METHODS

Cell culture

Cells were cultured at 37ºC in 5% CO2 atmosphere with 100% humidity. hTERT-immortalized, p53-/- RPE-1 cells and U2OS Flp-In T-REx cells were grown in DMEM/F12 (1:1) or DMEM respectively, and supplemented with 10% FBS, 100 IU/ml penicillin, and 100 µg/ml streptomycin. For cell lines containing doxycycline-inducible transgene constructs, tetracycline-free FBS (Takara) was used in all culture media, and 1 µg/ml of doxycycline was used to induce transgene expression. U2OS cells were provided by Dr. Jeremy Stark (1), and RPE-1 cells were provided by Dr. David Pellman (2).

Plasmid construction and cell line generation

We used several plasmid constructs to induce the expression of a codon-optimized sequence of human L1 (ORFeus) in cells. For the Tet-On expression system in RPE-1 cells, we exploited the Sleeping Beauty DNA transposon system to stably deliver a Tet-On L1 (ORFeus) cassette, which we previously validated (3). For this, we transfected hTERT RPE-1 p53-/- cells using Viafect (Progema) with an expression vector for the Sleepy Beauty transposase (pCMV(CAT)T7-SB100), and the donor plasmid containing Sleepy Beauty inverted terminal repeats flanking the inducible Tet-On L1 cassette, and a constitutively expressing rtTA-T2A-NeoR cassette (pCMD20D). Cells were selected with G418, and single cells were sorted to generate a monoclonal cell line (Tet-On L1 RPE-1 p53-/- cells), which we confirmed via immunoblotting of the L1-encoded proteins. Similarly, we also generated a monoclonal cell line containing a Tet-On Luciferase cassette (pCMD26A, Tet-On Luc RPE-1 p53-/-). For transient expression of L1 in RPE-1 cells or U2OS cells, we used a pCEP4 episomal expression vector that was modified to contain a Puromycin resistance gene to express under a CMV promoter a codon-optimized sequence of human L1 containing a GFP reporter for retrotransposition (pMT527, pCEP4 L1-ORFeus), which we previously validated (4). We also cloned retrotransposition mutant versions of the L1 reporter

plasmid containing mutations in the encoded ORF2p, including a reverse transcriptase mutant (pLD631, D702Y) (4) and two endonuclease mutants (pCMD104A, E43A:D145A & pCMD103A, D205G:H230A). To establish the Tet-On expression system in U2OS Flp-In T-REx cells, we removed the GFP cassette from pcDNA5 FRT/TO GFP (Addgene #19444) and cloned in our L1 expression cassette derived from our pCEP4 vectors aforementioned: pCMD111az (pcDNA5 FRT/TO L1-ORFeus), pCMD112az (pcDNA5 FRT/TO RTmutant D702Y), pCMD113az (pcDNA5 FRT/TO L1-ORFeus ENmutant D205:D145A) and pCMD114az (pcDNA5 FRT/TO L1-ORFeus ENmutant E43A:D145A). These Tet-On L1 expression plasmids were integrated into U20S Flp-In T-Rex cells by co-transfection with the PGK-Flp recombinase vector using FugeneHD (Promega), as previously described (1). Integrated clones were selected using hygromycin (0.2 ug/ uL) and subsequently screened by immunoblotting for both L1-encoded proteins.

L1 retrotransposition reporter assay

We performed the L1 GFP reporter assay for retrotransposition in RPE-1 p53-/- cells as previously described (3, 5). Briefly, we seeded 0.8 x10^5 cells in 6-well plates (day 1, d 1). The following day we transfected each well using Viafect (Promega) with 1 ug of pCEP4-Puromycin L1 GFP reporter plasmid (d 2): pMT527 (WT), pLD631 (RTmutant-D702Y), pCMD103A (ENmutant-D205G:H230A) or pCMD104A (ENmutant-E43A:D145A). A pCEP4-Puromycin vector expressing a GFP cassette (MT498, pCEP4-GFP) was used to monitor transfection efficiency by assaying the percentage of GFP+ cells on d 4 by flow cytometry (BD LSRFortessa). Media was changed 12 hours after transfection (d 3). Two days after transfection (d 4), the media for cells transfected with the L1 reporter plasmids were supplemented with 5 ug/ mL of Puromycin. Cells were then incubated until d 7 when cells were collected and assayed for the percentage of GFP+ cells by flow cytometry (BD LSRFortessa). Singlets were gated on side-scatter versus forward scatter and GFP+ cells were gated on GFP versus autofluorescence (PE). Cells with autofluorescence were detected on a diagonal line, whereas cells showing increased green fluorescence (GFP+) were

gated above the autofluorescence diagonal line. We normalized the percentage of GFP+ cells from the cells transfected with the L1 reporter assay to the percentage of GFP+ cells from cells transfected with the GFP plasmid.

Induction of L1 expression

Tet-On L1 p53-/- RPE-1 cells were exposed to Dox for five days to induce L1 expression. Tet-On Luc (also p53-/-) RPE-1 with the same treatment and Tet-On L1 p53-/- RPE-1 cells treated with DMSO for five days were used as control.

Immunoblotting

For Tet-On L1 or Luc RPE-1 p53-/- cells, cells were treated with DMSO or Dox (1 μg/mL) for 5 d and collected for protein extraction. As a control, parental RPE-1 p53-/- cells were treated with DMSO, or 1 μM mitomycin C (MMC, SC-3514) for 48 h. For U2OS Flp-In T-REx cells transfected with L1-expressing pCEP4 vectors, 2x10^5 cells were transfected with 1 ug of plasmid and collected two days after for protein extraction. For Tet-On L1 U20S cells, cells were treated with Dox (1 μg/mL) and collected three days later. Protein was extracted from cells using radioimmunoprecipitation assay buffer (Boston BioProducts BP-115) supplemented with protease and phosphatase inhibitors (Cell Signaling, 5872S). Gel electrophoresis was performed on protein extracts using 4 to 20% Mini-PROTEAN TGX gels (Bio-Rad, 456-1095). Proteins were then transferred to low fluorescence polyvinylidene difluoride membranes using Trans-Blot Turbo (Bio-Rad). Membranes were blocked using EveryBlot Blocking Buffer (Bio-Rad, 12010020) or Intercept Blocking Buffer (Li-COR, 927-60001), and probed with primary antibodies for pKAP1-S824 (Abcam, ab84077), KAP1 (Abcam, ab22553), RAD50-S635 (Cell Signaling, 14223S), RAD50 (Cell Signaling, 3427T), ORF2p (abcam, ab263071), ORF1p (Millipore Sigma, MABC1152), beta-tubulin (Cell Signaling, 2128S), gH2AX (Cell Signaling, 2577S), H3 (abcam, ab1791), CHK1 (Cell Signaling, 2360S), pCHK1-S345 (Cell Signaling, 2348S), RPA32 (Bethyl, A300-244), pRPA32-

S4/S8 (Bethyl, A300-245), or pRPA32-S33 (Bethyl, A300-246), followed by secondary antibodies (IRDye 800CW goat anti-mouse IgG, 925-32210; IRDye 680RD goat anti-rabbit IgG, 925-68071; and anti-rabbit IgG horseradish peroxidase (HRP), 7074S). ECL substrate (Thermo Scientific, 34580) was used to develop HRP signals. Immunoblotting signals were detected using the ChemiDoc imaging system (Bio-Rad).

Immunofluorescence

For Tet-On L1 cell lines, cells were seeded on #1.5 coverslips and treated with DMSO or Dox (1 ug/ uL) for five days (RPE-1 cells) or three days (U20S cells). U20S cells were washed with PBS and treated with pre-extraction (20 mM HEPES, 50 mM NaCl, 1 mM EDTA, 3 mM MgCl2, 300 mM sucrose, 0.25% Triton-X 100) prior to fixation. RPE-1 cells were washed with PBS and immediately fixed with 4% paraformaldehyde for 20 minutes at room temperature. Cells were then washed with 0.1 M glycine followed by PBS prior to permeabilization with 0.5% Triton X-100 in PBS for 15 minutes. Cells were washed with PBS three times and then blocked with 3% BSA in PBS for 1 hour. Cells were washed two times with PBS, followed by a 1-hour incubation with primary antibodies: yH2AX (Cell Signaling, 2577S) or 53BP1 (Cell Signaling, NB-100-904). Cells were washed three times with 0.05% Triton X-100 in PBS. Species-specific secondary antibodies were added to the cells for 1 hour followed by three washes with 0.05% Triton X-100 in PBS. 2.5 µg/mL Hoechst 33342 in PBS was added to the cells for 10 minutes. After three washes with PBS, Prolong Diamond Antifade (Life Technologies, P36961) was used for mounting the samples on glass slides. Imaging was performed on a Leica THUNDER Imager at 40X magnification using the Leica LAS X software. Quantification of yH2AX or 53BP1 nuclear foci per cell and the frequency of cells with micronuclei was performed using ImageJ.

Viability assay

Tet-On L1 or Tet-On Luc RPE-1 p53-/- cell lines or Tet-on L1 U20S cell lines ($1x10^3$ cells) were seeded in 48-well plates in media containing DMSO as control or varying concentration of Dox as indicated in triplicates. After eight days, the plates were washed with PBS and fixed with 0.5% crystal violet, 20% methanol solution for 20 mins. Plates were rinsed with H20 and air-dried and then imaged on a scanner. Once fully dried, 100 uL of methanol was added to each well to dissolve the crystal violet, and 90% of the solution was transferred to a clear 96-well plate to measure the absorbance at 570 nm from each treatment. Absorbance from the DMSO treatment per cell line was set to 1 and the rest of the absorbance measurement under Dox treatment was normalized to DMSO. This protocol was adapted from (6).

RNA-Seq data generation and analysis

Total RNA was purified from cells with either Tet-On L1 or Tet-On Luc after five days Dox treatment using RNAeasy Mini Kit including DNase treatment. Libraries were prepared using Roche Kapa mRNA HyperPrep strand-specific sample preparation kits from 200 ng of purified total RNA according to the manufacturer's protocol on a Beckman Coulter Biomek i7. The finished dsDNA libraries were quantified by Qubit fluorometer and Agilent TapeStation 4200. Uniquely dual indexed libraries were pooled in an equimolar ratio and shallowly sequenced on an Illumina MiSeq to further evaluate library quality and pool balance. The final pool was sequenced on an Illumina NovaSeq X Plus Instrument to generate 40 million 150bp read pairs per library at the Dana-Farber Cancer Institute Molecular Biology Core Facilities.

Sequencing reads were aligned to human genome reference (GRCh38) using STAR v2.7.11b (7) with the parameter "--outSAMstrandField intronMotif". Gene-level and transcript-level read counts were obtained using Stringtie (v2.2.3). DESeq2 (1.44.0) was used to perform differential expression analysis between Tet-On LINE-1 RPE-1 p53-/- and Tet-On Luc RPE-1 p53-/- cells. Differentially expressed genes were defined as those with an adjusted p-value < 0.05. Genes with fewer than 10 read counts in five or more samples (including replicates) were

excluded from analysis. Gene function enrichment analysis was performed using DAVID, supplying differentially expressed genes as defined above. Volcano plots and violin plots were generated using the results from DESeq2 and DAVID.

Generation of single cells and single-cell progeny clones for sequencing analysis

Cells with induced L1 expression (Tet-On L1 under 5 days of Dox treatment) and control (Tet-On L1 with 5 days of DMSO treatment) were sorted into 96-well plates. To generate whole-genome libraries from single cells, the sorted cells were immediately lysed and underwent whole-genome amplification using the REPLI-g Single Cell Kit (Qiagen, 150345), following a previously described protocol (8, 9). The amplified DNA was purified using ethanol precipitation, quantified using Qubit, and underwent library construction. 64 single cells with L1 induction and 32 control single cells (treated with DMSO) were sent for low-pass whole-genome sequencing (0.1x median depth). 78 samples (26 Dox treated; 52 DMSO treated) with uniform coverage were analyzed for large copy-number alterations. We further generated deep whole-genome sequencing data (30x median depth) on 28 cells with L1 induction and 12 control cells (including all with detectable large segmental copy-number alterations from the low-pass data).

To generate single-cell derived clones, single cells (with or without L1 induction) were sorted into 96-well plates containing media with 20% FBS. Ten plates were collected from each treatment, and the number of progeny clones was recorded to determine the clonogenicity of cells after exposure to L1 expression (Fig. 1F). Progeny clones were expanded in 6-well plates until confluency to generate a frozen vial. Cell pellets were used for genomic DNA extraction using PureLink Genomic DNA kit (Invitrogen) and for library construction. 60 Dox clones (derived from cells with L1 induction) and 32 control clones (derived from cells treated with DMSO) underwent low-pass whole-genome sequencing. 31 Dox clones and 10 control clones were sent for deep whole-genome sequencing (20x median depth). For 25 Dox clones and one control clones, we further generated PacBio long-read sequencing data (15x median depth).

Finally, clones were also derived from GFP(+) and GFP(-) RPE-1 cells with transient expression of the L1 GFP reporter as described above (seven days of L1 GFP expression; three independent experiments). A total of 62 clones derived from GFP(+) cells and 30 clones derived from GFP(-) cells underwent low-pass whole-genome sequencing. 29 GFP(+) clones and 5 GFP(-) clones were sent for deep whole-genome sequencing (20X median depth). For 13 GFP(+) clones, we generated PacBio long-read sequencing data (15x median depth).

Shotgun whole-genome sequencing data generation

200 ng of amplified gDNA from single cell samples or gDNA from progeny clones was fragmented to ~ 500 bp on a Covaris R220 instrument using a 96 microTUBE plate (Covaris, 520078). DNA libraries were prepared using reagents from LTP Library Preparation Kit (KAPA, KK8232) for multiplexed next-generation sequencing using Unique Dual-Indexed Adapters (KK8726). Finished libraries were quantified by Qubit fluorometer, and the fragment size distribution was evaluated by Agilent Bioanalyzer 2100 or Agilent TapeStation 4200. DNA libraries were pooled from each experimental condition and subjected to low-pass whole-genome sequencing (~0.1x mean coverage) on the MiSeq (Illumina) platform with paired-end 150bp reads to assess library quality and to estimate haplotype DNA copy-number for identifying samples with genomic alterations. Selected samples with DNA copy-number alterations were subsequently selected for deep sequencing (20x mean coverage) on the NovaSeq S4 (Illumina) platform with paired-end 150bp reads. Cells with Tet-On L1: 12 control single cells, 28 single cells with L1 induction, 10 progeny clones derived from control cells, 31 progeny clones derived from cells with L1 induction. Cells with L1 GFP reporter, 5 progeny clones derived from GFP(-) cells and 29 GFP (+) progeny clones derived from GFP(+) cells.

Shotgun sequencing data analysis

Shotgun sequencing reads were processed by the same workflow as described previously (8, 9), but with alignment (using bwa mem) to a customized reference consisting of both the primary sequences of GRCh38 and either transgene reference (Tet-On L1 or L1 GFP reporter). Haplotype-specific DNA copy-number calculation, identification of rearrangement junctions, and detection of short sequence changes (substitutions and insertion/deletion) were performed using the same workflow as described previously (8, 9).

Detection of de novo L1 insertions from shotgun sequencing data

Because the L1 transgene uses a codon-optimized L1 sequence (ORFeus) that is different from endogenous L1 sequences, insertions of L1-ORFeus can be determined directly from uniquely aligned reads. De novo L1 insertions were detected using three independent methods.

First, the junctions between inserted L1s and flanking genomic DNA were identified as part of the rearrangement junction analysis (the transgene being treated as a separate contig).

Second, insertion junctions were identified from reads aligned to the L1 transgene (either Tet-On L1 ORFeus or L1 ORFeus GFP) with split subsequences or discordant pairmates aligned to the human genome. Candidate junctions were considered when there were two or more reads mapped to the L1 transgene whose pairmates or soft-clipped subsequences were mapped to genomic loci within 1kb. Candidate insertion sites with two junctions flanking an insertion (i.e., with breakpoints on opposite sides) were then manually reviewed for the presence of poly-A/T sequences at each junction. Unpaired candidate junctions were used to identify insertion-mediated rearrangements (to be described later). For RPE-1 cells with genomically integrated Tet-On L1 ORFeus, the genomic integration sites were identified by a similar approach and further validated by long reads and long-read assembly.

Third, we used xTea (10) to identify candidate sites of L1 or pseudogene insertion from reads aligned to the human genome that had discordant and soft-clipped subsequences. A candidate L1 insertion site needed to meet two criteria: 1) at least one soft-clipped or discordant

read was mapped to the L1 transgene; 2) at least 60% of all soft-clipped reads were mapped to L1. Candidate sites passing these two filters were then assessed for either an insertion outcome, when a pair of sites were found within 200 bps and clipped on opposite sides, or a rearrangement junction with an L1 insertion, when unpaired. Both candidate sites of insertions (with paired junctions) or insertion-mediated rearrangements (single junctions) were further filtered by their proximity to endogenous L1 sequences. Candidate sites were removed if they met one of the following conditions: 1) each of two paired breakpoints was within 20bps of an endogenous L1; 2) both breakpoints were located in regions marked as “Simple_repeat”, “Low_complexity”, or “Other” in by repeatmasker; 3) either breakpoint of a candidate insertion site was located within 10bps of a region marked as “Simple_repeat”, “Low_complexity” or “Other” by repeatmasker.

In each of the three analyses, candidate insertion junctions and singleton junctions were detected from individual samples (including control samples), but the supporting reads were collected from all samples. Junctions identified in more than one sample or having supporting reads from more than one sample (indicating that they were either ancestral alterations or recurrent technical artifacts) were excluded.

Detection of de novo pseudogene insertions

Insertions of processed pseudogenes were also detected using three independent approaches.

First, insertions were identified as part of the rearrangement detection workflow from both discordant/split reads spanning exonic junctions at the source gene locus and from discordant/split reads at the insertion sites.

Second, SideRetro(11) was used to identify candidate insertion sites using the parameter ‘mc-m3-x200k’.

Third, xTea (10) was run to identify candidate sites of pseudogene insertions from reads aligned to the human genome that have discordant and soft-clipped subsequences. Candidate pseudogene insertion junctions were selected based on the following criteria: 1) there were two

adjacent breakpoints (within 200bps) consistent with an insertion; 2) neither breakpoint resided within repeats (same as above for detecting L1 insertions); 3) no other breakpoint was detected within 10bps from the candidate breakpoints; 4) the soft-clipped sequences and the discordant pairmates of supporting reads at each breakpoint must have aligned to locations within 20kb in the human genome; 5) the mapping locations of clipped parts of supporting clipped reads at two ends of a candidate insertion were within 5kb and the mapping locations of mates of discordant supporting reads at two ends of a candidate insertion were within 10kb.

Candidate sites output by SideRetro (11) and using xTea (10) with filtering were intersected to generate the list of candidate insertions, and the list of candidate insertions was merged with the list of junctions revealed from the rearrangement analysis. After excluding insertions with read support from more than one sample, the remaining insertions were manually reviewed and curated with help from long-read data and RNA-Seq data.

PacBio HiFi long-read sequencing analysis

To construct PacBio long-read sequencing libraries, high molecular weight (HMW) genomic DNA was first purified using the MagAttract HMW DNA kit (Qiagen) or PureLink Genomic DNA kit (Invitrogen). At least 4 μg HMW genomic DNA (> 50% of fragments ≥ 40 kb) was sheared to ~15 kb using the Megaruptor 3 (B06010003; Diagenode), followed by DNA repair and ligation of PacBio adapters using the SMRTbell Prep Kit 3.0 (102-141-700). Each library was subsequently size-selected for 10 kb ± 20% using the PippinHT with 0.75% agarose cassettes (Sage Science). After quantification with the Lunatic (Unchained Labs), libraries were diluted to 250 pM per single molecule, real-time (SMRT) cell, hybridized with PacBio standard sequencing primer, and bound with SMRT sequencing polymerase using the Revio polymerase kit (102-739-100).

Long-read sequencing was performed on the Revio instrument using 25M SMRT Cells (102-202-200) and Revio Sequencing Plate (102-587-400), with a 2-hour pre-extension time and 24-hour movie time per SMRT cell. Quality filtering, base calling, and adapter marking were done

automatically on the Revio instrument. Error correction for reads generated in circular consensus sequencing (CCS) mode was performed on-board the PacBio Revio with the vendor's ccs software (https://github.com/PacificBiosciences/pbccs) and with the following parameters:

--all --subread-fallback --num-threads 232

–streamed <movie_name>.consensusreadset.xml --bam <movie_name>.reads.bam. With these settings, all reads from the instrument (including those failing error correction) were presented in a single BAM file for downstream analysis. Multiplexed, barcoded libraries were demultiplexed automatically on-instrument. CCS reads for each barcode were separated into individual BAMs respectively. Reads that failed the CCS correction were separated from those that were successfully corrected.

Error-corrected (Hi-Fi) reads were aligned to the same references (GRCh38 + either L1 transgene) using minimap2 (version 2.26-r1175) with the following parameters "-a -k19 -w19 -U50,500 -g10k -A1 -B4 -O6,26 -E2,1 -s40 -Y -c." The only difference from the preset "--map-hifi" as suggested for the alignment of PacBio Hifi reads was '-s40' instead of '-s200.' The addition of '-c' and '-Y' was used for downstream processing of aligned reads.

We performed haplotype-resolved (diploid) assembly of long reads from three Dox clones (with Tet-On L1 ORFeus) and previously published Hi-C data using hifiasm (8, 9). We used minimap2 (-asm5) to align the assembled contigs to the Tet-On L1 transgene to identify contigs with genomically integrated Tet-On transgenes, and then determine the GRCh38 coordinates of the insertion sites by aligning these contigs to the human genome reference.

Manual curation of insertions and insertion-mediated rearrangement junctions

We employed the following criteria/strategies to validate/refine L1/pseudogene insertions or insertion-mediated rearrangements identified from short reads: (1) The insertion was supported by at least one long read when long-read data were available (see e.g., **Figure 3A**). (2) When long-read data were unavailable, two breakpoints need to be identified at the target site, including

one with soft-clipped poly-A/T sequences (the 3'-junction of the insertion); moreover, no copy-number changes were permitted at the breakpoints (based on 90kb-level haplotype-specific DNA copy number data as well as local sequence coverage). (3) For pseudogene insertions, reads derived from the inserted sequence (both short and long reads) were often misaligned to endogenous retrocopies of the same source gene. To resolve such ambiguity, we looked for the source gene and verified the 5' and 3' junctions from the short read data. We further validated expression of the source gene using the RNA-Seq data. (4) For L1-mediated reciprocal translocations, the features were similar to (2) except that the two junctions had different translocation partners. (5) For L1-mediated rearrangement junctions with unbalanced translocations, the breakpoint orientation (+) or (-) needed to be consistent with the directionality of copy-number change as assessed from the sequence coverage. (6) All copy-number changepoints between segments of 1Mb or higher were manually reviewed to identify additional breakpoints that were missed by the automatic rearrangement and insertion detection method.

Identification of 5'-inverted insertions was based on the orientation of the split/discordant subsequences to the L1 transgene reference and the split alignments of the insertion sequence resolved by long reads. The internal junction of a 5'-inverted insertion was either resolved by long reads or identified based on proximity to the breakpoint of the twin-primed reverse transcription (see **Fig.2C**).

Microhomology and untemplated insertions at junctions

Microhomology or untemplated insertions at any rearrangement junction (including insertion junctions) were calculated based on the lengths of aligned subsequences of soft-clipped reads. Microhomology/untemplated insertions at the 3'-end (poly-A) of retrocopied sequences were not assessed.

Calculation of insertion lengths and target site deletion/duplication sizes

For L1 insertions without 5' inversions, the insertion length was calculated from the 5' and 3'-breakpoints in the L1 reference (excluding the poly-A sequence). For 5'-inverted L1 insertions, the insertion length was calculated as the sum of two inverted insertions. For pseudogene insertions, the insertion length was calculated from the number of bases from the inserted sequence (resolved by long reads) that were aligned to the source gene locus. When long-read data were unavailable, the insertion length was calculated similarly as for L1 insertions except that the introns were excluded.

The length of target site duplication or deletion was calculated (1) from the breakpoints at the target site, or (2) from the insertion sequence resolved by long reads.

Mutation signature analysis

Single-nucleotide substitutions and short insertion/deletion changes were called by GATK as described previously (9). To filter false mutation calls, we first assessed the distribution of minor allele reads at sites on chromosome 2 with the homozygous reference genotype in the RPE-1 genome (homozygous sites were selected from common single-nucleotide polymorphic sites where half or more samples showed the same homozygous genotype). Among the homozygous sites, 95% showed minor/alternate allele fraction (VAF) less than 0.25 and minor allelic depth (ALT_AD) of three or less. We therefore considered variants with VAF and ALT_AD above these thresholds to be true variants. We then selected private mutations that only passed these thresholds in one sample and performed substitution and indel signature analyses using SigProfilerExtractor (version 1.1.23, python version 3.9.19) with default parameters.

Methods References:

1. Kelso, A. A., Lopezcolorado, F. W., Bhargava, R., & Stark, J. M. (2019). Distinct roles of RAD52 and POLQ in chromosomal break repair and replication stress response. *PLoS genetics*, *15*(8), e1008319. https://doi.org/10.1371/journal.pgen.1008319
2. Zimmermann, M., Murina, O., Reijns, M. A. M., Agathanggelou, A., Challis, R., Tarnauskaitė, Ž., Muir, M., Fluteau, A., Aregger, M., McEwan, A., Yuan, W., Clarke, M., Lambros, M. B., Paneesha, S., Moss, P., Chandrashekhar, M., Angers, S., Moffat, J., Brunton, V. G., Hart, T., … Durocher, D. (2018). CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature*, *559*(7713), 285–289. https://doi.org/10.1038/s41586-018-0291-z
3. Ardeljan, D., Steranka, J. P., Liu, C., Li, Z., Taylor, M. S., Payer, L. M., Gorbounov, M., Sarnecki, J. S., Deshpande, V., Hruban, R. H., Boeke, J. D., Fenyö, D., Wu, P. H., Smogorzewska, A., Holland, A. J., & Burns, K. H. (2020). Cell fitness screens reveal a conflict between LINE-1 retrotransposition and DNA replication. *Nature structural & molecular biology*, *27*(2), 168–178. https://doi.org/10.1038/s41594-020-0372-1
4. Tao, J., Wang, Q., Mendez-Dorantes, C., Burns, K. H., & Chiarle, R. (2022). Frequency and mechanisms of LINE-1 retrotransposon insertions at CRISPR/Cas9 sites. *Nature communications*, *13*(1), 3685. https://doi.org/10.1038/s41467-022-31322-3
5. Kopera, H. C., Larson, P. A., Moldovan, J. B., Richardson, S. R., Liu, Y., & Moran, J. V. (2016). LINE-1 Cultured Cell Retrotransposition Assay. *Methods in molecular biology (Clifton, N.J.)*, *1400*, 139–156. https://doi.org/10.1007/978-1-4939-3372-3_10
6. Feoktistova, M., Geserick, P., & Leverkus, M. (2016). Crystal Violet Assay for Determining Viability of Cultured Cells. *Cold Spring Harbor protocols*, *2016*(4), pdb.prot087379. https://doi.org/10.1101/pdb.prot087379
7. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
8. Zhang, C. Z., Spektor, A., Cornils, H., Francis, J. M., Jackson, E. K., Liu, S., Meyerson, M., & Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. *Nature*, *522*(7555), 179–184. https://doi.org/10.1038/nature14493
9. Umbreit, N. T., Zhang, C. Z., Lynch, L. D., Blaine, L. J., Cheng, A. M., Tourdot, R., Sun, L., Almubarak, H. F., Judge, K., Mitchell, T. J., Spektor, A., & Pellman, D. (2020). Mechanisms generating cancer genome complexity from a single cell division error. *Science (New York, N.Y.)*, *368*(6488), eaba0712. https://doi.org/10.1126/science.aba0712
10. Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., & Park, P. J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nature communications, 12(1), 3836. https://doi.org/10.1038/s41467-021-24041-8
11. Miller, T. L. A., Orpinelli Rego, F., Buzzo, J. L. L., & Galante, P. A. F. (2021). sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. Bioinformatics (Oxford, England), 37(3), 419–421. https://doi.org/10.1093/bioinformatics/btaa689