Chromosomal rearrangements and instability caused by the LINE-1 retrotransposon

Carlos Mendez-Dorantes[1,4,6] *, Xi Zeng[3,5,7] *, Jennifer A. Karlow[1,2,4,6] *, Phillip Schofield[1], Serafina Turner[2], Jupiter Kalinowski[1], Danielle Denisko[3,5], Eunjung Alice Lee[3,5,6†], Kathleen H. Burns[1,4,6†], Cheng-Zhong Zhang[2,4,6†]

1. Department of Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA;

2. Department of Data Science, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA:

3. Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, 02115, USA;

4. Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA;

5. Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA;

6. Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142, USA

7. Department of Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, PRC

* equal contribution, † co-senior authors.

ABSTRACT

LINE-1 (L1) retrotransposition is widespread in many cancers, especially those with a high burden of chromosomal rearrangements. However, whether and to what degree L1 activity directly impacts genome integrity is unclear. Here, we apply whole-genome sequencing to experimental models of L1 expression to comprehensively define the spectrum of genomic changes caused by L1. We provide definitive evidence that L1 expression frequently and directly causes both local and long-range chromosomal rearrangements, small and large segmental copy-number alterations, and subclonal copy-number heterogeneity due to ongoing chromosomal instability. Mechanistically, all these alterations arise from DNA double-strand breaks (DSBs) generated by L1-encoded ORF2p. The processing of ORF2p-generated DSB ends prior to their ligation can produce diverse rearrangements of the target sequences. Ligation between DSB ends generated at distal loci can generate either stable chromosomes or unstable dicentric, acentric, or ring chromosomes that undergo subsequent evolution through breakage-fusion bridge cycles or DNA fragmentation. Together, these findings suggest L1 is a potent mutagenic force capable of driving genome evolution beyond simple insertions.

Long interspersed element 1 (LINE-1, L1) retrotransposons are the only protein-coding mobile genetic elements in the human genome (*1, 2*). L1 produces bicistronic transcripts (~6kb) encoding two proteins essential for retrotransposition: open reading frame 1 protein (ORF1p), an RNA-binding protein (*3, 4*), and ORF2p, a protein with endonuclease (EN) (*5*) and reverse transcriptase (RT) activities (*6, 7*). In addition to making L1 insertions, the L1 encoded machinery mediates retrotransposition in *trans* of RNA transcripts derived from non-coding mobile elements, including *Alu* (*8*) and SVA (*9*), and mRNAs as pseudogenes (retrocopied mRNA) (*10-12*). Hence, L1 activity has generated over a third of our genome.

L1 retrotransposons are epigenetically silenced in somatic cells but are active in many human malignancies (*13, 14*). This is demonstrated by pervasive hypomethylation of the L1 promoter and expression ORF1p in malignant tissues (*15-17*). Unequivocal genetic evidence of L1 mutagenesis is provided by widespread somatically-acquired L1 insertions in pre-cancerous lesions and advanced cancers (*18-24*). Although genomic analyses have revealed that L1 insertions accrue throughout tumor evolution (*20, 25*), L1 insertions rarely directly disrupt tumor suppressor genes in cancers (*26-28*). Thus, whether L1 expression is largely epiphenomenal or directly impacts cancer inception and evolution remains elusive. Several studies have shown that L1 expression can cause DNA damage and contribute to chromosomal rearrangements (*29-33*). However, the full impact of L1 activity on genome integrity beyond the formation of canonical L1 insertions remains incompletely understood.

L1 retrotransposition occurs via target-primed reverse transcription (TPRT) (*34*): L1 ORF2p nicks genomic DNA (3′-AA/TTTT-5′) to form a primer-template structure between single-stranded DNA (ssDNA) and the poly-A sequence of the RNA template, and then reverse transcribes the RNA to create a L1 cDNA that is eventually converted to double-stranded DNA (dsDNA) as an insertion. Inserted sequences by ORF2p are enriched at EN cleavage sequences, contain poly-A sequences, and are flanked by short target site duplications (TSDs). Recently, analyses of cancer genomes have linked L1 retrotransposition to large-scale genome

rearrangements (*24*), including large segmental copy-number alterations (CNAs). Here, the connection between retrotransposition and the rearrangements is inferred by the presence of de novo L1 insertions at the breakpoint junctions, suggesting that intermediates of retrotransposition may be prone to rearrangement. However, the full scope of L1 mutagenic outcomes and their mechanistic connections to TPRT are unknown.

To address these questions, we have developed an experimental system with inducible L1 expression to investigate the molecular biology of L1-induced DNA damage and interrogate the genomic consequences of L1-mediated chromosomal instability. Using a combination of shotgun and long-read whole genome sequencing (WGS), we identified a wide range of long-range genome rearrangements that can be attributed to three mechanisms initiated by L1: (1) Reciprocal balanced translocations formed by end-joining between distal DSB ends created by ORF2p; (2) Complex rearrangements of unstable dicentric or acentric chromosomes generated by L1-mediated translocations; (3) Replication of DSB ends generated by ORF2p prior to ligation leading to duplications and foldback rearrangements. These rearrangements can lead to large (>1Mb) segmental copy-number alterations and subclonal copy-number heterogeneity. We further identified diverse sequence alterations at the insertion sites including deletions, duplications, inversions, and templated insertions. These findings, together with our analysis of insertional mutagenesis, demonstrate the recombinogenic potential of DNA ends created by ORF2p-mediated DSB, including both the DNA end extended by TPRT (leading to rearrangement junctions with L1 insertions) and the reciprocal DNA end (leading to rearrangement junctions without sequence evidence of L1 retrotransposition). Hence, our data generate new insight into the mechanism of L1 retrotransposition, significantly expand the scope of genomic consequences of L1-mediated chromosomal instability, and suggest a profound role of L1 activation in the evolution of chromosomal complexity and copy-number heterogeneity in cancer cells.

RESULTS

L1 expression causes extensive DNA damage including double-strand breaks.

Prior studies indicated p53 loss is essential for cell proliferation after L1 expression (*33*). We therefore established a Tet-On system containing a codon-optimized sequence of human L1 in p53-/-, hTERT immortalized RPE-1 cells to address both the short- and the long-term impact of L1 expression on genome integrity (Fig. 1A). We confirmed the specific expression of L1 encoded ORF1p and ORF2p by doxycycline (Dox) (Fig. 1B, Fig. S1A).

We found that L1 induction resulted in significant DNA damage. This was demonstrated by the elevation of markers of DNA double-strand breaks (DSBs) including γH2AX and downstream targets of ATM including pRAD50 (S635) and pKAP1 (S824), but not ATR targets such as pCHK1(S345) and pRPA (S4/S8) or pRPA (S33) (Fig. 1B, Fig. S1B). Transcriptome analyses of cells upon L1 induction revealed an enrichment of differential gene expression related to DNA damage response and repair (Fig. 1C, Fig. S1C, Table S1). Generation of DSBs was further supported by the increase of γH2AX and 53BP1 nuclear foci (Fig. 1D, Fig. S1D-E) indicating nascent DSBs, and the formation of micronuclei (Fig. 1E) indicating acentric lagging chromosomes from unrepaired DSBs. Finally, using catalytically dead L1 mutants lacking endonuclease (EN) or reverse transcriptase (RT) activity, we found that both ORF2p EN and RT, which are required for proficient retrotransposition (*35, 36*), contribute to L1-mediated DNA damage, with ORF2p EN being the dominant contributor (Fig. S1F-G). Together, these results demonstrate that L1 expression causes DNA damage including DSBs.

To determine whether L1-mediated DNA breakage causes chromosomal instability, we next performed low-pass shotgun WGS (0.1x median coverage) on both single cells and their progeny clones after L1 expression to assess the burden of large segmental copy-number alterations (CNAs). We calculated haplotype-specific DNA copy number using the parental haplotype of RPE-1 cells as described previously (*37, 38*) (**Methods**). We observed a significant increase of large segmental CNAs (5Mb or above) in both single cells harvested directly after L1

induction and their progeny clones in comparison to control cells or clones (Fig. 1F-G, Fig. S2). Notably, a higher CNA burden was observed in L1-exposed single cells in comparison to progeny clones. This likely reflects negative selection of cells with a high burden of DNA damage that is supported by reduced clonogenicity of cells after L1 induction (Fig. 1F, Fig. S1A, Fig. S1F).

As an alternative approach to induce L1 expression, we used a L1 GFP reporter assay for retrotransposition (*39*) in parental p53-/-, hTERT immortalized RPE-1 cells. In this assay, the GFP cassette can only be expressed after it has undergone retrotransposition in cells (Fig. S3A-C), and hence, GFP marks a subpopulation of cells with de novo L1 insertions. We sorted GFP (+) or GFP (-) cells to generate single cell progeny clones. Low-pass shotgun WGS of these clones revealed a higher CNA burden in the GFP (+) clones than the GFP (-) clones (Fig. S3D-E), confirming the observations in the Tet-On L1 system.

Together, these results demonstrate that L1 expression causes DNA DSBs and leads to the acquisition and accumulation of large segmental CNAs. We next sought to comprehensively characterize genomic changes induced by L1 expression to investigate the mechanistic connections between L1 retrotransposition, L1-dependent DNA breakage, and L1-mediated rearrangements.

Identification of L1-induced genomic alterations by whole-genome sequencing

To determine the full spectrum of L1-induced genomic alterations, we performed 20x shotgun WGS on 75 progeny clones derived from single cells with (31 with Tet-On L1, 29 with L1 GFP reporter) and without L1 expression (10 with Tet-On L1, 5 with L1 GFP reporter). To determine the complete sequence spanning both insertions and their junctions, we generated PacBio HiFi sequencing (median 15x) on 38 L1 clones and one control clone (**Methods**). We also performed 30x shotgun WGS on 28 cells with L1 induction and 12 control cells (both from the Tet-On system) to detect genomic alterations that could be lost during clonal expansion due to negative selection (**Methods**). We performed comprehensive mutation discovery (single-nucleotide substitutions,

short insertions/deletions, chromosomal rearrangements, DNA copy-number alterations) from shotgun sequencing data using bioinformatic workflows described previously (*38, 40*). L1-mediated retrotranspositions were detected using xTea (for L1, other mobile DNA insertions and processed pseudogenes) (*40*) and sideRETRO (for processed pseudogenes) (*41*). Both types of insertions were also independently identified as part of DNA rearrangement discovery (**Methods**). We then used long read data to verify and determine the inserted sequence (L1 or pseudogene insertions) and to identify DNA breakpoints at copy-number transitions that were not resolved by short reads (**Methods**). Finally, we also performed de novo assembly of the long reads and determined the precise locations of eleven integrated copies of the Tet-On L1 construct (Fig. S4).

Landscape of canonical L1-mediated insertions

We first analyzed insertional mutagenesis caused by L1 expression. L1 ORF2p can generate insertions of L1 cDNA and other reverse transcribed sequences (in *trans*) including non-coding retrotransposons (*Alu* and SVA) (*8, 42*) and endogenous mRNAs (*10, 11*). The sequence structures of ORF2p-mediated insertions can be full-length, 5′ truncated, or 5′ inverted (*43, 44*). We detected de novo ORF2-mediated insertions only in single cells harvested after L1 expression and their progeny clones (Fig. 2A, Fig. S5A-C, Table S2), but not in control cells or clones, thus validating the specificity of the L1 expression experimental systems. Among 31 clones induced with L1 expression using the Tet-On system, 28 acquired one or more L1 insertions and 30 acquired one or more insertions of processed pseudogenes (Fig. 2A). In contrast to L1 insertions that were mostly 5′ truncated (76/99), many pseudogene insertions were full-length (144/235) (Fig. 2A), and the median length of pseudogene insertions (2559 bp) was also longer than L1 insertions (712 bp) (Fig. 2B). A subset of L1 (13/99) and pseudogene (20/235) insertions exhibited 5′ inversions (Fig. 2A, Fig. 2C). Finally, all L1 and pseudogene insertions showed enrichment of the integration target sequence (5′-TT/AAAA-3′) (Fig. 2D, Fig. S5D) consistent with the preferred cleavage motif of ORF2p EN during TPRT (*45-47*).

The number of pseudogene insertions was modestly correlated to L1 insertions (Fig. S5E). The source or parental genes of inserted pseudogenes were significantly enriched for highly expressed genes (Fig. 2E), and a subset of inserted pseudogenes contain *Alu* sequences in the 3′ UTR (Fig. 2F), implicating a potential role of *Alu* sequences in mediating ORF2p-mRNA interactions (*48*). We further identified four instances of pseudogenes derived from mRNA transcripts with endogenous L1 sequences at or near the 3′-ends, suggesting a potential bias for ORF2p for mRNAs with 3′ L1 elements (Fig. S6). While the high frequency of pseudogene insertions may be attributed to sequence changes of the L1 RNA or L1 overexpression inherent to our experimental system, there may also be more tendency for ORF2p to act on mRNAs in *trans* than previously appreciated (*49, 50*). The generation of many de novo, full-length pseudogene insertions allowed us to analyze and compare the sequence features of full-length insertions to 5′ truncations and 5′ inversions.

Duplication of the target site (target site duplication, or TSD) flanking the inserted sequences is a signature of ORF2p-mediated retrotransposition. We observed the most discrete 12-18bp TSDs flanking full-length pseudogene insertions (Fig. 2G, top) and 5′-inverted insertions (Fig. 2G, bottom). These observations are consistent with ORF2p generating staggered nicks on the genomic DNA, resulting in two sticky DSB ends with 3′ overhangs. For full-length insertions, the T-rich DNA end (primary RT end) is extended with a fully reverse-transcribed cDNA and then ligated to the reciprocal end containing a short 3′ overhang with little or no alteration, thus generating relatively consistently sized TSDs. The ligation is most likely generated by canonical non-homologous end joining (c-NHEJ), which results in junctions without microhomology or insertions/deletions (indels) (*51-53*). Consistent with this model, most 5′ junctions of full-length insertions typically contained an extra G base reflecting 5′-mRNA capping and little or no microhomology (Fig. 2H, top). By contrast, 5′ truncated insertions were associated with a broader range of TSDs, including short TSDs (<10bp) and short target site deletions (0-20 bp) (Fig. 2G, middle). This could result from processing or removal of the 3′ flap of the reciprocal DNA end

during joining with the DNA end extended by RT via microhomology-mediated end-joining (MMEJ), which results in deletions with microhomology (*52-54*). This interpretation was supported by the presence of microhomology (≥2bp) at the 5′ junctions of 5′ truncated insertions (Fig. 2H).

For 5′-inverted insertions, we observed both discrete 12-18bp TSDs and microhomology at the 5′ junctions (Fig. 2G, bottom, Fig. 2H, bottom). These observations are consistent with the twin priming model (*43*), which proposes a second RT extending the reciprocal end by the genomic DNA 3′ OH priming internally to the RNA template. The distinct TSD feature suggests that the secondary RT reaction is primed on the reciprocal end before any DSB end processing, therefore limiting genomic DNA loss and preserving the TSD. The observation of microhomology (Fig. 2H, 3rd row) at the 5′ junction (71% with ≥2bp MH) suggest that the secondary RT was primed using microhomology-mediated annealing. In addition, the inversion junctions between inverted RT sequences (Fig. 2H, 4th row) and the 5′ junctions of 5′-truncated insertions (Fig. 2H, 2nd row) showed similar distributions of microhomology, indicating similar end-joining processes (c-NHEJ or MMEJ) resolve these insertions. Interestingly, the internal breakpoints within the RNA template of inverted RT sequences were often in close proximity containing short (1-10bp) deletions (22/32) or duplications (10/32) (Fig. 2C), suggesting the secondary RT is primed adjacent to the end of the primary RT. Together, these observations suggest c-NHEJ completes full-length insertions, microhomology-mediated annealing mediates twin priming, and c-NHEJ or MMEJ resolve 5′-truncated and 5′-inverted insertions.

Non-canonical L1-mediated insertions revealed by long reads

PacBio long-read sequencing enabled us to resolve the complete sequences of insertions that included several non-canonical patterns. In one special example, we identified an insertion containing three adjacent, but non-overlapping subsequences derived from the 3′ UTR of the *GREM* mRNA (Fig. 3A). The distal 5′ and 3′ subsequences (red and green lines) were in inverted orientation that is consistent with the twin priming model that produces 5′-inverted insertions. The

orientation of the reverse transcribed sequences implied that the middle subsequence (blue) was most likely generated from a template switch of the secondary RT (green) before end-joining/ligation to the primary RT (red).

We further identified nine insertions containing sequences derived from two RNA transcripts (Fig. 3B and Fig.S7). We can classify these insertions based on the relative orientation of the inserted sequences (represented by arrows reflecting the 5′ to 3′ direction of RT). When the inserted sequences have the same orientation, they may arise from one continuous RT with template-switching between different mRNAs (*6*) or from RNA ligation preceding RT as was previously described for U6-3′ L1 chimeric insertions (*55*). When the two inserted sequences are in opposite orientations, they are likely generated by ligation between opposite DNA ends that have been extended by RT from different templates. The presence of microhomology or untemplated insertions at these two types of junctions (Fig. 3C) is consistent with these models. Lastly, we detected one insertion containing two inverted subsequences derived from the *SEMA3C* mRNA that contain two separate RT sequences at their internal junction (Fig. S7), suggesting that DNA ends extended by RT can also incorporate additional cDNA products before their ligation.

In addition to these template rearrangements, we identified three instances of pseudogene insertions (3/235) containing clustered T>C substitutions (Fig. S8). These substitutions were restricted to inverted *Alu* repeats found in the 3′ UTR of the inserted mRNA sequences, consistent with ADAR editing (A>I) of double-stranded RNA mediated by inverted *Alu* repeats (*56, 57*). Notably, ADAR1 is a known molecular dependency when L1 expression is induced in this system (*33*).

Finally, we detected three insertions accompanied by inversions of genomic DNA sequences at the reciprocal DNA end (Fig. S9) and three insertions containing additional sequences templated from genomic DNA near the integration site (Fig. S10). We attributed these alterations to various types of DSB end processing of the reciprocal DNA ends prior to their

ligation to DNA ends extended by TPRT. See Fig. S9 and Fig. S10 and their captions for a more detailed explanation.

Taken together, the observations of non-canonical insertion outcomes highlight the diversity of sequence alterations that can occur during TPRT before DNA end ligation that produces a stable insertion outcome.

End-joining between L1-induced DSBs creates reciprocal translocations

The diverse insertional outcomes of L1 retrotransposition suggest that ORF2p-generated DSB ends can be recombinogenic before their ligation to form stable dsDNA. Moreover, our observation of many γH2AX foci in cells after L1 induction (Fig. 1D) suggested that L1 can generate multiple DSBs. Taken together, we predicted that end-joining between distal DSB ends generated by ORF2p can generate chromosomal translocations (Fig. 4A). For instance, when ORF2p creates two DSBs on different chromosomes, a reciprocal exchange between the DSB ends can lead to two possible outcomes of reciprocal translocations. In the first scenario (middle), the translocations produce two stable derivative chromosomes. In the second scenario (right), the translocations produce two unstable chromosomes, one dicentric chromosome and one acentric chromosome. In either case, the inserted sequences by TPRT may be retained either in reciprocal translocation junctions or in a single translocation junction; in the second scenario, the reciprocal junction will show no inserted sequence despite being a direct outcome of ORF2p mediated DSB.

Consistent with these predictions, we observed both stable and unstable derivative chromosomes generated from L1-mediated translocations. In the first example shown in Fig. 4B, we inferred two stable derivative chromosomes with no DNA copy-number loss or gain were formed by reciprocal translocations between two pairs of DSB ends generated by ORF2p. The origin of chromosomal breakage at both loci by L1 retrotransposition is supported by (1) the presence of polyadenylated sequences at one DNA end from each locus at ORF2p EN cleavage sequences (Tc/AAAA on chr16 and TaT/AAAA on chr17); and (2) the duplication of “target sites”

(16bps duplicated from chr16 and 14 bps duplicated from chr17) between the primary RT ends and the reciprocal ends preserved in two derivative chromosomes. Therefore, all four DNA break ends displayed the hallmark features of TPRT except that their illegitimate ligation formed translocations instead of insertions (Fig. 4B). Notably, if the derivative chromosome der(16) were lost during clonal expansion, the insertion footprints of TPRT would have been lost, leaving only indirect evidence relating the translocation to retrotransposition from the presence of an ORF2p EN target sequence *adjacent* to the breakpoints within the range of TSDs (Fig. 4B, bottom). The observation of translocations between reciprocal DNA break ends demonstrates that L1-mediated translocations may not contain direct evidence of TPRT but can be recognized indirectly based on ORF2p EN motifs adjacent to the breakpoints.

In another example shown in Fig. S11A, we identified four-way translocations. Among all six breakpoints, we inferred that two breakpoints on chr18 were generated by ORF2p since these ends showed evidence of a twin-primed 5′ inversion of *OSBPL8* mRNA that failed to cause an insertion and instead caused translocations. Therefore, the intermediates of 5′ inversions can source translocations in addition to forming insertions. We further inferred that two breakpoints on chr11 were likely generated by ORF2p based on the ORF2p EN motif near one breakpoint [76199662(-)]; the absence of RT insertion could be explained by cleavage of the ssDNA with RT extension (Fig S10A). Together, these observations highlight the capacity of ORF2p-generated DSB ends to form translocations in addition to insertion outcomes.

In addition to stable chromosomes, we observed examples of dicentric and acentric chromosomes generated by L1-mediated translocations. In Fig. 4C, we inferred a dicentric chromosome from an L1-mediated translocation between chr5 and chr18. The origin of both translocation breakpoints from L1 was established by TPRT features at both breakpoints. The inference of a dicentric derivative chromosome was based on (1) absence of additional breakpoints on the translocated chromosome; and (2) presence of subclonal copy-number losses of DNA between the two centromeres, and minor subclonal losses of the 18p arm (highlighted in

blue), both consistent with chromosome-type breakage-fusion-bridge (BFB) cycles following the formation of a dicentric chromosome (*38, 58, 59*). A similar example is shown in Fig. S11B where a dicentric chromosome was inferred to arise from translocations between three chromosomes, with all breakpoints displaying either insertions from TPRT signatures or having adjacent ORF2p EN cutting sequences.

Finally, the reciprocal translocation model (Fig. 4A) also predicts the formation of an acentric chromosome from the terminal segments in addition to the dicentric chromosome as shown in Fig.4C. This prediction was supported by the observation that L1 expression led to more frequent formation of micronuclei (Fig.1E) containing acentric chromosomes. Because acentric chromosomes are mitotically unstable and typically lost during clonal expansion unless reintegrated into centric chromosomes, we expect them to be lost in the progeny clones. However, we previously reported that acentric chromosomes in micronuclei can acquire massive DNA damage causing complex rearrangements that can be detected in single cells (*60*). Accordingly, we observed one such example in a single cell exposed to L1 expression (shown in Fig. 4D). The presence of clustered DNA rearrangements within and between terminal fragments from both chr7 and chr15 indicates chromothripsis on an acentric chromosome trapped in a micronucleus. The origin of this acentric chromosome from L1-mediated translocation was established based on the presence of (1) an inserted polyadenylated sequence at an ORF2p EN cut site at the chr15 breakpoint; (2) the presence of plausible ORF2p EN cutting sequences near the chr7 breakpoint.

In summary, we provide definitive genomic evidence for L1-mediated chromosome translocations that can lead to both stable and unstable derivative chromosomes. The involvement of retrotransposition in these translocations is established by three sequence features attributed to TPRT: (1) insertions of polyadenylated sequences; (2) presence of ORF2p EN cutting sequences at or near the breakpoints; and (3) short TSDs between breakpoints from different translocation junctions that are inferred to be the DNA ends extended by TPRT and the reciprocal DNA ends generated by ORF2p. Notably, the translocation junctions between

reciprocal DNA ends only preserve the second feature (proximity to ORF2p EN cutting motifs). Therefore, definitive TPRT signatures (insertions and TSDs) may be present in only a fraction of all chromosomal rearrangements that are directly incited by retrotransposition.

Segmental copy-number alterations from L1-mediated chromosomal rearrangements

L1 retrotransposition has previously been linked to small genomic deletions (1-10kb) at the insertion site (*30, 32*). In addition to small deletions, we also identified small tandem duplications (Fig. 5A,B) and inversions (Fig. S9). These small-scale alterations at the insertion sites could result from DSB end resection or processing of the reciprocal DNA ends prior to their ligation to the DNA ends extended by TPRT.

We further observed large (1Mb or above) internal or terminal segmental deletions and duplications with one or both breakpoints displaying features of L1 TPRT (Fig. 5D-G). One interesting example was a Chr12 ring chromosome (Fig. 5F) inferred to have been generated by two DSB ends on the p- and q-arms joined together by an ORF2p-mediated insertion. The large distance between distal breakpoints of large segmental CNAs suggests that these breakpoints originate from DSB ends created by two independent events, one or both incited by ORF2p.

Our findings of L1-mediated translocations without inserted polyadenylated sequences suggest that not all rearrangement junctions or breakpoints display evident TPRT features. For example, we observed three instances of dicentric chromosomes with unbalanced breakpoints that are adjacent to near-perfect ORF2p EN cutting sequences (Fig. S12). The processing of DSB ends as shown in Fig. S9 and S10 could also explain translocation breakpoints without adjacent ORF2p EN cutting sequences (Fig. S13). Consistent with these observations, we observed large CNAs with breakpoints that sometimes do not display evident features of TPRT (Fig. S14). We further observed many terminal deletions accompanied by subclonal segmental losses between the deletion breakpoint and the centromere (“sloping copy-number variation”) and/or subclonal loss of the broken chromosome (Fig. S15); these patterns are consistent with BFB cycles of

dicentric chromosomes as discussed in Fig. 4C. Importantly, large segmental CNAs or subclonal copy-number heterogeneity were rarely present in control clones and appeared at a much lower frequency in control cells (Fig. S2, S3, Table S3). Therefore, we expect many large CNAs to arise either directly from L1-mediated DNA breakage or from the downstream evolution of unstable chromosomes.

In summary, the prevalence of long-range DNA rearrangements, large segmental CNAs, and copy-number heterogeneity in progeny clones from cells exposed to L1 indicates that L1 mutagenesis both directly generates stable rearrangements and CNAs and drives subsequent copy-number evolution by generating unstable chromosomes.

Foldback rearrangements arise from replication of L1-induced DNA ends prior to ligation

In this and the next section, we present genomic findings that highlight the interplay between L1 retrotransposition and other mechanisms of chromosomal instability, including DNA replication and chromothripsis. An interesting observation from the PCAWG of L1 retrotransposition was the detection of L1 insertions at foldback junctions (*24*). The breakpoints of foldback junctions were thought to originate from independently generated DSB ends on sister chromatids. In a recent study, we suggested that a single ancestral DSB end can be converted to two DNA ends by DNA replication, which can then fuse together to form a foldback junction (*61*). Based on this model, we suggest two processes that can generate L1-mediated foldback rearrangements (Fig. 6A). In the first model (top), a DSB end is either directly generated by ORF2p EN or bound by ORF2p and extended via reverse transcription (i). In either scenario, a hairpin can be formed when the 3′-end of the DSB is ligated to the 5′-end on the opposite strand following RNA- or cDNA-mediated self-annealing. Replication then converts the hairpin into a foldback junction containing reverse transcribed sequences (iii). In the second model (bottom), a DSB end is first converted to a pair of DNA ends on sister chromatids by DNA replication (i), the sister DNA ends are then tethered by reverse transcribed cDNA (ii), which is then converted to foldback junctions containing

cDNA/L1 insertions after 'fill-in' synthesis of the second cDNA strand and ligation (iii). In both models, the two breakpoints of a foldback junction arise from replication of a single DSB end (*62*), but do not require independent breaks on the sister chromatids or a preceding BFB cycle.

The first model was supported by two examples of foldback junctions on chr5 in a Dox clone (Fig. 6B). The 5A homolog (copy number shown in red) had a triplication flanked by two foldback junctions. The telomeric foldback was formed between two adjacent breakpoints (147445104 and 147438833) with a full-length insertion (7704 bps) of the *CHML* mRNA (Fig. 6B right). The distal breakpoint (chr5:147445104) was located at an ORF2p EN cutting sequence (TgAAAgAA) (Fig. 6B, right), indicating its origin from the ancestral DSB end that was extended by TPRT. The proximal breakpoint of this junction (chr5:147438833) could have been generated from hyper-resection of the 5′ end of the same ancestral DSB end (6271 bp resection), allowing self-annealing of the 3′ cDNA end to form a hairpin (see Figure Caption for more details). On the 5B homolog (copy number shown blue), we found a separate foldback junction at the boundary of terminal deletion. Although there was no insertion at the junction, both breakpoints had adjacent ORF2p EN cutting motifs, suggesting ORF2p activity as the plausible origin of these breaks.

In support of the second model (Fig. 6A), we identified a foldback junction containing a chimeric insertion containing two RT sequences in inverted orientations (Fig. 6C): one from the *PTMA* gene and one from L1. The tail-to-tail orientations of the inserted sequences indicated an annealing between the poly-A sequence of the L1 mRNA and a TTT sequence in the UTR of the *PTMA* mRNA; presumably, this RNA duplex (or its cDNA sequences) tethered two adjacent DNA ends on sister chromatids to create the foldback junction. These observations are consistent with previous findings of L1 ORF2p RT exploiting endogenous DNA breaks such as unprotected telomeres to generate chromosome fusions (*63*).

In addition to foldbacks with insertions, we identified many examples of foldback junctions without insertions but with breakpoints adjacent to ORF2p EN cutting sequences (Fig. S16). By contrast, no foldback was detected in control clones (Table S3). We expect that breakpoints that

are adjacent to ORF2p EN cutting sequences may still be related to ORF2p-dependent DNA breakage. This hypothesis was supported by the example shown in Fig. 6D. Here, we identified a foldback junction with adjacent breakpoints on chr1q (chr1:208242622 and chr1:208240872), neither of which has an adjacent ORF2p EN cutting sequence. However, we further identified two pieces of DNA sequences (chr1:208297605-639 and chr1:208297658-915) that were inserted into a complex insertion junction in chr10 (Fig. S10A). Notably, these two short pieces displayed signatures of TPRT including a polyadenylated sequence derived from the *SH3BP4* mRNA. Based on this observation, we inferred that the DNA insertions were derived from cleaved ssDNA flaps during retrotransposition (*64*). Based on the proximity between these short insertions and the foldback breakpoints (50kb), we suggest that the foldback junction resulted from a DSB end that was originally generated by TPRT but acquired secondary deletions. This observation provides a plausible mechanism for DNA rearrangements including foldbacks in L1-exposed cells that do not display sequence hallmarks of TPRT.

In summary, we identified examples suggesting different mechanisms that can generate foldback rearrangements involving L1 retrotransposition and DNA replication.

L1 retrotransposition can cause chromothripsis or tether chromosome fragments

Prior studies from us and others have shown that chromosomes partitioned into abnormal nuclear structures including micronuclei (*60, 65*) and bridges (*38, 66*) can undergo DNA fragmentation and clustered DNA rearrangements, which are signatures of chromothripsis (Fig. 7A). Based on these results, we predict that unstable dicentric and acentric chromosomes generated by L1-mediated translocations can also acquire chromothripsis. One such example of chromothripsis on an acentric chromosome was described in Fig. 4D.

We identified several additional examples of chromothripsis with features of chromosome bridge breakage. In the example shown in Fig. 7B, the 3B homolog displayed p-terminal deletion with an adjacent terminal duplication and multiple clustered rearrangement junctions within a

region of subclonal copy number loss on the p-arm. These features are consistent with a broken bridge chromosome (*38, 66*). The presence of an L1 insertion at the deletion boundary suggests that the bridge chromosome was generated by an L1-mediated translocation. A similar pattern of terminal deletion, subclonal DNA loss, and complex rearrangement was observed in two more examples (Fig. S17) that indicate chromothripsis from bridge resolution. In both examples, we could not determine the junction of the deletion breakpoint but identified multiple junctions near the deletion boundary containing both genomic DNA breakpoints and L1 inserted sequences, which suggest ORF2p may be tethering DNA break ends via reverse transcription. Interestingly, we identified two short genomic DNA sequences originating from the terminal deletion on chr4 (Fig. S17B) in a complex insertion junction on chr1 (Fig. S10C). These observations highlight the multiple roles of ORF2p and TPRT in promoting chromosomal instability and rearrangement.

Finally, we identified one instance of chromothripsis where L1 insertions were detected at three breakpoint junctions (Fig. 7C). Although the retained fragments showed different copy-number states, we inferred that most, if not all the rearrangements were generated all-at-once based on the proximity between breakpoints. We further identified two L1 junctions near the 6q terminus that were likely in *cis*, connecting two 6q-ter fragments (138.15Mb-qter. and 155.91Mb-qter.) This example suggests that L1-mediated retrotransposition can occur to both sister chromatids at the same time and generate rearrangements involving both sister chromatids.

In summary, our data suggest that L1 retrotransposition can create unstable chromosomes that subsequently acquire complex rearrangements including chromothripsis, tether DNA ends from broken chromosomes, and capture DNA fragments at insertion junctions. Thus, L1 retrotransposition can not only directly incite but also compound chromosomal instability.

CONCLUSIONS

Since the discovery of DNA transposition by McClintock (*59, 67*), transposable element activity has become a well-recognized source of heritable genetic variation (*68*) and somatic mosaicism (*69*) across species. Although canonical transposition results in sequence insertions, bursts of transposition are associated with karyotypic changes in speciation and in malignancy (*24, 70*). In cancer genomes, the association between L1 retrotransposition and chromosomal instability is suggested both by the positive correlation between somatic insertions and rearrangements and by the presence of L1 sequence insertions at rearrangement junctions (*24*). Although cancer genome analyses provide valuable snapshots of these outcomes, they offer limited ability to discern direct and indirect contributions of L1 retrotransposition to genome rearrangements and chromosomal instability.

In this study, we provide, to our knowledge, the first comprehensive characterization of genomic alterations caused by L1 retrotransposition by shotgun and long-read whole-genome sequencing analyses of experimental models of L1 expression. We find that even a short period of L1 expression can produce a wide range of chromosomal rearrangements. By contrast, we observed no appreciable enrichment of single nucleotide substitutions or short insertion/deletion changes after such exposure (Fig. S18). Our whole-genome sequencing analyses identify several categories of insertion and rearrangement outcomes that significantly expand the repertoire of L1-mediated genomic alterations beyond insertional mutagenesis (*30, 32, 46, 47*). These newly identified genomic consequences further provide new insight into the molecular processes of retrotransposition (Fig 8).

Complex insertions from local rearrangements of ORF2p-generated dsDNA ends

Besides full-length, 5′ truncated, or 5′-inverted insertions, we find many insertions consisting of two or more sequences derived from different RNA transcripts in addition to local sequence changes at the insertion site. These outcomes display sequence features of canonical TPRT (e.g.,

inserted sequences containing poly-A tails, found at an ORF2p EN cleavage sequence, and flanked by TSDs), suggesting mechanistic overlap with canonical retrotransposition. Based on the 5′-to-3′ orientation of reverse transcribed sequences, we infer that such complex insertions may arise from either template-switching RT (when concatenated RT sequences show the same orientation) or an annealing/end-joining between two distinct RT products (when the RT sequences show opposite orientations). We infer local rearrangements at the target site, including deletions, inversions, and inverted duplications, are often caused by processing of ORF2p-mediated DSB ends (Fig. 8A). These complex events highlight that chromosomal breakage is a step in retrotransposition or a commonly-occurring risk of retrotransposition, consistent with our measurements and previous reports of markers of DNA damage in cells induced with L1 expression (*29, 33*).

Translocations from illegitimate recombination between distal dsDNA ends generated by ORF2p

We have uncovered a variety of long-range chromosomal rearrangements arising from ‘erroneous’ repair of DSB ends caused by ORF2p. The first class of rearrangements encompass reciprocal translocations between distal DSB break ends generated by L1 retrotransposition, resulting in balanced, structurally stable translocations. We found that these L1-mediated reciprocal translocations display a mix of sequence features of TPRT. DNA ends extended by TPRT have polyadenylated sequences which may become incorporated at the breakpoint junctions, thus preserving unequivocal evidence of L1 activity as the cause of the DSB. In contrast, the reciprocal genomic DNA ends with short 3′ overhangs of ORF2p-mediated DSBs recombine without inserted sequences at the rearrangement junctions. For these events, if the 3′ overhangs are preserved, the breakpoints of these translocations will be proximal to an EN cut motif sequence within the range of TSDs, revealing a cryptic signature of L1-mediated translocations. Our findings indicate that the reciprocal DNA ends of ORF2p-mediated DSBs are

equally or more commonly substrates for translocations than DNA ends with extended cDNA flaps by ORF2p RT. We infer that these genomic DNA ends with a short 3′ overhangs are readily ligatable and prone to chromosomal rearrangements. Together, all these rearrangement events demonstrate the recombinogenic potential of both DNA ends cause by ORF2p-mediated DSBs.

The second class of rearrangements are foldbacks created by L1-mediated fusions between sister DNA ends. Foldback junctions are typically attributed to such fusions at a broken chromosome after chromosome bridge resolution (*66*). Here, we suggest that foldbacks can arise directly from the replication/fusion of L1-induced DSBs, where ORF2p can cause DNA breakage and DNA end tethering via reverse transcription. Together, L1-mediated translocations and L1-mediated foldbacks are two direct rearrangement outcomes of retrotransposition when two DNA ends generated by ORF2p fail to join each other to form an insertion.

Chromosomal instability from L1-mediated translocations

In addition to translocations and foldback junctions from ORF2p-generated DSB ends, we identified a significant number of long-range rearrangements and segmental copy-number alterations without conspicuous features of TPRT at the breakpoints. We demonstrate that many of these alterations were acquired during the downstream evolution of unstable chromosomes (both dicentric and acentric) initially generated by L1-mediated translocations. These unstable chromosomes can undergo BFB cycles or chromosome fragmentation to acquire complex rearrangements. These observations illustrate how L1 can precipitate chromosomal instability and drive acquisition of genome complexity and heterogeneity. It is noteworthy that McClintock's original discovery of DNA transposition was based on similar outcomes of unligated DNA ends generated by the self-excision of DNA transposons, including large terminal deletions (Fig. 5C) and dicentric chromosomes (Fig. 4C) that undergo BFB cycles. It is tempting to hypothesize that transposable elements drive genome evolution in part by their capacity to generate DNA breaks that can result in large-scale genome rearrangements.

Mechanistic implications for retrotransposition and retrotransposition-mediated rearrangements

Our findings both corroborate key steps of TPRT and provide new insights. To generate an insertion outcome, the two DSB ends generated by ORF2p must be ligated together. For full-length insertions, the discrete TSDs (12-18bp) and the lack of microhomology at their 5′ junctions support a model wherein the reciprocal DNA end is protected from processing or illegitimate recombination before its ligation with the DNA end extended by TPRT via c-NHEJ.

The sequence features of 5′ inverted insertions, including an example where two inverted sequences from an RNA template were present at two distal translocation junctions (Fig. S11A), support the model of twin priming since the 5′ junctions are enriched for microhomology and suggest an end-joining step resolving the inverted sequences since inversion junctions show features of c-NHEJ or MMEJ repair. The discrete TSD feature in 5′ inverted insertions also suggests that the reciprocal DNA end is protected by twin-priming (secondary RT) from end processing after or concurrent with the second-strand nick.

Finally, the broader range of TSDs and the prevalence of target site deletions/rearrangements at 5′-truncated insertions in comparison to full-length or 5′-inverted insertions suggests that the reciprocal DNA ends can undergo various forms of DSB end processing. The sequence features of 5′ truncations also support c-NHEJ or MMEJ involved in resolving the insertions. How and when the reciprocal end is protected, the order of events including second-strand nicking, RNA degradation, second-strand cDNA synthesis, and ligation, and whether the order of these events differ between full-length and other insertion outcomes require further investigation.

The factors involved in promoting the resolution of ORF2p-mediated DSB ends as either insertions or translocations remain unknown. The junctions between distal ORF2p-mediated DSB ends show similar sequence as the junctions in insertion outcomes, suggesting that they are

generated by similar end-joining processes mediating translocations detected in cancer genomes (*71-73*). In addition to translocation junctions consistent with end joining between distal ORF2p-generated DSB ends, we have also identified rearrangement junctions consistent with EN-independent RT events (i.e., reverse transcription at pre-existing DNA ends) (*36*), suggesting that ORF2p can bridge DNA break ends via reverse transcription. Finally, consistent with recent biochemical studies showing that ORF2p can DNA ends using a DNA template (*6*), we observed multiple examples of insertions containing templated genomic DNA insertions. Future studies are required to clarify the contributions of ORF2p and host factors in the formation of these complex insertions and rearrangements.

The impact of L1 retrotransposition on cancer evolution

Our findings have several implications for L1 retrotransposition and chromosomal rearrangements in cancer evolution. First, we detect L1-mediated translocations and foldback rearrangements with breakpoint junctions containing TPRT genomic signatures, including reverse transcribed sequences, like those previously described in cancer genomes (*24*). Second, we observe long-range rearrangements formed between reciprocal DNA ends upstream of the extended TPRT DNA ends from distant loci (Fig. 8), leading to junctions that do not contain inserted transposon sequence. Here, the mechanistic connection to L1 may be recognized by the presence of EN cutting motifs proximal to the breakpoints. This discovery predicts a new sequence feature that can be used to detect L1-induced translocations in cancer genomes. This also indicates that the contribution of L1 to genome instability may be underestimated in our current cancer genome analyses. Third, the identification of complex rearrangements and copy-number heterogeneity from the downstream evolution of unstable chromosomes generated by L1-mediated translocations highlights the compounding potential of L1 mutagenesis with other mechanisms of chromosomal instability such as the formation of chromosome bridges or micronuclei. This suggests that L1 expression in cancer (*15*) may directly promote tumor

development by creating genomic diversity and complexity, providing a plausible explanation of the positive correlation between L1 insertions and chromosome rearrangements in p53-mutated cancers (*24, 74*).

Our results also raise a question about the dynamics of L1 mutagenesis during cancer evolution. It is striking that L1-mediated rearrangements involving co-existing retrotransposition lesions on different chromosomes are observed in both our experimental system and in cancer genomes (*75*). This implies either that multiple retrotransposition intermediates are co-localized in nuclear space consistent with mechanisms of clustering of DSBs (*76, 77*), or that large numbers of intermediates co-occur in a single cell cycle, which are either removed or resolved into canonical insertions or genome rearrangements. For cancer genomes with hundreds of somatically acquired L1 copies, are these insertions accumulated over a period spanning many generations, or in episodic bursts of L1 activity? What host defense mechanisms are breached to cause this? Do numerous L1 insertions occur simultaneously with a myriad of de novo translocations? Addressing these questions will require multisite sampling and/or longitudinal analyses exploiting long-read sequencing and single-cell analysis of primary tumor samples and experimental models of L1 retrotransposition.

In summary, this work underscores that L1 activity does not only result in insertion mutagenesis, which infrequently drives tumorigenesis by mutating tumor suppressor genes (*26-28*). Rather, L1 frequently introduces DSBs, promotes structural chromosomal rearrangements and incites genome instability that evolves in progeny clones. Our findings indicate diverse outcomes of L1 expression at a genomic level and provide the most comprehensive picture to date of L1 mutagenesis. They also predict a coupling between retroelement dysregulation and L1 overexpression in cancerous precursors with chromosomal complexity and genome instability in these lesions (*78-80*).

Data Availability

DNA sequencing data are available from the Sequencing Read Archive (SRA) under BioProject PRJNA1197453.

Code Availability

Scripts used for the genomic analyses are available at https://github.com/zengxi-hada/Chromosomal-rearrangements-and-instability-caused-by-the-LINE-1-retrotransposon

AUTHOR CONTRIBUTIONS

C.-Z.Z and K.H.B. conceived the project. C.M-D, K.H.B., and C-Z.Z. designed the experiments. C.M.D. performed all the experiments and analyses of experimental data with support from P.S. and J.C.K. C.M-D. and J.A.K. performed the DNA copy-number analysis of low-pass whole-genome sequencing data. X.Z. and J.A.K. performed the insertion analysis with contributions from D.D. S.T. performed the analysis of local sequence changes and mutational signatures. C.-Z.Z performed the analysis of DNA rearrangements with help from X.Z. and J.A.K. K.H.B. supervised the experimental analyses. C.-Z.Z. and E.A.L. supervised the genomic analyses. C.M-D. K.H.B and C-Z.Z. wrote the manuscript with input from all authors.

FUNDING SOURCES

This work was supported by the Innovations Research Fund from Dana-Farber Cancer Institute (to K.H.B. and C.-Z.Z.) and by the National Cancer Institute (R01CA276112, to K.H.B., E.A.L and C-Z. Z.). C.M-D is a Fellow of The Jane Coffin Childs Fund for Medical Research. J.A.K. was supported by a postdoctoral fellowship from the American Cancer Society (PF-22–123–01-DMC). E.A.L was supported by the National Institute of Health (NIH) (DP2 AG072437) and the Suh Kyungbae Foundation. K.H.B. is supported by the National Institutes of Health (R01CA240816, R01CA289390, UG3NS132127). C.-Z.Z. was also supported by the Claudia Adams Barr Program for Innovative Cancer Research.

References

1. K. H. Burns, Repetitive DNA in disease. *Science (New York, N.Y.)* **376**, 353-354 (2022).
2. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
3. S. E. Holmes, M. F. Singer, G. D. Swergold, Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *J Biol Chem* **267**, 19765-19768 (1992).
4. E. Khazina *et al.*, Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nature structural & molecular biology* **18**, 1006-1014 (2011).
5. O. Weichenrieder, K. Repanas, A. Perrakis, Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986 (2004).
6. E. T. Baldwin *et al.*, Structures, functions and adaptations of the human LINE-1 ORF2 protein. *Nature* **626**, 194-206 (2024).
7. A. Thawani, A. J. F. Ariza, E. Nogales, K. Collins, Template and target-site recognition by human LINE-1 in retrotransposition. *Nature* **626**, 186-193 (2024).
8. M. Dewannieux, C. Esnault, T. Heidmann, LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41-48 (2003).
9. E. M. Ostertag, J. L. Goodier, Y. Zhang, H. H. Kazazian, Jr., SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**, 1444-1451 (2003).
10. H. H. Kazazian, Jr., Processed pseudogene insertions in somatic cells. *Mob DNA* **5**, 20 (2014).
11. C. Esnault, J. Maestre, T. Heidmann, Human LINE retrotransposons generate processed pseudogenes. *Nature genetics* **24**, 363-367 (2000).
12. W. Wei *et al.*, Human L1 retrotransposition: cis preference versus trans complementation. *Molecular and cellular biology* **21**, 1429-1439 (2001).
13. C. Mendez-Dorantes, K. H. Burns, LINE-1 retrotransposition and its deregulation in cancers: implications for therapeutic opportunities. *Genes & development* **37**, 948-967 (2023).
14. K. H. Burns, Transposable elements in cancer. *Nature reviews. Cancer* **17**, 415-424 (2017).
15. N. Rodic *et al.*, Long interspersed element-1 protein expression is a hallmark of many human cancers. *The American journal of pathology* **184**, 1280-1286 (2014).
16. M. S. Taylor *et al.*, Ultrasensitive Detection of Circulating LINE-1 ORF1p as a Specific Multicancer Biomarker. *Cancer discovery* **13**, 2532-2547 (2023).
17. K. Chalitchagorn *et al.*, Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene* **23**, 8841-8846 (2004).
18. E. Lee *et al.*, Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-971 (2012).
19. E. Helman *et al.*, Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome research* **24**, 1053-1063 (2014).

20. N. Rodić *et al.*, Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nature medicine* **21**, 1060-1064 (2015).
21. T. T. Doucet-O'Hare *et al.*, Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma. *Hum Mutat* **37**, 942-954 (2016).
22. C. H. Nam *et al.*, Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* **617**, 540-547 (2023).
23. S. Solyom *et al.*, Extensive somatic L1 retrotransposition in colorectal tumors. *Genome research* **22**, 2328-2338 (2012).
24. B. Rodriguez-Martin *et al.*, Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature genetics* **52**, 306-319 (2020).
25. J. M. C. Tubio *et al.*, Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science (New York, N.Y.)* **345**, 1251343 (2014).
26. T. Cajuso *et al.*, Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival. *Nature communications* **10**, 4022 (2019).
27. E. C. Scott *et al.*, A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome research* **26**, 745-755 (2016).
28. Y. Miki *et al.*, Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research* **52**, 643-645 (1992).
29. S. L. Gasior, T. P. Wakeman, B. Xu, P. L. Deininger, The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of molecular biology* **357**, 1383-1393 (2006).
30. D. E. Symer *et al.*, Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327-338 (2002).
31. N. Gilbert, S. Lutz, T. A. Morrish, J. V. Moran, Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-7795 (2005).
32. N. Gilbert, S. Lutz-Prigge, J. V. Moran, Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315-325 (2002).
33. D. Ardeljan *et al.*, Cell fitness screens reveal a conflict between LINE-1 retrotransposition and DNA replication. *Nature structural & molecular biology* **27**, 168-178 (2020).
34. D. D. Luan, M. H. Korman, J. L. Jakubczak, T. H. Eickbush, Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605 (1993).
35. J. V. Moran *et al.*, High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).
36. T. A. Morrish *et al.*, DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature genetics* **31**, 159-165 (2002).
37. R. W. Tourdot, G. J. Brunette, R. A. Pinto, C. Z. Zhang, Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol* **22**, 139 (2021).
38. N. T. Umbreit *et al.*, Mechanisms generating cancer genome complexity from a single cell division error. *Science (New York, N.Y.)* **368**, (2020).

39. S. H. Rangwala, H. H. Kazazian, Jr., The L1 retrotransposition assay: a retrospective and toolkit. *Methods* **49**, 219-226 (2009).

40. C. Chu *et al.*, Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nature communications* **12**, 3836 (2021).

41. T. L. A. Miller, F. Orpinelli Rego, J. L. L. Buzzo, P. A. F. Galante, sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. *Bioinformatics* **37**, 419-421 (2021).

42. D. C. Hancks, J. L. Goodier, P. K. Mandal, L. E. Cheung, H. H. Kazazian, Jr., Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human molecular genetics* **20**, 3386-3400 (2011).

43. E. M. Ostertag, H. H. Kazazian, Jr., Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome research* **11**, 2059-2065 (2001).

44. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics* **12**, 187-215 (2011).

45. Q. Feng, J. V. Moran, H. H. Kazazian, Jr., J. D. Boeke, Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916 (1996).

46. T. Sultana *et al.*, The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular cell* **74**, 555-570.e557 (2019).

47. D. A. Flasch *et al.*, Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* **177**, 837-851.e828 (2019).

48. V. Ahl, H. Keller, S. Schmidt, O. Weichenrieder, Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Molecular cell* **60**, 715-727 (2015).

49. S. L. Cooke *et al.*, Processed pseudogenes acquired somatically during cancer development. *Nature communications* **5**, 3644 (2014).

50. L. Han *et al.*, The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications* **5**, 3963 (2014).

51. R. Bhargava *et al.*, C-NHEJ without indels is robust and requires synergistic function of distinct XLF domains. *Nature communications* **9**, 2484 (2018).

52. M. Cisneros-Aguirre, X. Ping, J. M. Stark, To indel or not to indel: Factors influencing mutagenesis during chromosomal break end joining. *DNA repair* **118**, 103380 (2022).

53. H. H. Y. Chang, N. R. Pannunzio, N. Adachi, M. R. Lieber, Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature reviews. Molecular cell biology* **18**, 495-506 (2017).

54. N. Bennardo, A. Cheng, N. Huang, J. M. Stark, Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS genetics* **4**, e1000110 (2008).

55. J. B. Moldovan, Y. Wang, S. Shuman, R. E. Mills, J. V. Moran, RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 20612-20622 (2019).

56. A. Athanasiadis, A. Rich, S. Maas, Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS biology* **2**, e391 (2004).

57. D. D. Kim *et al.*, Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome research* **14**, 1719-1725 (2004).

58. B. McClintock, The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. *Proc Natl Acad Sci U S A* **28**, 458-463 (1942).

59. B. McClintock, Chromosome organization and genic expression. *Cold Spring Harbor symposia on quantitative biology* **16**, 13-47 (1951).

60. C. Z. Zhang *et al.*, Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179-184 (2015).

61. C.-Z. Zhang, D. Pellman, Chromosome breakage-replication/fusion enables rapid DNA amplification. *bioRxiv*, 2024.2008.2017.608415 (2024).

62. A. M. Al-Zain, M. R. Nester, I. Ahmed, L. S. Symington, Double-strand breaks induce inverted duplication chromosome rearrangements by a DNA polymerase delta-dependent mechanism. *Nat Commun* **14**, 7020 (2023).

63. T. A. Morrish *et al.*, Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**, 208-212 (2007).

64. N. Bona, G. P. Crossan, Fanconi anemia DNA crosslink repair factors protect against LINE-1 retrotransposition during mouse development. *Nat Struct Mol Biol* **30**, 1434-1445 (2023).

65. P. Ly *et al.*, Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat Genet* **51**, 705-715 (2019).

66. J. Maciejowski, Y. Li, N. Bosco, P. J. Campbell, T. de Lange, Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163**, 1641-1654 (2015).

67. B. McClintock, The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**, 344-355 (1950).

68. M. J. Curcio, K. M. Derbyshire, The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**, 865-877 (2003).

69. H. H. Kazazian, Jr., J. V. Moran, Mobile DNA in Health and Disease. *N Engl J Med* **377**, 361-370 (2017).

70. L. Carbone *et al.*, Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195-201 (2014).

71. H. Ghezraoui *et al.*, Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Mol Cell* **55**, 829-842 (2014).

72. M. Achom *et al.*, A genetic basis for sex differences in Xp11 translocation renal cell carcinoma. *Cell* **187**, 5735-5752 e5725 (2024).

73. D. F. Robbiani *et al.*, AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol Cell* **36**, 631-641 (2009).

74. W. McKerrow *et al.*, LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint. *Proceedings of the National Academy of Sciences of the United States of America* **119**, (2022).

75. S. Zumalave *et al.*, Synchronous L1 retrotransposition events promote chromosomal crossover early in human tumorigenesis. *bioRxiv*, 2024.2008.2027.596794 (2024).
76. C. Arnould *et al.*, Chromatin compartmentalization regulates the response to DNA damage. *Nature* **623**, 183-192 (2023).
77. V. Roukos *et al.*, Spatial dynamics of chromosome translocations in living cells. *Science (New York, N.Y.)* **341**, 660-664 (2013).
78. T. R. Pisanic, 2nd *et al.*, Long Interspersed Nuclear Element 1 Retrotransposons Become Deregulated during the Development of Ovarian Cancer Precursor Lesions. *The American journal of pathology* **189**, 513-520 (2019).
79. Z. Xia *et al.*, Expression of L1 retrotransposon open reading frame protein 1 in gynecologic cancers. *Human pathology* **92**, 39-47 (2019).
80. S. Sato *et al.*, LINE-1 ORF1p as a candidate biomarker in high grade serous ovarian carcinoma. *Sci Rep* **13**, 1537 (2023).

Chromosomal rearrangements and instability caused by L1 retrotransposition

1 L1 expression causes extensive DNA breakage and large segmental copy-number alterations . . . 1
S1 Further evidence of DNA damage from L1 expression . . . 2
S2 Large de novo CNAs after L1 induction with the Tet-On system . . . 4
S3 Large de novo CNAs in cells with L1 retrotransposition identified using the L1 GFP reporter . . . 5
S4 Sites of integration of the Tet-On L1 transgene . . . 6
2 Landscape of canonical retrotranspositions after L1 expression . . . 8
S5 Additional data on insertions from retrotransposition . . . 9
S6 Retrocopied pseudogene insertions of aberrant mRNA transcripts . . . 10
3 Complex insertions from template-switching RT or end-joining between RT DNA ends . . . 11
S7 Additional examples of complex insertions . . . 12
S8 Retrocopied pseudogene insertions with substitutions due to ADAR editing . . . 14
S9 Insertions with inversions of genomic DNA sequences near the insertion site . . . 15
S10 Insertions containing both RT sequences and templated genomic sequences . . . 17
4 Reciprocal translocations between DNA ends generated by L1 retrotransposition . . . 19
S11 Additional examples of L1-mediated translocations . . . 19
S12 Dicentric chromosomes in Dox clones . . . 21
S13 Dicentric chromosomes in GFP+ clones . . . 22
5 Segmental copy-number alterations from retrotransposition-mediated rearrangement . . . 24
S14 Additional examples of large segmental deletions in clones after L1 activation . . . 25
S15 Sloping copy-number variation in clones after L1 activation . . . 26
6 Foldback junctions with retrotransposition insertions . . . 28
S16 Additional instances of foldback junctions after L1 activation . . . 29
7 Chromothripsis and retrotransposition . . . 31
S17 Additional instances of chromothripsis in clones after L1 activation . . . 32
S18 Single-base substitutions in clones after L1 activation . . . 33
8 Insertion and rearrangement outcomes of retrotransposition . . . 34

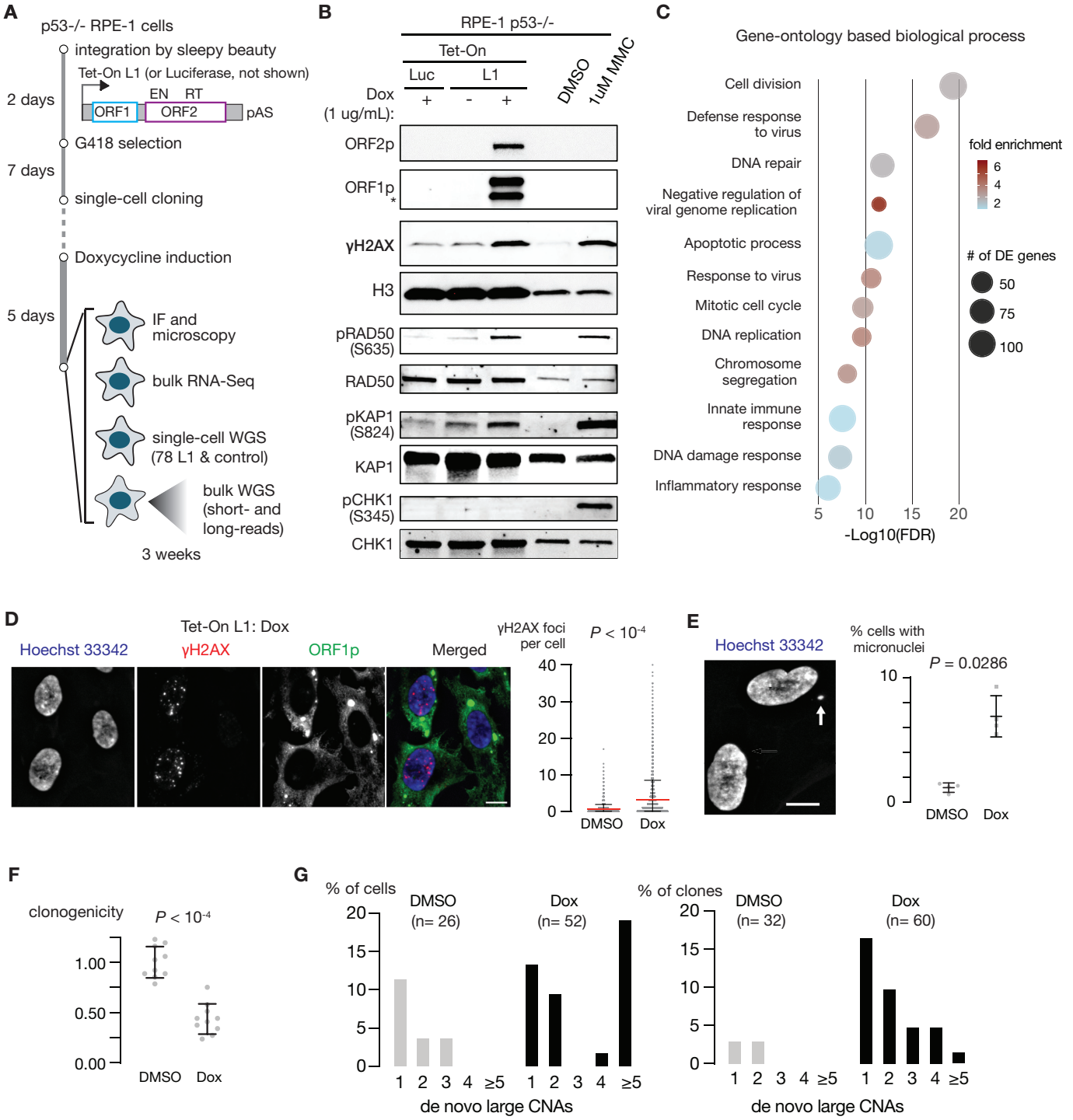


Figure 1 | L1 expression causes extensive DNA damage including double-strand breaks.

A. A schematic workflow of L1 induction and sequencing analysis. Expression of a codon-optimized L1 is induced in p53-null RPE-1 cells for five days using a Tet-On expression system. Tet-On L1 transgene is integrated at 11 locations in the genome (see **Figure S4**). Cells with L1 induction, with induction of luciferase expression, and without induction (treated with DMSO) were used for subsequent analyses.

B. Immunoblots of L1 encoded proteins ORF1p and ORF2p, γ-H2Ax, a histone marker of DNA damage, and markers of the DNA damage response including pRAD50 (S635), pKAP1 (S824), pCHK1 (S345) from whole cell lysates with or without L1 induction. Whole cell lysates of cells treated with 1 μM MMC were used as a positive control.

C. Biological processes inferred from Gene Ontology (GO) analysis of differential gene expression due to L1 expression. Differentially expressed (DE) genes were identified by a comparative analysis of the RNA-Seq data of p53-/- RPE-1 cells with Tet-On L1 and p53-/- RPE-1 cells with Tet-On Luciferase (Tet-On Luc) under treatment by Doxycycline (N = 3 technical replicates) using a threshold of adjusted $P < 0.05$. The size of each circle reflects the number of DE genes associated with each process; the color shade reflects the enrichment of DE genes (fold change of gene count) in each process. See also **Figure S1C**.

D. DNA damage in cells after L1 induction reflected in the significant increase of γ-H2Ax foci in cells with L1 induction.
Left: Representative images of γ-H2AX foci in cells with L1 expression; Bar scale: 10μm.
Right: Quantification of γ-H2AX foci per cell with (n = 2889 cells) and without (n = 2274 cells) L1 expression (N = 2 experiments). $P < 0.0001$; two-tailed Mann-Whitney U-test.

E. L1 induction leads to more frequent micronucleation. *Left*: an example of micronucleus in a cell after L1 induction (Bar scale: 10μm). *Right*: Quantification of cells with micronuclei after L1 induction (N = 4 independent experiments; DMSO: 1,820 cells; Dox: 1,223 cells; P = 0.0286; two-tailed Mann-Whitney U-test).

F. Reduced clonogenicity of single cells after L1 induction (5-day treatment of Doxycycline) in comparison to control cells (5-day treatment of DMSO). $P < 0.0001$; two-tailed Mann-Whitney U-test. See also **Figure S1A**.

G. L1 induction causes large segmental copy-number alterations in both single cells (left) and single-cell derived clones (right). Shown are the percentage of single cells or single-cell derived clones that harbor 1, 2, 3, 4 or ≥ 5 large de novo DNA copy-number alterations assessed from 0.1× whole-genome sequencing data. P = 0.0211 for single cells, P = 0.0011 for single-cell derived clones; Fisher's exact test. See also **Figure S2**.

Figure S1 | (*Figure on next page.*) Further evidence of DNA damage from L1 expression.

A. *Left*: Immunoblot of ORF1p and ORF2p expression in p53-null RPE-1 cells with Tet-On L1 after treatment with Doxycline at different concentrations. *Middle*: Reduced colony formation in cells with induced L1 expression. *Right*: Quantification of the survival fraction of cells under L1 expression in comparison to the control (luciferase).

B. Immunoblots of pRPA (S4/S8) and pRPA (S33) from whole cell lysates after L1 induction. Similar to **Figure 1B**.

C. Upregulated genes in the DNA damage response and DNA repair pathways (based on Gene-Ontology). Related to **Figure 1C**.

D. Quantification of 53BP1 foci per cell in p53-null RPE-1 cells with induced L1 expression (Dox, 612 cells) and control cells (DMSO, 538 cells) from two independent experiments. $P < 0.0001$; two-tailed Mann-Whitney U-test.

E. Representative immunofluorescence images of 53BP1, ORF1p, and Hoechst 33342 in p53-null RPE-1 with (Dox) or without (DMSO) L1 induction. Bar size: 10μm.

F. Validation of L1 induction in U2OS cells with either wild-type or mutant L1 with a Tet-On promoter integrated at an FRT locus. *Left*: Immunoblots similar to **A**. *Right*: Reduced cell survival under the induction of wildtype or EN-proficient L1 in comparison to the induction of L1 with inactivated EN. N = 6 replicates except for ENmut (D205G:H230A).

G. Quantification of γH2AX foci (top) and micronucleation (bottom) in U2OS FRT cells with Dox-induced expression of wildtype and mutant L1. Two independent experiments for each condition. Induction of wildtype L1 (first group) produces the most significant increase in γH2AX foci and micronucleation. Induction of L1 with proficient EN but inactive RT (second group) produces reduced but significant increase in γH2AX foci and a similar increase in micronucleation as wildtype L1. By contrast, induction of L1 with proficient RT but inactive EN produces no noticeable change in either γH2AX or micronucleation relative to control. P-values are calculated using one-way ANOVA with Tukey test.

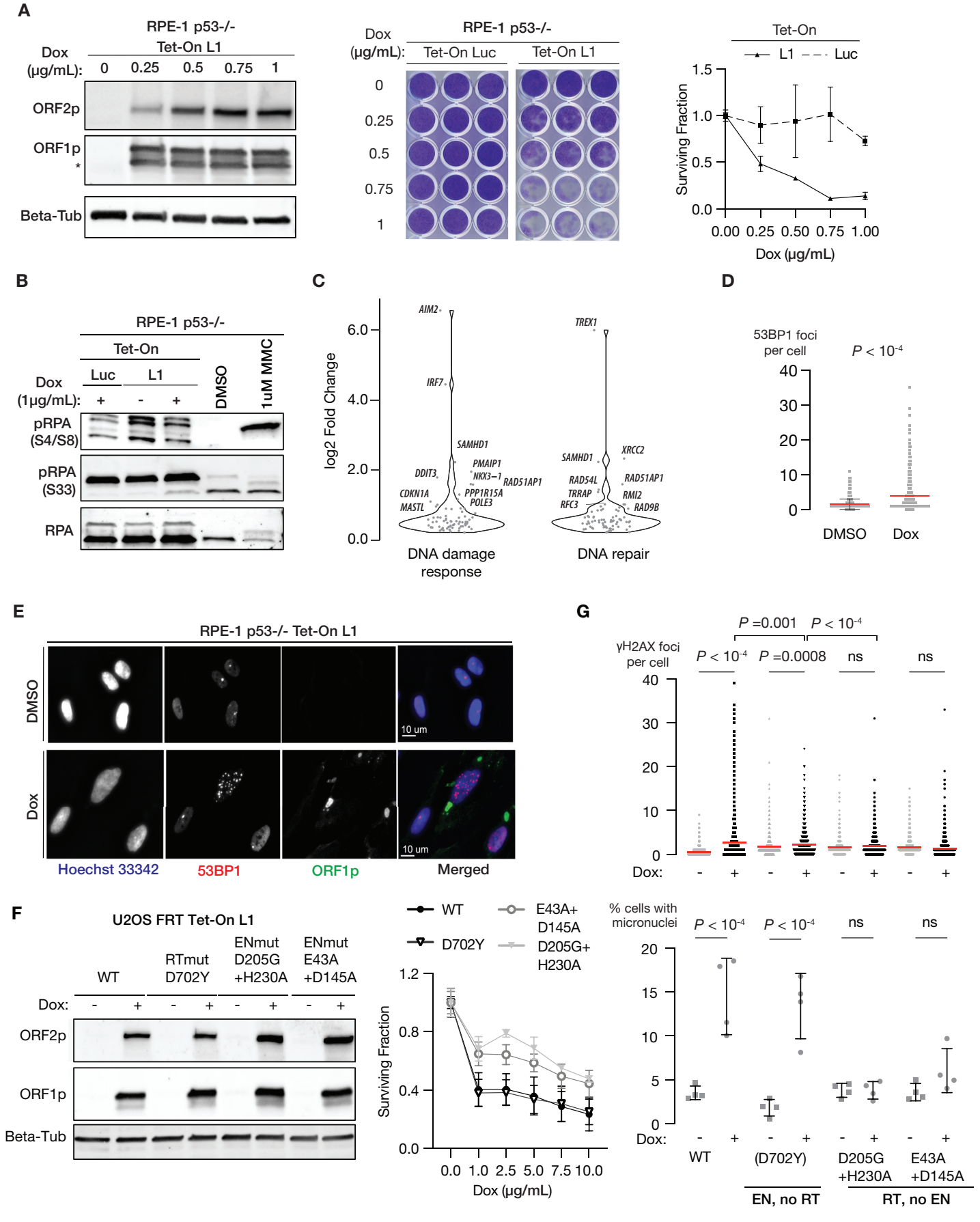

A
RPE-1 p53-/-
Tet-On L1
Dox (µg/mL):
0
0.25
0.5
0.75
1
ORF2p
ORF1p
*
Beta-Tub
RPE-1 p53-/-
Dox (µg/mL):
Tet-On Luc
Tet-On L1
Tet-On
L1
Luc
Surviving Fraction
Dox (µg/mL)
B
RPE-1 p53-/-
Tet-On
Luc
L1
Dox (1µg/mL):
DMSO
1uM MMC
pRPA (S4/S8)
pRPA (S33)
RPA
C
log2 Fold Change
AIM2
IRF7
SAMHD1
DDIT3
PMAIP1
NKX3–1
RAD51AP1
CDKN1A
PPP1R15A
MASTL
POLE3
TREX1
XRCC2
RAD54L
TRRAP
RFC3
RMI2
RAD9B
DNA damage response
DNA repair
D
53BP1 foci per cell
P < 10-4
DMSO
Dox
E
RPE-1 p53-/- Tet-On L1
DMSO
Dox
10 um
Hoechst 33342
53BP1
ORF1p
Merged
G
P =0.001
P < 10-4
γH2AX foci per cell
P < 10-4
P =0.0008
ns
ns
Dox:
F
U2OS FRT Tet-On L1
WT
RTmut D702Y
ENmut D205G +H230A
ENmut E43A +D145A
Dox:
ORF2p
ORF1p
Beta-Tub
WT
D702Y
E43A+ D145A
D205G+ H230A
Surviving Fraction
Dox (µg/mL)
% cells with micronuclei
P < 10-4
P < 10-4
ns
ns
Dox:
WT
(D702Y)
D205G +H230A
E43A +D145A
EN, no RT
RT, no EN

Figure S2 | Large copy-number alterations (CNAs) after L1 induction assessed from 0.1× whole-genome sequencing data.

A. Quantification of large CNAs in single cells (left) and single-cell derived clones (right) after L1 induction using the Tet-On system. Only alterations of segments ≥5Mb are counted; CNAs on different parental chromosomes are evaluated separately.

B. Heatmap of large CNAs (grouped by chromosome arms) in single-cell derived clones. Only clones with detectable large CNAs are shown.

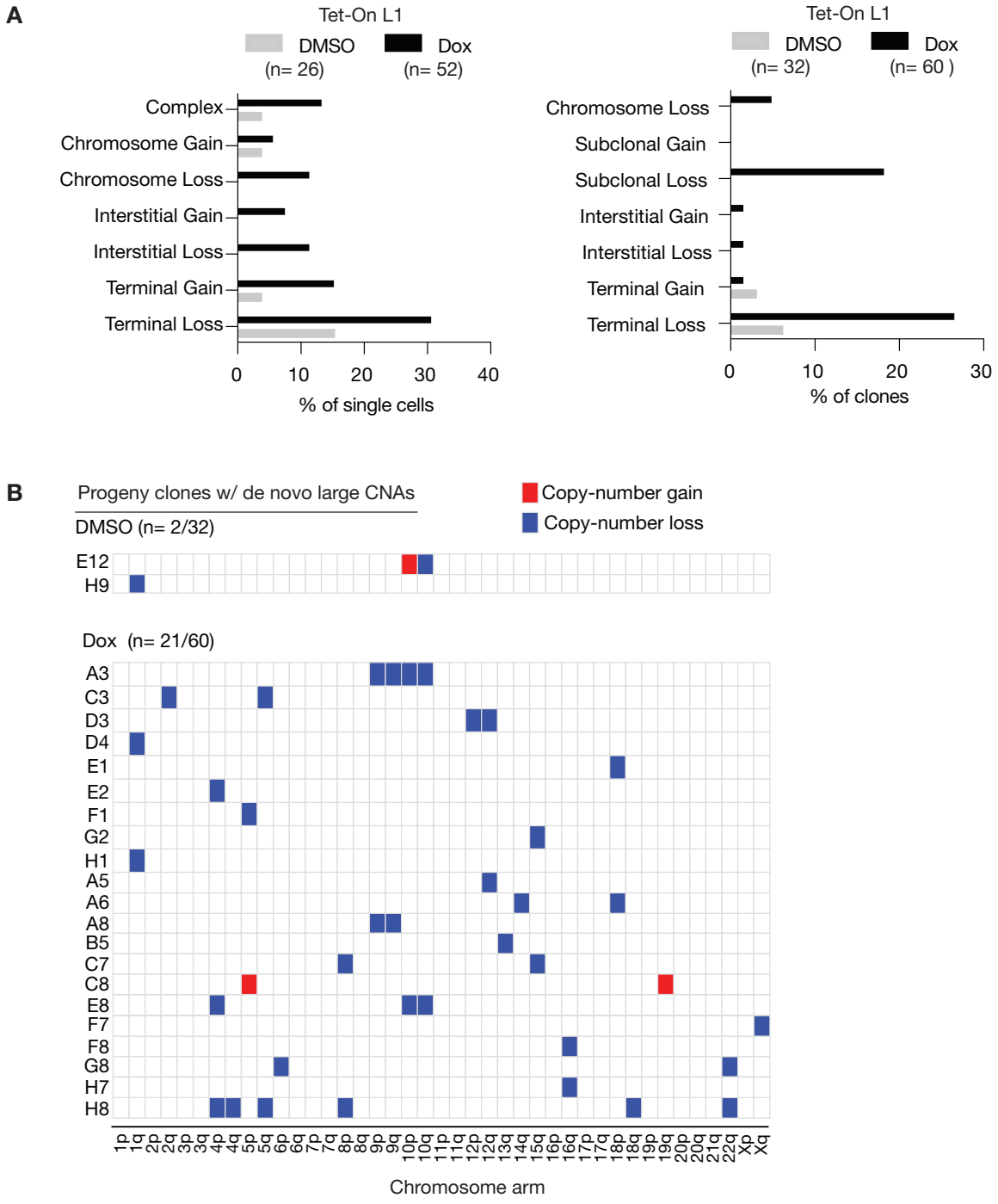


Figure S3 | Large de novo CNAs in cells with L1 retrotransposition identified using the L1 GFP reporter.

A. Schematic diagram of the L1 GFP reporter consisting of a codon-optimized human L1 and an anti-sense split GFP gene in the 3'-UTR. Expression of GFP only occurs with the integration of the split GFP gene by retrotransposition. Therefore, GFP+ cells must have had one or multiple retrotranspositions. However, GFP- cells may also contain one or multiple truncated copies of the L1 GFP reporter. Therefore, GFP- cells may also have undergone retrotransposition. See **Figure S5B**.

B.Schematic diagram of the experimental workflow.

C. Frequency of L1 retrotransposition in p53-null RPE-1 cells assessed using the L1 GFP reporter. Inactivation of reverse transcriptase activity completely abolishes retrotransposition, whereas inactivation of endonuclease activity suppresses but does not eliminate retrotransposition. Three replicates in each condition. *P*-values are calculated using one-way ANOVA with Tukey test.

D. and **E**. Quantification of large CNAs in clones derived from GFP+ cells similar to **Figure S2**.

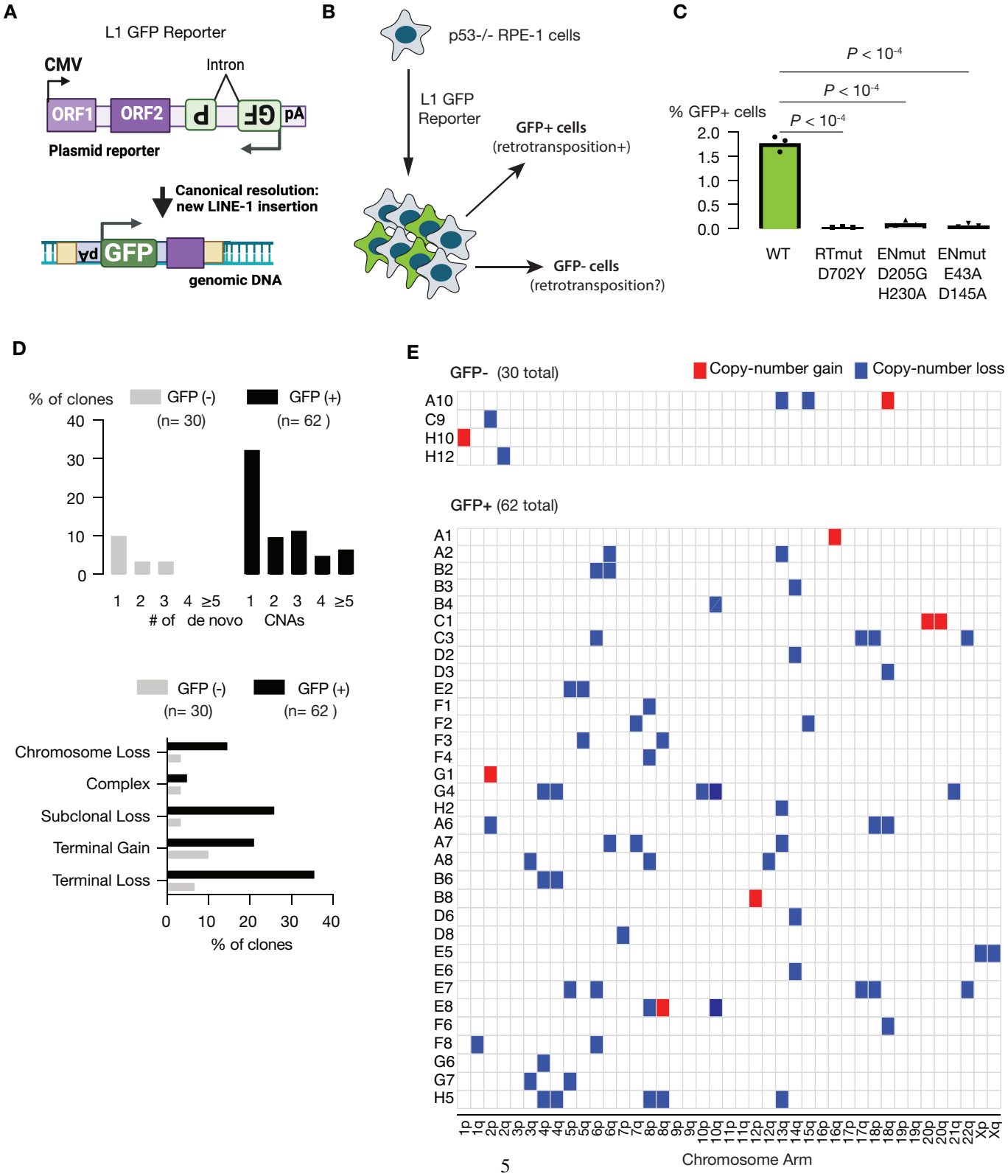
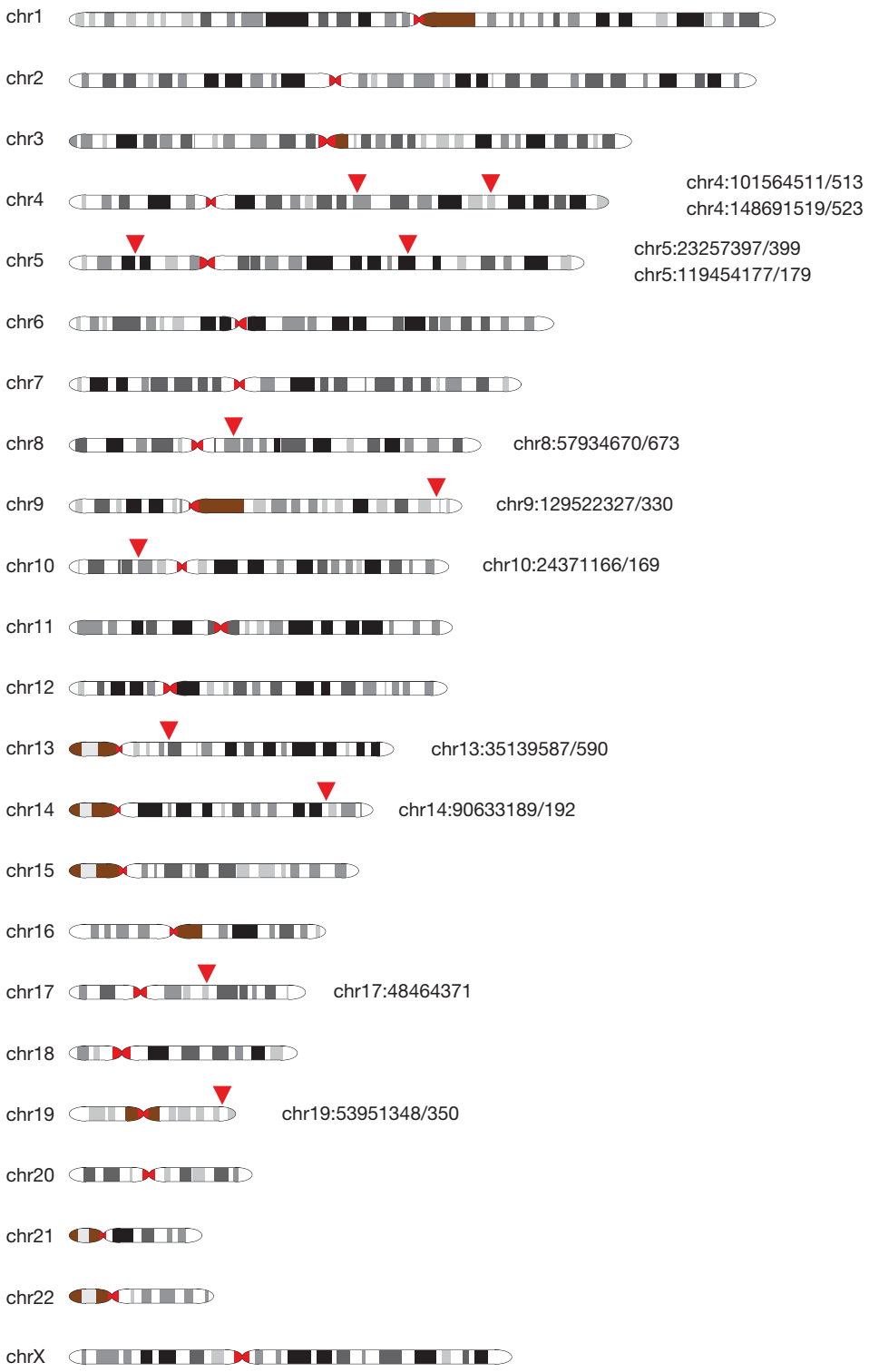


Figure S4 | Chromosomal locations and GRCh38 coordinates of 11 integrated Tet-On L1 transgene (red arrows) determined from breakpoints in reads (both short and long) and de novo assembled contigs (from PacBio long reads) with split alignments to both the transgene sequence and the human genome.



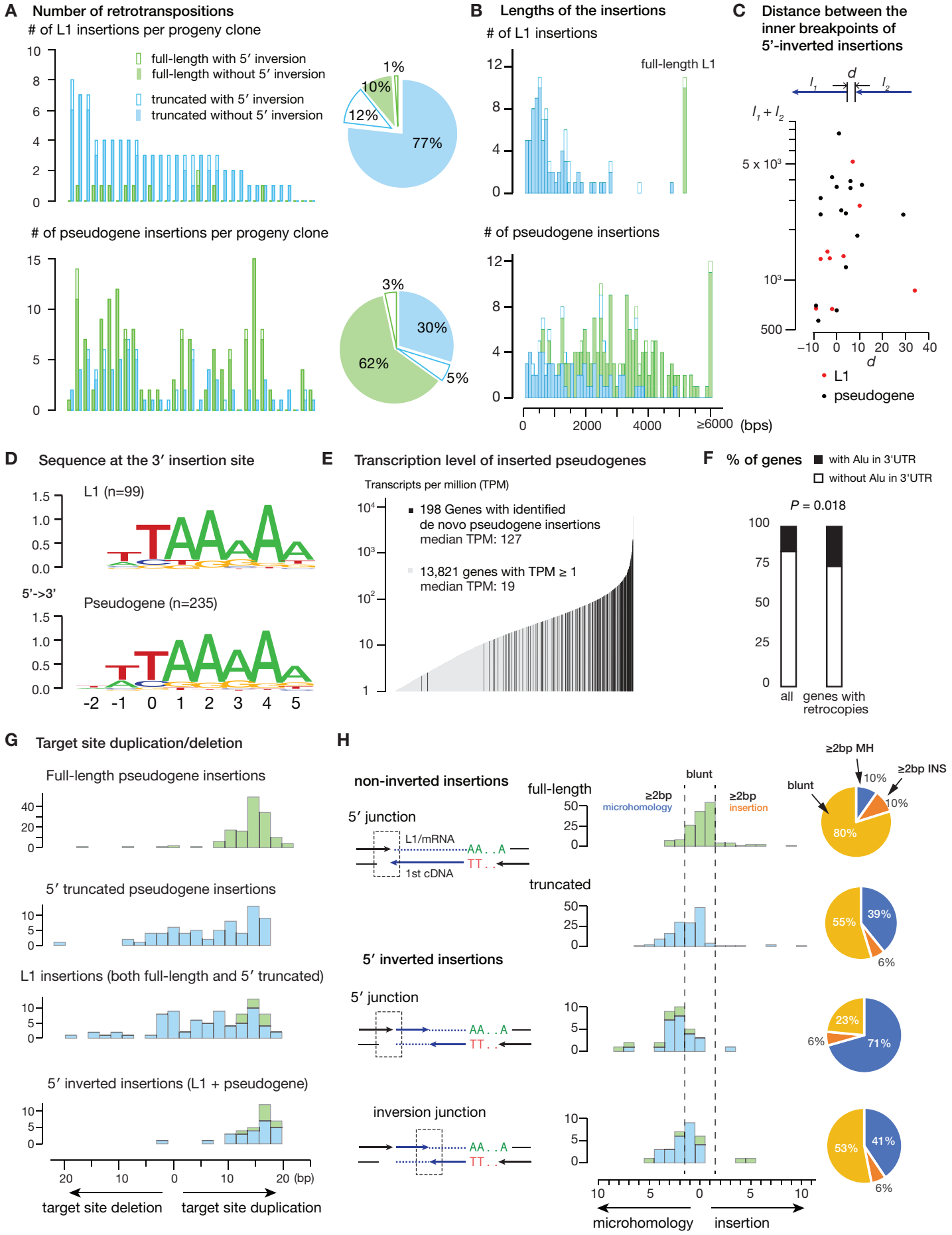

A Number of retrotranspositions
# of L1 insertions per progeny clone
full-length with 5′ inversion
full-length without 5′ inversion
truncated with 5′ inversion
truncated without 5′ inversion
1%
10%
12%
77%
# of pseudogene insertions per progeny clone
3%
30%
62%
5%
B Lengths of the insertions
# of L1 insertions
full-length L1
# of pseudogene insertions
0
2000
4000
≥6000
(bps)
C Distance between the inner breakpoints of 5'-inverted insertions
d
l_1
l_2
$l_1 + l_2$
5 x 10³
10³
500
−10
0
10
20
30
40
d
L1
pseudogene
D Sequence at the 3′ insertion site
L1 (n=99)
5'->3'
Pseudogene (n=235)
-2
-1
0
1
2
3
4
5
E Transcription level of inserted pseudogenes
Transcripts per million (TPM)
198 Genes with identified de novo pseudogene insertions median TPM: 127
13,821 genes with TPM ≥ 1 median TPM: 19
F % of genes
with Alu in 3'UTR
without Alu in 3'UTR
P = 0.018
all
genes with retrocopies
G Target site duplication/deletion
Full-length pseudogene insertions
5′ truncated pseudogene insertions
L1 insertions (both full-length and 5′ truncated)
5′ inverted insertions (L1 + pseudogene)
20
10
0
10
20 (bp)
target site deletion
target site duplication
H
non-inverted insertions
5′ junction
L1/mRNA
1st cDNA
AA . . A
TT . .
5′ inverted insertions
5′ junction
inversion junction
full-length
truncated
≥2bp microhomology
blunt
≥2bp insertion
≥2bp MH
≥2bp INS
blunt
10%
10%
80%
55%
39%
6%
23%
6%
71%
53%
41%
6%
microhomology
insertion

Figure 2 | Landscape of de novo full-length, 5'-truncated, and 5'-inverted insertions identified in the progeny clones of single cells with 5-day Dox-induced L1 expression.

A. Number of full-length, 5'-truncated, or 5'-inverted insertions of L1 (upper) or processed pseudogenes (lower) detected in 31 single-cell derived clones.

B. Length distribution of the inserted sequences of L1 (upper, $n = 99$) or pseudogene (lower, $n = 235$) insertions. The insertion length (excluding the poly-A sequence) is both calculated from breakpoints in the source sequence (L1 or mRNA) and further validated by long reads.

C. The total size (y-axis, $l_1 + l_2$) and distance between inner breakpoints (x-axis, d) of 5'-inverted insertions.

D. Sequence logo plots of the genomic DNA sequence at the 3'-end of insertions (starting site of TPRT) of L1 (upper) and pseudogenes (lower).

E. Pseudogene insertions are enriched for highly expressed genes. Shown are the transcripts-per-million (TPM) values of endogenous genes with (black) and without pseudogene insertions. Except for one insertion of *PLCL1* with TPM=0.61, all the remaining insertions are derived from endogenous genes with TPM >1. The median TPM value for 198 genes used as source for one or more pseudogene insertions is 127 in comparison to the median TPM of 19 for the remaining 13,821 genes with TPM ≥1.

F. Enrichment of pseudogene insertions from source genes with *Alu* in the 3'-UTR. $P = 0.018$; χ^2 test.

G. Length distribution of deleted or duplicated sequences at the target site for different categories of L1 or pseudogene insertions.

H. Length distribution of microhomology or untemplated sequences at the 5'-junctions of non-inverted insertions (top two panels) and at the 5' (middle bottom) and the internal junctions (bottom) of 5'-inverted insertions. The locations of junctions are schematically shown on the left. The percentages of junctions with ≥2bp microhomology, ≥2bp untemplated insertions, and near blunt (otherwise) are shown on the right as piecharts. We use thick arrows to represent the 3'-ends and thin lines to represent the complementary 5'-ends of dsDNA, colored arrows to represent the first cDNA strand and dotted lines to represent the L1/mRNA template/second cDNA strand. These conventions are used throughout the remaining figures.

Figure S5 | Additional data on insertions from retrotransposition.

A. Consistency between insertions detected from short reads without manual curation and from long reads plus manual review. The comparison is done for clones with Dox-induced L1 expression (**Dox clones**) for which both short- and long-read data are available. The manually curated results are used for the final analysis .

B. Number of L1 and pseudogene insertions in clones expanded from GFP+ cells with the L1-GFP reporter (**GFP+ clones**). We identified seven truncated insertions in 1/5 GFP- clone. The predominance of L1 insertions over pseudogene insertions in GFP+ clones is likely due to selection for retrotransposition of the L1-GFP reporter.

C. L1 and pseudogene insertions in single cells after Dox-induced L1 expression. The panels are arranged similarly as in **Figure 2A,B,G** except that pseudogene insertions are excluded from the assessment of insertion length and target site sequence changes.

Note: As only a few GFP+ clones and none of the single cells have long-read data, we derive the main findings largely from the **Dox clones** for which the complete insertion/rearrangement junctions can be determined. We present examples from the single cells and from the GFP+ clones as complementary evidence for these findings.

D. Sequence logo plots of the genomic DNA sequence at the 5'-end of insertions of L1 (upper) and pseudogenes (lower), similar to **Figure 2D**.

E. Positive correlation between the number of pseudogene insertions and L1 insertions.

F. Positive correlation between the number of insertions (L1 and pseudogene) and large segmental copy-number alterations.

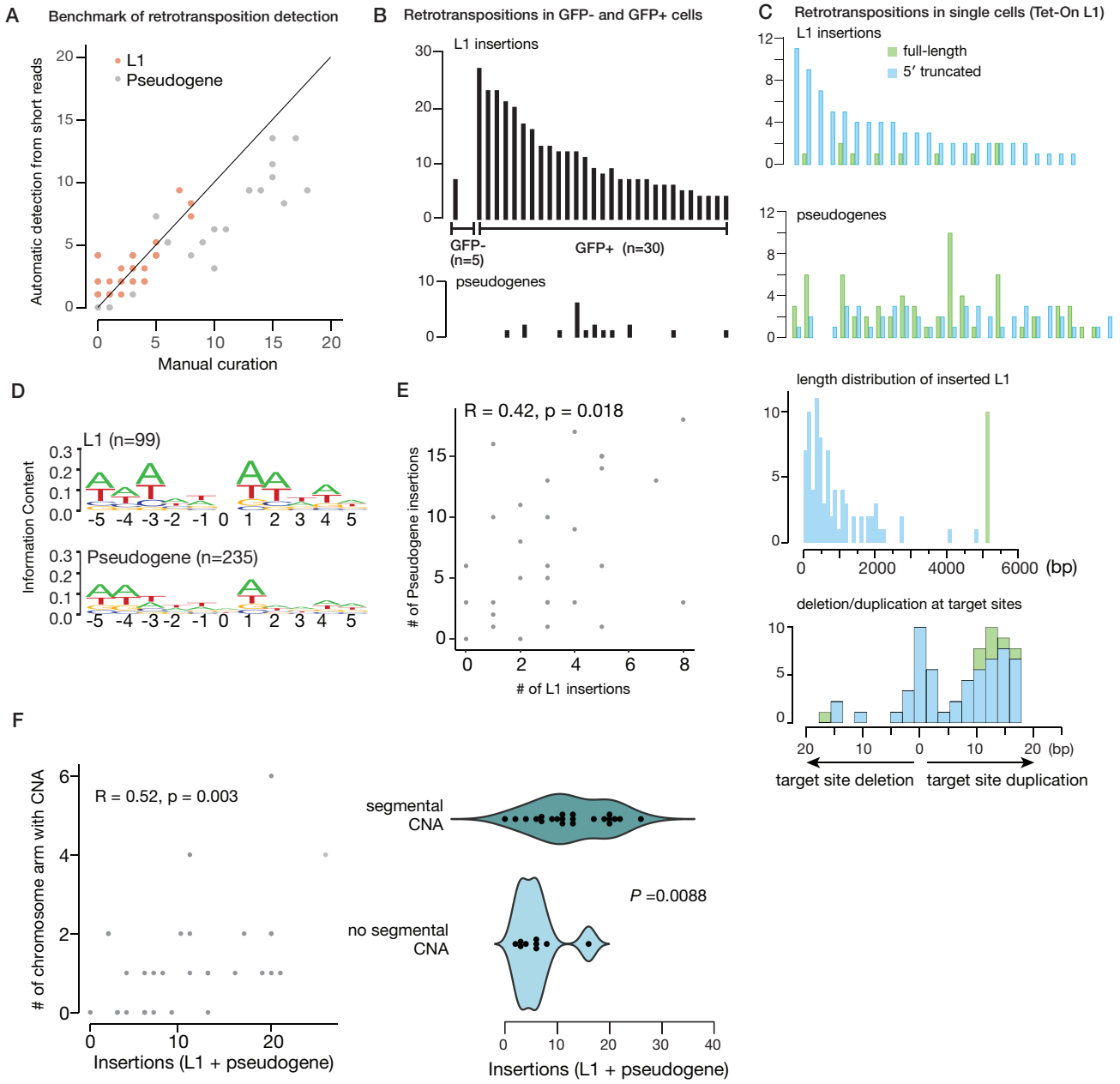
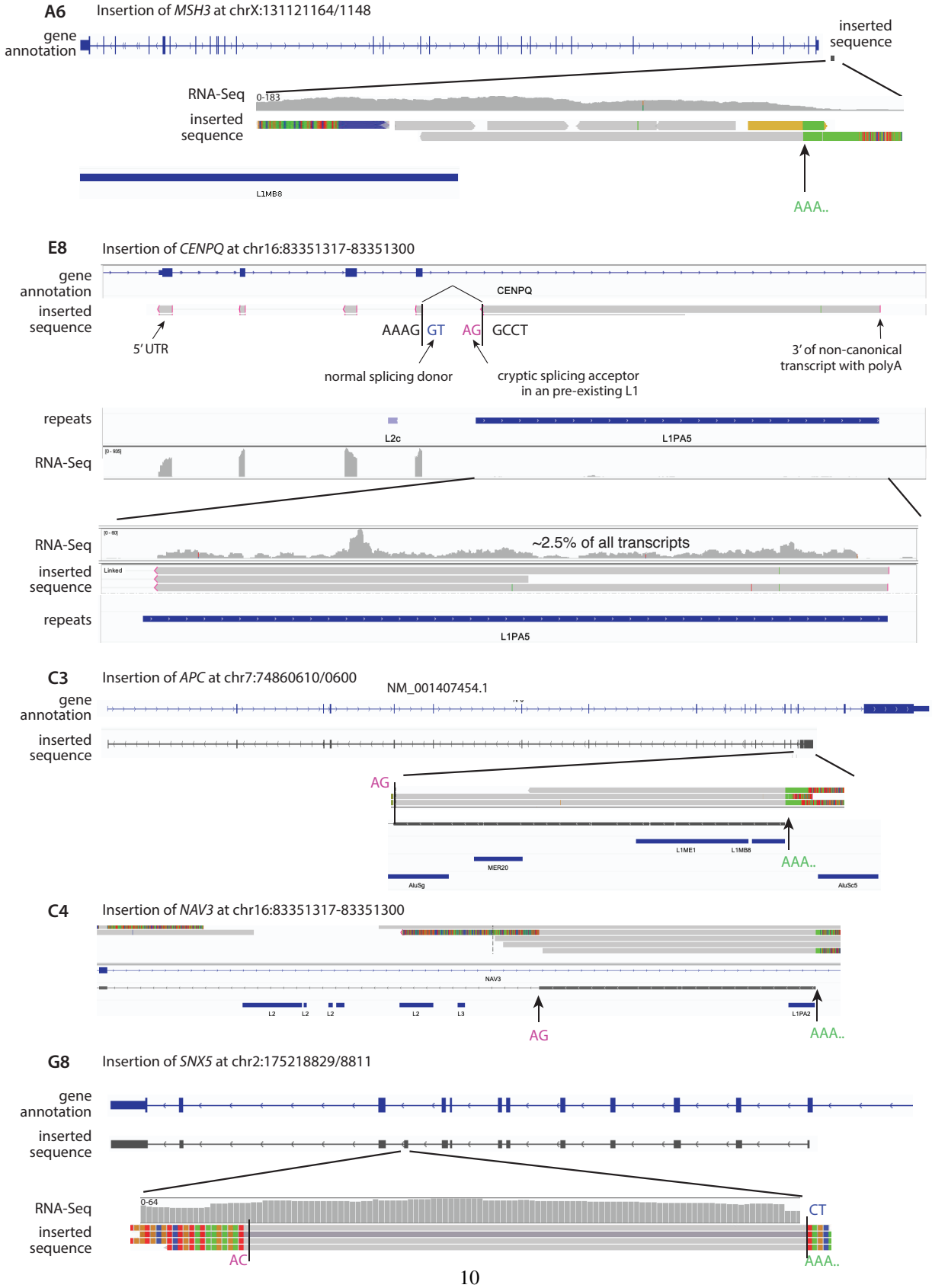


Figure S6 | L1-mediated insertions of non-canonical transcripts. All the insertions are identified in **Dox clones** and are validated by long reads. The RNA-Seq data are from parental RPE-1 cells without L1 induction.

In the first example (**A6**), the inserted sequence is from the 3'-UTR region of a rare transcript that spans a truncated endogenous L1. In the next three examples (**E8**, **C3** and **C4**), the inserted sequences contain intronic sequences flanked by cryptic splicing acceptor (`AG`) and endogenous L1 elements. For these four cases, the insertion of these rare transcripts instead of the canonical transcripts suggests a preferred interaction between ORF2p and mRNAs with subsequences from L1 at the 3' end.

In the last example (**G8**), the retained intronic sequence is flanked by cryptic `AC` and `CT` sequences that are identical to the donor and acceptor sequences of the up- and down-stream exons.



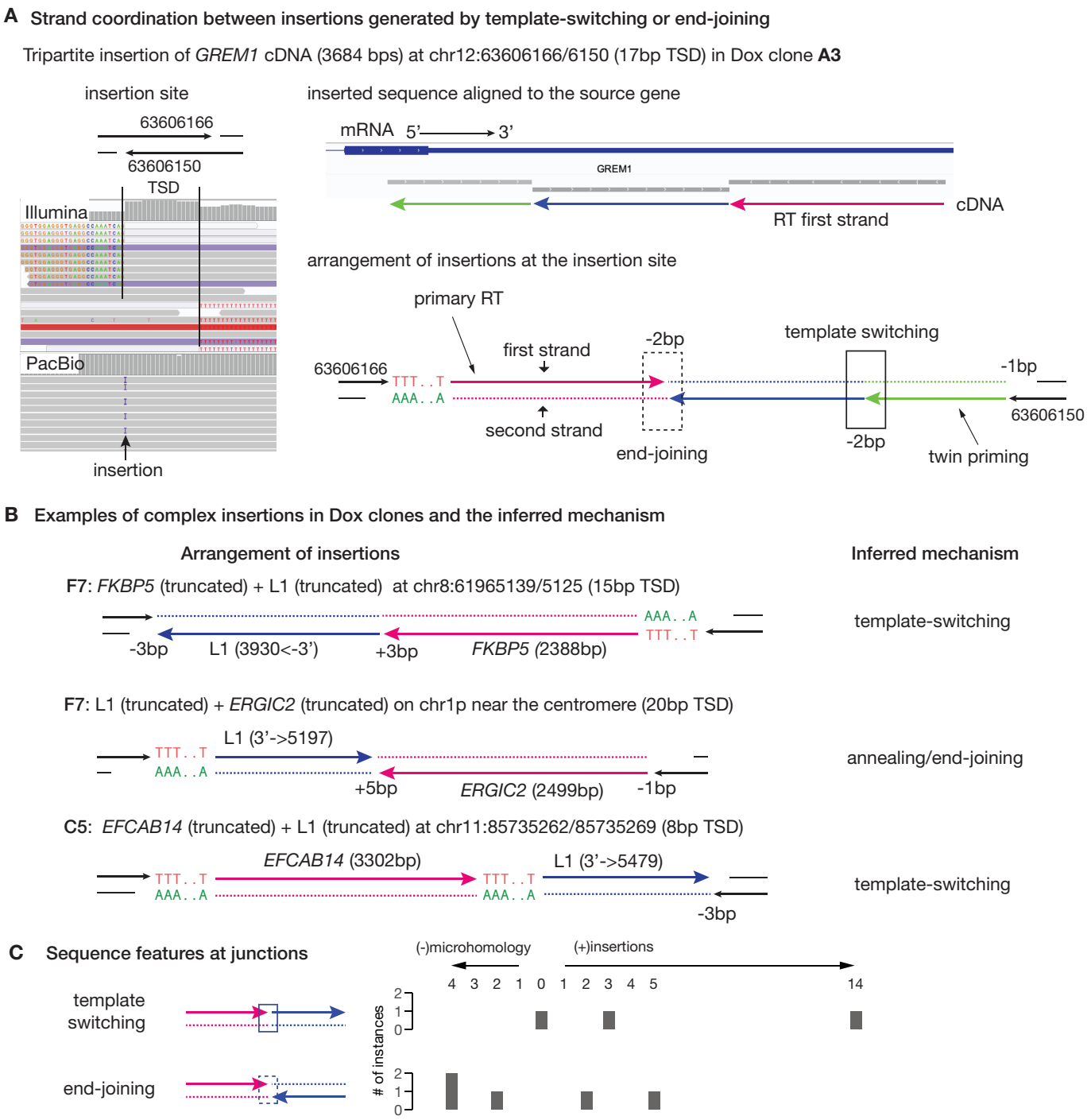


Figure 3 | Examples of complex insertions indicating template-switching during reverse transcription or annealing/ligation of two DNA ends each having undergone independent reverse transcription.

A. Strand coordination between insertions generated by template-switching during RT or from ligation/annealing of two cDNA ends illustrated by a tripartite insertion of the *GREM1* cDNA in **Dox clone A3**.

Left: Screenshot of short (top) and long (bottom) reads at the insertion site, showing hallmark features of TPRT including poly-A and target-site duplication in the short reads, and a single insertion in the long reads.

Right: (*Top*) Alignment of the inserted sequence to the source gene (*GREM1*) reveals three pieces of reverse-transcribed sequences (green, blue, and magenta arrows); (*Bottom*) arrangement of the three RT sequences at the insertion site determines that the magenta sequence with poly-T is generated by the primary RT extending the DNA end on the forward strand (chr12:63606166); the blue and the green sequences are generated by twin-primed RT extending the DNA end on the reverse strand (chr12:63606150). The parallel orientation between the blue and green insertions indicates a template-switching event during RT, whereas the opposite orientation between the red and blue insertions implies an annealing between ssDNA ends or a ligation between dsDNA ends. Microhomology (-) and untemplated insertions (+) are annotated at each junction by the same convention as in **Figure 2H**.

B. Selected examples of complex insertions containing sequences from more than one source. All three examples display TSD and poly-A features as conventional full-length or truncated insertions. The arrangement of insertions are shown on the left, with the inferred mechanism (template-switching or annealing/end-joining) annotated on the right. The insertions in **F7** are completely resolved by long reads; the insertion in **C5** is assembled based on junctions detected from short reads. The presence of microhomology (e.g., '-2bp' for two basepair microhomology) or untemplated insertions ('+5bp' for an insertion of five base pairs) is annotated at all junctions except the poly-A junctions. See **Figure S7** for additional examples.

C. Summary of microhomology/untemplated insertions at template-switching or annealing/end-joining junctions.

Figure S7 | (*Figure on next page.*) Additional examples of complex insertions and their mechanistic interpretations.

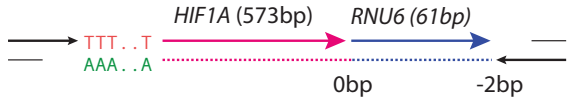
A. Additional examples of complex insertions reflecting template-switching RT or annealing/end-joining between twin-primed RT ends. Except for the insertion in **Dox clone D6**, which is assembled based on junctions detected from short reads, all the other insertions are completely resolved by long reads. Note the proximity between the two RT sequences of *SEMA3C* in sample **D6** that recapitulates the feature of 5'-inverted insertions as shown in **Figure 2C**.

B. Two alternative outcomes of twin priming suggested by the insertion rearrangements shown in **A**. In the first model, the reciprocal end is primed to the 3' poly-A sequence, resulting in a 3'-inverted insertion; this model explains insertions with poly-A/T on both sides. In the second model, the primary and the twin-primed RT ends are joined by ligation with DNA or DNA/RNA duplexes. The insertion of two retrocopied sequences between two DNA ends is seen in another example in **Figure 6C**.

A Additional examples of complex insertions containing RT sequences from two sources, all **Dox clones**

Arrangement of insertions | **Inferred mechanism**

A5: *HIF1A* (5′ truncated) + *RNU6* at chr15:75729346/9331 (16bp TSD)

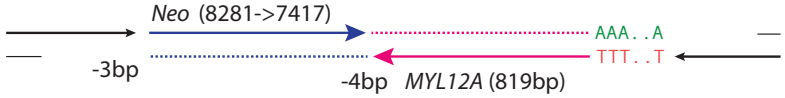


template-switching or chimeric mRNA

G5: L1 (truncated) + *UBP1* (full-length) at chr17:10737280/7277 (4bp TSD)

template-switching

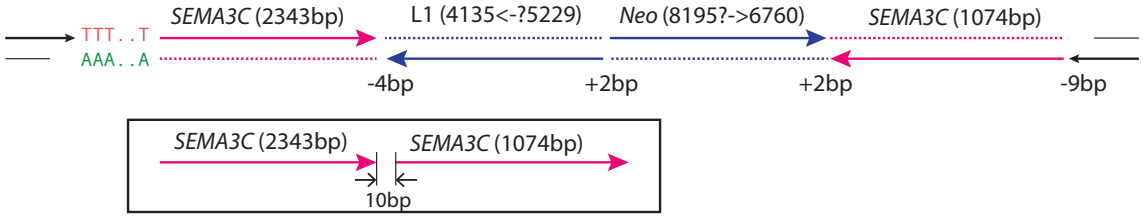
H8: *Neo* (truncated) + *MYL12A* (truncated) at chr9:128700830/128700128 (703bp TSD)



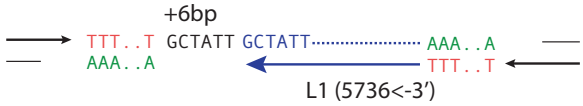
annealing/end-joining

D6: L1, *Neo*, & *SEMA3C* at chr8:29667387/7377 (11bp TSD)

twin-priming & annealing/end-joining

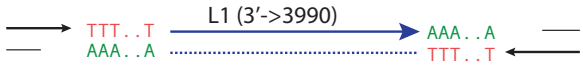


A7: L1 + poly-A at chr5:32883965/3952 (14bp TSD)



annealing/end-joining

D4: L1 + poly-A at chr1:56064833/4820 (14bp TSD)



annealing/end-joining

B Alternative outcomes of twin priming

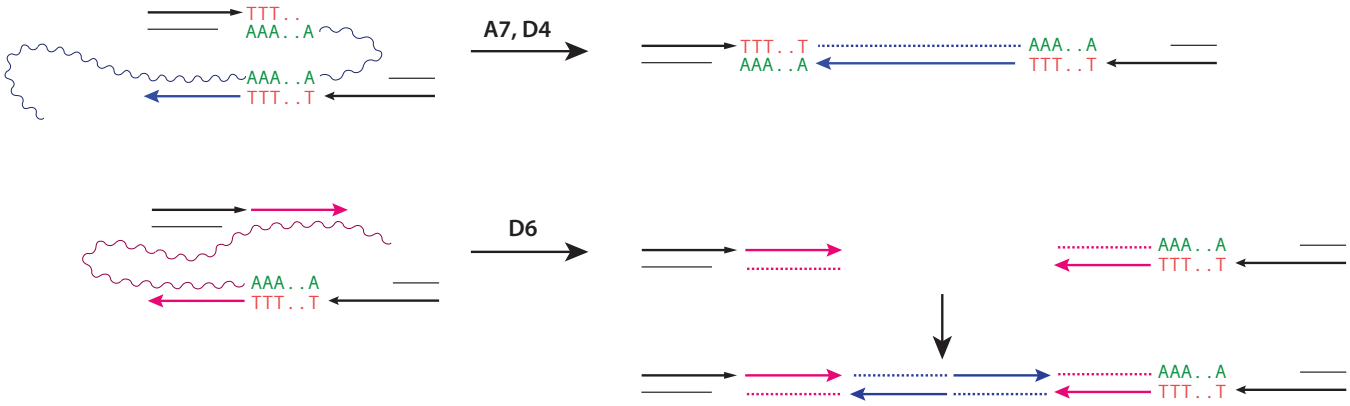
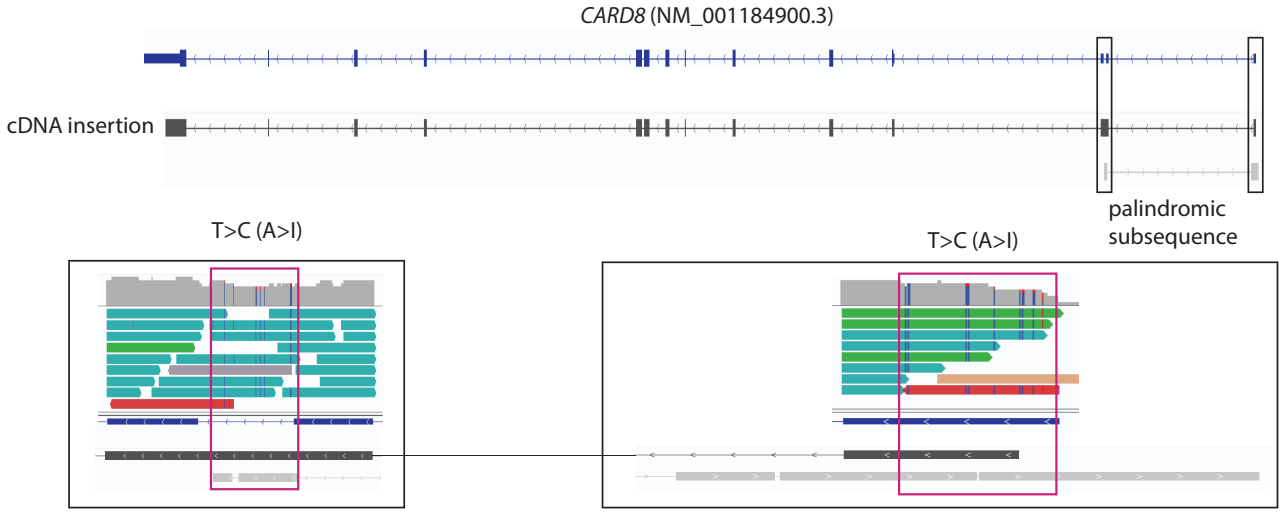
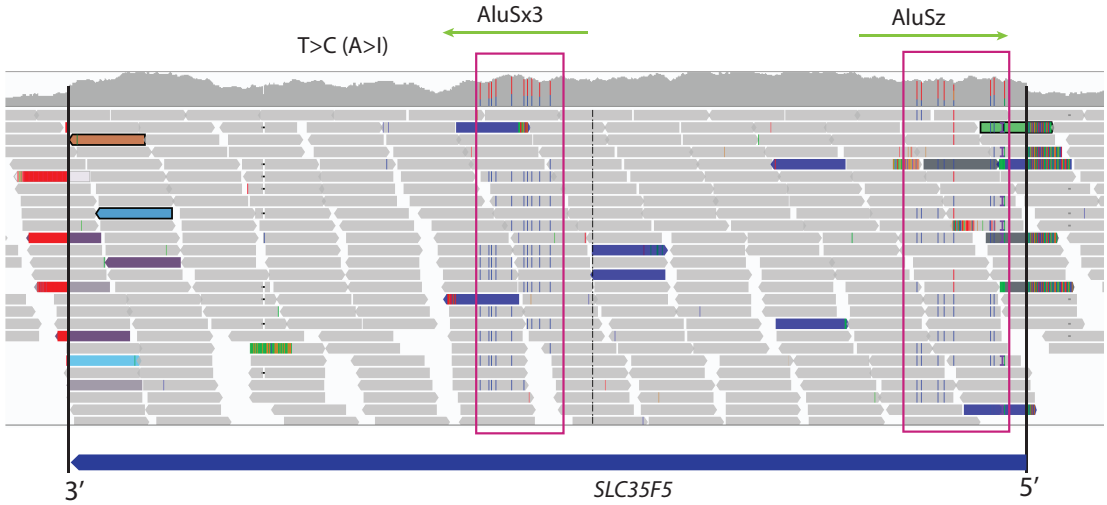


Figure S8 | Retrocopied pseudogene insertions containing substitutions due to ADAR editing. ADAR editing (A>I) results in T:A>C:G substitutions. All the examples are from **Dox clones** and are validated by long reads. In **A**, the clustered substitutions are restricted to a palindromic sequence near the 3'-end of an aberrantly spliced *CARD8* transcript (gray bars showing alignment of the inserted sequence). In **B**, the substitutions are restricted to a pair of inverted *Alu* sequences. In **C**, the substitutions are restricted to a single *Alu* sequence that likely forms a duplex with an inverted *Alu* that was not retrocopied due to incomplete reverse transcription. The insertion in **A** is identified at an insertion junction; the insertions in both **B** and **C** are identified at rearrangement junctions instead of insertion junctions.

A Insertion of a non-canonical transcript with a palindromic subsequence at chr19:34,282,753(-)/34282739(+) in Dox clone **A7**



B Insertion between junctions chr17:18595750(-) and chr16:89836576(-) in Dox clone **C3**



C Insertions between chr5:134,158,757(-) and chr18:48,623,479(-) in Dox clone **H8**

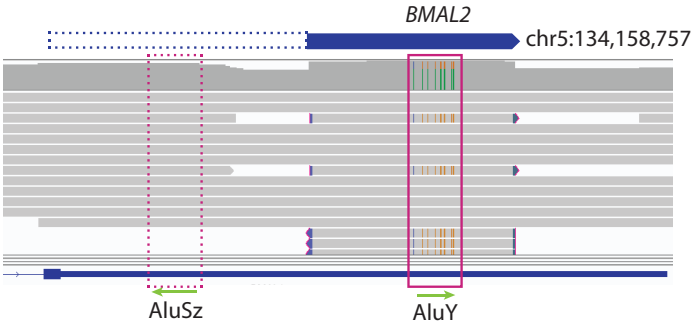


Figure S9 | (*Figure on next page.*) Insertions with inversions of genomic DNA sequences near the insertion site.

A. An inversion (chr2:201691621-2537) next to an L1 insertion between chr2:201691555(-) and chr2:201692543(+) identified in a **GFP+ clone**. A plausible mechanism for the inversion is shown below. Starting from the right: (i) ORF2p creates two sticky DNA ends with 3'-ends at 201692537 and 201692543 and extends 201692543(+) by RT; (ii) the reciprocal end (201692543) is resected, creating a long ssDNA overhang; (iii) the ssDNA overhang folds back to itself and forms a hairpin; (iv) cleavage of the top strand results in an inversion of the ssDNA overhang from the top strand to the bottom (highlighted by the gray arrow), creating a 5'-overhang; (v) after fill-in synthesis, the left DNA end joins the primary RT end to complete the insertion. The self-annealing step is supported by the observation of 14bp homology between the two breakpoints. By contrast, the junction between the L1 insertion and the inverted DNA end shows no homology.

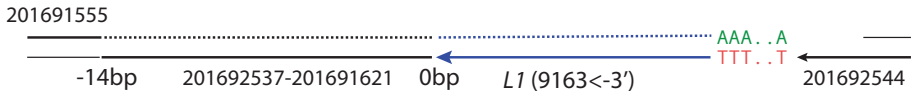
B. An inversion of genomic DNA next to a 5'-inverted L1 insertion. The inferred process is similar to **A**, except with the additional step of twin priming that gives rise to a secondary 5'-inverted insertion. The large gap between the two inverted RT sequences is distinct from most 5'-inverted insertions (see **Figure 2C**). The self-annealing step is supported by a 6bp homology (`TTGTTT`) between 24871375-1381 (reverse strand) and 24871508-1513 (forward strand).

C. An inverted duplication next to a L1 insertion. The first three steps (i-iii) of the inferred process are similar to **A** and **B**. The inverted duplication is generated by the displacement of newly synthesized DNA from the inverted 3'-end (red bases), which is then ligated to the RT DNA end.

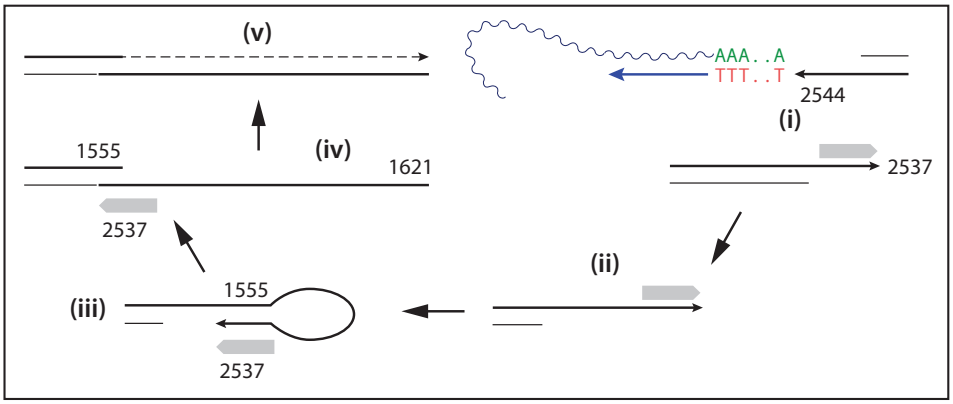
Insertions with inversions of genomic DNA sequences near the insertion site

A GFP+ clone H7: L1(truncated) plus inversion of sequence near the insertion site at of chr2:201692543/2537

insertion outcome

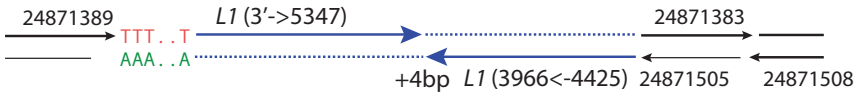


Inferred mechanism

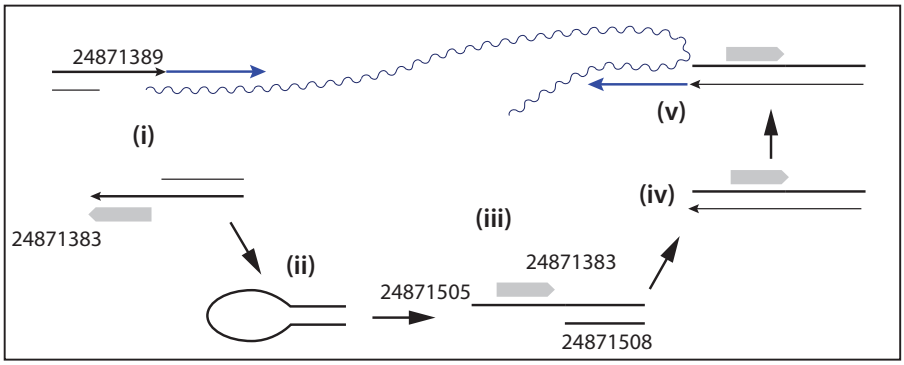


B Dox clone C8: 5′-inverted L1(truncated) plus inversion of sequence near the insertion site at chr10:24871389/1382

insertion outcome

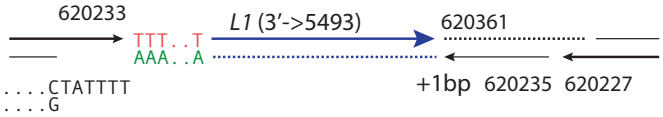


Inferred mechanism



C Dox clone E1: L1(truncated) plus inverted duplication of sequence near the insertion site at chr10:620233/227

insertion outcome



Inferred mechanism

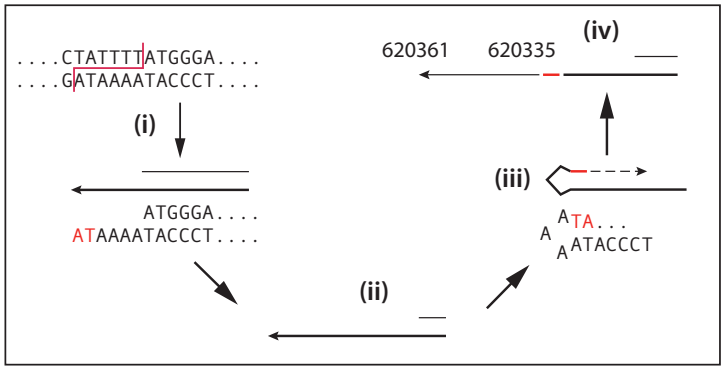


Figure S10 | Insertions containing both RT sequences and templated genomic sequences.

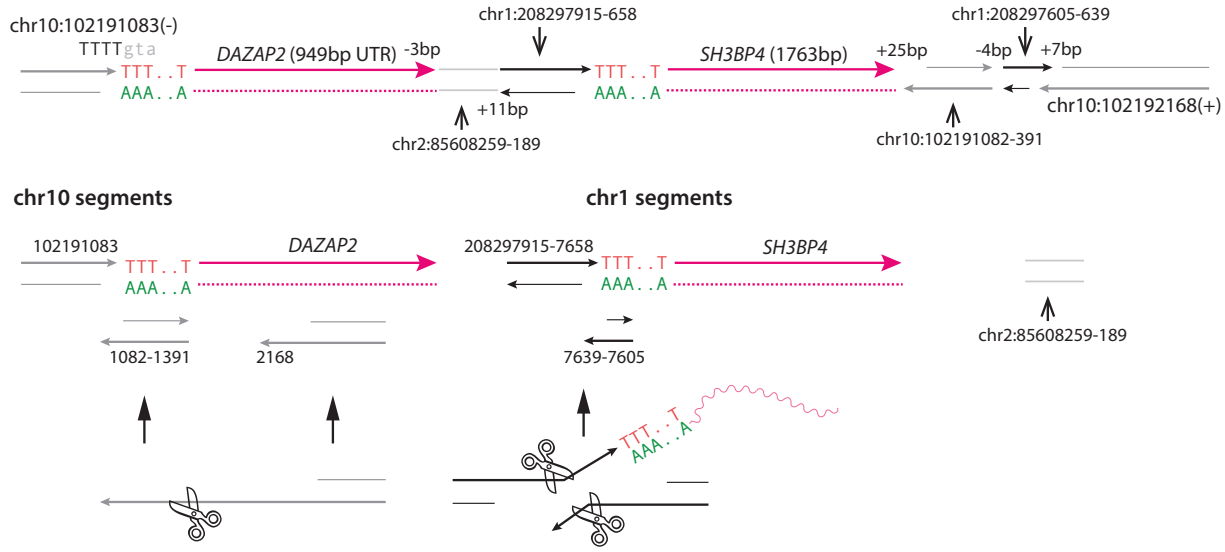
A. An insertion containing two retrocopied sequences and four genomic DNA sequences. The two sequences from chr1 (in black) are consistent with an origin from retrotransposition followed by cleavage of ssDNA fragments (also see **Figure 6**); the same mechanism can explain the short chr10 sequence (102191082-1391). The origin of the chr2 sequence cannot be determined.

B. An insertion junction containing four short sequences mapped to regions near the insertion site. The two pieces 12108205-8328 and 12108331-8467 reflect a deletion of one of two nearly identical tandem copies (capital letters for the retained sequence):
(12108274-8330) `CCATAATTGAAGCCCTTGGACAAAGTTTGTTTACTGTGATTTAAGATTTTGGTTAcT` and
(12108331-8387) `CCATAATTGAAGCCCTTGGACAAAGTaTGTTTACTGTGATTTAAGATTTTGGTTATT`.
This deletion may be explained by a slippage during DNA synthesis from a ssDNA template (12108205-8467).

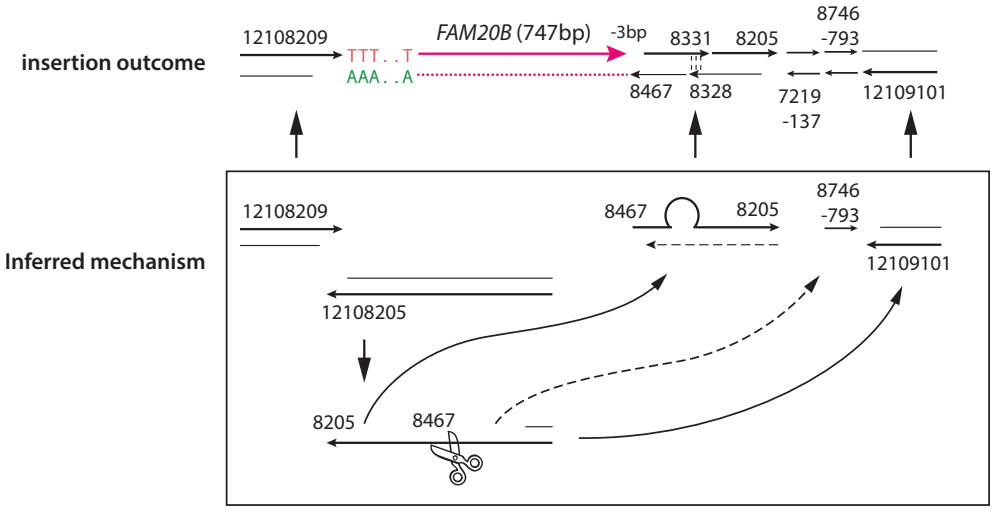
C. An insertion containing two genomic DNA sequences plus an L1 insertion. The L1 insertion is not completely resolved from short reads. The two genomic DNA insertions are mapped to regions on chr4 with local DNA fragmentation. See **Figure S17B**.

Insertions containing both RT sequences and genomic sequences originating from insertion sites/other breakpoints

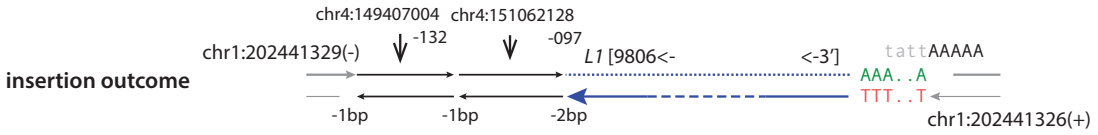
A **Dox clone H1**: Complex insertion at chr10:102191083/102192168



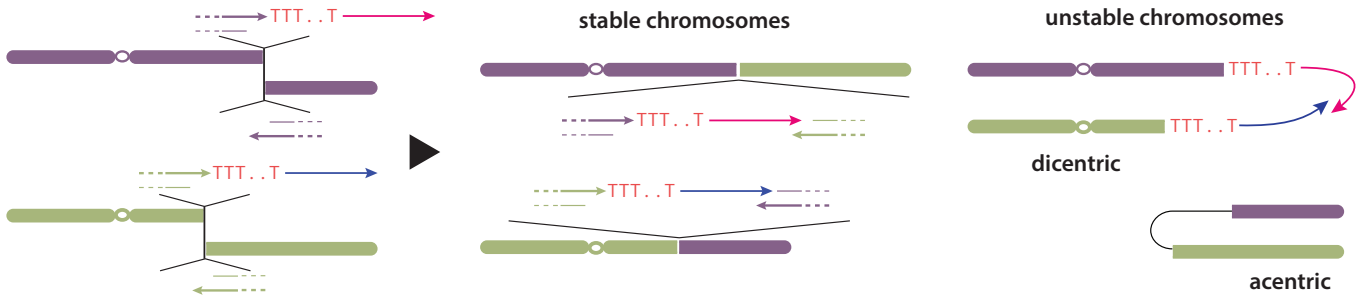
B **Dox clone C8**: Insertion of *FAM20B* plus local rearrangement at chr17:12108209/12108205



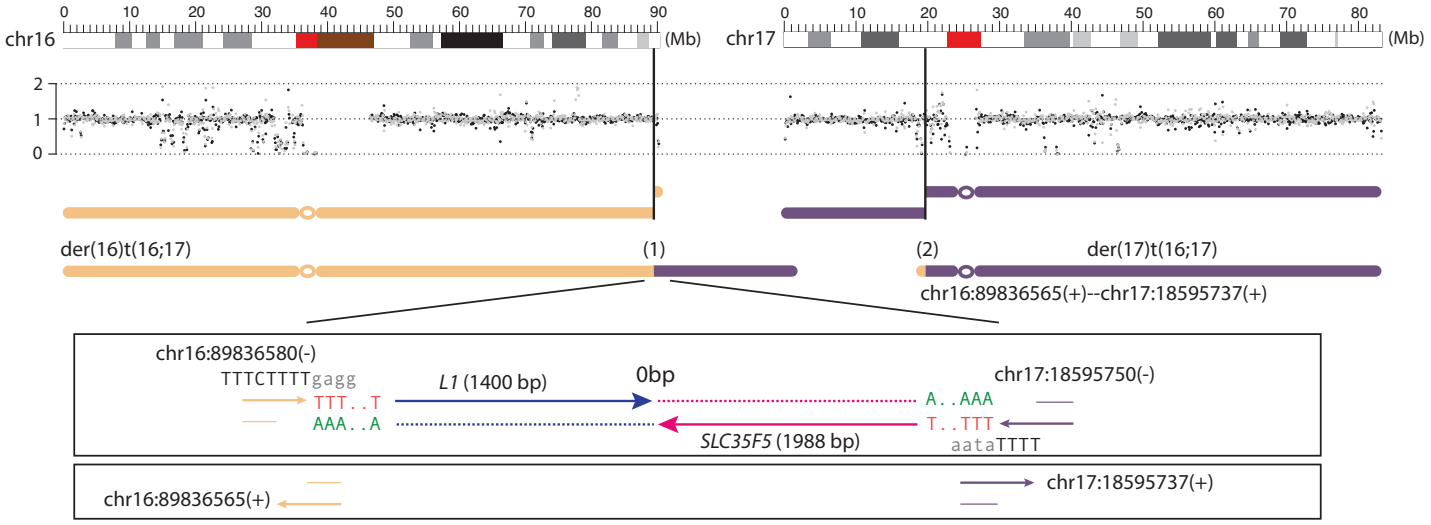
C **GFP+ clone H7:** Insertion of L1 and genomic DNA sequences at chr1:202441329/326



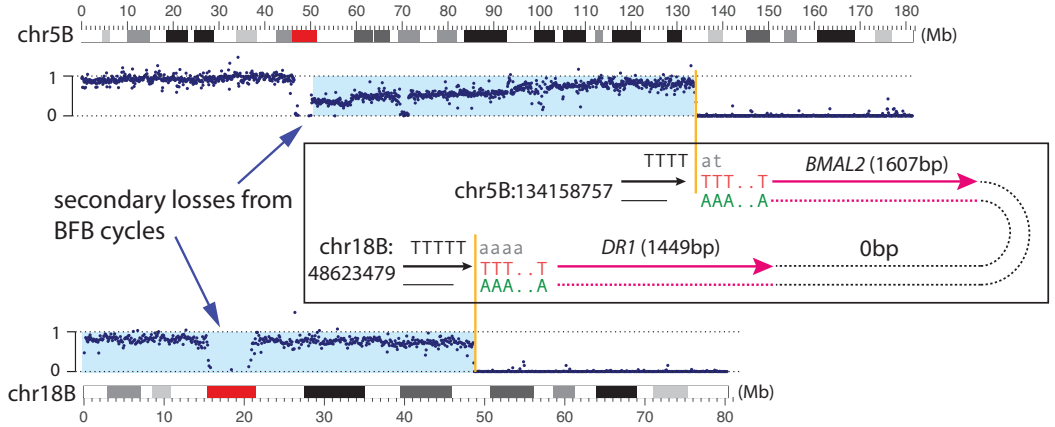
A Reciprocal translocations from end-joining between two DSBs generated by ORF2p-mediated RT



B Balanced translocations between chr16 and chr17 in Dox clone **C3**



C A dicentric chromosome **dic(5;18)** inferred to be present in Dox clone **H8**



D An acentric derivative chromosome inferred to present in a Dox-treated single cell **b1**

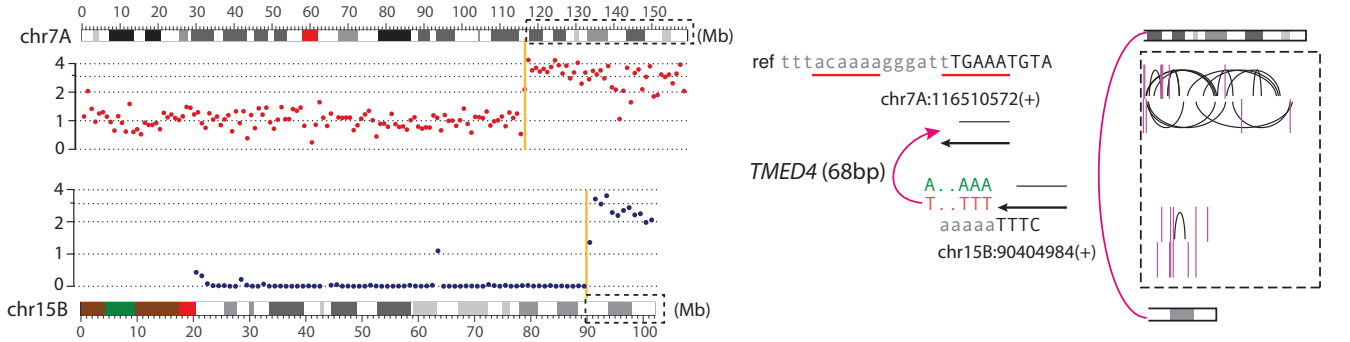


Figure 4 | Reciprocal translocations between DNA ends generated by L1 retrotransposition.

A. A schematic diagram of reciprocal translocations between DNA ends generated by independent retrotranspositions at two different loci. ORF2p generates two DNA ends at each locus: one end resides within an ORF2p EN cutting sequence and is extended by RT, hereafter referred to as the primary RT end; the other end has a partial overlap with the primary RT end and does not undergo RT except with twin priming; we refer to this end as the reciprocal end. Reciprocal translocations arise from a two-by-two exchange between two pairs of RT ends and reciprocal ends, and can generate either two stable chromosomes, or two unstable chromosomes.

B. An example of balanced translocations between chr16 and chr17 in the **Dox clone C3**. Gray and black dots represent normalized DNA copy number (90kb bins) of each parental haplotype, showing no copy-number alteration throughout each chromosome as expected for balanced translocations. Breakpoints at both loci [chr16:89836580(-)/6565(+) and chr17:18595750(-)/5737(+)] display TSD. The primary RT ends give rise to the breakpoints at chr16:89836580(-) and at chr17:18595750(-): both are within sequences suitable for ORF2p EN cutting and are connected with poly-A/T and truncated insertions, reflecting ORF2p mediated TPRT. Although both insertions are retained in one derivative chromosome der(16)t(16;17), we determine that the reciprocal translocation in der(17)t(16;17) is generated when the reciprocal ends are joined together. Therefore, both translocations are the direct outcome of L1 retrotransposition. Note: We have used '-' to denote breakpoints for which the 3'-end is on the forward strand, and '+' for breakpoints for which the 3'-end is on the reverse strand. For junction sequences, we use black, uppercase letters for nucleotides that are retained in the rearrangement junction and gray, lowercase letters for sequences in the reference.

C. An example of dicentric chromosome inferred to have been generated by L1-mediated translocation in the **Dox clone H8**. Shown are the haplotype-specific DNA copy number of the translocated homolog (5B and 18B). The inference of dic(5;18) is based on three pieces of evidence: (1) a single breakpoint on each arm of the translocated chromosome; (2) deletion of sequences telomeric to the breakpoints; and (3) subclonal loss of the dicentric chromosome, including segmental losses between the two centromeres (highlighted in blue), as expected for the chromosome-type breakage-fusion-bridge cycles. Both breakpoints are inferred to have descended from the primary RT ends based on TPRT signatures.

D. An example of acentric chromosome inferred to have been generated by L1-mediated translocation in a single cell **b1**. The inference of an acentric chromosome is based on (1) the structure of the segmental gain (on 7q) and retention (on 15q) based on haplotype-specific DNA copy number (7A and 15B); (2) the presence of intra- and inter-chromosomal rearrangements restricted to these two regions, indicating chromothripsis of an acentric chromosome partitioned in a micronucleus. The origin of the chr15 breakpoint from retrotransposition is directly established by the presence of a truncated pseudogene insertion, the poly-A/T sequence, and the ORF2p EN motif. The presence of two plausible ORF2p EN cutting sites (underlined) near the chr7 breakpoint suggests that this breakpoint could also have descended from a primary RT end that did not undergo RT or had the RT sequence cleaved.

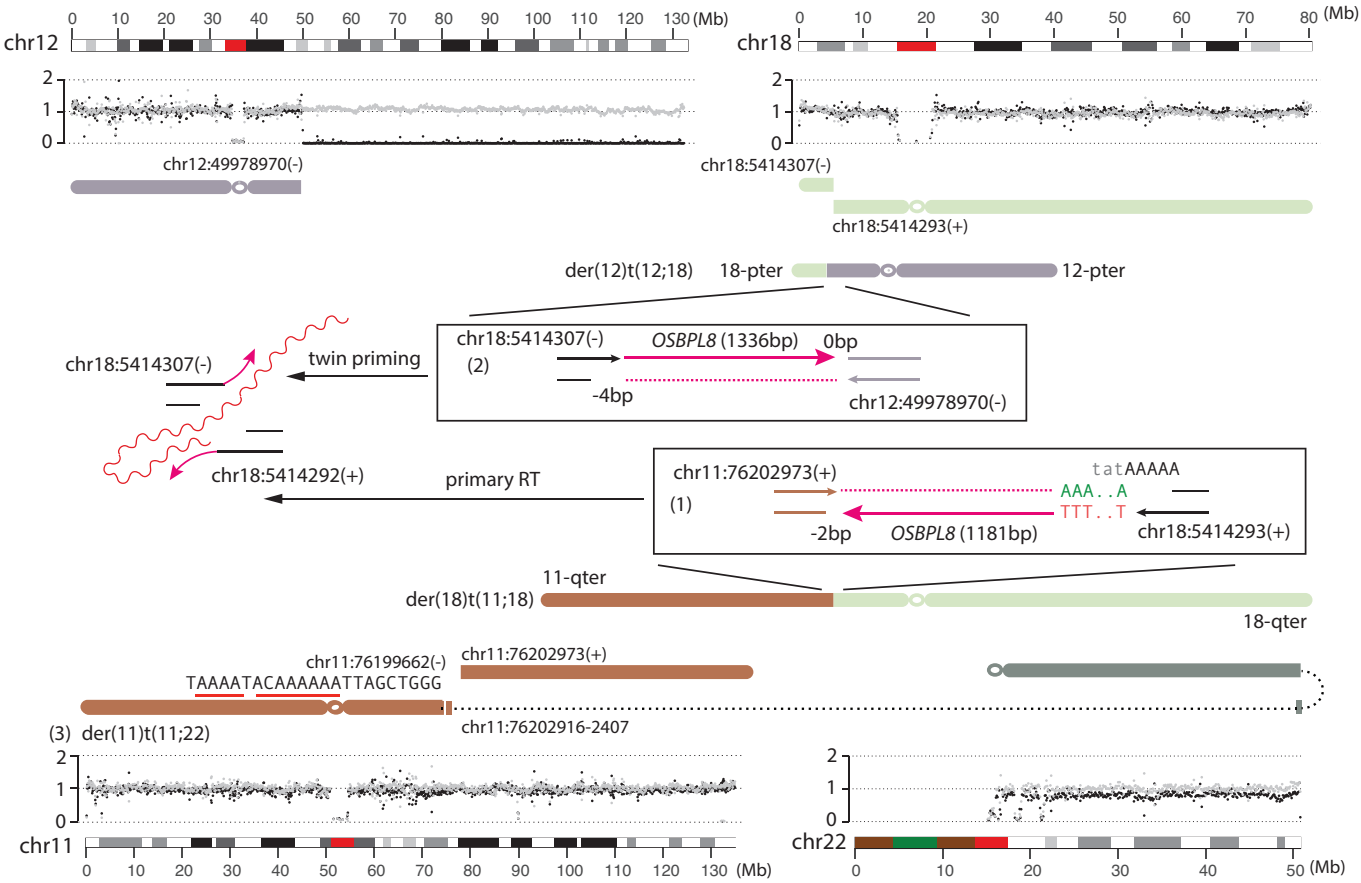
Figure S11 | (*Figure on next page.*) Additional examples of L1-mediated translocations.

A. Four-way translocations between chr11, chr18, chr12, and chr22. Copy-number data are similar to **Figure 4**. Two inverted RT sequences derived from the *OSBPL8* mRNA, chr12:76351797-2977 and chr12:76352990-4325 are identified at primary and twin-primed ends on chr18 [5414307(-)/5414292(+)], displaying all features of 5'-inverted retrotransposition except that the two ends are ligated to distal DNA ends. Although there is no evidence of retrotransposition at either breakpoint from chr11, the breakpoint at chr11:76199662(-) is adjacent to two plausible ORF2p EN cutting sites (underlined); it is therefore possible that this breakpoint originates from the reciprocal end generated by ORF2p. The breakpoint chr11:76202973(+) could have been generated by cleavage of the RT DNA as shown in **Figure S10A**; the short insertion mapped to chr11:76202407-916 could have been generated in the same event. The 11p arm is ligated to the chr22q terminus (with a small inverted terminal duplication, see **Figure 6C**), therefore forming a dicentric chromosome; there is subclonal copy-number loss between the two centromeres (black dots).

B. A dicentric chromosome inferred to be present in **Dox clone H8** with two translocations between chr10, chr4, and chr8. The junction between the 4q breakpoint [chr4:136923571(+)] and [chr8:37006129(+)] contains two independent RT insertions, indicating both breakpoints originating as primary ends generated by ORF2p. The reciprocal breakpoint chr4:136923575(-) (see IGV screenshot of a 5bp target-site duplication) originates as the reciprocal end and is extended by RT using a different mRNA (*EIF4G2*). The extension of both DNA ends generated by ORF2p is essentially the same as shown in **Figure 3B**, except that the two ends are ligated to distal DNA ends. Finally, the breakpoint at chr4:140628115(-) is to the right of a plausible ORF2p EN cutting site (`TAAAT|gac`), suggesting a plausible origin as the reciprocal end generated by ORF2p. Taken together, all four breakpoints (three on chr4 and one on chr8) are directly attributed to retrotransposition.

The 4p breakpoint chr4:39911656(+) is extended by RT of the *BCL2L2* mRNA from within the 3'-UTR (no poly-A); it also does not have any adjacent ORF2p EN cutting site. This junction could result from reverse transcription at a DNA end generated independent of ORF2p.

A Translocations between chr11, chr18, chr12 and chr22 in **Dox clone A5**



B **dic(4;8)** with **translocations to chr10** inferred to be present in **Dox clone H8**

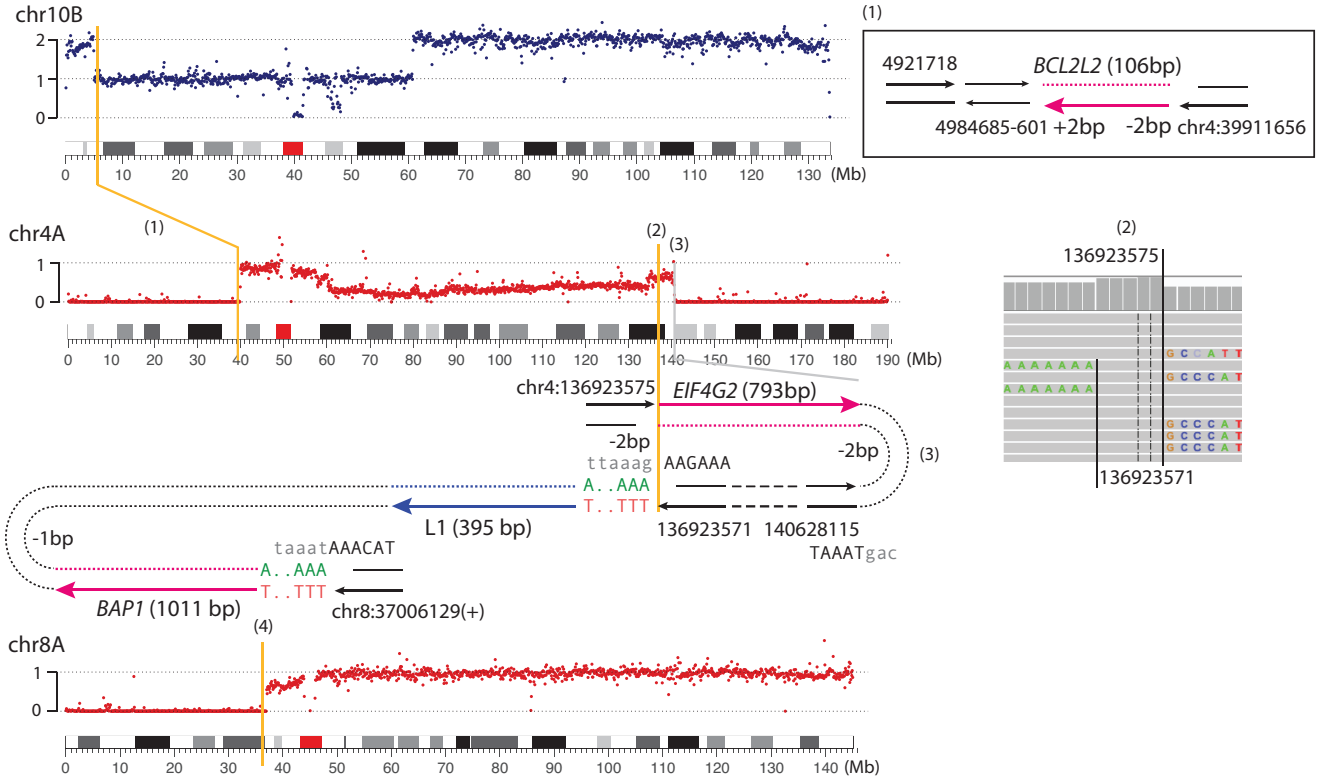


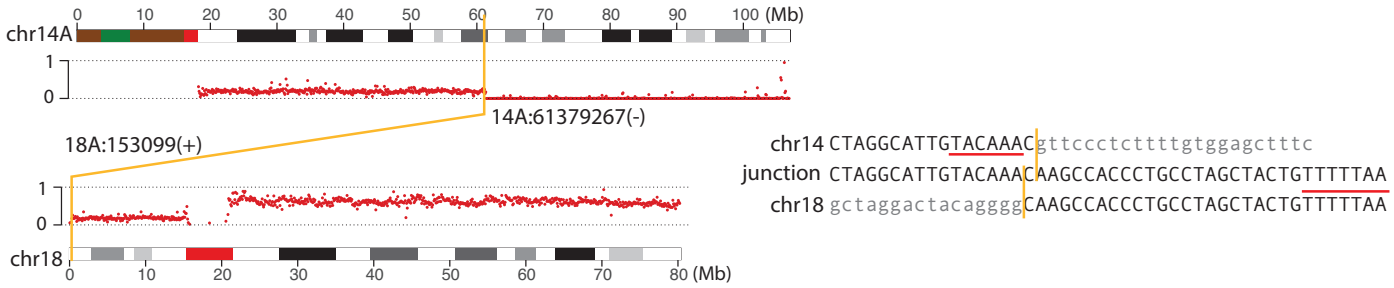
Figure S12 | Dicentric chromosomes in **Dox clones**. The inference of dicentric chromosomes is based on similar evidence as discussed in **Figure 4C**.

A. A dicentric chromosome inferred to be present in **Dox clone A6**. Both breakpoints are adjacent to ORF2p EN cutting sites (underlined/overlined) consistent with these breakpoints originating as the reciprocal DNA ends generated by ORF2p.

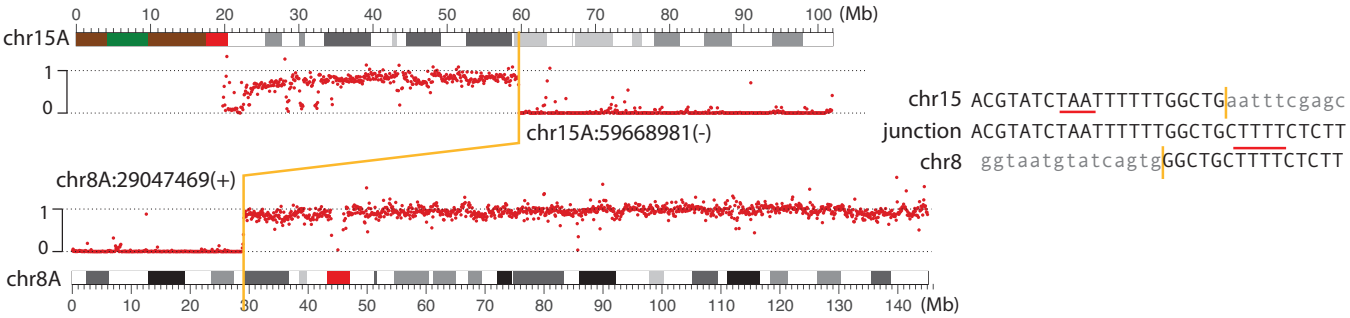
B. Another example similar to **A**.

C. A dicentric chromosome with both translocation breakpoints adjacent to ORF2p EN cutting sites. Notably, the homeology between the junction sequence and the genomic DNA sequence suggests that the joining between DNA ends involves some form of error-prone DNA synthesis.

A **dic(14;18)** inferred to be present in **Dox clone A6**



B **dic(8;15)** inferred to be present in **Dox clone C7**



C **dic(9;10)** inferred to be present in **Dox clone A3**

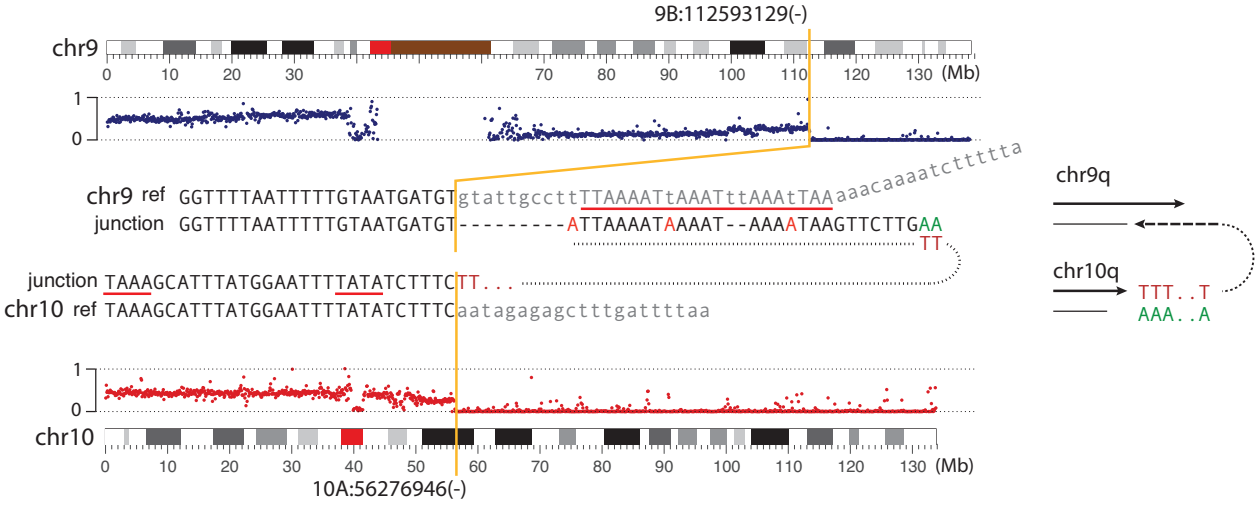
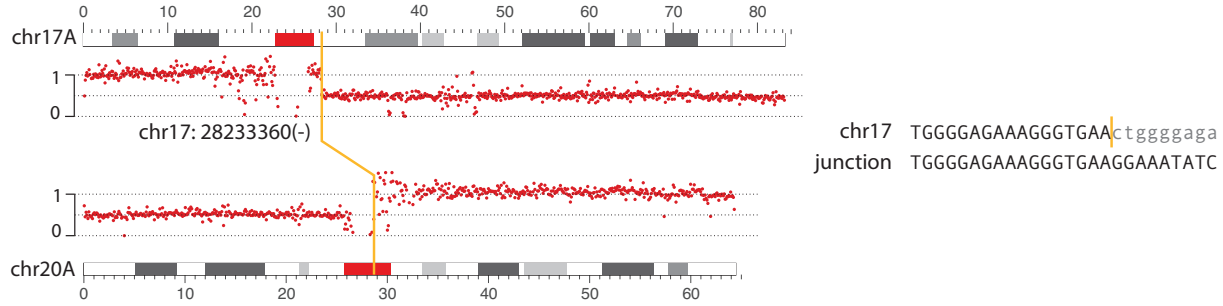
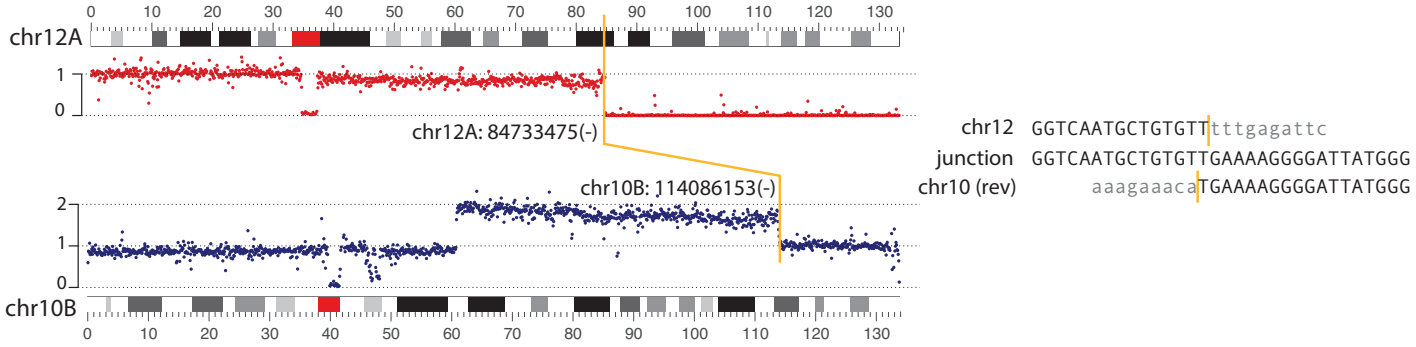


Figure S13 | Dicentric chromosomes in GFP+ clones. The **C3** clone is derived from a tetraploid ancestor; the **E8** and **F2** are derived from diploid ancestors. As none of the breakpoints is adjacent to an ORF2p EN cutting site, these translocations may have arisen independent of or downstream of ancestral DNA breaks generated by retrotransposition.

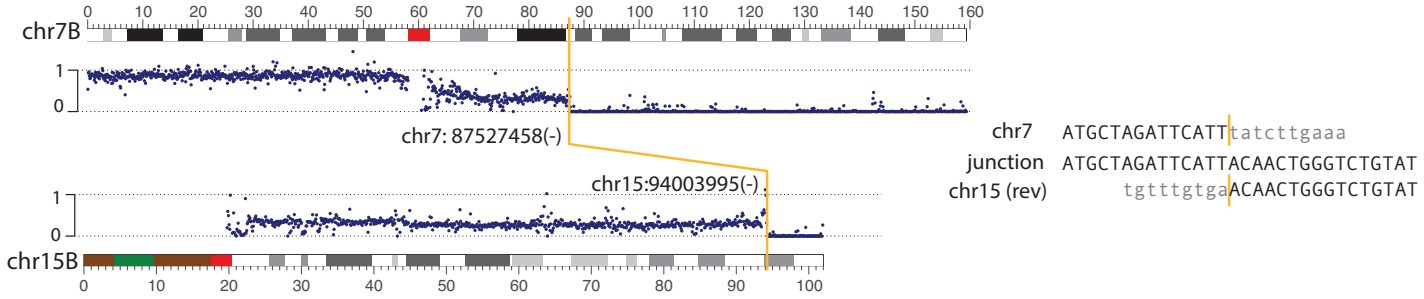
der(17)t(17;20) or **dic(17;20)** inferred to be present in **GFP+ clone C3**



dic(10;12) inferred to be present in **GFP+ clone E8**

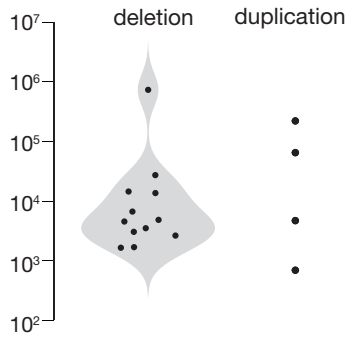


dic(7;15) inferred to be present in **GFP+ clone F2**

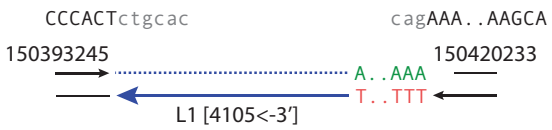


Segmental copy-number alterations from retrotransposition-mediated rearrangements

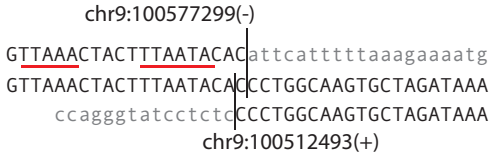
A Length distribution of deletions/duplications (<1Mb)



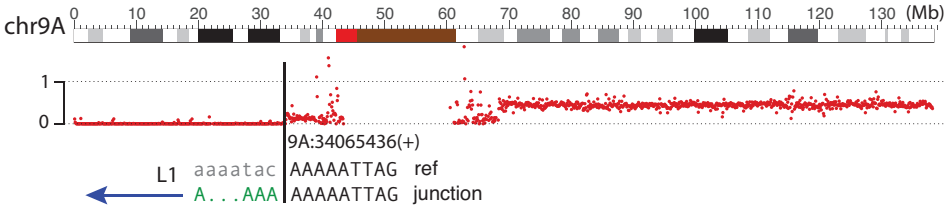
B 27kb deletion on chr5 in Dox **clone H1**



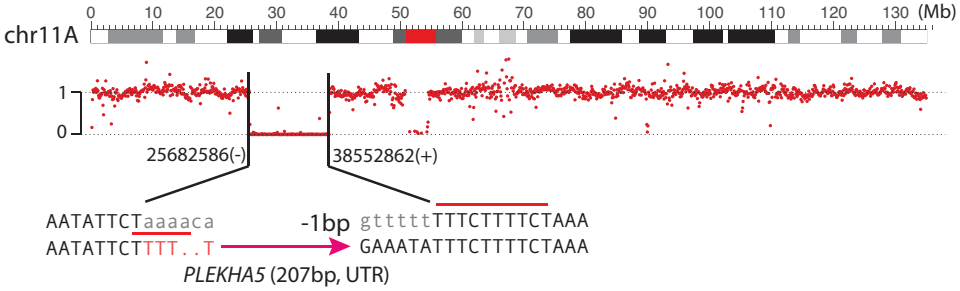
C 64kb tandem duplication on chr9 in Dox **clone C4**



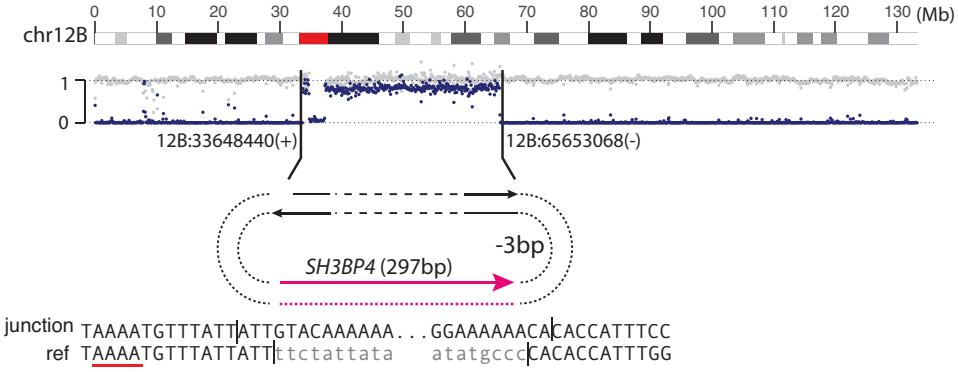
D Terminal deletion on chr9 in Dox **clone A8**



E 13Mb paracentric deletion on chr12p with cDNA insertion in Dox clone **G5**



F Pericentric fusion of chr12 generating an unstable ring chromosome in Dox **clone D3**



G 16q-terminal duplication (70Mb-qter.) in **GFP+ clone E6**

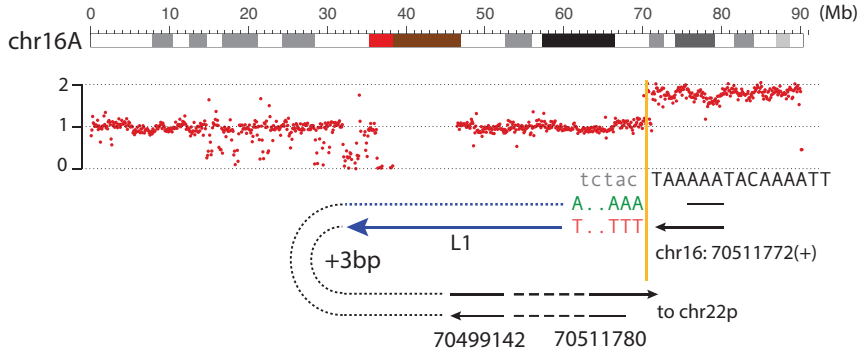


Figure 5 | Segmental copy-number alterations resulting from retrotransposition-induced chromosomal rearrangement.

A. Size distribution of sub-megabase segmental deletions and duplications.

B. An example of L1-mediated short (27kb) deletion.

C. A 64kb tandem duplication with one breakpoint (chr9:100577299) adjacent to two ORF2p EN cutting sites.

D-G. Examples of Large segmental CNAs **D** A 9p-terminal deletion with an L1 insertion junction. The complete junction is undetermined.

E. A 13Mb internal deletion on 12p with an insertion of the *PLEKHA5* cDNA joining two breakpoints both with adjacent ORF2p EN cutting sequences.

F. A junction containing an insertion of the *SH3BP* cDNA between breakpoints on 12p and 12q that results in a ring chromosome. The inference of a ring chromosome is supported by the subclonal loss of this chromosome (relative to the intact homolog shown in gray), in contrast to the preservation of chromosomes with large internal deletions in **B** and **C**.

G. A large terminal duplication generated by retrotransposition. Two breakpoints are detected near the duplication boundary: the primary RT end gives rise to the breakpoint at chr16:70511772(+), whereas the reciprocal end gives rise to the breakpoint at chr16:70511780(-). The reciprocal breakpoint is retained in a 12kb DNA sequence that is inserted at the translocation junction between the reverse transcribed L1 and the translocation partner.

Figure S14 | Additional examples of large segmental deletions in Dox clones and GFP+ clones.

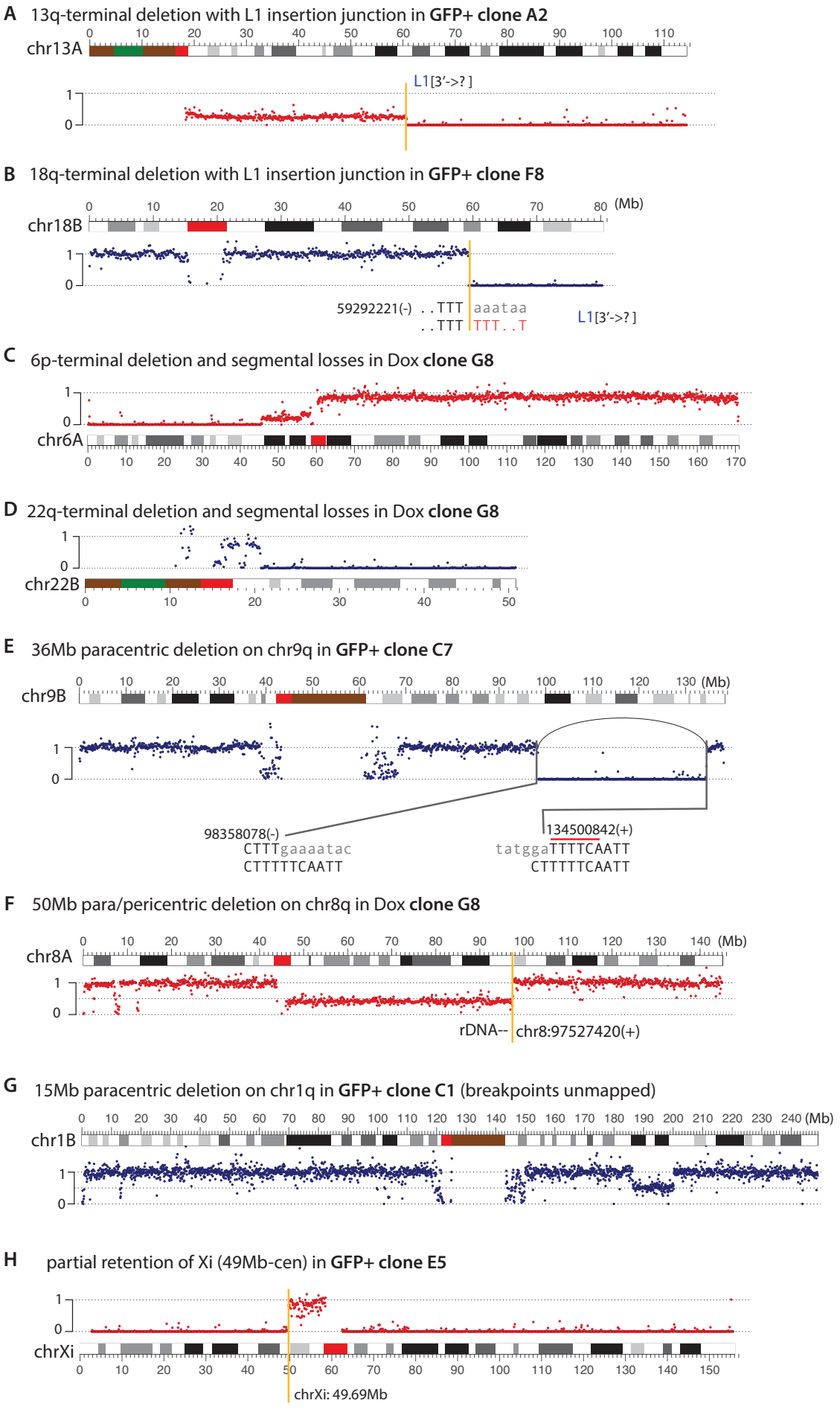
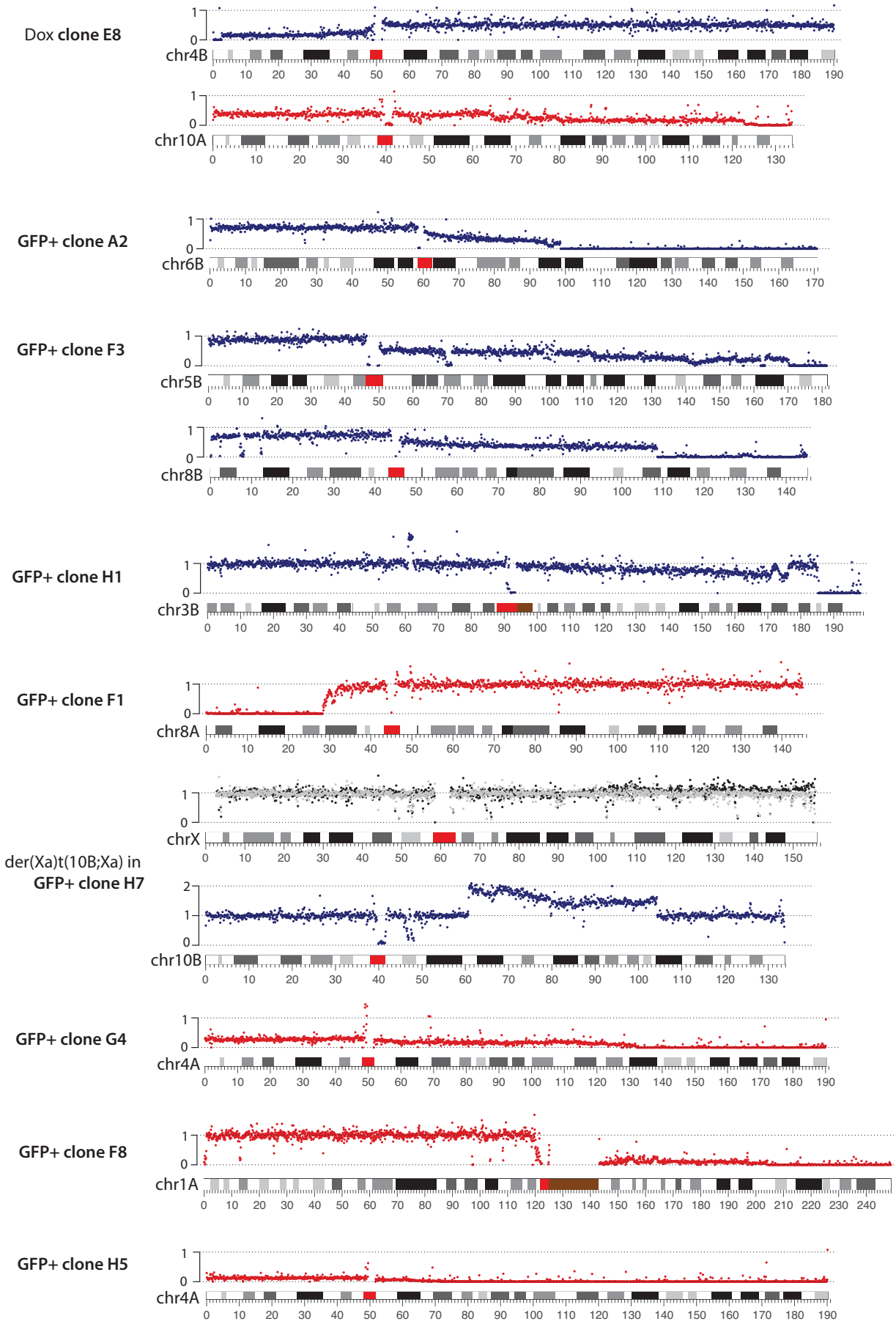
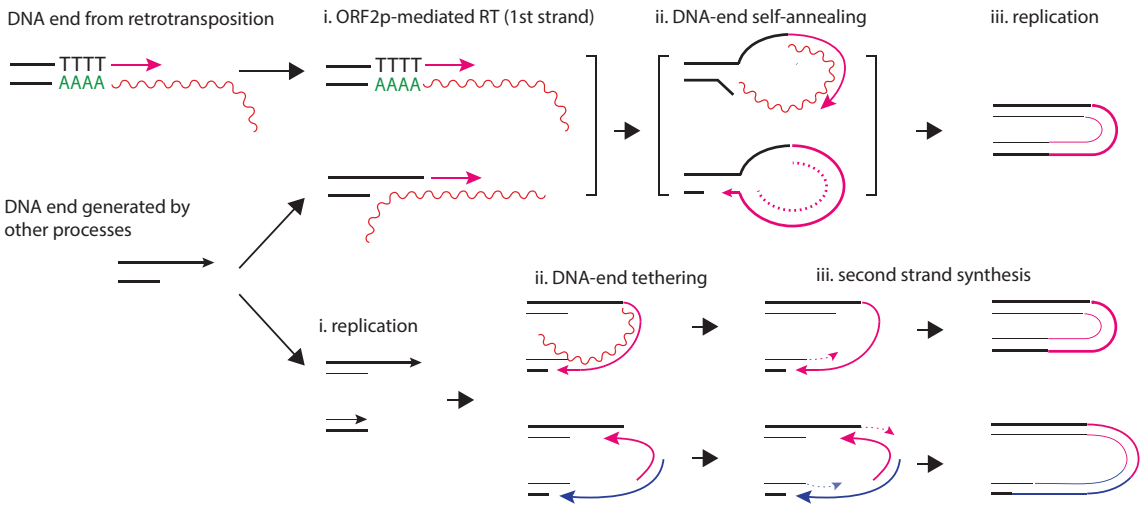


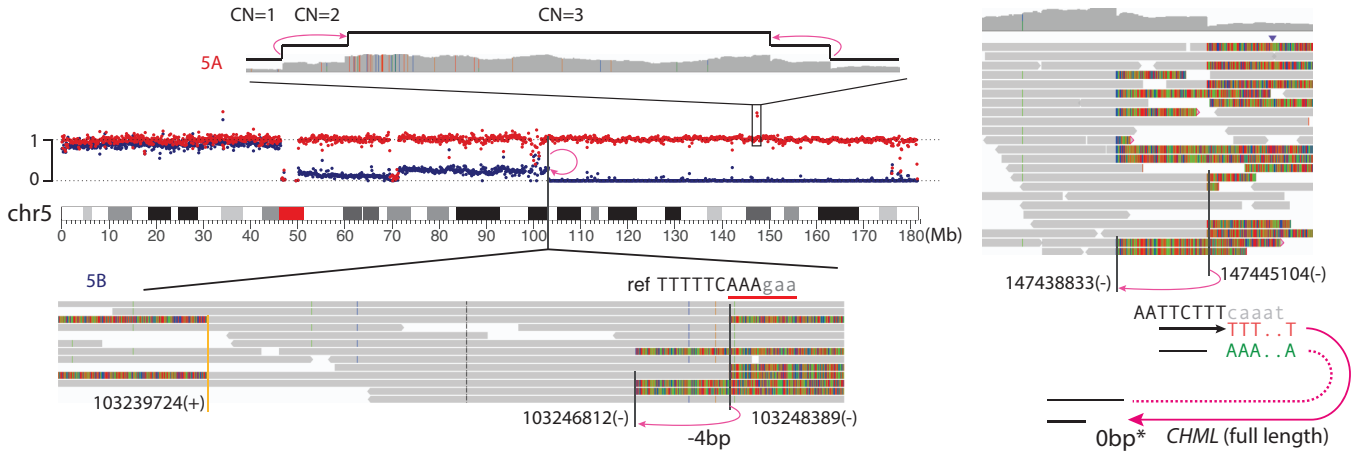
Figure S15 | Examples of sloping copy-number variation indicating BFB cycles in Dox clones and GFP+ clones. For the example in **GFP+ clone H7**, the sloping copy-number variation is on the extra copy of the 10q segment that is appended to the active X. Note the minor copy-number gain in Xa (black dots) but copy-number losses in the 10q arm relative to trisomy in the parental line.



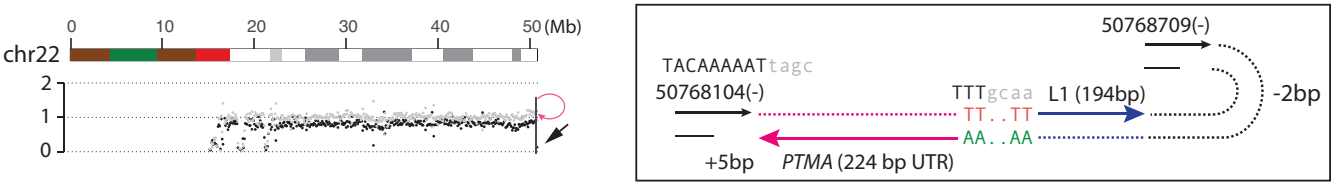
A Processes that can generate foldback junctions with insertions due to retrotransposition



B Foldback junctions on on chr5 in **Dox clone C3**



C A foldback junction containing a **pseudogene insertion** and a **insertion** on chr22q in **Dox clone A5**



D A foldback junction on chr1q in **Dox clone H1** inferred to have been a downstream consequence of retrotransposition

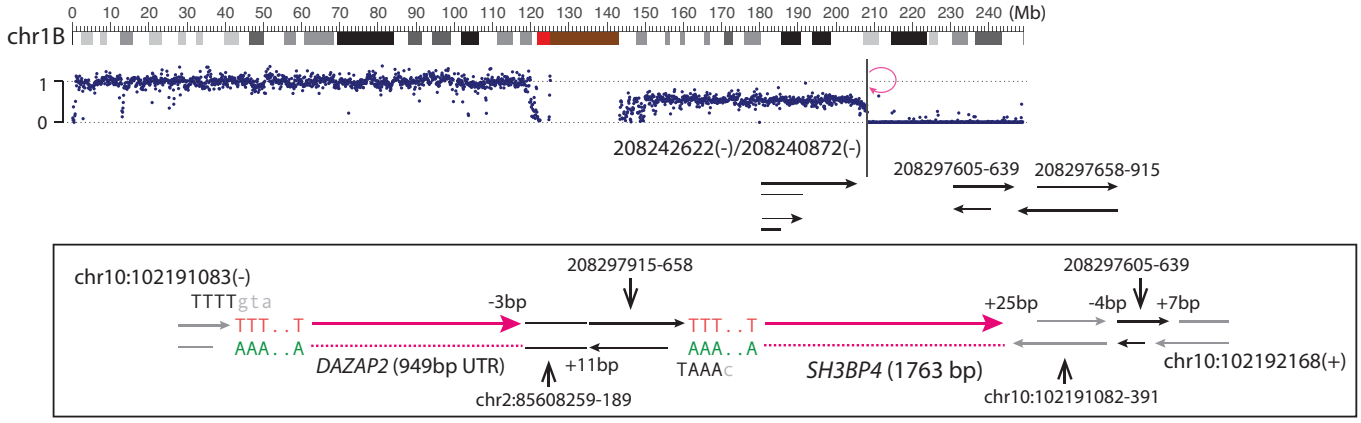


Figure 6 | Foldback junctions with retrotransposition insertions.

A. Two processes that can generate foldback junctions with insertions of reverse transcribed sequences.

Top: In the first row, a DNA end is generated and extended by ORF2p; in the second row, a DNA end generated independent of retrotransposition is extended by ORF2p using the RNA template (wiggly line). In both scenarios, ligation between the extended 3'-DNA end (red solid arrow) and the 5'-end on the complementary strand can be initiated by mRNA tethering or by microhomology-mediated self-annealing. The ssDNA ligation creates a hairpin, which can be converted into a foldback junction by DNA replication. Note that a similar mechanism can produce a foldback junction at the reciprocal DNA end generated by ORF2p.

Bottom: A foldback junction with retrotransposition insertions can also arise when two replicated DNA ends are tethered by a RNA (wiggly line) or cDNA (solid lines with arrowheads), with second-strand synthesis completed by ORF2p or DNA polymerases.

B. Two examples of foldbacks on chr5q in **Dox clone C3** that are related to retrotransposition.

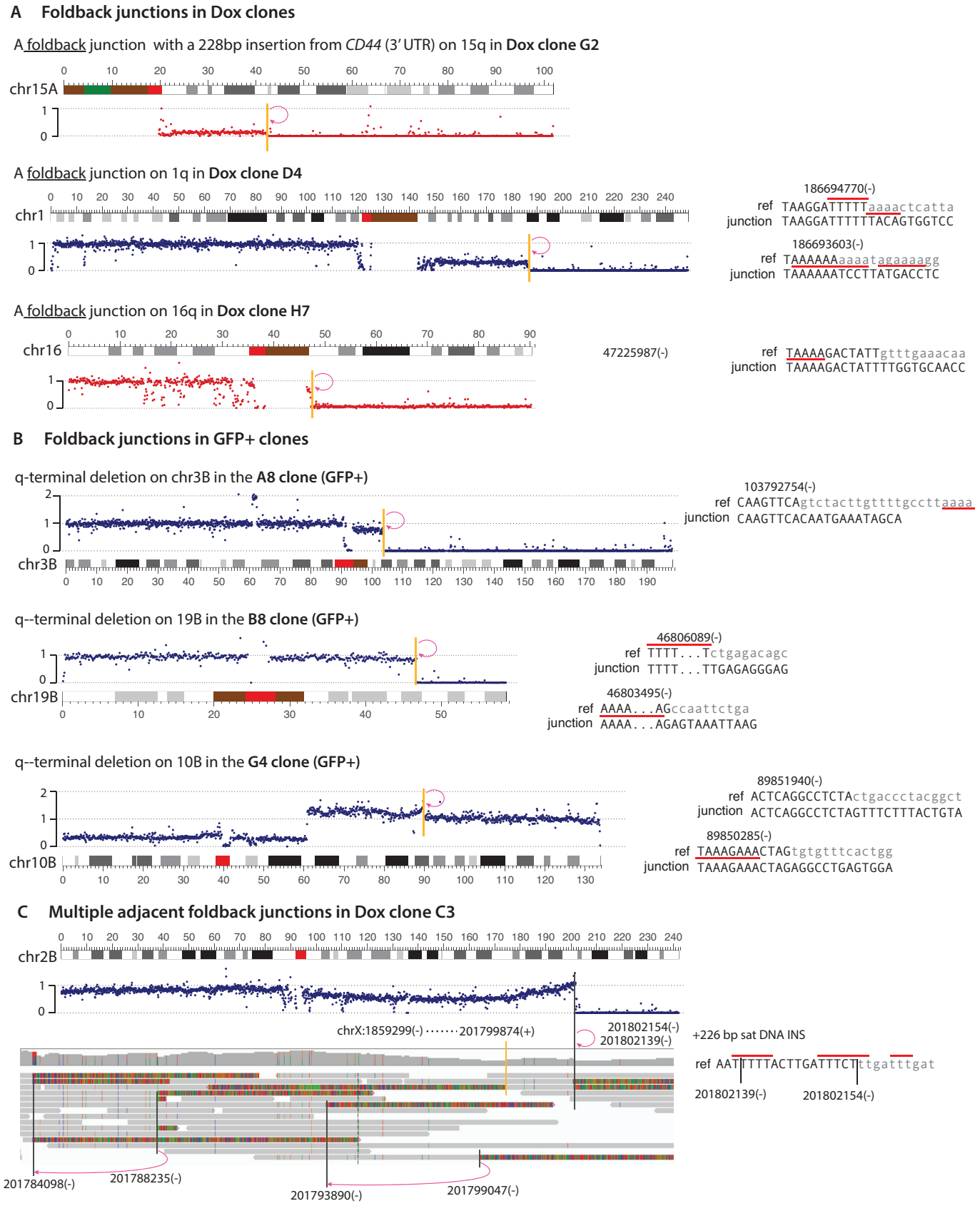
On chr5A (copy number shown in red), there is a foldback junction between chr5:147445104(-) and 147438833(-) that contains a full-length insertion of the *CHML* cDNA (7.7kb). The breakpoint at 147445104 is located in an ORF2p EN cutting site (`TTT|cAAA`) and extended from the 3'-end of the *CHML* transcript with poly-A. Although there is no apparent microhomology between 147438833(-) and the 5'-end of the inserted *CHML* sequence (0bp*), there is a 2bp microhomology (`TG`) when including the extra 'G' base from mRNA capping.

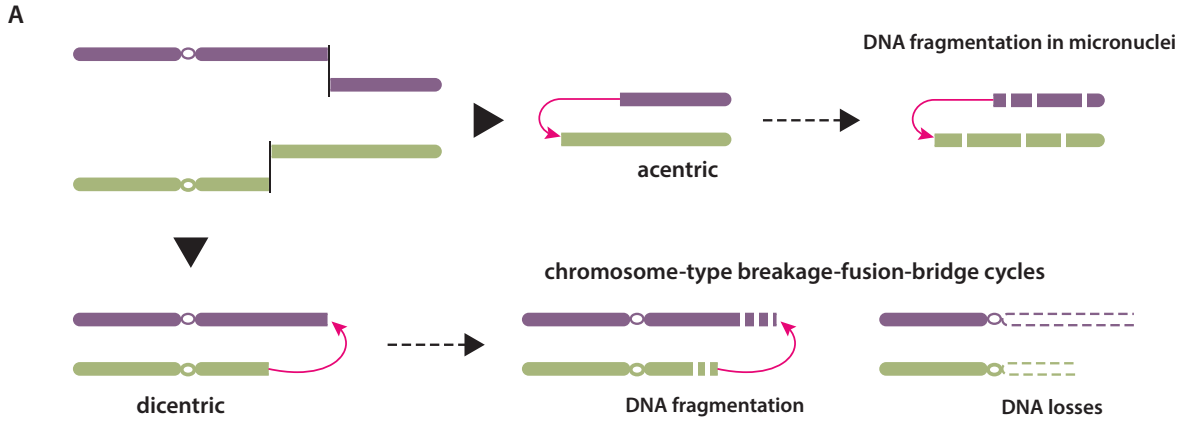
On chr5B (copy number shown in blue), there is a foldback junction between chr5:103248389(-) and 103246812(-). Although the junction does not contain any insertion, breakpoint 103248389(-) is located within an ORF2p EN substrate (TTTTTcAAA|gAA) and may have descended from a reciprocal end generated by ORF2p EN. The ancestral DNA end may be near blunt but becomes staggered after 5'-resection; the resected 5'-end produces the breakpoint at 103246812(-) that is then ligated to the breakpoint at 103248389 to create the foldback junction.

C. The foldback junction at the end of chr22 in **Dox clone A5** clone as shown in **Figure 4**. The presence of two insertions in the foldback junction is consistent with the model depicted in the last row of panel **A** and also similar to the complex insertion junction shown in **Figure S7A**.

D. An example of foldback junction at the end of chr1q in **Dox clone D4** clone that is inferred to be a downstream consequence of retrotransposition. Although the junction between two breakpoints chr1:208242622(-) and 208240872(-) does not contain any insertion, the identification of two short DNA sequences (chr1:208297605-639 and chr1:208297658-915) at another insertion junction (**Figure S10A**) indicates an ancestral DNA breakage due to retrotransposition: We infer that these two short DNA fragments originate from ssDNA fragments that are cleaved from a retrotransposition intermediate, which produces two dsDNA ends without any footprint of retrotransposition. One of the dsDNA end subsequently gives rise to the pair of DNA ends in the foldback junction. Notably, both short DNA pieces and the cDNA of a truncated transcript of the *SH3BP4* gene are inserted into a complex insertion junction on chr10 containing another truncated pseudogene insertion (*DAZAP2*); this observation highlights the dynamic interplay between retrotransposition and endogenous DNA repair.

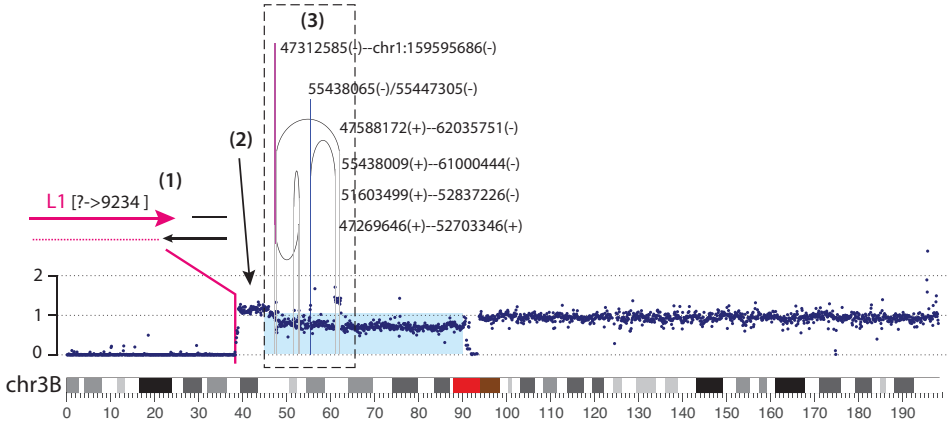
Figure S16 | Additional instances of foldback junctions in **Dox clones** (**A**) and **GFP+ clones** (**B**). **C**. An example of multiple adjacent foldback junctions. Potential ORF2p EN cutting sites near the breakpoints are highlighted.





B Complex rearrangements on the p-arm of chr3B in the **F8 clone (GFP+)**

RT-mediated translocation: 1. p-terminal deletion (0-38.5Mb);
BFB cycles and downstream evolution: 2. terminal duplication; 3. regional chromothripsis; 4. subclonal DNA losses



C Complex rearrangements on chr6B in the **F8 clone (GFP+)**

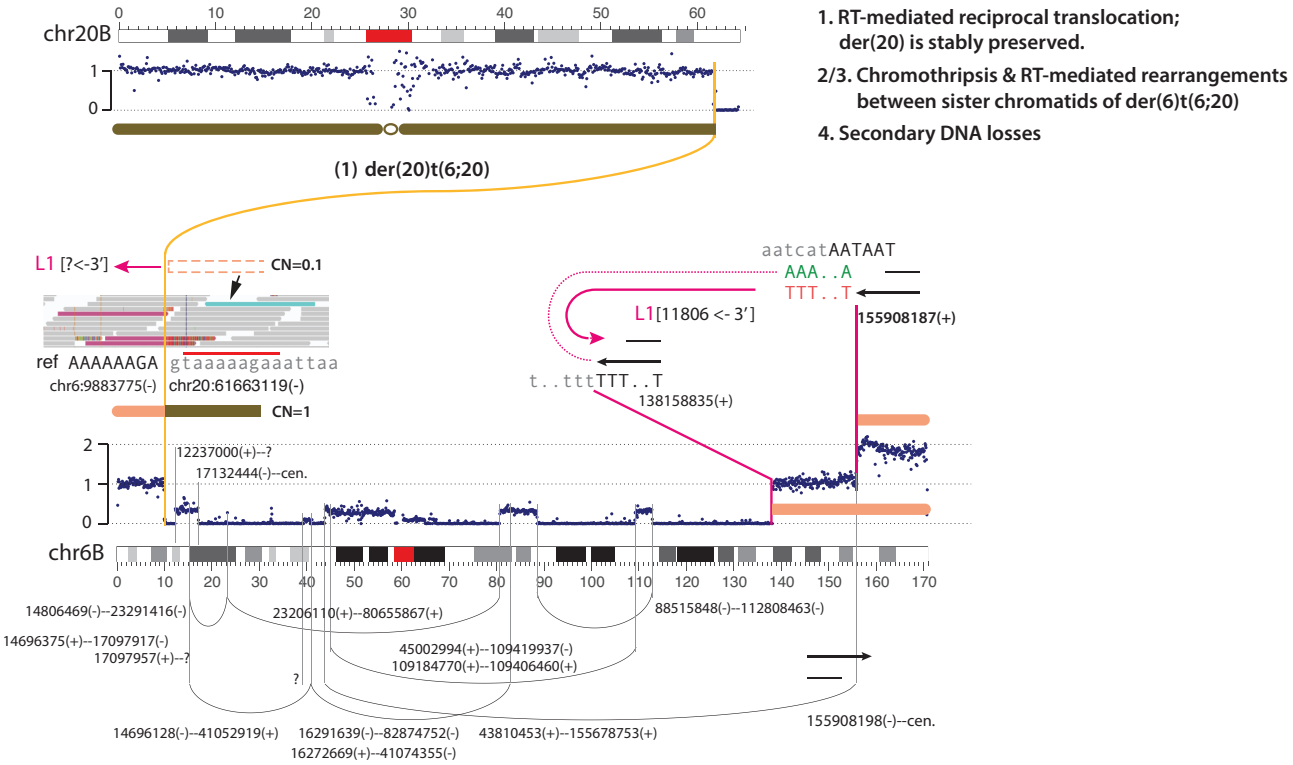


Figure 7 | Chromothripsis and retrotransposition.

A. Two mechanisms by which unstable chromosomes generated by translocations can lead to chromothripsis.

B. An example of complex rearrangements detected in the **GFP+ clone F8**. A retrotransposition-mediated translocation leads to the p-terminal deletion, followed by one or multiple BFB cycles, creating complex rearrangements on the 3p arm. The L1-mediated translocation is inferred to be the initiating event of complex rearrangements.

C. An example of chromothripsis in the same sample as in **B**. Two junctions contain L1 insertions. The first on the p-terminus is inferred to have generated reciprocal translocations between chr6B and chr20B. The second junction joins two distal breakpoints on the q-arm [138158835(+) and 155908187(+)]; based on the duplication of the q-terminal segments, we infer the two breakpoints to originate from DNA ends on sister chromatids. We identify a breakpoint [155908198(-)] that descends from the reciprocal end of the RT end [155908187(+)] with a 12bp TSD; the TSD feature suggests that the ORF2p created two DSB ends in this chromatid.

Figure S17 | Additional instances of chromothripsis in **GFP+ clones**.

A. An example of complex rearrangement/CNAs consistent with the outcome of a BFB cycle leading to regional chromothripsis. L1 insertions are found at three junctions with the following breakpoints: 217713213(+), 217715699(-), and 221044734(-).

B. A similar example as in **A** identified in a different clone. L1 insertions are found at three junctions involving the following breakpoints: 144917551(+), 148637312(+), 148748795(-). These breakpoints and their reciprocal breakpoints are highlighted in bold. We also identify a junction involving the breakpoint chr4:128140824(-) that is joined to a poly-T sequence.

C. Three examples of chromothripsis indicated by oscillating DNA deletion and retention.

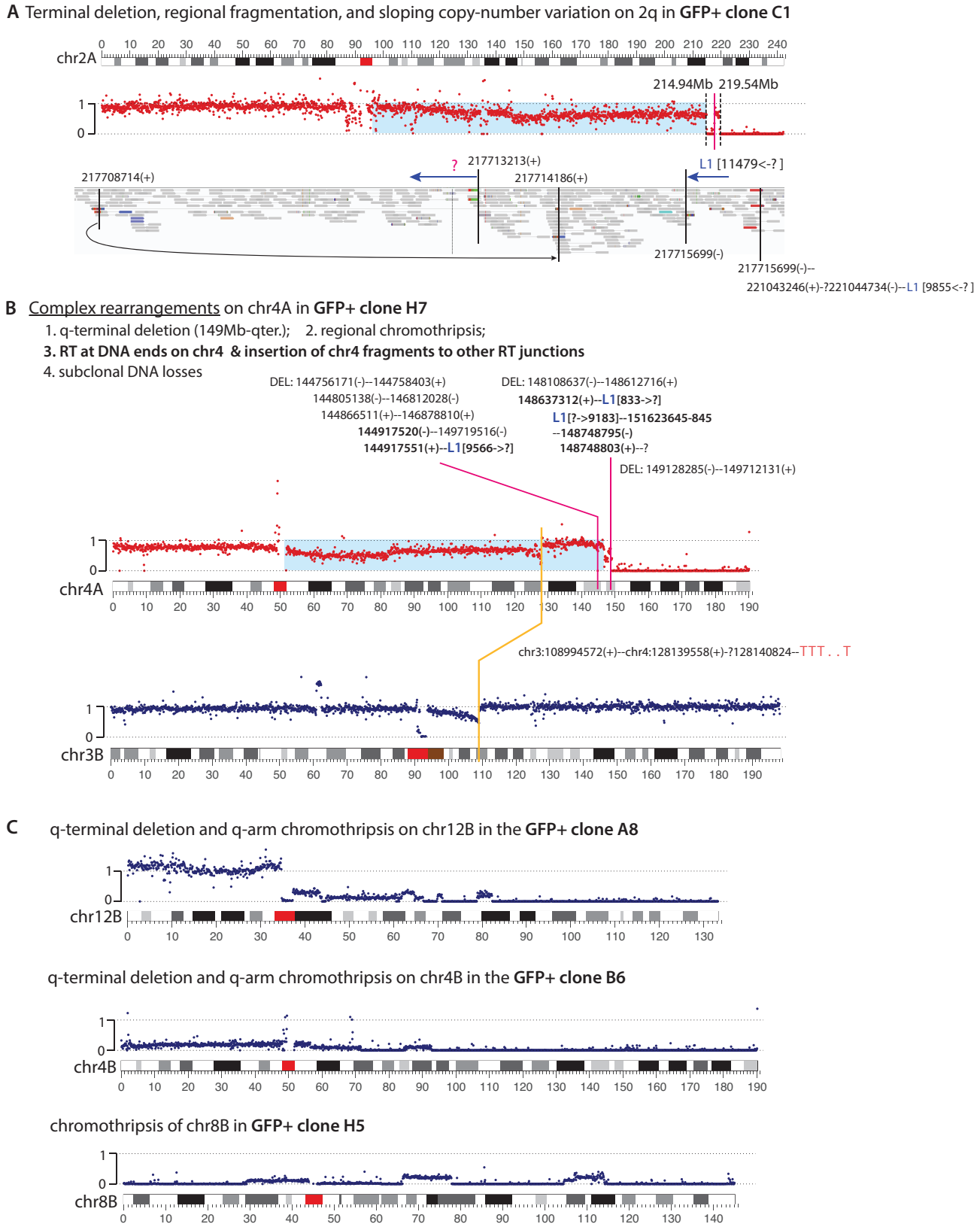
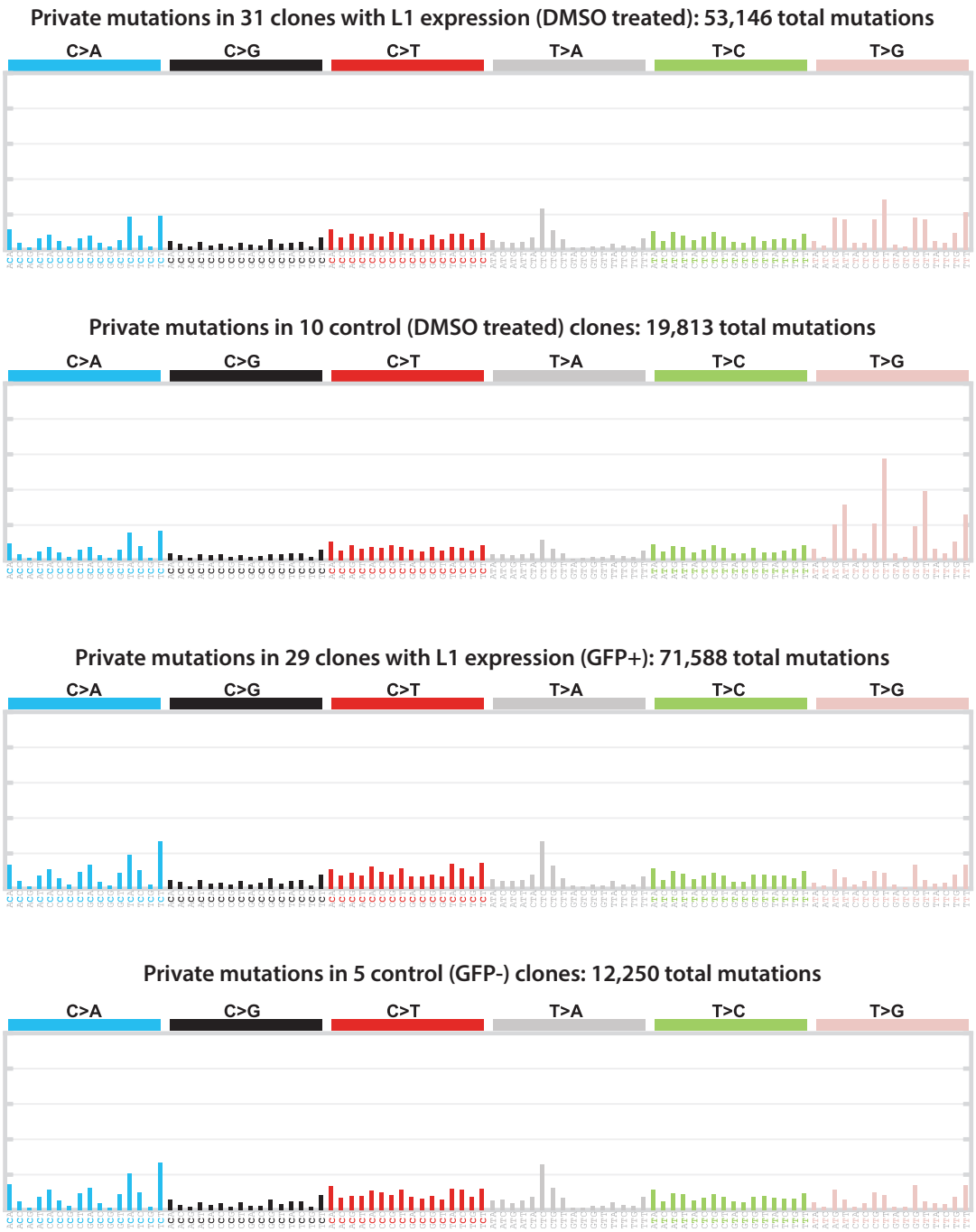


Figure S18 | Spectra of single-base substitutions (SBS) in clones with L1 expression (Tet-On L1 with Dox treatment and GFP+) and control clones (Tet-On L1 with DMSO treatment and GFP-) suggest no apparent SBS signature associated with L1 expression.



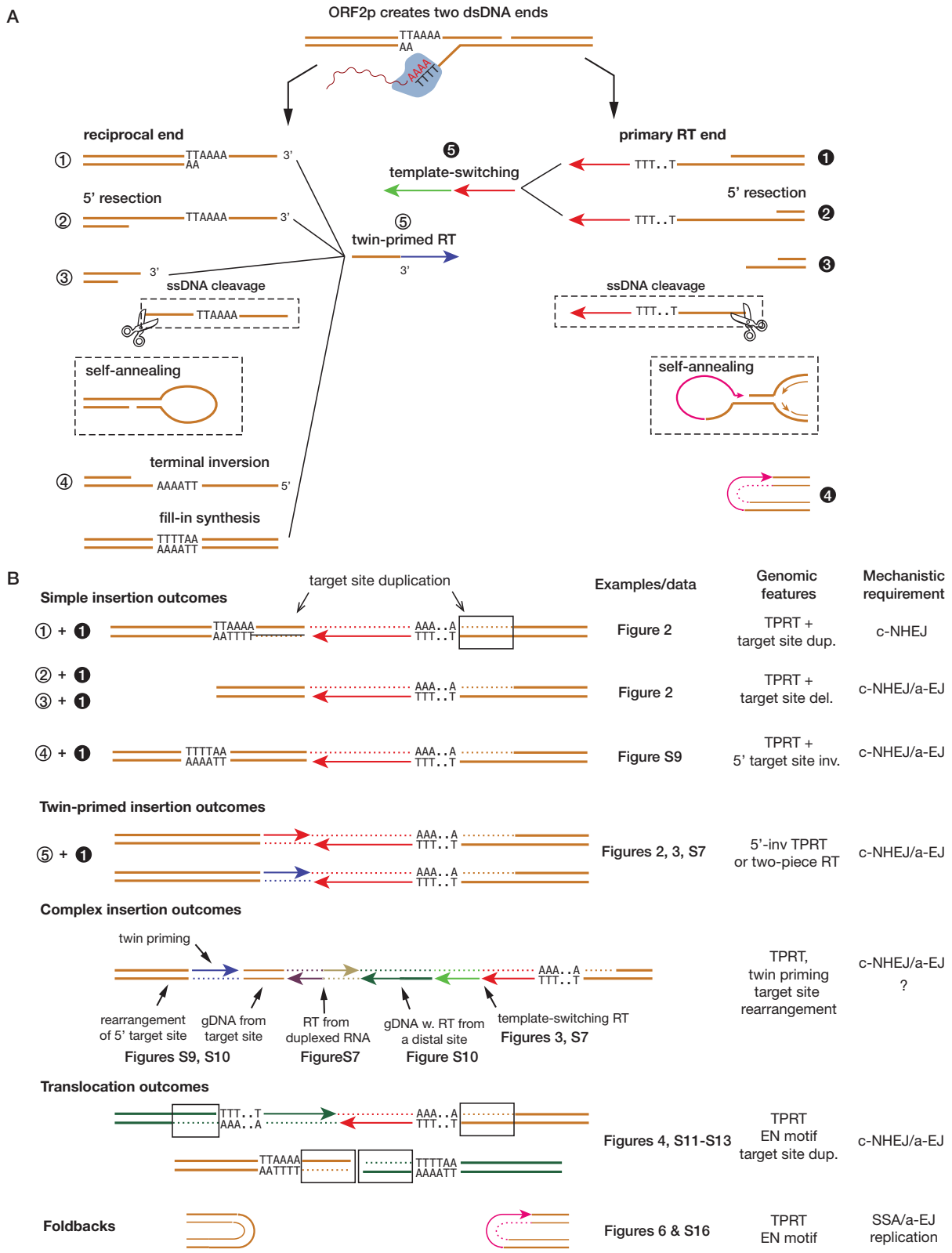


Figure 8 | Summary of insertion and rearrangement outcomes of L1 retrotransposition

A. Different processes that can alter dsDNA ends generated by ORF2p. Both the primary RT end and the reciprocal end can undergo 5'-resection (2), ssDNA flap removal (3), or self annealing. Replication through a hairpin formed by self-annealing can generate foldback junctions (4 on the right). The 3'-flap of the reciprocal end can be inverted (4 on the left). The reciprocal end can undergo twin-primed RT (5 on the left); the primary RT end can undergo template switching RT (5 on the right).

B. Insertion or rearrangement outcomes from different combinations of dsDNA ends joined together. Insertion outcomes are generated by ligation between the primary RT end and the reciprocal end with one or multiple sequence insertions including RT. Translocation outcomes are generated by illegitimate recombination between dsDNA ends from distal loci. Foldbacks arise from the replication/fusion of unligated dsDNA ends. Examples or data related to different types of insertion/rearrangement outcomes and the implicated mechanisms of DNA end-joining are listed.