA allows for stochastic inclusion of uniquely mapping, non-repeated genomic regions flanking the captured repeats; we further joined paired reads to generate longer contigs and improved the global mapping quality. The probes:target hybridization time was reduced substantially from several days in previous capture-seq

protocols iterations (*34*, *35*) down to 5 minutes. This allowed a 1-day library production time while retaining a low number of post-enrichment PCR cycles (n=12) and optimal enrichment efficiency (Fig. S1A, B).

We applied our capture-seq workflow to a panel of post-mortem tissues from 10 donors showing no obvious disease at the time of death (Table S2). For each donor, we selected available tissues derived from the 3 developmental germ layers: kidney (mesoderm), liver (endoderm) and 3 cortical brain regions (frontal cortex, temporal cortex, parietal cortex; ectoderm). The brain samples were stained with neuron-specific antibody NeuN and underwent fluorescence-activated nuclei sorting (FANS) to separate the neuronal and non-neuronal fractions (*36*, *37*) (Fig. 1A and Fig. S2). We sequenced 78 capture-seq libraries yielding ~960 millions raw reads (Table S3). Quality control performed on uniquely mapping reads confirmed the consistent and efficient capture of designated targets for a panel of L1HS (*28*, *35*) and reference AluY elements (Fig. S3A, Fig. S4A). In protocols for enrichment of repeat elements based on capture probes, the repeated nature of the genomic targets overpowers the specificity of the probes. For instance, in our dataset we also detected a comprehensive and highly reproducible enrichment of L1 and Alu elements that were not originally targeted by experimental design (Fig. S3B, C; Fig. S4B-D). We took advantage of the richness and complexity of our capture-seq libraries by extending downstream analyses to all Alu and primate-specific L1 subfamilies annotated in the Dfam database (*32*) (Table S4). The median capture rate across the whole dataset was 94% ± 0.7% for annotated AluY elements, 83% ± 5% for AluS elements and 45% ± 7% for AluJ elements with an overall Alu capture rate of 75% ± 3% (Fig. S3A-D). For L1, we divided the enriched elements into 4 groups (Table S4). As expected, L1HS showed the highest capture rate (94% ± 1%), followed by L1PA2-L1PA7 (85% ± 1%), L1PA8-L1PA10 (63% ± 3%) and L1PA12-L1PA17 (31% ± 6%). The collective capture rate for L1 elements was 64% ± 2% (Fig. S4A-E).

Genome-wide discovery of Alu and L1 NAHR events with TE-reX

To find NAHR events in Alu and L1 capture-seq libraries we developed TE-reX, a new bioinformatic pipeline based on LAST (*38*). TE-reX was designed to identify recombination events from split reads that join repeat elements at homologous positions (Fig. 1B). Using TE-reX on capture-seq data we retrieved thousands of putative Alu and L1 NAHR events. The number of Alu and L1 recombination events per sample was higher in kidney and liver compared with brain, whereas we did not detect any difference between the neuronal and non-neuronal fractions (Fig. 1C, D; Table S5). The number of recombination events per chromosome was highly correlated with the number of repeats annotated per chromosome in RepeatMasker (cumulative median across chromosomes for Alu dataset: r = 0.98, p = 2.2e-16; for L1 dataset: r = 0.95, p = 7.7e-13; Fig. S5A, B). Within each library the relative abundance of Alu recombination events exceeded that of L1 recombination events by several folds (Fig. S6A) as expected from the higher similarity among the Alu elements examined than among the L1 elements. We confirmed the TE-reX results by a thorough validation based on polymerase chain reaction (PCR) followed by Sanger sequencing. To avoid cross-amplification of homologous repeat sequences, we focused the validation on those recombination events where we could identify non-repeat sequences flanking the recombined repeats. We designed forward and reverse oligonucleotides for 112 recombination events (79 inter-chromosomal, 33 intra-chromosomal) found across all libraries; of these, 101 (90%) were supported by a single contig, 10 (9%) by 2 contigs and 1 (1%) by 4 contigs (Table S6). The low number of supporting contigs is strongly indicative of the somatic quality of the selected targets and it directly implies that templates for the recombination events, while present in the capture libraries, will most likely be depleted from the starting genomic material. Hence, using as input the same capture libraries subjected to sequencing we obtained clean amplicons for 103/112 targets. Sanger

sequencing verified the target identity for 93/103 amplicons and confirmed that NAHR events detected by TE-reX can be validated with a sensitivity of 92% and specificity of 90% (illustrative PCR results in Fig. S7).

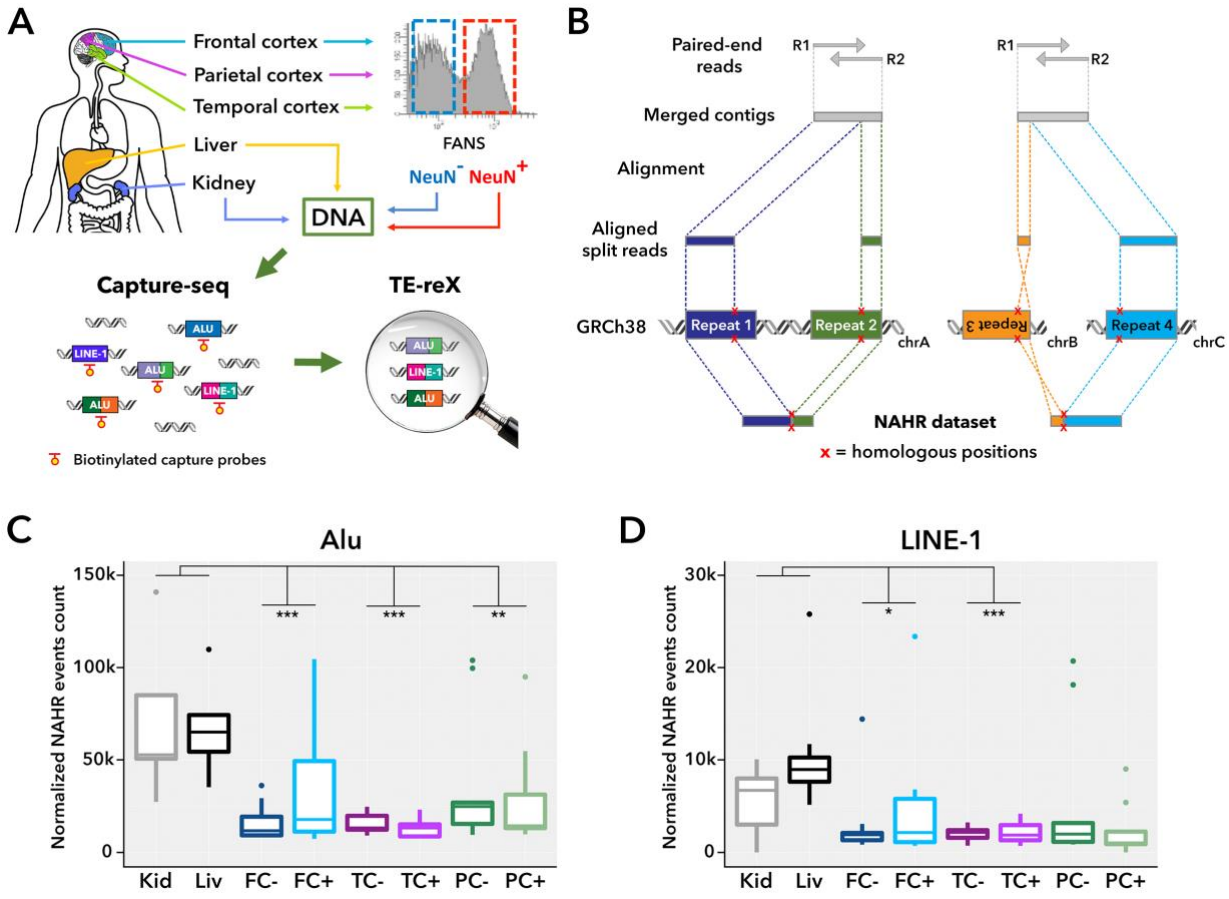

Fig. 1 Discovery of Alu and L1 NAHR by capture-seq and TE-reX pipeline.

A) Schematic of the production of the capture-seq dataset. Genomic DNA samples purified from bulk tissues and sorted nuclei were enriched for Alu and L1 elements by using biotinylated RNA capture probes spanning the entire sequence of young AluY elements and the 5-' and 3'- regions of L1HS. gDNA: genomic DNA.

B) The TE-reX pipeline identifies NAHR from split-alignments of contigs with breakpoints that are mapped within repeats of the same family, and are located in homologous positions with respect to repeat model sequences.

C, D) Box plots of NAHR event counts for Alu (C) and L1 (D) elements in capture-seq libraries from post-mortem samples; counts are normalized by sequencing depth. FC, frontal cortex; Kid, kidney; Liv, liver; PC, parietal cortex;TC, temporal cortex. PC, parietal cortex; +, neuron-specific antibody (NeuN) positive; –, NeuN, negative. *P < 0.05; **P < 0.01; ***P < 0.001 (single factor ANOVA).

Genome-wide annotation of Alu and L1 NAHR and tissue-specific characteristics of somatic recombination

To confidently annotate NAHR events genome-wide we relied solely on recombination breakpoints with extremely low mismapping probability ($p \leq 1e10\text{-}5$), reported as "mismap" in LAST alignments (*38*). The median relative abundance of high mapping confidence (HMC) events in the total Alu and L1 events across the capture-seq dataset were 54% and 42%, respectively (Fig. S8A, B; Table S7). Interestingly, within the HMC dataset the relative abundance of somatic NAHR events involving mobile, young AluY and L1HS elements was higher in brain samples compared with kidney and liver (Fig. S9A, B). In addition, among all L1HS elements the relative abundance of recombined full-length L1HS elements (>6 kb) was higher in brain samples compared to kidney and liver (Fig. S9C).

We investigated the enrichment and depletion of HMC Alu and L1 NAHR events in gene and regulatory regions by comparing the real dataset (O, observed) to datasets comprising random permutations of the genomic coordinates of NAHR breakpoints (E, expected) (Fig.S10-S11). We did not observe any substantial enrichment or depletion within the gene body, except for a mild depletion of L1 NAHR events in 3' untranslated regions (UTR)**.** L1 events, but not Alu events, were significantly enriched in transcription start sites (TSS) of genes. A similar trend of enrichment, but to a larger extent, was observed in promoters (*39*)**,** in particular across the brain tissues (log2 O/E ratio 1.5 to 2.9). In contrast, we observed an overall mild but significant

depletion of both Alu and L1 events within enhancers (*39*) across all tissues and samples (log2 O/E ratio –0.4 to –1.2, $P < 0.05$, Student's *t*-test). Intriguingly, separating the regulatory regions according to their cell-type and tissue activity (*39*) (Fig. S12-S13) showed that the enrichment of L1 events in promoters was prominently attributed to the promoter active in stem cells, independent of the tissues in which the NAHR events were observed (Fig. S13B). We did not observe consistent enrichment or depletion of NAHR events in regulatory regions active in the matched tissues (e.g. liver NAHR events in liver active promoters). These observations implied the occurrence of NAHR events might be attributed to genomic (i.e. cell-type independent, e.g. sequence composition) and/or epigenomic (i.e. cell-type dependent, e.g. chromatin modification) factors.

The brain samples showed a consistent and significantly higher rate of intra-chromosomal recombination compared with kidney and liver samples; among the 3 tested cortical brain regions the temporal cortex samples had the highest rate of intra-chromosomal recombination (Fig. 2A, B). To deepen this observation, we compared the chromosome-to-chromosome recombination rate of each sample with a dataset of random Alu and L1 pairs generated *in silico*. This analysis confirmed the enrichment of Alu and L1 intra-chromosomal NAHR events in all brain samples over the random dataset, whereas kidney and liver samples showed only a very marginal enrichment (Fig. 2C, D; enrichment of intra-chromosomal recombination rates of merged kidney and liver vs. merged brain: $p = 2.6e\text{-}24$, two-tailed Student's *t*-test. Unmerged data, Fig. S14-S15).

The number of Alu NAHR events in the dataset was sufficient to repeat this analysis separating the data for male and female donors, to eliminate the potential confounding factor of sex chromosome ploidy. After adjusting the random background dataset for sex chromosome ploidy, we confirmed that there was no depletion of intra-chromosomal NAHR for chromosome X in either male or female donors (Fig. S16). Conversely, the depletion of intra-

chromosomal NAHR for chromosome Y in male donors remained substantial even after correcting for ploidy (Fig. S16). According to Umap data (*40*), a possible explanation for this observation is the lower mappability score of chromosome Y compared with all other chromosomes.

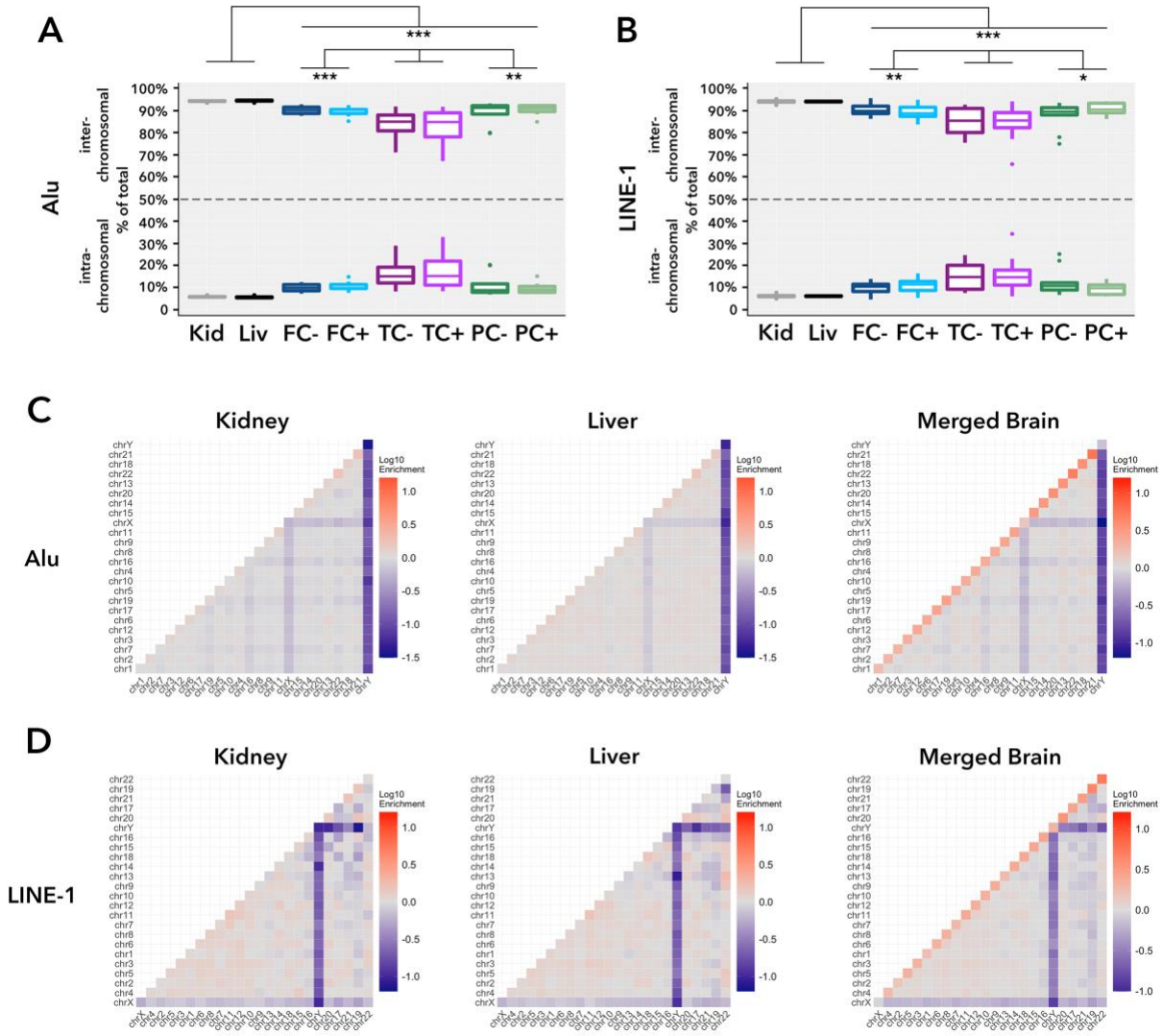

Fig. 2 Tissue-specific profiles of somatic recombination of Alu and L1 elements.

A, B) Intra-chromosomal and inter-chromosomal recombination rates for Alu (A) and L1 (B) somatic recombination. For tissue and sample abbreviations, see Fig. 1B. *P < 0.05; **P < 0.01; ***P < 0.001 (single factor ANOVA).

C, D) Chromosome-to-chromosome matrices of somatic NAHR rates for Alu (C) and L1 (D). Chromosomes are ranked by Alu and L1 RepeatMasker content in (C) and (D) panels, respectively. Colors show folds enrichment of recombination rates for each individual chromosome compared with aggregated recombination rates of 10 random Alu and L1 recombination datasets of comparable size. Merged brain, combined data for the neuronal and non-neuronal fractions of the 3 tested cortical regions.

Profiling of intra-chromosomal recombination and tissue-specific recombination landscapes

Next we explored the landscape of intra-chromosomal recombination by calculating the genomic distance (d) between members of each pair of Alu and L1 elements that we found recombined in the HMC dataset, based on their RepeatMasker annotation. After binning the distances in four intervals (d<25 kb, 25 kb<d<250 kb, 250 kb<d<2.5 Mb, d>2.5 Mb) we observed that, in all samples, the majority of intra-chromosomal recombination events were established between retroelements that are either proximal to (d<25 kb) or far away (d>2.5 Mb) from each other, suggesting different NAHR mechanisms for close and distant repeats (Fig. 3A). The distance profiles for kidney and liver samples were overall similar, with ~95% of the intra-chromosomal recombination involving repeat >2.5 Mb apart. In contrast, somatic NAHR in brain samples showed a significantly higher relative abundance of intra-chromosomal events in the d<25 kb interval; this was more pronounced in the temporal cortex (~75%) than the frontal or parietal cortex (~50%). Most Alu recombination events in the d<25 kb interval involved pairs distanced <5 kb, with a higher relative abundance in non-brain samples compared with brain samples (Fig. S17A).

The human genome sequence is depleted of inverted proximal Alu pairs, an evolutionary consequence of the genomic instability of close Alu elements in this configuration (*41*, *42*).

Analysis of strand orientation for proximal recombined repeats in our data revealed a distinct recombination bias towards repeats in inverted configuration over those in direct configuration (Fig. 3B). For both Alu and L1 NAHR events the bias was significantly stronger in brain samples compared with other tissues. Regardless of the orientation, in all tissues the bias was dependent on the distance of recombined elements and it was absent in NAHR of repeats >2.5 Mb apart (Fig. S18A).

Previous analyses of deletions mediated by Alu in human genome evolution and *in vitro* recombination analyses of Alu pairs have found a similar enrichment of breakpoints in the 5'-region of Alu model sequence (*13*, *25*). In our dataset the breakpoints frequency profile along the Alu model sequence for inter-chromosomal recombination events was similar in all samples (Fig. S18C); however, intra-chromosomal recombination profiles showed an enrichment of breakpoints frequency in the 5'-region of brain samples (Fig. S18D). When separating the intra-chromosomal breakpoints frequency profiles according to genomic distance and directionality of recombined Alu pairs we observed a clear difference between the profiles of proximal recombined Alu in direct or inverted configuration. Inverted proximal recombined Alu pairs showed strong breakpoints frequency enrichment in their 5'-region in all samples, while proximal recombined Alu in direct configuration lacked any 5'-enrichment (Fig. 3C). Profiles for distant recombined Alu pairs were very similar in all samples and devoid of any enrichment in their 5'-regions, regardless of the orientation (Fig. S18E, F).

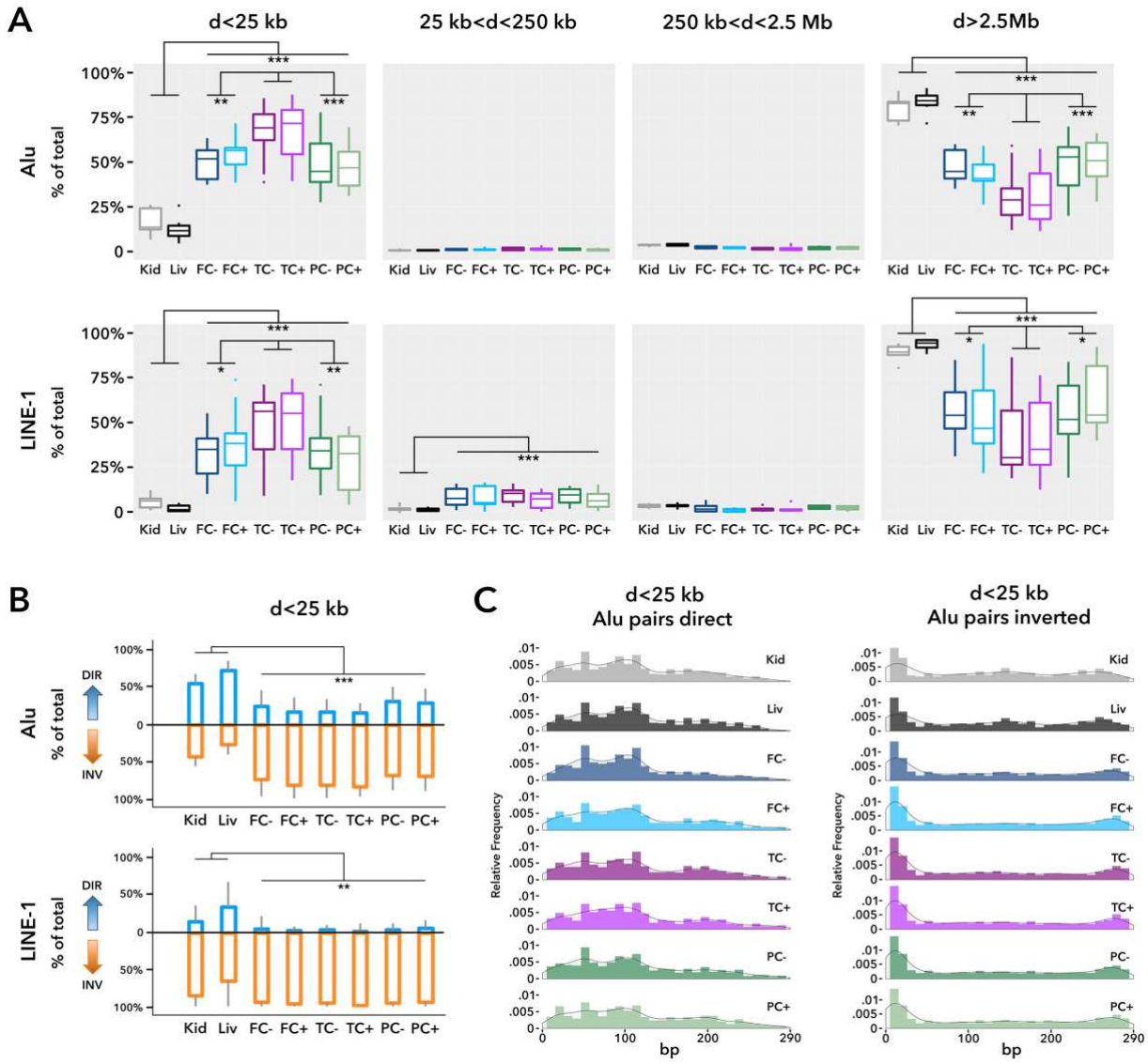

Fig. 3 Profiling of intra-chromosomal recombination and directionality bias of proximal somatic recombination.

A) Boxplots of somatic intra-chromosomal NAHR profiles of Alu and L1; data are binned into non-overlapping consecutive intervals of genomic distance (d) between members of the recombined pair.

B) Analysis of directionality for recombined proximal Alu and L1 elements shows a strong bias against recombination of elements in inverted configuration. No directionality bias was

observed for recombined pairs distanced more than 2.5Mb. DIR, direct configuration; INV, inverted configuration. *P < 0.05; **P < 0.01; ***P < 0.001 (single factor ANOVA).

C) Breakpoints frequency displayed along Alu model sequence showing different profiles for somatic NAHR events involving proximal Alu elements in direct and inverted configurations. For tissue and sample abbreviations, see Fig. 1B.

Detection of Alu and L1 NAHR in capture-free and PCR-free libraries

The above characterization of somatic NAHR of Alu and L1 has established its tissue-specificity and sharp divergence from randomness. To verify that somatic NAHR events are not a byproduct of the capture-seq workflow, we prepared capture-free libraries for 3 kidney samples and 3 temporal cortex NeuN+ samples by omitting the steps for enrichment and capture (RNA:DNA hybridization, on-beads capture and post-capture PCR). These samples therefore exclusively underwent random genomic DNA fragmentation and the minimum 3 PCR cycles required to introduce the platform-specific sequencing linkers. These capture-free, quasi-PCR-free libraries were sequenced on Illumina Miseq at 300 bp reads in paired-end mode, yielding in total ~30 millions of raw reads. After data processing with TE-reX and downstream analyses, we readily detected recombination events in these libraries (Table S7). Genome-wide annotation of NAHR events detected in capture-free conditions returned results consistent with those from the capture-seq dataset (Fig. 4A, B).

Relatively short reads sequenced on Illumina platforms rarely extend to non-repeat flanking regions, thus making it impossible to understand the complete anatomy of the identified recombination events. To overcome this issue and to further substantiate our findings, we again sequenced the 3 kidney and 3 temporal cortex NeuN+ samples on the Oxford Nanopore Technologies (ONT) MinION platform. Library construction was performed with the 100% PCR-free Rapid Barcoding Kit; sequencing the 6 pooled samples yielded in total ~750.000 raw

reads.

Using TE-reX we identified respectively 62 Alu and 5 L1 NAHR events with high mapping confidence, the majority of which were intra-chromosomal (57/67). Most Alu pairs (44/62) and L1 pairs (3/5) detected by ONT were detected also in the capture-seq dataset. Notwithstanding ONT platform limitations in sequencing depth, the ultra-long reads offered a wider view around the recombination breakpoints (Fig. 4C-J). The median length of the split reads for the 67 HMC events was 5328 bp (longest 45580 bp, shortest 68 bp). The majority of intra-chromosomal recombination events between repeats in direct configuration (48/57) caused deletions of intervening genomic regions ranging from 176 bp to 1.6 Mb (median, 1037 bp). Nine recombinations between repeats in inverted configuration caused intra-chromosomal inversions of genomic fragments sized from 68 bp to 8625 bp (median 228 bp). The 10 inter-chromosomal recombination events confirmed by ONT sequencing caused exchange of fragments sized between 85 bp and 3079 bp (median 197 bp). Forty-seven percent of recombined elements were located in introns of RefSeq protein-coding genes, and 59% of the recombined intra-chromosomal pairs were within the same gene; none of these recombination events affected intervening exons.

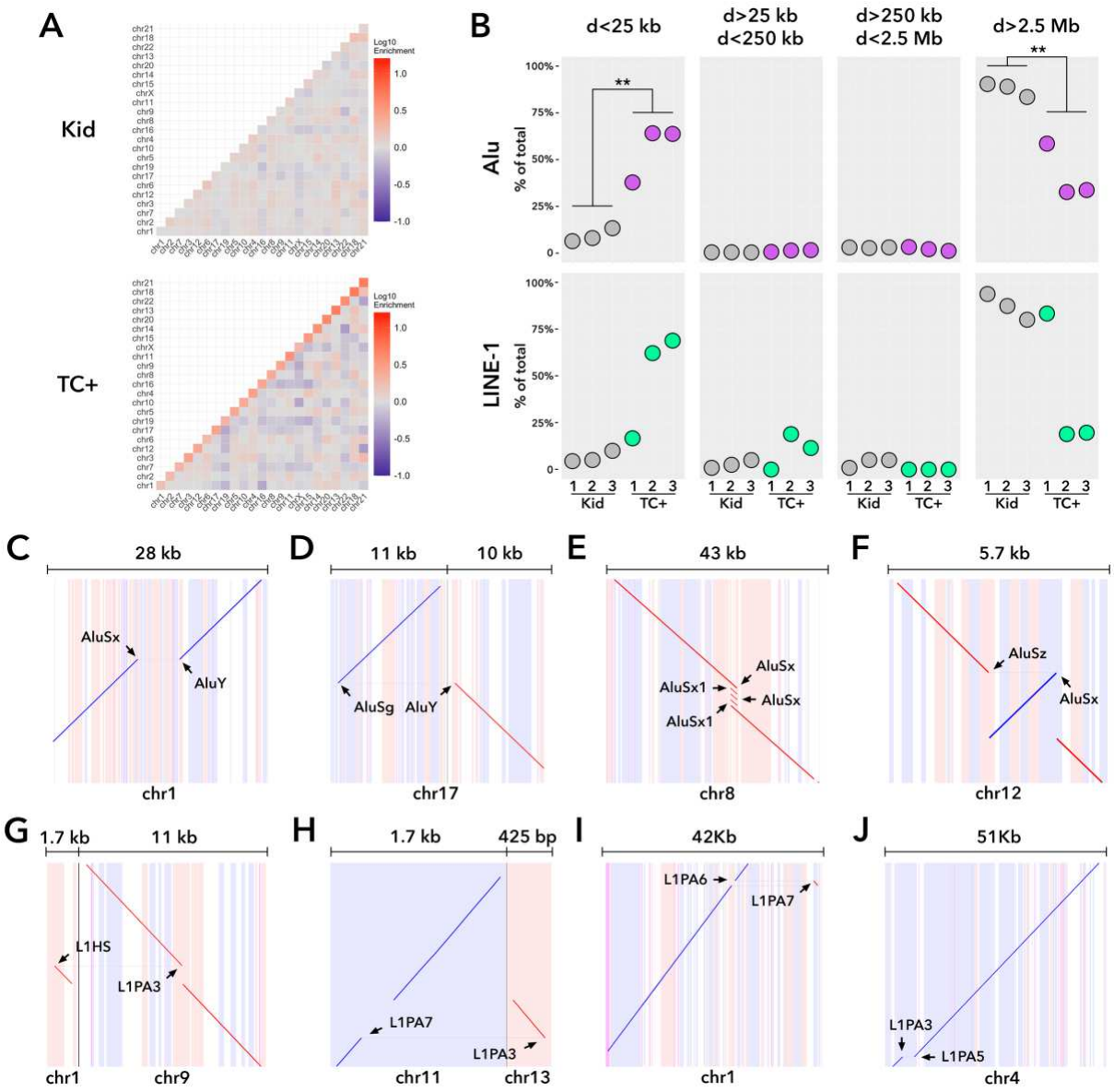

Fig. 4 NAHR of Alu and L1 elements detected in capture-free Illumina libraries and in PCR-free ONT libraries.

A) Merged chromosome-to-chromosome Alu recombination matrices for kidney and temporal cortex NeuN+ samples of 3 donors in capture-free short-reads libraries sequenced on Illumina Miseq. For each tissue, data for the 3 control donors are merged. Kid: kidney; TC+: Temporal Cortex NeuN+ fraction.

B) Genomic distance (d) of intra-chromosomal Alu and L1 recombinaed pairs for kidney and temporal cortex NeuN+ samples of 3 control donors in capture-free short-reads libraries. Dots represent values for individual samples.

C-J) Representative examples of Alu (C-F) and L1 (G-J) recombination events detected in PCR-free ultra-long DNA reads libraries sequenced on the ONT MinION platform. The vertical stripes are repeat annotations in the reference genome: pink = forward-oriented repeat elements, blue = reverse-oriented repeat elements. *P < 0.05; **P < 0.01; ***P < 0.001 (single-factor ANOVA).

Annotation of Alu and L1 elements with high recombinogenic activity

Alu and L1 elements have frequently been found at the boundaries of structural variants associated with genetic disorders, however the genomic prevalence of somatic NAHR hotspots in normal physiological conditions has never been addressed systematically. To gain additional statistical power, we merged the whole dataset and screened for Alu and L1 elements involved in recombination events recurring more than expected from random genomic distribution. Although the vast majority of Alu and L1 pairs in the dataset were non-recurrent (i.e., detected once) we identified hundreds of Alu and L1 pairs recurring across the whole dataset more than expected (n≥5 times, see methods for definition of thresholds) (Fig. S19). Interestingly, recurrently recombined pairs were predominantly intra-chromosomal (Fig. S20A-C). In addition, recurrent intra-chromosomal NAHR was strongly biased towards proximal repeats, in contrast with non-recurrent intra-chromosomal recombination (Fig. S20D-F). These differences between recurrent and non-recurrent recombination events may indicate a mutational bias or a purifying selection of inter-chromosomal and long-range intra-chromosomal recombination.

To investigate the recombinogenic activity of individual Alu and L1 elements, we disjoined the Alu-Alu and L1-L1 pairing information and calculated the number of individual recombination events per each repeat across the whole dataset ("recombination index", RI). About 50% of Alu and L1 elements had RI exceeding the threshold of random genomic distribution and were flagged as hotspots ("hot" elements) (Fig. S21 and Table S8).

We then explored the potential relevance of particularly hot Alu and L1 elements (RI≥40) in genomic instability; the most represented subfamilies in this subset were the youngest Alu (AluY) (7792/8073) and L1PA3/L1PA4/L1PA5 (cumulatively 2860/3570). These highly recombinogenic Alu and L1 elements were overall enriched in genomic intervals encompassing structural variations annotated in the Structural Variants Database of the 1000 Genomes Project (Fig 5A, B) (*43*). For Alu elements the enrichment was dependent on the genomic distance from the structural variants suggesting that highly recombinogenic Alu elements may be involved in genomic rearrangements underlying the annotated variants (Fig. 5C). Alu and L1 are responsible for recurrent mutations in several cancer types (*16*, *44–47*). Hence, we analyzed the genomic relationship of recombined retroelements with oncogenes and tumor suppressor genes included in the COSMIC database (*48*). Overall, the average RI of all recombined Alu elements annotated in COSMIC cancer genes (RI=10.48) was higher than the average RI of recombined Alu elements annotated in cancer-unrelated RefSeq protein-coding genes (RI=10.34) ($p=0.003$, single factor ANOVA). Hot Alu elements with RI≥40 were found more frequently in COSMIC cancer genes than in cancer-unrelated RefSeq genes (Fig. 5D). Conversely, the RI of L1 elements in COSMIC genes and cancer-unrelated RefSeq genes were not statistically different from each other (data not shown), and hot L1 elements were depleted from cancer genes (Fig. 5E). In addition, hot Alu and L1 elements with RI≥40 localized in proximity to centromere and telomeres on several chromosomes (Fig. S22D, G, I, K, L, N, O, P, R, S, U, V). Figure 5F depicts a cluster of hot retroelements in the subtelomeric region of

chromosome 4, which were involved in 698 NAHR events genome-wide. This highly recombinogenic repeats cluster encompasses the 5'-region of the gene ZNF595, which exhibits copy number variation and overexpression in several cancers (*49*). Coinciding hot Alu and L1 clusters were observed in the peri-centromeric region of the acrocentric chromosomes 21 and 22 (Fig. 5G, H); interestingly, the chromosome 21 locus hosting this cluster of highly recombinogenic repeats is involved in recurrent structural variants including Robertsonian translocations underlying some Down Syndrome cases (*50*).

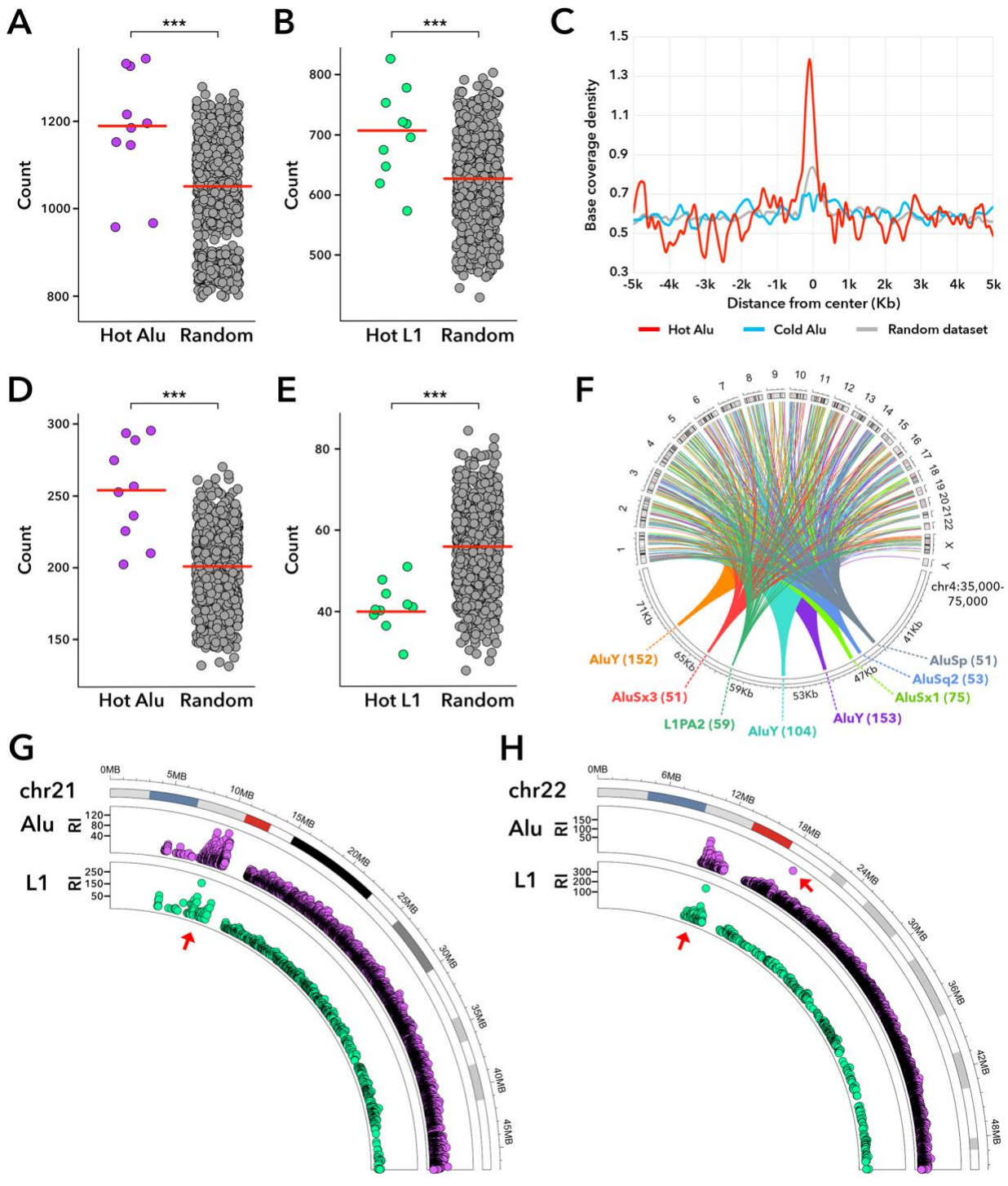

Fig. 5 Highly recombinogenic Alu and L1 elements are enriched in genomic features relevant to pathological contexts.

A, B) Hot Alu (A) and L1 (B) elements with Recombination Index (RI) ≥40 were enriched in Structural Variants (SVs) dataset of the 1000 Genomes Project Phase 3. Dots represent the number of hot elements per each donor that intersect the SVs intervals or 100x random datasets

with matching interval and sample sizes. Horizontal red bars indicate median. ***P < 0.001 (single-factor ANOVA).

C) The enrichment of Hot Alu elements with RI≥40 in 1000 Genomes Project Phase 3 SVs intervals was dependent on genomic distance from the center of SVs genomic intervals. In comparison, cold Alu elements (RI=1) and control Alu elements from 100x random datasets showed no significant enrichment.

D, E) Hot Alu (D) and L1 (E) elements with RI≥40 were enriched and depleted, respectively, in cancer genes annotated in the COSMIC database versus cancer-unrelated RefSeq protein-coding genes. Dots represent the count of hot elements per donor that intersected the coordinates of Alu or L1 annotated in COSMIC genes or in 100x equally sized random datasets obtained by shuffling RepeatMasker annotations of all Alu or L1 contained in cancer-unrelated Refseq protein-coding genes. Horizontal red bars indicate the median. ***P < 0.001 (single-factor ANOVA).

F) A cluster of hot Alu and L1 elements in a subtelomeric region of chromosome 4 is responsible for 698 individual recombination events across the merged dataset. Numbers in parentheses indicate the respective RI values. Chromosome 4 (bottom) is shown expanded in positions 35,000-75,000.

G, H) Representative examples of hot Alu and L1 elements and hot clusters (red arrows) in peri-centromeric regions of chromosome 21 and chromosome 22. Dots represent individual repeat elements found recombined in the merged dataset, irrespective of the RI.

Differentiation of induced pluripotent stem cells to neurons triggers emergence of cell-specific recombination profiles

To gain insights into the origins of tissue-specific recombination profiles observed in post-mortem tissues, we exploited an *in vitro* model of neuronal differentiation (*51*). This protocol

allows for the differentiation of induced pluripotent stem cells (iPSCs) into medial ganglionic eminence (MGE)-progenitor cells within 26 days, after which the specific induction towards GABAergic interneurons is started and prolonged for an additional 24 days (Fig. S23). We applied our capture-seq workflow to 3 biological replicas of iPSCs and differentiated neurons ("iNEU"; induced from iPSCs) and paired-end sequenced the libraries on Illumina Miseq platform at 300 bp reads yielding a total of ~32 millions of raw reads. TE-reX analysis and annotation of NAHR events in iPSCs and iNEUs revealed that the differentiation triggered significant changes in the recombination profiles of the induced neurons. Although the total number of recombination events did not differ, the intra-chromosomal recombination rates of iPSCs and iNEUs were marginally but significantly different (median intra-chromosomal rate iPSCs=6.51%, median intra-chromosomal rate iNEUs=6.76%, p=0.037, two-tailed Student's *t*-test). The analysis of intra-chromosomal recombination distance intervals showed significantly higher recombination of proximal Alu and L1 pairs in iNEUs compared with iPSCs (Fig. 6A), reminiscent of the difference between postmortem samples of brain and non-brain tissues. These results show that cell-fate commitment is accompanied by changes in recombination profiles and suggest that tissue-specific recombination profiles may be established during early developmental stages.

Somatic recombination is altered in neurodegeneration

The observation that somatic NAHR is pervasive in normal physiological conditions provokes questions about how the tissue-specific, complex network of genome-wide recombinations described so far is affected in disease. We probed the dynamics of Alu and L1 NAHR in the two most common forms of neurodegeneration, sporadic Parkinson's disease (PD) and sporadic Alzheimer's disease (AD). We obtained post-mortem tissue samples with equal composition to the main dataset from an equal number of PD and AD donors (Table S2).

Sequencing of 159 PD and AD capture-seq libraries yielded ~960 and ~990 millions of raw reads, respectively; quality control of PD and AD libraries showed capture efficiency and enrichment of Alu and L1 elements comparable to that of control donor libraries with no major differences (Fig. S1C-F; Fig. S3E-L; Fig. S4F-O; Fig. S5C-F; Fig. S6B, C; Fig. S8C-F; Fig. S9D-I). Genome-wide annotation of somatic NAHR events in PD and AD datasets recapitulated the findings of the control dataset in most aspects regarding the high counts of NAHR events (Fig. S24), genomic distribution (Fig. S10-S13), inter- vs intra-chromosomal recombination rates (Fig. S25), distance profiles of intra-chromosomal NAHR, recombination bias towards proximal inverted elements (Fig. S17C-F; Fig. S26-S27,) and enrichment or depletion of hot elements in COSMIC genes and structural variants (Fig. S28). A comparison of the total number of recombination events per sample revealed that the temporal cortex of AD donors was characterized by a significantly higher number of Alu and L1 recombination events compared with control donors (Fig. 6B); no differences were observed in other brain regions for AD samples or in PD samples versus controls. Moreover, analysis of chromosome-to-chromosome recombination matrices in PD and AD revealed a significant enrichment of intra-chromosomal recombination specific for the NeuN+ fraction of parietal cortex samples, for both Alu and L1 NAHR, compared with the respective control samples. A similar result was observed also for the frontal cortex samples of AD, while in PD both the NeuN- and NeuN+ fractions showed an increase of intra-chromosomal NAHR compared with the control dataset (Fig 6C, D: illustrative results for parietal cortex in PD and AD; complete list of panels in Fig. S29-S36). These findings suggest that pathological processes related to neurodegeneration can affect genome-wide NAHR profiles in a cell- and tissue-specific fashion.

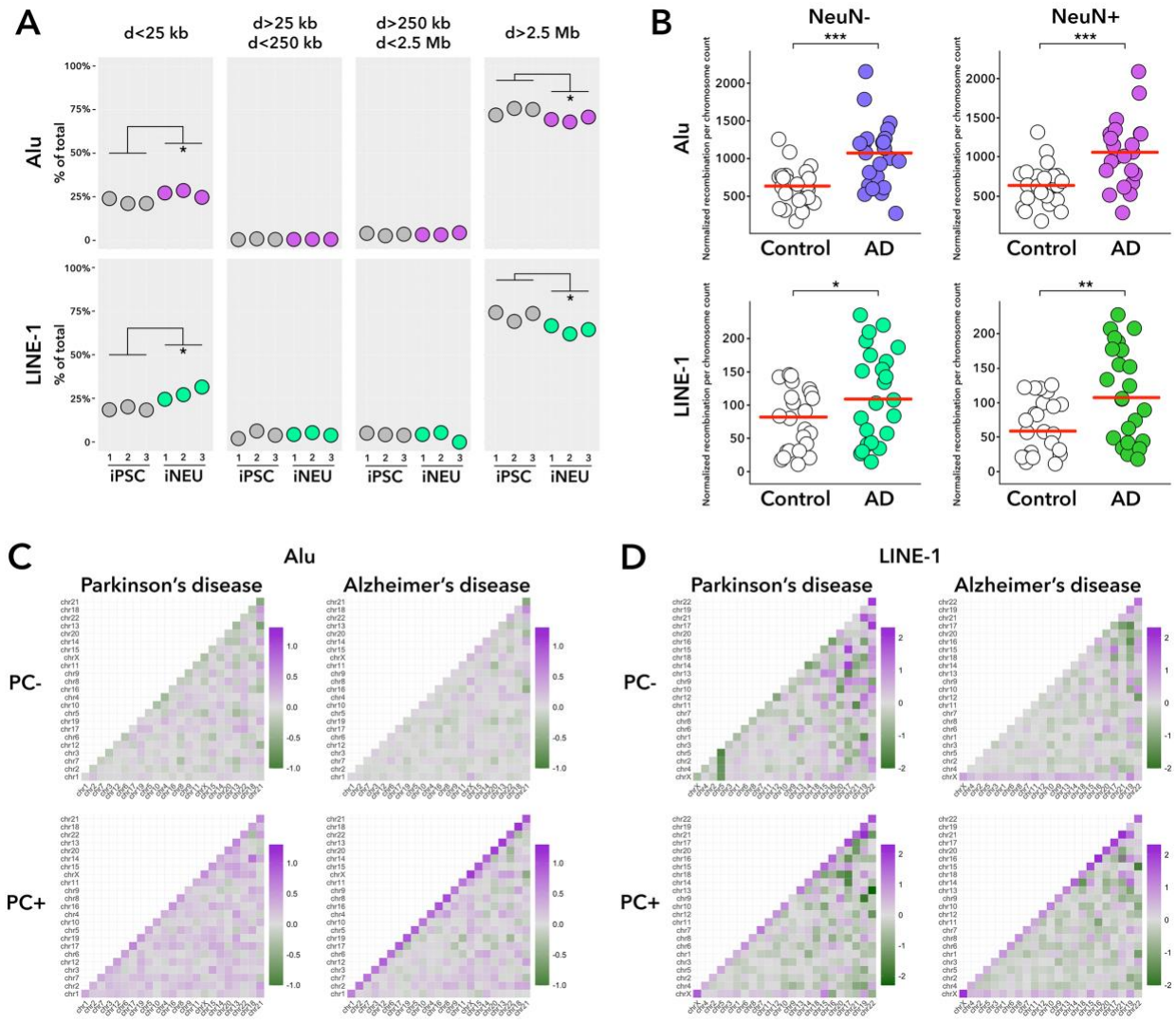

Fig. 6 Alu and L1 recombination profiles are shaped by in vitro neuronal differentiation and are altered in neurodegeneration.

A) GABAergic cortical interneurons (iNEU) derived from human iPSCs were specifically enriched in proximal intra-chromosomal recombination of Alu and L1 elements compared with iPSCs. Dots represent the 3 biological replicates per condition. *P < 0.05 (single-factor ANOVA).

B) Somatic NAHR of Alu and L1 elements is higher in temporal cortex samples of sporadic AD donors compared with control donors. Each dot represents the median of the recombination event count per chromosome for 10 AD or 10 control donors; the horizontal red bar denotes

the median across all chromosomes. *P < 0.05; **P < 0.01; ***P < 0.001 (single-factor ANOVA).

C, D) Chromosome-to-chromosome recombination matrices of parietal cortex samples normalized to values of control donors show a NeuN+ specific enrichment of intra-chromosomal recombination rates in sporadic Parkinson's disease (PD) and Alzheimer's disease (AD) donors.

Colors show Log2 fold-enrichment of recombination rates for each individual chromosome compared with recombination rate values of respective samples in the control dataset.

Data for chromosome Y not shown. PC: parietal cortex; -/+: NeuN- and NeuN+ fractions.

Discussion

One of the most puzzling discoveries in the first draft of the human genome sequence was that roughly half of the bases in our genomes belonged to repeated sequences, for long considered as inert remnants of our evolution. However, evidence accumulated during the intervening 20 years shows that many repeats are responsible for genomic diversity and mechanisms fundamental for life, bringing an evolutionary gain that has its trade-off in sporadic diseases generated or promoted by repeat element activity. Here, we show that somatic mosaicism caused by recombination of repeats in the human genome is extensive and complex, adding a new page to the developing story of how a myriad of genomic variants coexist in the same individual. A conservative estimate from our capture-seq libraries is that there are ~5 and ~1.2 Alu recombination events per cell in non-brain and brain tissues respectively, whereas for L1 recombination the count is ~10-fold lower (~7 events per 10 cells in kidney or liver, ~1.6 events per 10 cells in brain samples). These figures are coherent with estimates of somatic retrotransposition in the brain (*30*, *31*) and suggest an important contribution of NAHR to somatic genome diversity; however, dedicated and more sensitive technical approaches will be

required to confirm our data at the single-cell level. Besides being pervasive, somatic NAHR exhibits tissue-specific characteristics that distinguish the brain regions from other tissues assayed. Brain-specific NAHR is characterized by a higher rate of intra-chromosomal recombination and by higher recombination rates between close repeat elements. In addition, close-range recombination in the brain exhibits a strong directionality bias in favor of repeats in inverted configuration. These differences could arise from several factors, including cell- and tissue-specific DSBR pathways, chromatin architecture, epigenetic modifications, and developmental factors. Our analyses of recombination in iPSC and differentiated neurons suggest that the specific recombination profiles observed in post-mortem samples may be generated in progenitor cells during early developmental stages; this notion is supported by the homogeneity of intra-sample profiles and in the lack of fundamental differences in the NeuN- and NeuN+ fractions in the control donor samples. It is nevertheless an open question whether tissue-specific recombination profiles are a cause or consequence of cell-fate determination. It is tempting to speculate that recombination of repeat elements is actively engaged in genome remodeling during developmental programs; another possibility is that differentiation from stem cells into a given progenitor cell type may trigger specific changes in chromatin conformation accompanied by reproducible patterns of DNA damage, resulting in discrete recombination profiles. Similarly, the higher enrichment of intra-chromosomal recombination in the NeuN+ fraction than the NeuN- fraction in our comparison of AD and PD versus control sample may be the consequence of cell-type specific neurodegenerative processes causing differential DNA damage and cell-type specific alterations of recombination profiles.

The extent of somatic recombination detected in this study poses the problem of compatibility of the observed rearrangements with genome fitness. Ultra-long read sequencing showed that Alu and L1 NAHR events are responsible for deletions, inversions and translocations spanning a wide range of sizes. While none of the NAHR events detected by ONT sequencing affected

intervening coding exons, inter- and intra-genic regulatory elements may be affected by NAHR-induced CNVs with consequences on close and distant chromatin architecture and gene expression in the affected cells (*52*). Spatial genome organization at single cell level is characterized by high variability (*53*); on the basis of its magnitude we speculate that somatic NAHR, as well as other structural variants coexisting in the same genomic milieu, may contribute to this heterogeneity. The enrichment of highly recombinogenic Alu and L1 elements in contexts of genomic instability and particularly in cancer genes suggests that somatic recombination of Alu and L1 may occasionally prime the genome of individual cells at vulnerable sites and drive the transition from healthy to pathological states. This scenario becomes even more plausible when considering possible complex interplay between different types of somatic mutation events; for example, a somatic retrotransposition event may be accompanied by a local genomic destabilization and consequent recombination if the newly inserted retroelement finds itself flanked by inverted homologous repeats (*54*). This is of particular interest also for estimates of the rate of retrotransposition in the human genome, because insertion of a young retroelement followed by recombination may create complex rearrangements masking the structural hallmarks of canonical retrotransposition events, namely target site duplications and a poly(A) tail, possibly resulting in underestimation of germ-line and somatic retrotransposition rates. A recurrent question in repeat elements research is, what possible benefit can counterbalance the risk of tolerating the presence in our genomes of such a high number of potentially mutagenic elements? Restricting the range of answers to the field of DNA repair, it is possible that when a DNA break occurs interspersed repeats with high homology offer an ideal repertoire of “rapid emergency kits" conveniently available at every genomic corner. Furthermore Alu, LINE-1 and LTR (Long terminal repeat) sequences can form G-quadruplex structures that may help to stabilize complexes formed with the DNA double-strand break repair components via liquid-liquid phase separation (*55–57*). The

thorough characterization of somatic recombination of Alu and L1 elements in this study paves the way to future experiments that will explore the dynamics of somatic NAHR events and their impact on the structure and function of our genomes.

References

1. A. Smith, R. Hubley, P. Green, RepeatMasker. *RepeatMasker Open-4.0*, (available at http://www.repeatmasker.org/).

2. P. Deininger, Alu elements: know the SINEs. *Genome Biol*. **12**, 236 (2011).

3. H. Khan, A. Smit, S. Boissinot, Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res*. **16**, 78–87 (2006).

4. C. Daniel, G. Silberberg, M. Behm, M. Öhman, Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biol.* **15**, R28 (2014).

5. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. **351**, 1083–1087 (2016).

6. J. W. Jachowicz, X. Bing, J. Pontabry, A. Bošković, O. J. Rando, M.-E. Torres-Padilla, LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.* **49**, 1502–1510 (2017).

7. X.-O. Zhang, T. R. Gingeras, Z. Weng, Genome-wide analysis of polymerase III-transcribed Alu elements suggests cell-type-specific enhancer function. *Genome Res.* **29**, 1402–1414 (2019).

8. J. Pontis, E. Planet, S. Offner, P. Turelli, J. Duc, A. Coudray, T. W. Theunissen, R. Jaenisch, D. Trono, Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell*. **24**, 724-735.e5 (2019).

9. C. R. Beck, P. Collier, C. Macfarlane, M. Malig, J. M. Kidd, E. E. Eichler, R. M. Badge, J. V. Moran, LINE-1 retrotransposition activity in human genomes. *Cell*. **141**, 1159–1170 (2010).

10. J. L. Goodier, Restricting retrotransposons: a review. *Mobile DNA*. **7**, 16 (2016).

11. C. Philippe, D. B. Vargas-Landin, A. J. Doucet, D. van Essen, J. Vera-Otarola, M. Kuciak, A. Corbin, P. Nigumann, G. Cristofari, Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife Sciences*. **5**, e13926 (2016).

12. G. J. Faulkner, V. Billon, L1 retrotransposition in the soma: a field jumping ahead. *Mobile DNA*. **9**, 22 (2018).

13. S. K. Sen, K. Han, J. Wang, J. Lee, H. Wang, P. A. Callinan, M. Dyer, R. Cordaux, P. Liang, M. A. Batzer, Human genomic deletions mediated by recombination between Alu elements. *Am. J. Hum. Genet.* **79**, 41–53 (2006).

14. W. Gu, F. Zhang, J. R. Lupski, Mechanisms for human genomic rearrangements. *Pathogenetics*. **1**, 4 (2008).

15. M. Sasaki, J. Lange, S. Keeney, Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol*. **11**, 182–195 (2010).

16. C. Robberecht, T. Voet, M. Z. Esteki, B. A. Nowakowska, J. R. Vermeesch, Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res*. **23**, 411–418 (2013).

17. B. B. Currall, C. Chiangmai, M. E. Talkowski, C. C. Morton, Mechanisms for structural variation in the human genome. *Curr Genet Med Rep*. **1**, 81–90 (2013).

18. A. Piazza, W.-D. Heyer, Homologous recombination and the formation of complex genomic rearrangements. *Trends Cell Biol*. **29**, 135–149 (2019).

19. A. Piazza, W.-D. Heyer, Moving forward one step back at a time: reversibility during homologous recombination. *Curr Genet*. **65**, 1333–1340 (2019).

20. J.-M. Chen, D. N. Cooper, N. Chuzhanova, C. Férec, G. P. Patrinos, Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).

21. E. Kolomietz, M. S. Meyn, A. Pandita, J. A. Squire, The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer*. **35**, 97–112 (2002).

22. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet*. **12**, 187–215 (2011).

23. F. Zhang, W. Gu, M. E. Hurles, J. R. Lupski, Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. **10**, 451–481 (2009).

24. B. Elliott, C. Richardson, M. Jasin, Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. *Mol. Cell*. **17**, 885–894 (2005).

25. M. E. Morales, T. B. White, V. A. Streva, C. B. DeFreece, D. J. Hedges, P. L. Deininger, The contribution of alu elements to mutagenic DNA double-strand break repair. *PLoS Genet.* **11**, e1005016 (2015).

26. M. M. Parks, C. E. Lawrence, B. J. Raphael, Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology*. **16**, 72 (2015).

27. M. Startek, P. Szafranski, T. Gambin, I. M. Campbell, P. Hixson, C. A. Shaw, P. Stankiewicz, A. Gambin, Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res*. **43**, 2188–2198 (2015).

28. G. D. Evrony, X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, P. J. Park, C. A. Walsh, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*. **151**, 483–496 (2012).

29. G. D. Evrony, E. Lee, B. K. Mehta, Y. Benjamini, R. M. Johnson, X. Cai, L. Yang, P. Haseley, H. S. Lehmann, P. J. Park, C. A. Walsh, Cell lineage analysis in human brain using endogenous retroelements. *Neuron*. **85**, 49–59 (2015).

30. G. D. Evrony, E. Lee, P. J. Park, C. A. Walsh, Resolving rates of mutation in the brain using single-neuron genomics. *eLife*. **5**, e12966 (2016).

31. J. A. Erwin, A. C. M. Paquola, T. Singer, I. Gallina, M. Novotny, C. Quayle, T. A. Bedrosian, F. I. A. Alves, C. R. Butcher, J. R. Herdy, A. Sarkar, R. S. Lasken, A. R.

Muotri, F. H. Gage, L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci*. **19**, 1583–1591 (2016).

32. R. Hubley, R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. A. Smit, T. J. Wheeler, The Dfam database of repetitive DNA families. *Nucleic Acids Res*. **44**, D81–D89 (2016).

33. J. K. Baillie, M. W. Barnett, K. R. Upton, D. J. Gerhardt, T. A. Richmond, F. De Sapio, P. M. Brennan, P. Rizzu, S. Smith, M. Fell, R. T. Talbot, S. Gustincich, T. C. Freeman, J. S. Mattick, D. A. Hume, P. Heutink, P. Carninci, J. A. Jeddeloh, G. J. Faulkner, Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*. **479**, 534–537 (2011).

34. R. Shukla, K. R. Upton, M. Muñoz-Lopez, D. J. Gerhardt, M. E. Fisher, T. Nguyen, P. M. Brennan, J. K. Baillie, A. Collino, S. Ghisletti, S. Sinha, F. Iannelli, E. Radaelli, A. Dos Santos, D. Rapoud, C. Guettier, D. Samuel, G. Natoli, P. Carninci, F. D. Ciccarelli, J. L. Garcia-Perez, J. Faivre, G. J. Faulkner, Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*. **153**, 101–111 (2013).

35. P. E. Carreira, A. D. Ewing, G. Li, S. N. Schauer, K. R. Upton, A. C. Fagg, S. Morell, M. Kindlova, P. Gerdes, S. R. Richardson, B. Li, D. J. Gerhardt, J. Wang, P. M. Brennan, G. J. Faulkner, Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mob DNA*. **7**, 21 (2016).

36. A. Matevossian, S. Akbarian, Neuronal nuclei isolation from human postmortem brain tissue. *J Vis Exp* (2008), doi:10.3791/914.

37. K. Iwamoto, M. Bundo, J. Ueda, M. C. Oldham, W. Ukai, E. Hashimoto, T. Saito, D. H. Geschwind, T. Kato, Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons. *Genome Res*. **21**, 688–696 (2011).

38. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).

39. A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning,

X. Wang, M. Claussnitzer, Yaping Liu, C. Coarfa, R. Alan Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. David Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. Scott Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

40. M. Karimzadeh, C. Ernst, A. Kundaje, M. M. Hoffman, Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res*. **46**, e120–e120 (2018).

41. K. S. Lobachev, J. E. Stenger, O. G. Kozyreva, J. Jurka, D. A. Gordenin, M. A. Resnick, Inverted Alu repeats unstable in yeast are excluded from the human genome. *EMBO J.* **19**, 3822–3830 (2000).

42. J. E. Stenger, K. S. Lobachev, D. Gordenin, T. A. Darden, J. Jurka, M. A. Resnick, Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. *Genome Res.* **11**, 12–27 (2001).

43. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M.

A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).

44. P. J. Hastings, J. R. Lupski, S. M. Rosenberg, G. Ira, Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).

45. K. Kitada, K. Kitada, S. Aikawa, S. Aida, Alu-Alu fusion sequences identified at junction sites of copy number amplified regions in cancer cell lines. *CGR*. **139**, 1–8 (2013).

46. J. Smida, H. Xu, Y. Zhang, D. Baumhoer, S. Ribi, M. Kovac, I. von Luettichau, S. Bielack, V. B. O'Leary, C. Leib-Mösch, D. Frishman, M. Nathrath, Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma. *International Journal of Cancer*. **141**, 816–828 (2017).

47. Y. Wang, A. J. Bernhardy, J. Nacson, J. J. Krais, Y.-F. Tan, E. Nicolas, M. R. Radke, E. Handorf, A. Llop-Guevara, J. Balmaña, E. M. Swisher, V. Serra, S. Peri, N. Johnson, BRCA1 intronic Alu elements drive gene rearrangements and PARP inhibitor resistance. *Nat Commun*. **10**, 5661 (2019).

48. Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, S. A. Forbes, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*. **18**, 696–705 (2018).

49. ZNF595 gene - Somatic mutations in cancer, (available at https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=ZNF595).

50. S.-W. Shaw, C.-P. Chen, P.-J. Cheng, T.-H. Wang, J.-W. Hou, C.-T. Lin, S.-D. Chang, H.-L. Hwa, J.-L. Lin, A.-S. Chao, Y.-K. Soong, F.-J. Hsieh, Gene dosage change of TPTE and BAGE2 and breakpoint analysis in Robertsonian Down syndrome. *J. Hum. Genet.* **53**, 136–143 (2008).

51. Y. Liu, H. Liu, C. Sauvey, L. Yao, E. D. Zarnowska, S.-C. Zhang, Directed differentiation of forebrain GABA interneurons from human pluripotent stem cells. *Nat Protoc*. **8**, 1670–1679 (2013).

52. M. Rigau, D. Juan, A. Valencia, D. Rico, Intronic CNVs and gene expression variation in human populations. *PLOS Genetics*. **15**, e1007902 (2019).

53. E. H. Finn, G. Pegoraro, H. B. Brandão, A.-L. Valton, M. E. Oomen, J. Dekker, L. Mirny, T. Misteli, Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*. **176**, 1502-1515.e10 (2019).

54. N. Gilbert, S. Lutz-Prigge, J. V. Moran, Genomic deletions created upon LINE-1 retrotransposition. *Cell*. **110**, 315–325 (2002).

55. M. Lexa, P. Steflova, T. Martinek, M. Vorlickova, B. Vyskot, E. Kejnovsky, Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics*. **15**, 1032 (2014).

56. A. C. Hall, L. A. Ostrowski, K. Mekhail, Phase separation as a melting pot for DNA repeats. *Trends Genet.* **35**, 589–600 (2019).

57. F. Pessina, F. Giavazzi, Y. Yin, U. Gioia, V. Vitelli, A. Galbiati, S. Barozzi, M. Garre, A. Oldani, A. Flaus, R. Cerbino, D. Parazzoli, E. Rothenberg, F. d'Adda di Fagagna, Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat. Cell Biol.* **21**, 1286–1299 (2019).

58. P. W. Laird, A. Zijderveld, K. Linders, M. A. Rudnicki, R. Jaenisch, A. Berns, Simplified mammalian DNA isolation procedure. Nucleic Acids Res. 19, 4293 (1991).

59. J. Sambrook, D. W. Russell, Cold Spring Harb. Protoc., in press, doi:10.1101/pdb.prot4455.

60. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 27, 2957–2963 (2011).

61. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 841–842 (2010).

62. M. Hamada, Y. Ono, K. Asai, M. C. Frith, Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics. 33, 926–928 (2017).

63. M. C. Frith, R. Kawaguchi, Split-alignment of genomes finds orthologies more accurately. Genome Biol. 16, 106 (2015).

64. A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, S. G. Rozen, Primer3—new capabilities and interfaces. Nucleic Acids Res. 40, e115 (2012).

65. Homer Software and Data Download, (available at http://homer.ucsd.edu/homer/).

Acknowledgements: The authors are indebted to Eric Arner (RIKEN IMS) and Charles Plessy (OIST) for help revising the manuscript and useful comments. **Funding:** This work was funded by a Research Grant from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, to the RIKEN Center for Integrative Medical Sciences. This work was partly supported by Japan Society for the Promotion of Science KAKENHI (CoBiA)(JP16H06277 to Murayama S.) and by Japan Agency for Medical Research and Development (JP19dm0107106 to Murayama S.). **Authors contributions:** P.C., M.F. and G.P. conceived the original idea and managed the project; G.P. and P.C. developed the capture-seq protocol and designed the experiments; M.F. developed the TE-reX pipeline and supervised the analyses of TE-reX data; G.P. and A.Bu. processed all samples and prepared the capture-seq libraries; G.P., K.H. and M.F. conceived and performed the computational analyses; J.L. performed the iPSC differentiation experiments; C.P. and Y.H.W. provided support for ONT and Illumina libraries preparation; C.C.H. performed computational analyses and provided insightful comments on data interpretation; A.K., F.A. and J.S. provided support for computational analyses; A.Bo. provided valuable support for interpretation of results; S.M. provided human post-mortem samples; S.G. provided critical feedback; G.P. produced the figures and wrote the manuscript with valuable contributions from M.F., A.Bo., F.A., K.H., P.C. and S.G. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** TE-reX and related documentation can be accessed at: https://gitlab.com/mcfrith/te-rex. Sequencing data for this project have been deposited in the NCBI Sequence Read Archive (SRA) database under accession number PRJNA636606, and are accessible at the following link: https://www.ncbi.nlm.nih.gov/sra/PRJNA636606.