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 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.

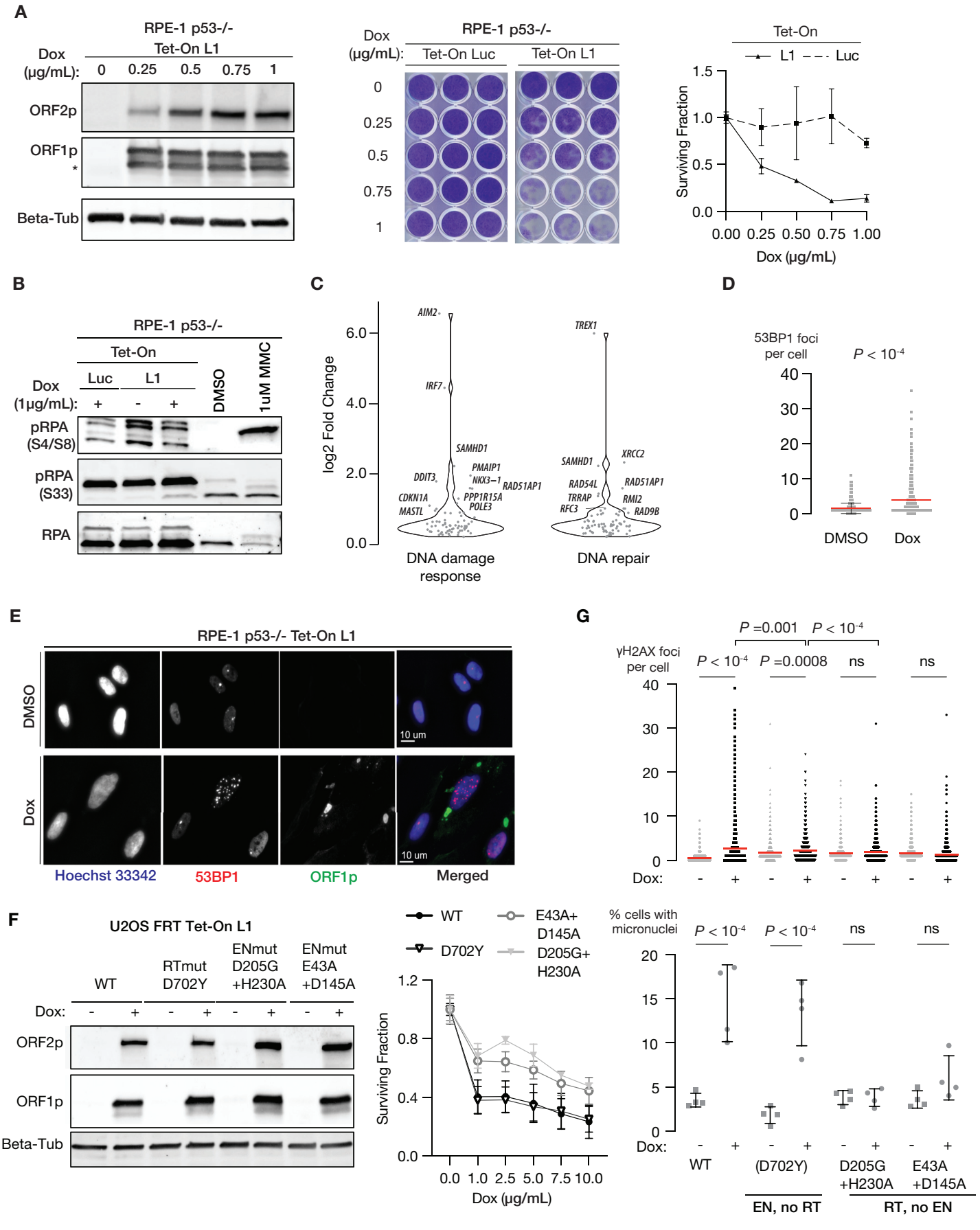


Figure S2 | Large copy-number alterations (CNAs) after L1 induction assessed from 0.1× whole-genome sequencing data.

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.

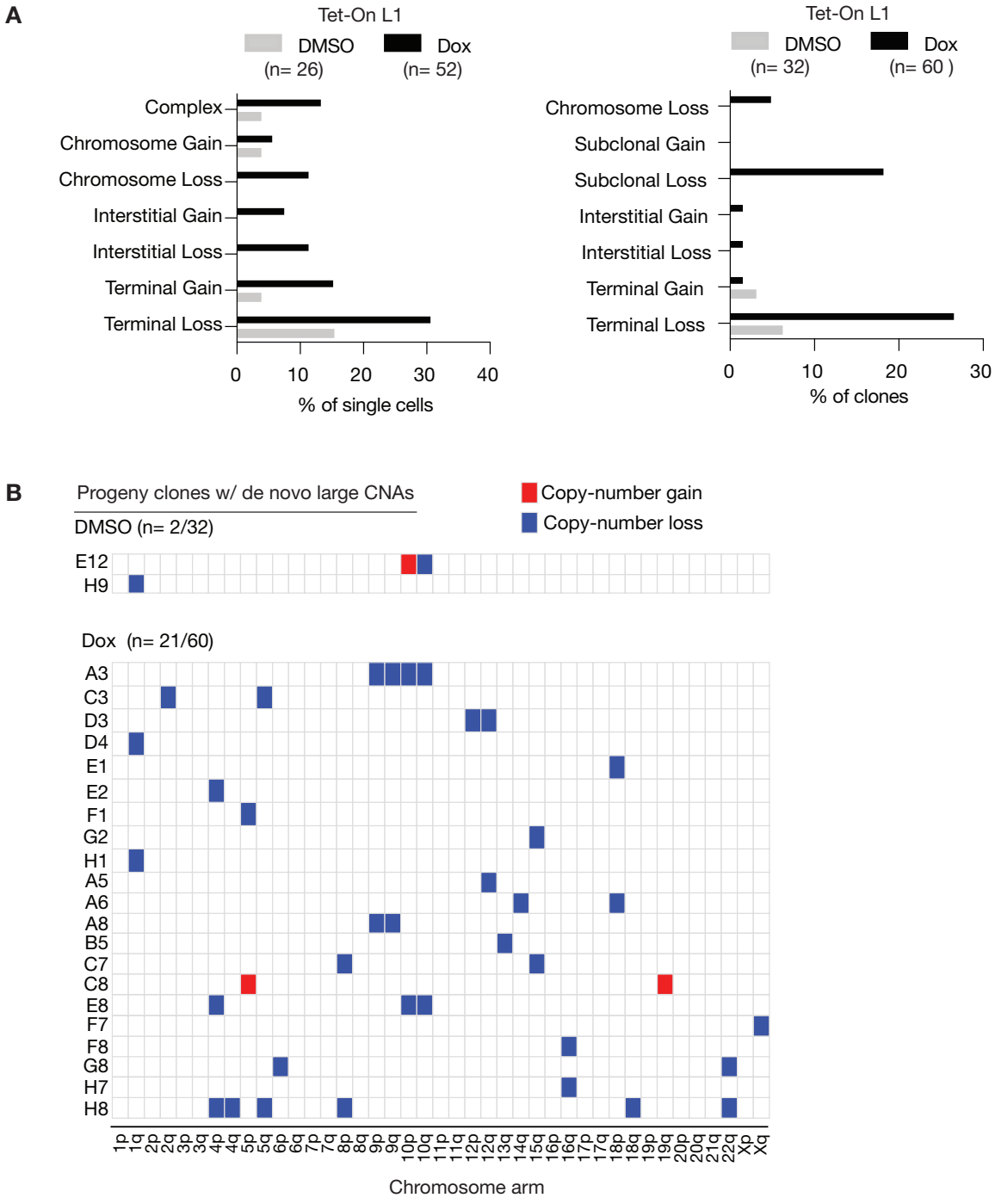


Figure S3 | Large de novo CNAs in cells with L1 retrotransposition identified using the L1 GFP reporter.

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**.

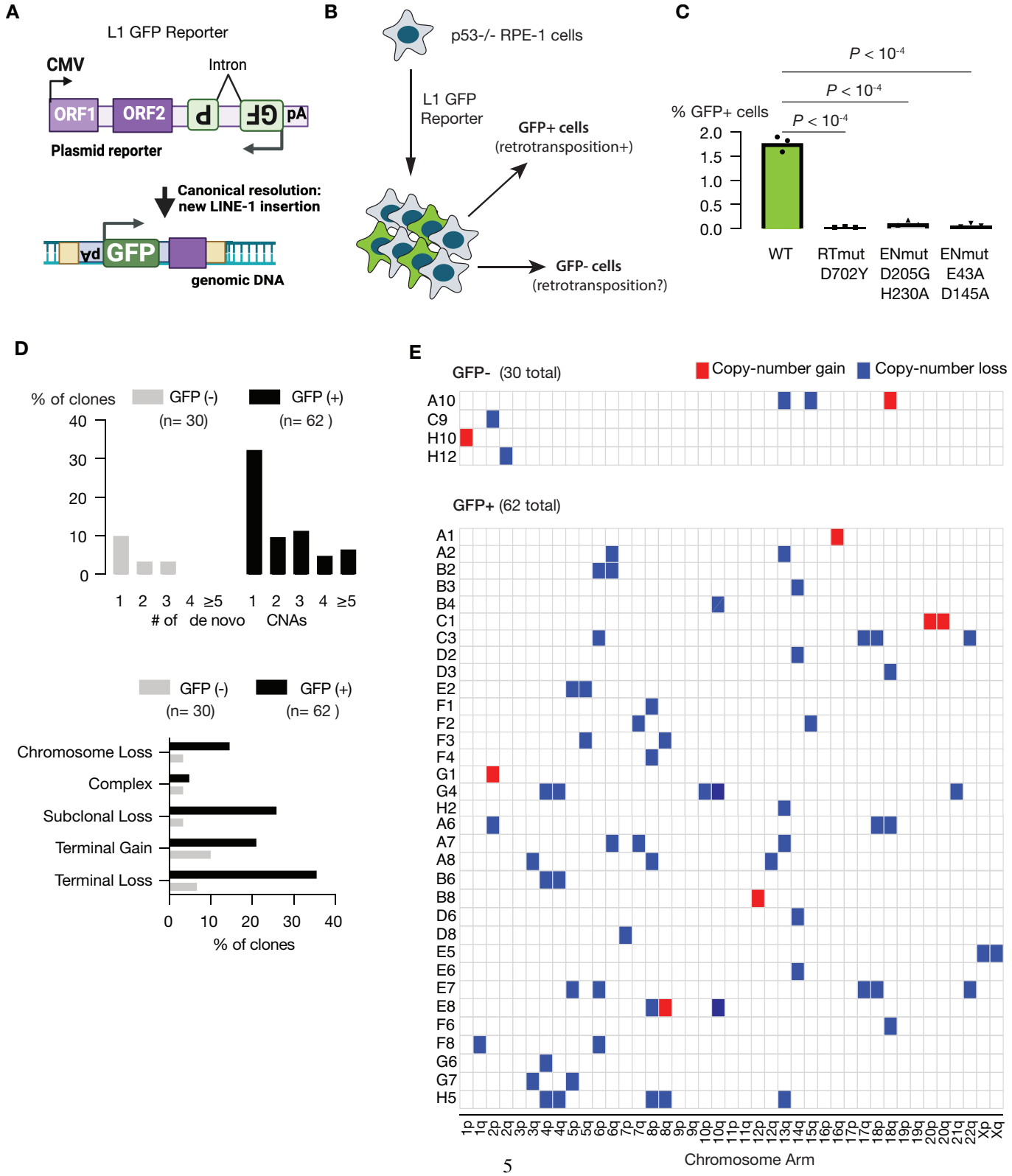


Figure S4 | Chromosomal locations and GRCh38 coordinates of 11 integrated Tet-On L1 transgene (red arrows) determined from breakpoints in reads (both short and long) and de novo assembled contigs (from PacBio long reads) with split alignments to both the transgene sequence and the human genome.

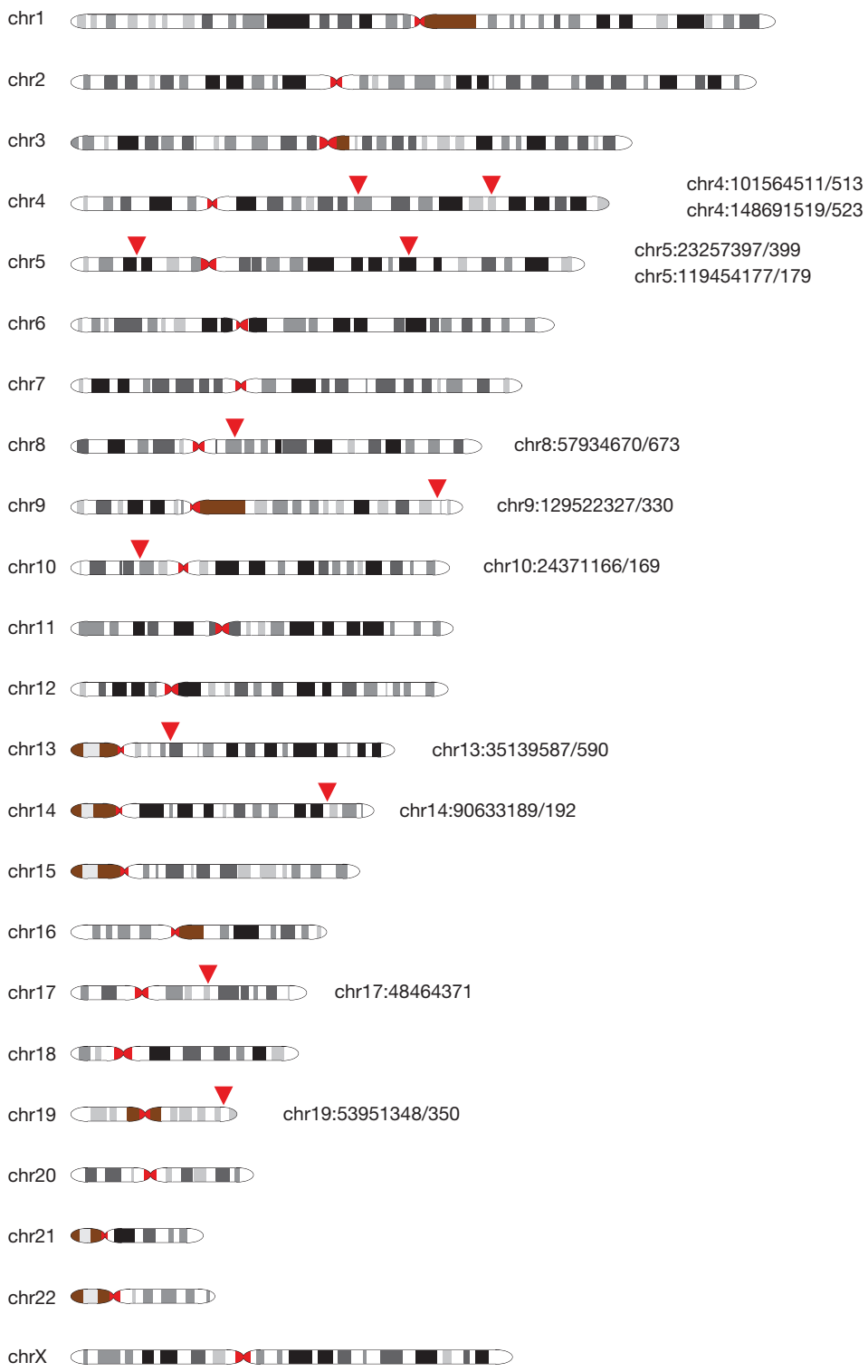


Figure S5 | Additional data on insertions from retrotransposition.

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.

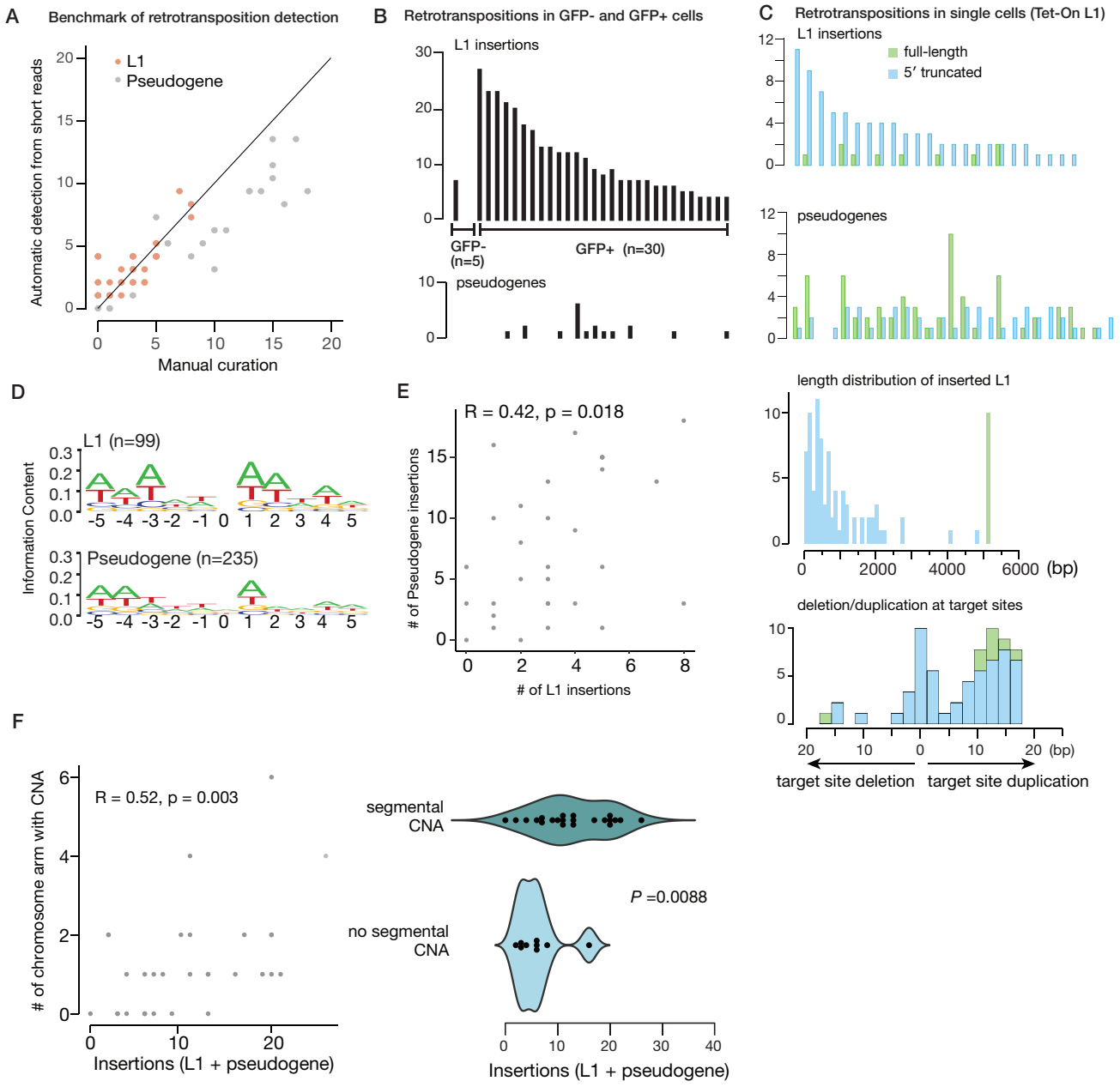


Figure S6 | L1-mediated insertions of non-canonical transcripts. All the insertions are identified in **Dox clones** and are validated by long reads. The RNA-Seq data are from parental RPE-1 cells without L1 induction.

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.

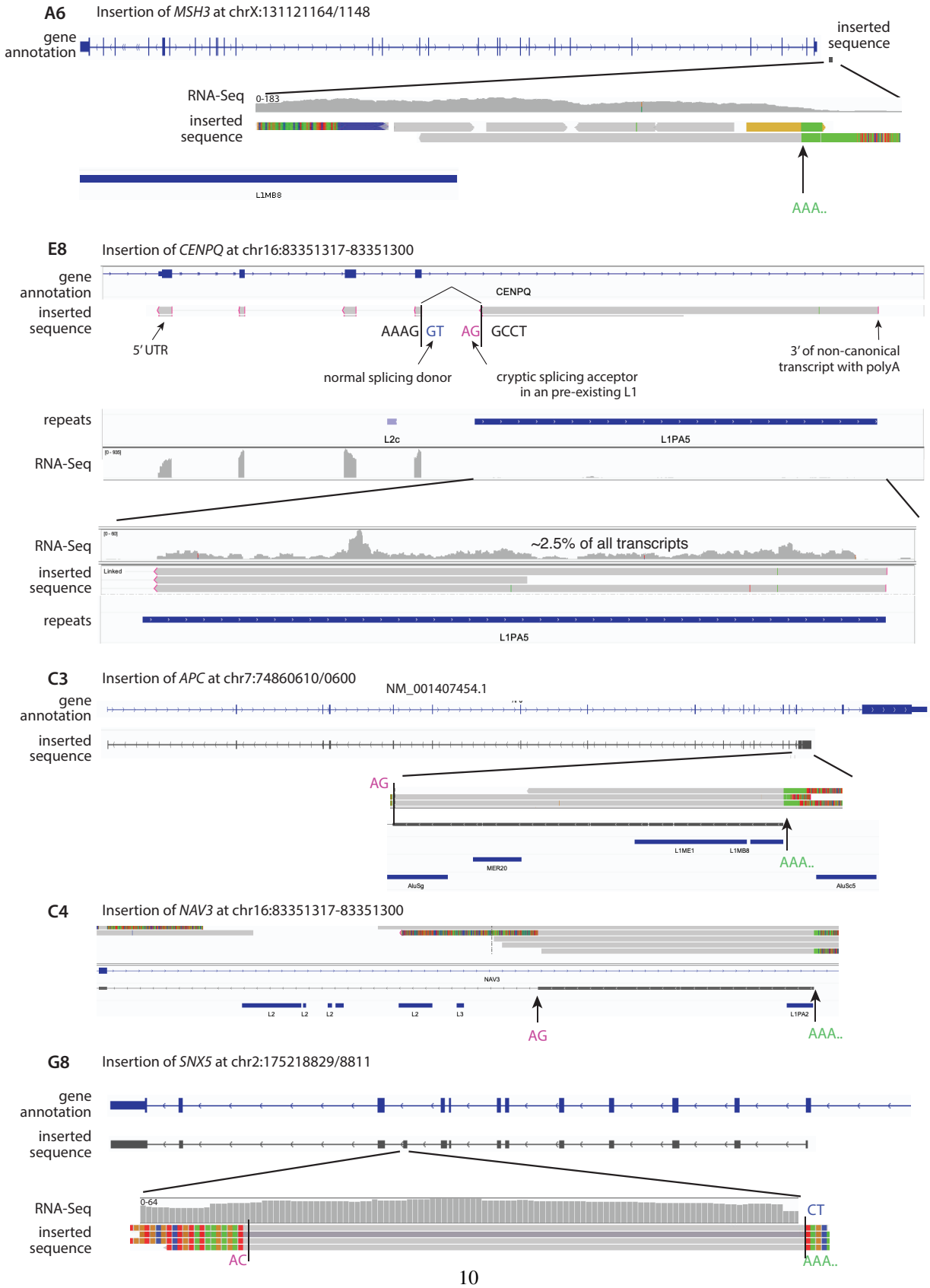


Figure S7 | (*Figure on next page.*) Additional examples of complex insertions and their mechanistic interpretations.

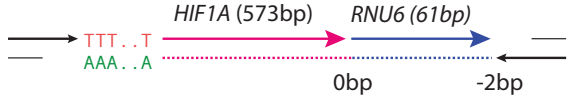
A. Additional examples of complex insertions reflecting template-switching RT or annealing/end-joining between twin-primed RT ends. Except for the insertion in **Dox clone D6**, which is assembled based on junctions detected from short reads, all the other insertions are completely resolved by long reads. 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)

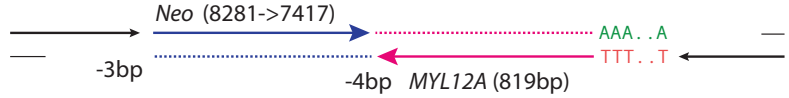


template-switching or chimeric mRNA

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

template-switching

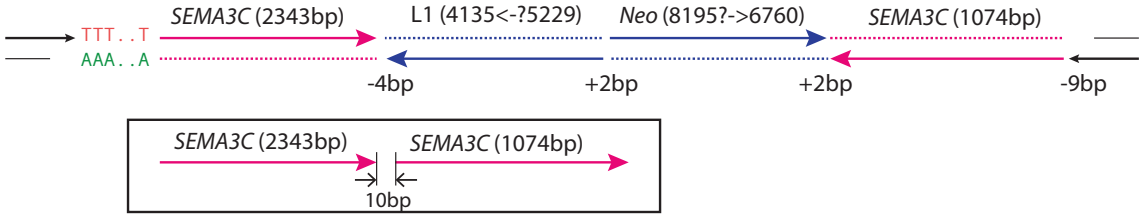
H8: *Neo* (truncated) + *MYL12A* (truncated) at chr9:128700830/128700128 (703bp TSD)



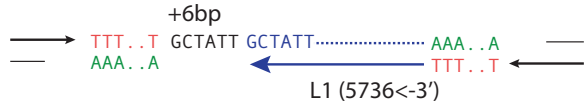
annealing/end-joining

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

twin-priming & annealing/end-joining

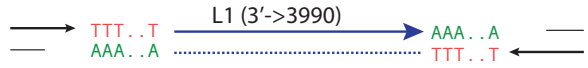


A7: L1 + poly-A at chr5:32883965/3952 (14bp TSD)



annealing/end-joining

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



annealing/end-joining

B Alternative outcomes of twin priming

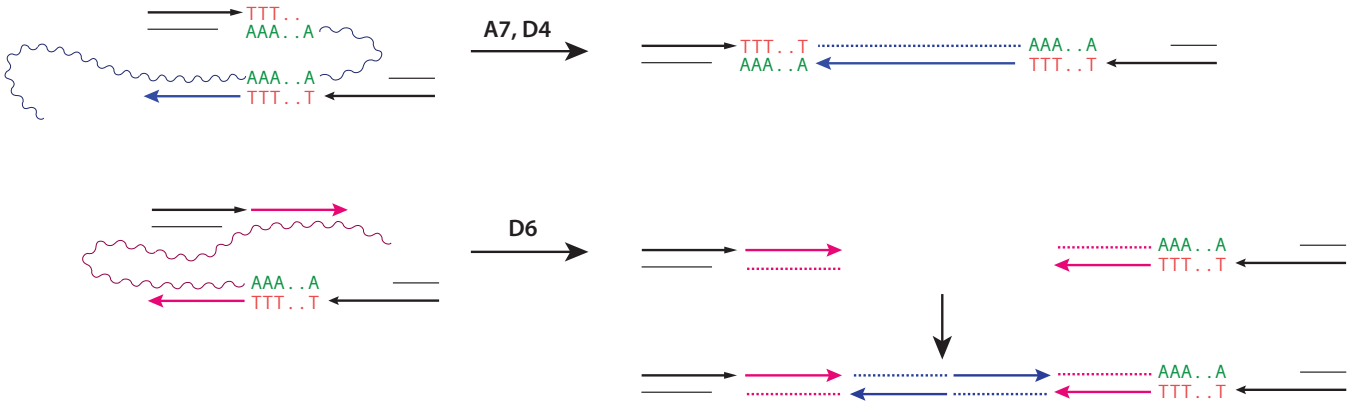
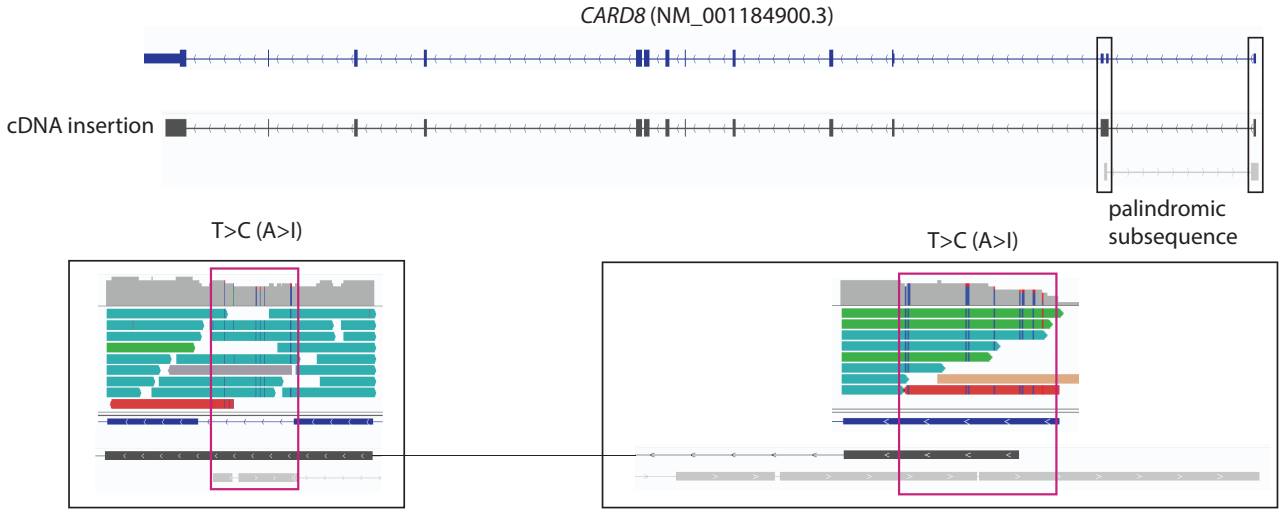
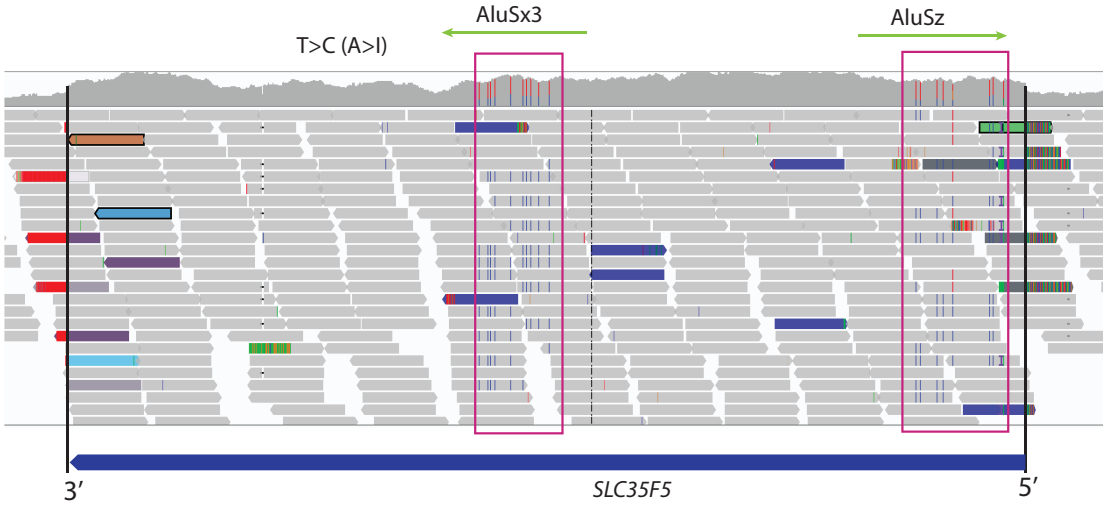


Figure S8 | Retrocopied pseudogene insertions containing substitutions due to ADAR editing. ADAR editing (A>I) results in T:A>C:G substitutions. All the examples are from **Dox clones** and are validated by long reads. In **A**, the clustered substitutions are restricted to a palindromic sequence near 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**

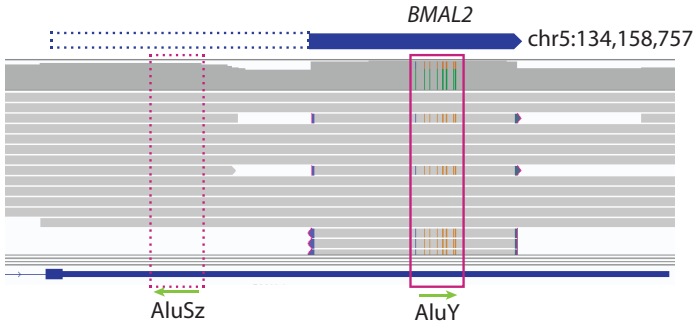


Figure S9 | (*Figure on next page.*) Insertions with inversions of genomic DNA sequences near the insertion site.

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.

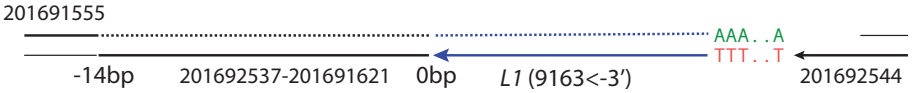
B. An inversion of genomic DNA next to a 5'-inverted L1 insertion. The inferred process is similar to **A**, except with the additional step of twin priming that gives rise to a secondary 5'-inverted insertion. The large gap between the two inverted RT sequences is distinct from most 5'-inverted insertions (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.

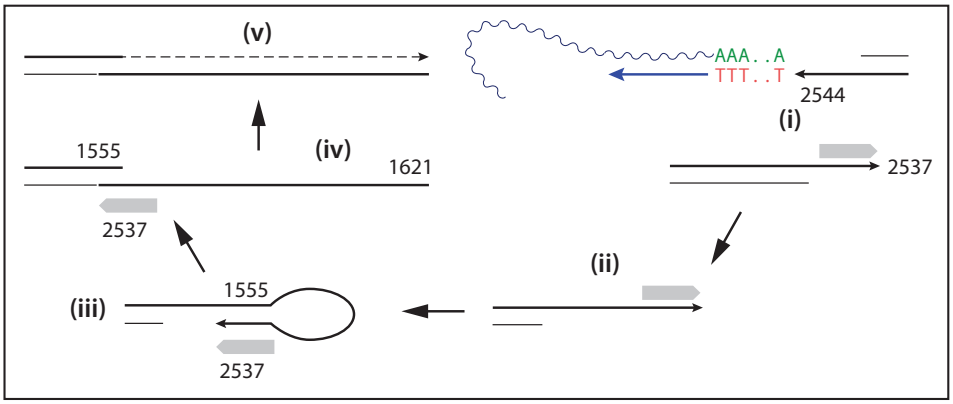
Insertions with inversions of genomic DNA sequences near the insertion site

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

insertion outcome

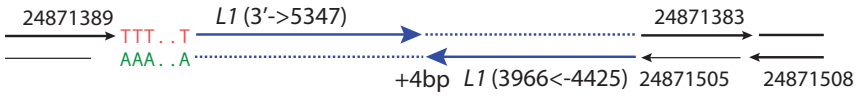


Inferred mechanism

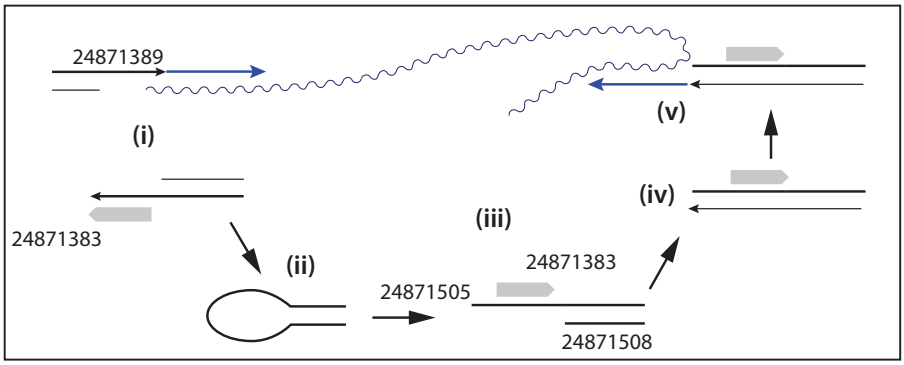


B Dox clone C8: 5'-inverted L1(truncated) plus inversion of sequence near the insertion site at chr10:24871389/1382

insertion outcome

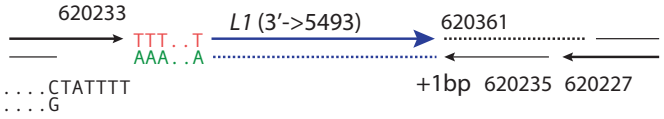


Inferred mechanism



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

insertion outcome



Inferred mechanism

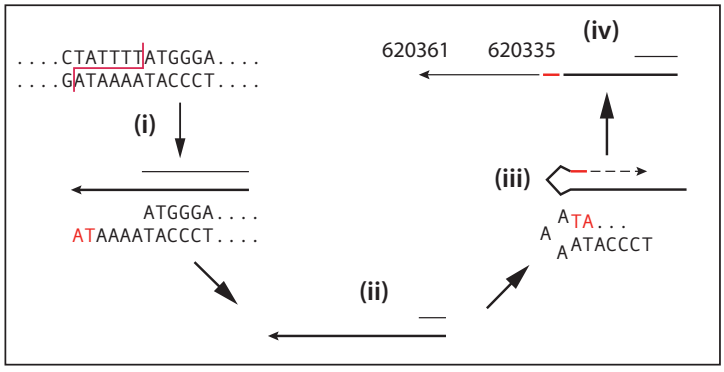


Figure S10 | Insertions containing both RT sequences and templated genomic sequences.

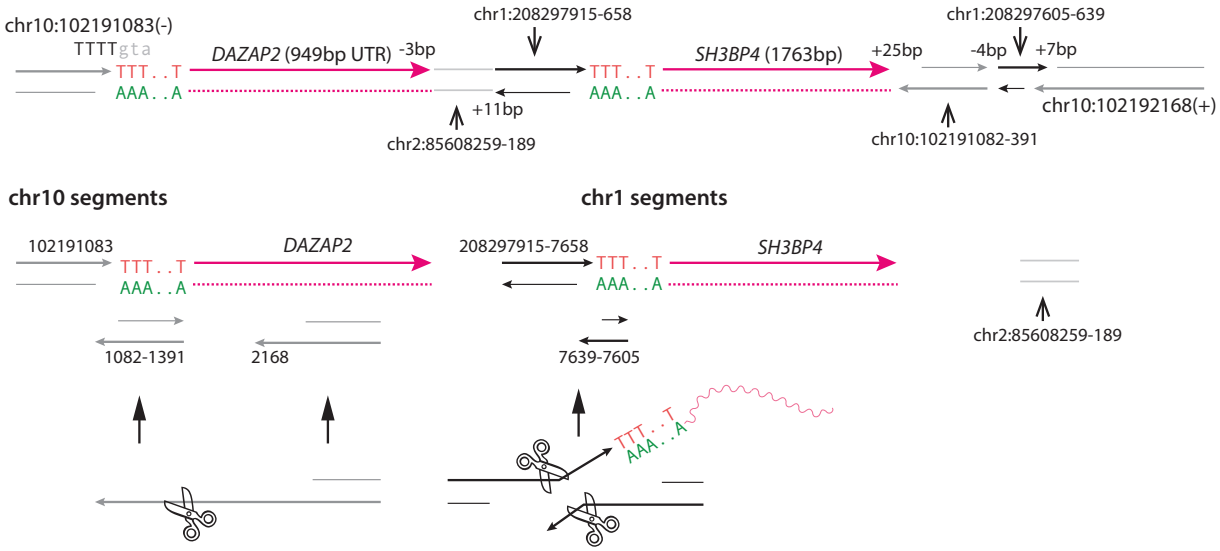
A. An insertion containing two retrocopied sequences and four genomic DNA sequences. The two sequences from chr1 (in black) are consistent with an origin from retrotransposition followed by cleavage of ssDNA fragments (also see **Figure 6**); the same mechanism can explain the short chr10 sequence (102191082-1391). 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).

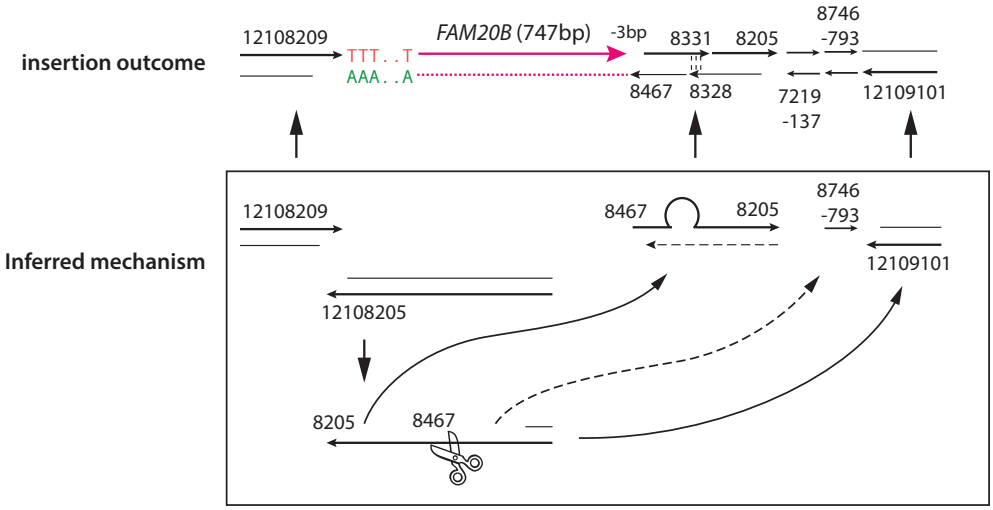
C. An insertion containing two genomic DNA sequences plus an L1 insertion. The L1 insertion is not completely resolved from short reads. The two genomic DNA insertions are mapped to regions on chr4 with local DNA fragmentation. See **Figure S17B**.

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

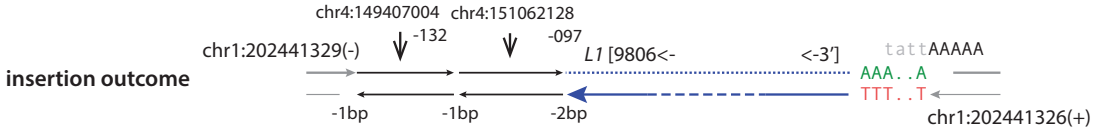


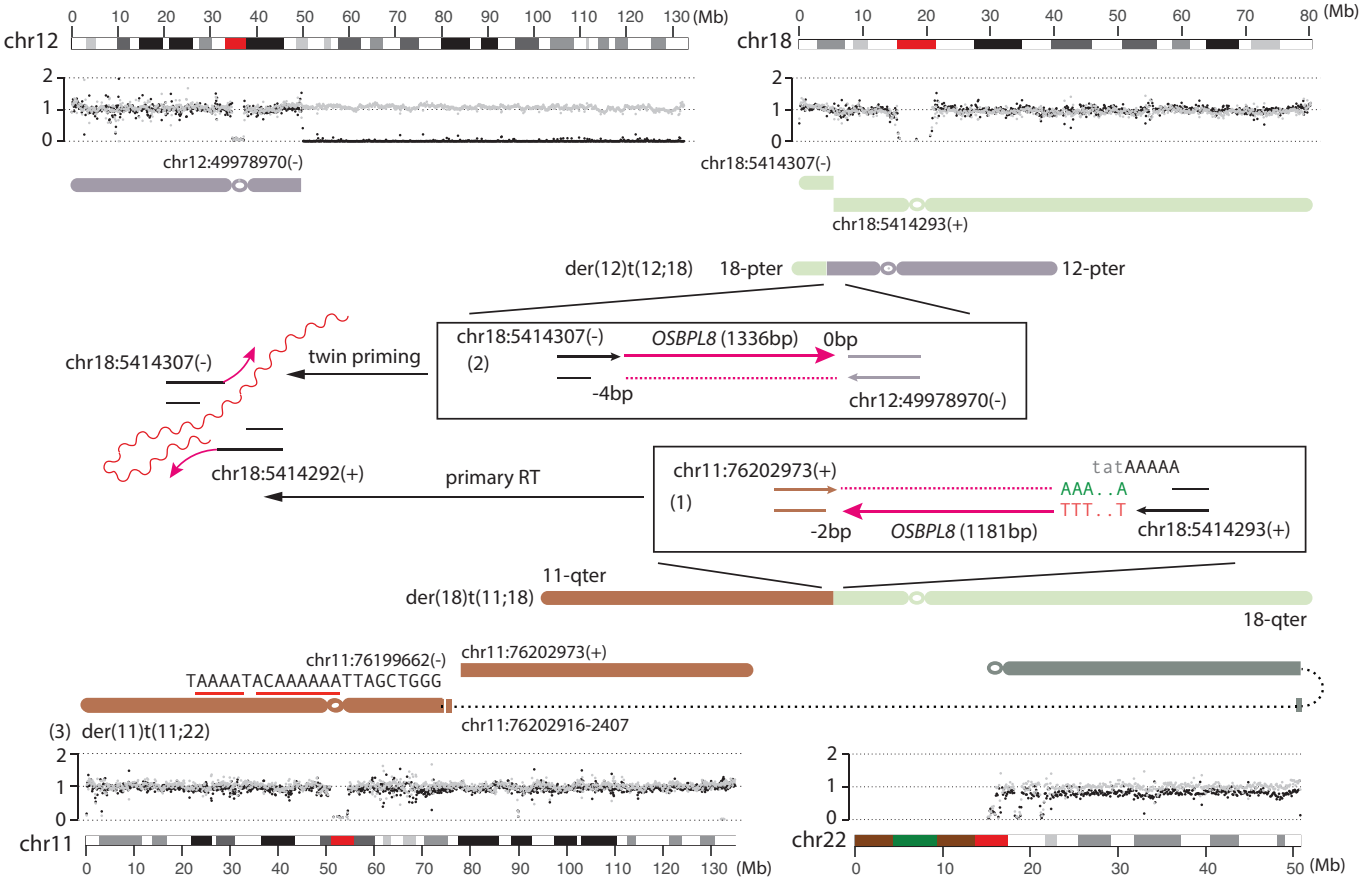
Figure S11 | (*Figure on next page.*) Additional examples of L1-mediated translocations.

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**

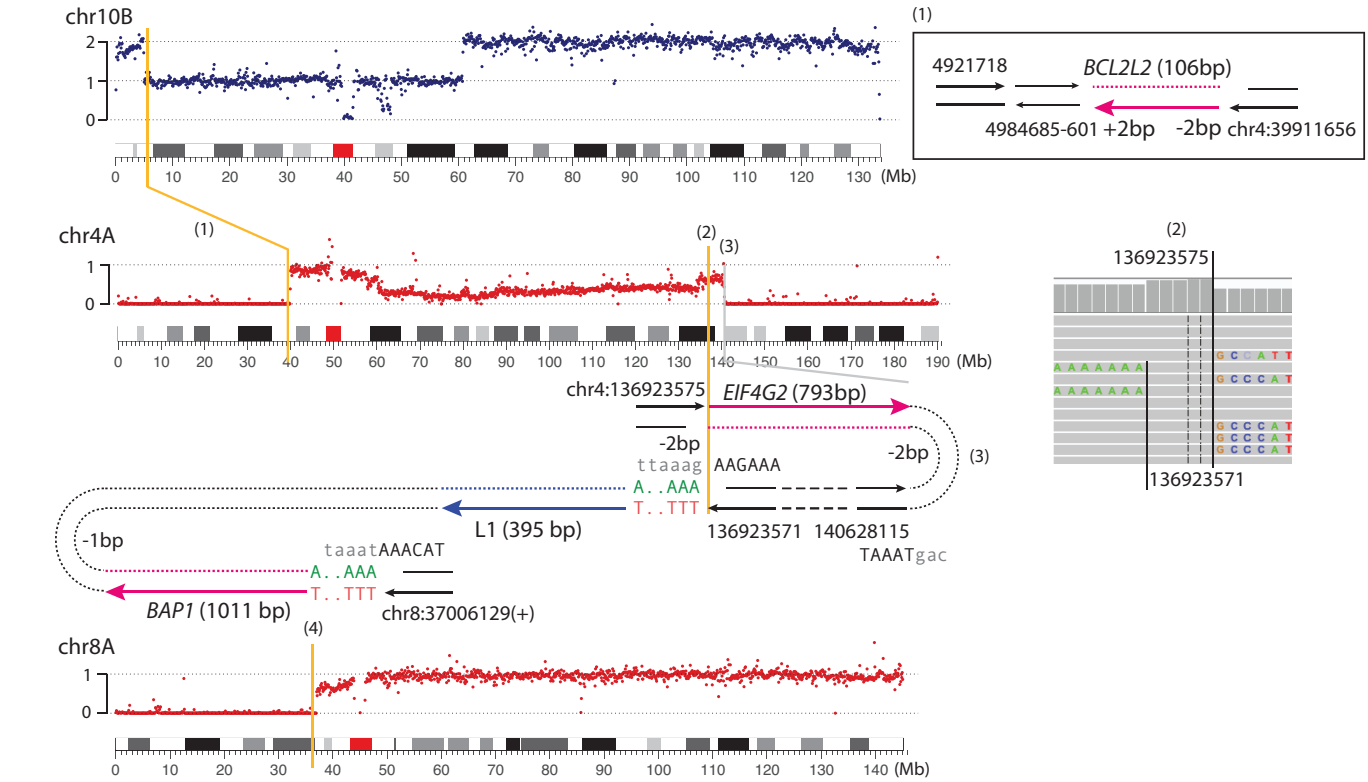


Figure S12 | Dicentric chromosomes in **Dox clones**. The inference of dicentric chromosomes is based on similar evidence as discussed in **Figure 4C**.

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.

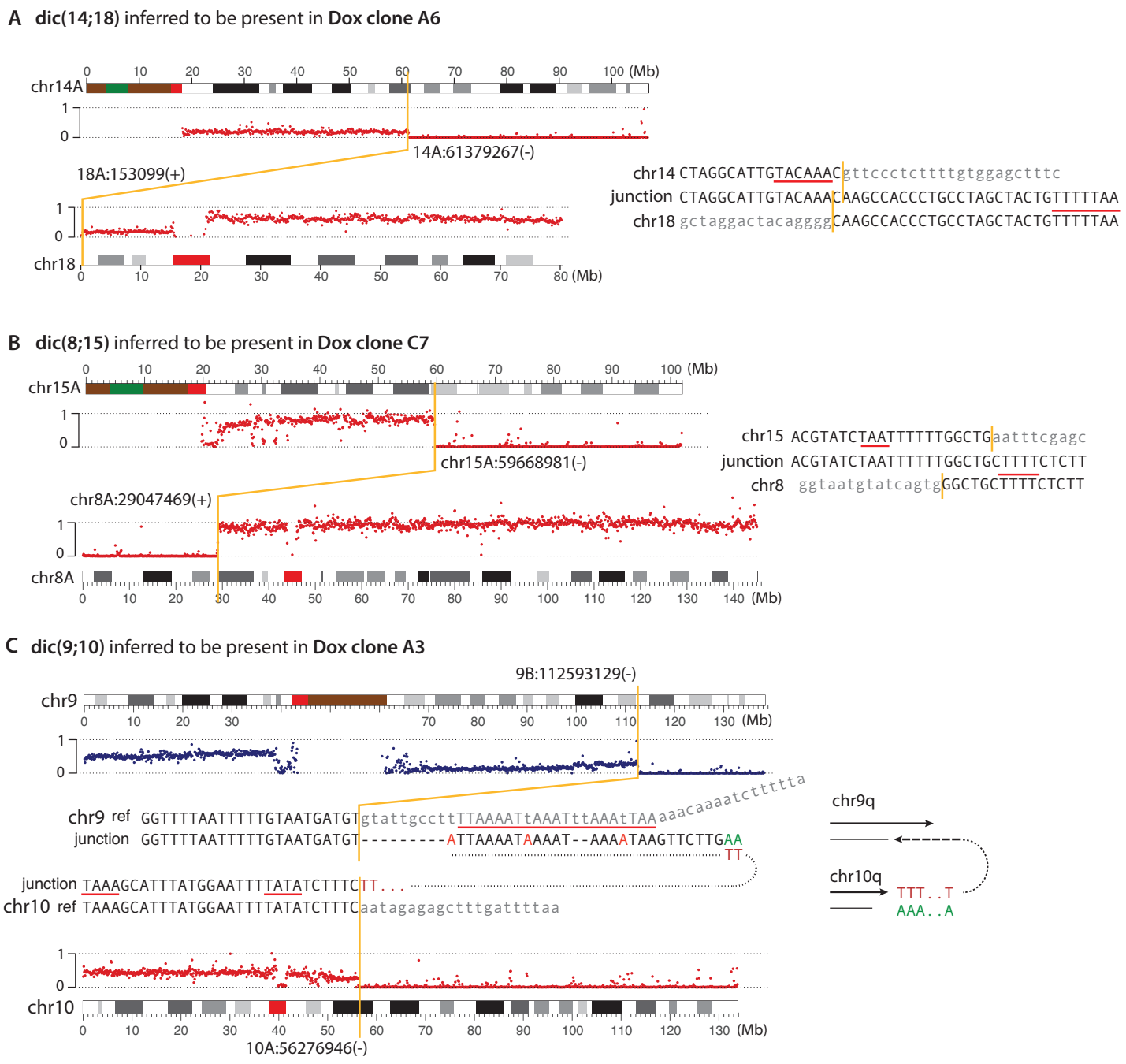
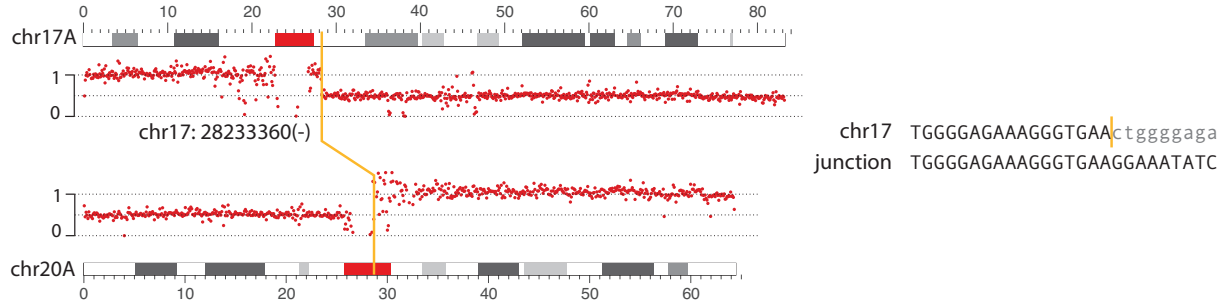
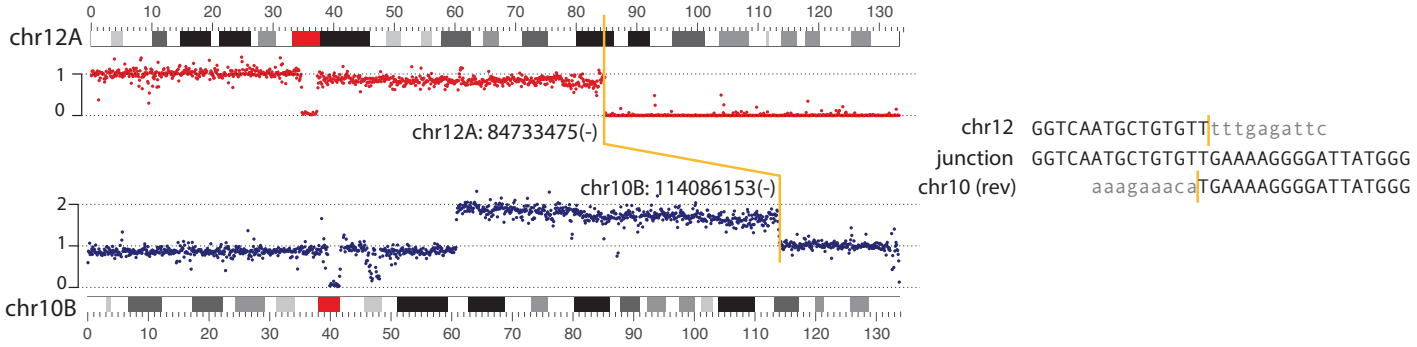


Figure S13 | Dicentric chromosomes in GFP+ clones. The **C3** clone is derived from a tetraploid ancestor; the **E8** and **F2** are derived from diploid ancestors. As none of the breakpoints is adjacent to an ORF2p EN cutting site, these translocations may have arisen independent of or downstream of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

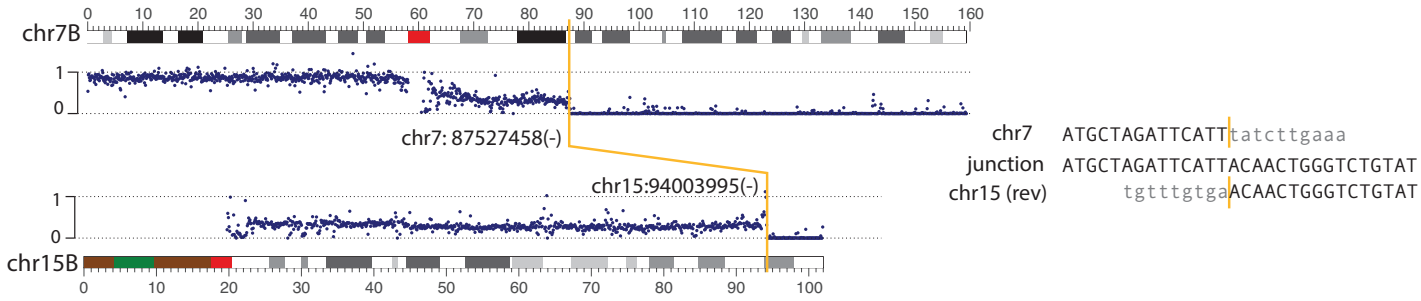


Figure S14 | Additional examples of large segmental deletions in Dox clones and GFP+ clones.

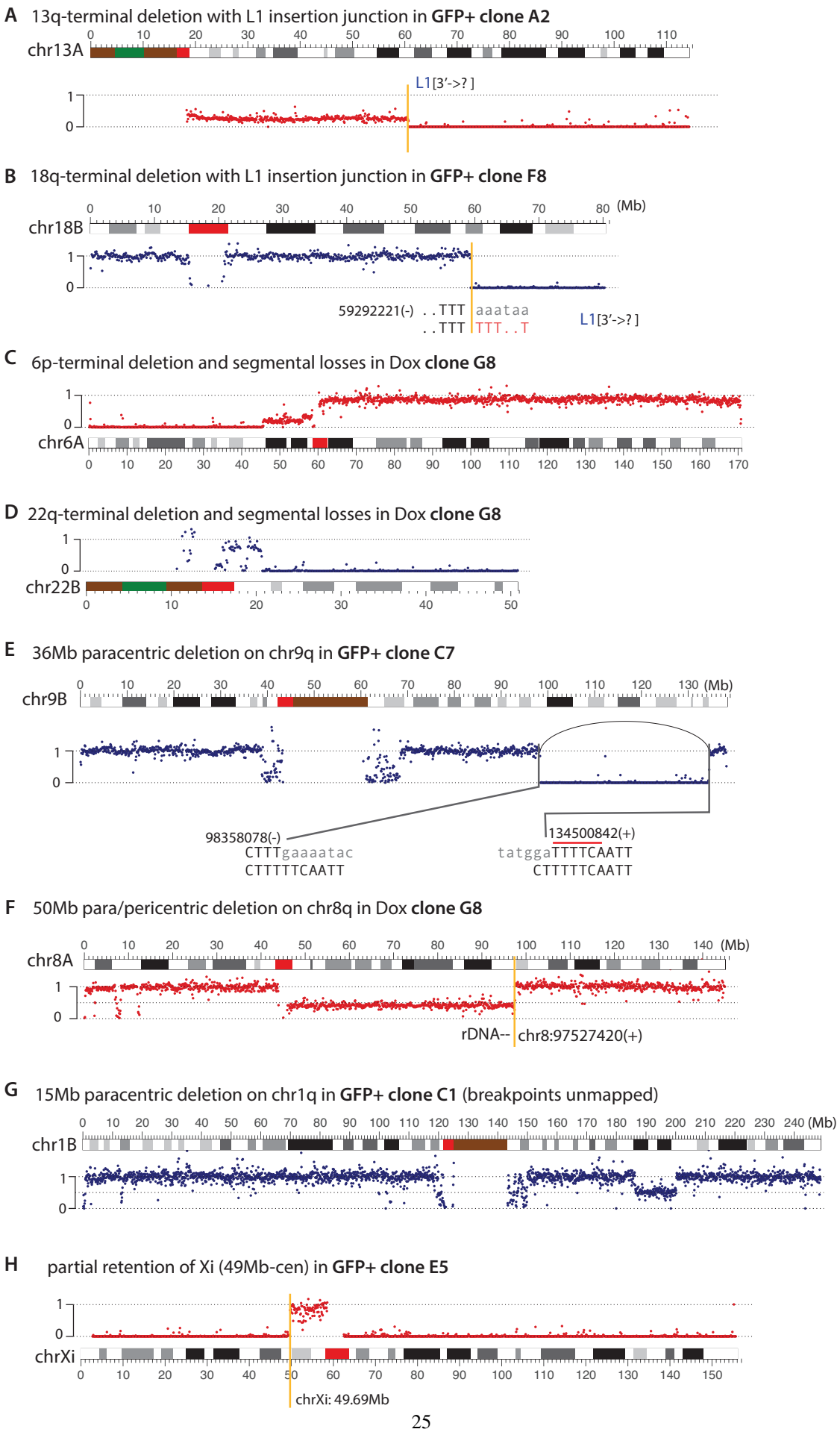


Figure S15 | Examples of sloping copy-number variation indicating BFB cycles in Dox clones and GFP+ clones. For the example in **GFP+ clone H7**, the sloping copy-number variation is on the extra copy of the 10q segment that is appended to the active X. Note the minor copy-number gain in Xa (black dots)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

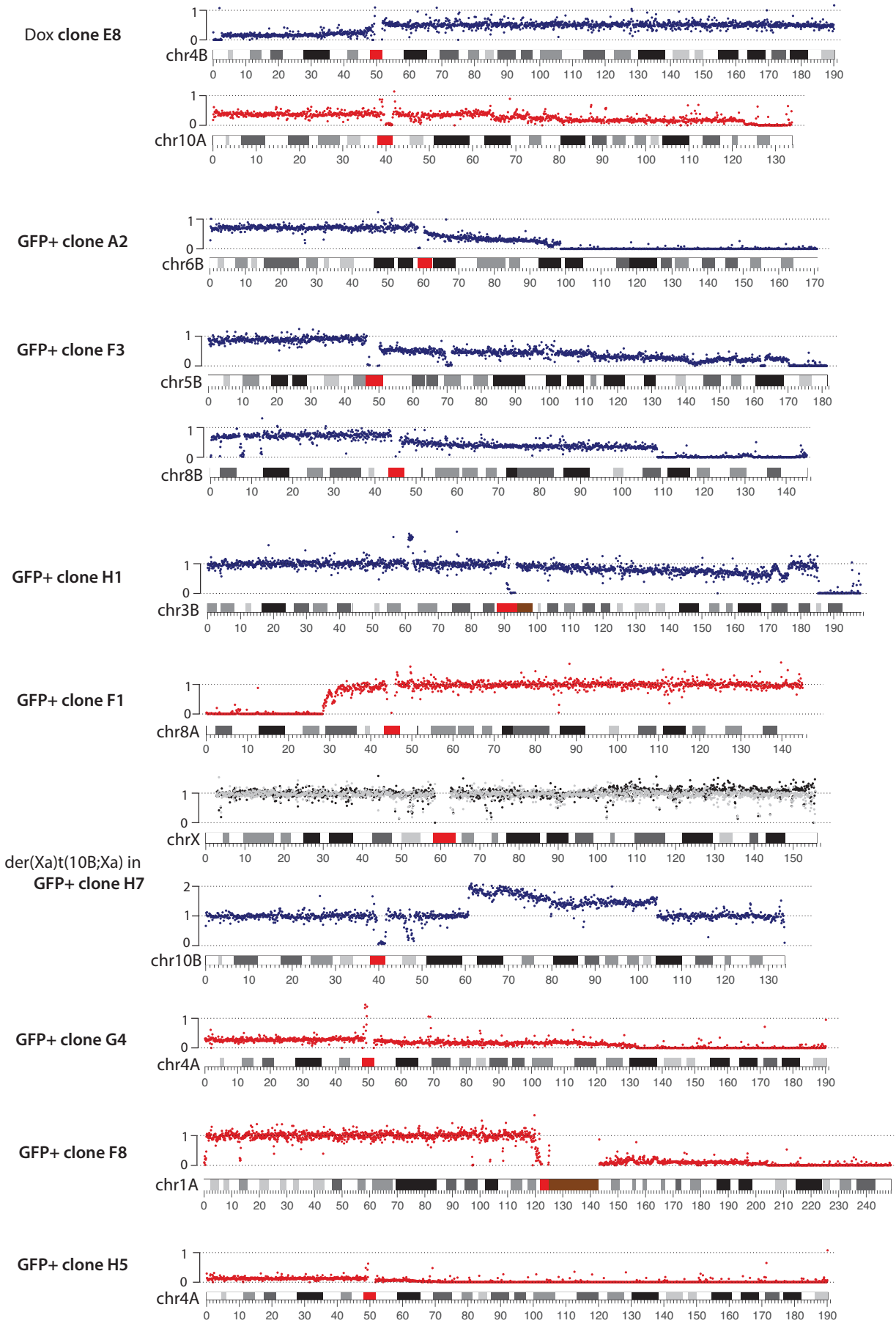


Figure S16 | Additional instances of foldback junctions in **Dox clones** (**A**) and **GFP+ clones** (**B**). **C**. An example of multiple adjacent foldback junctions. Potential ORF2p EN cutting sites near the breakpoints are highlighted.

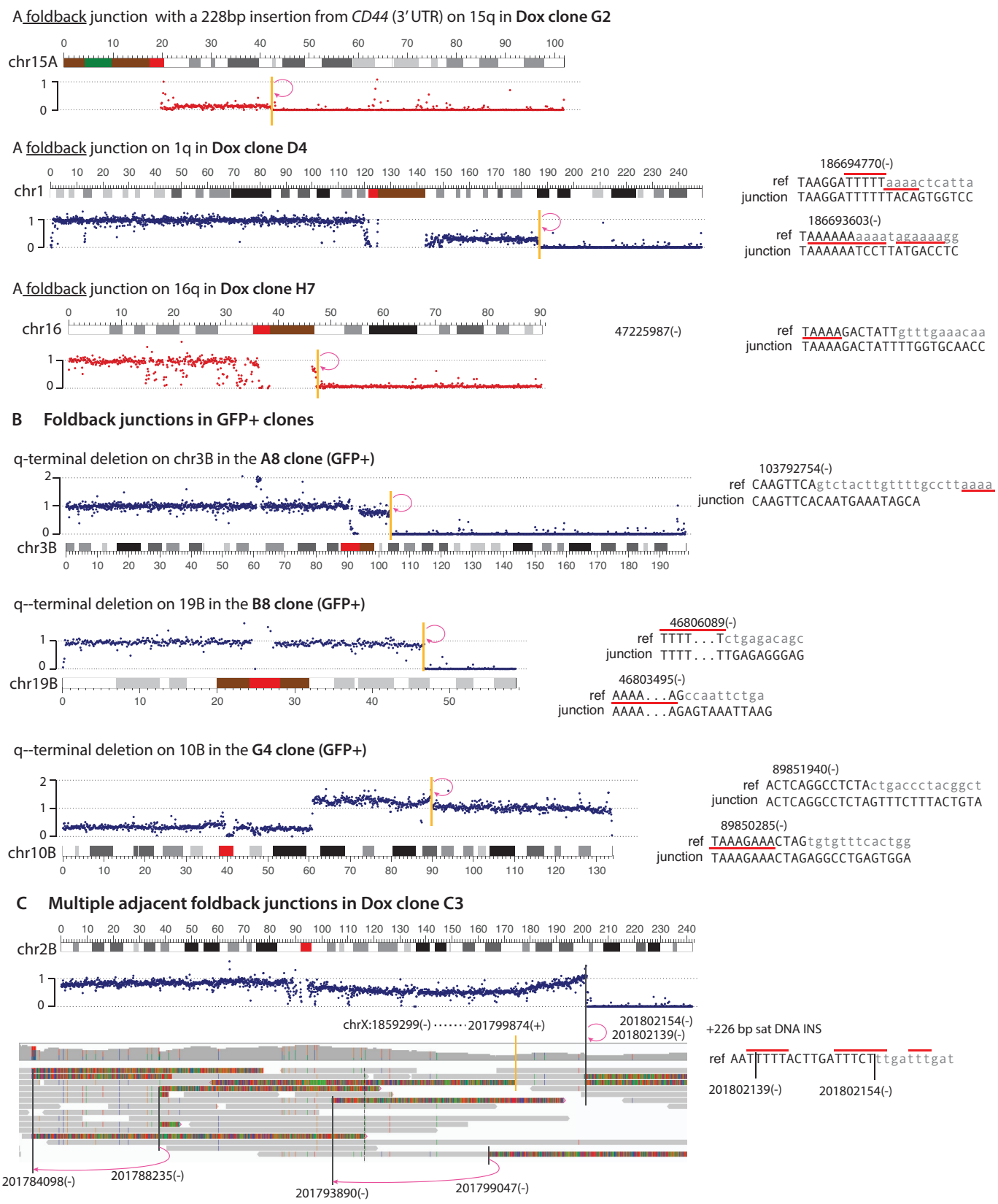


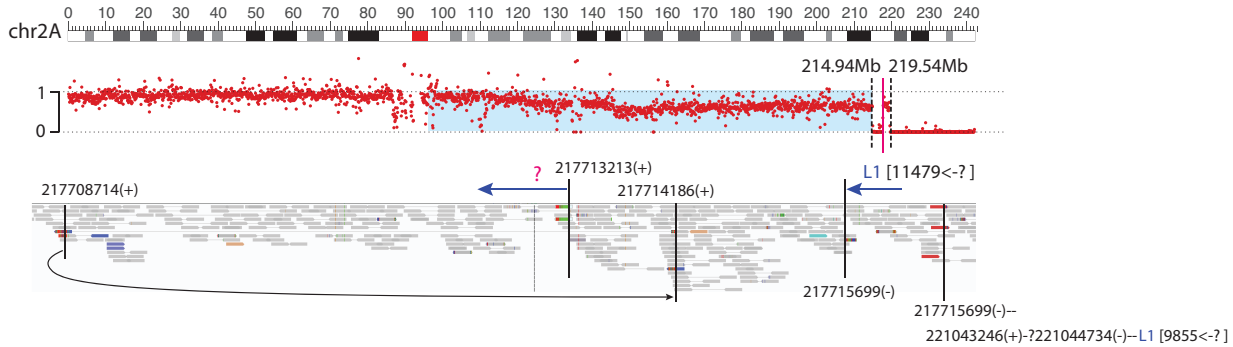
Figure S17 | Additional instances of chromothripsis in **GFP+ clones**.

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.

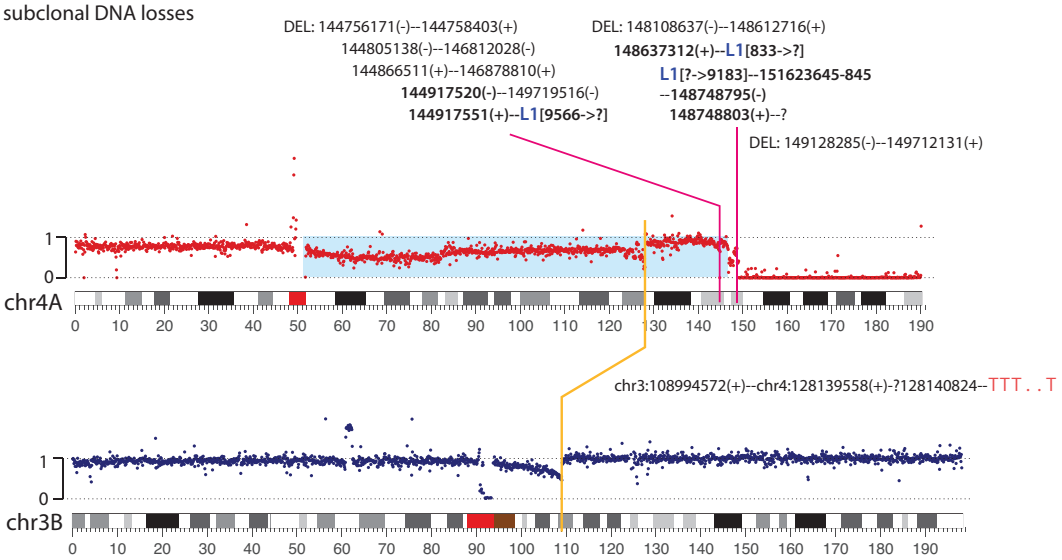
C. Three examples of chromothripsis indicated by oscillating DNA deletion and retention.

A Terminal deletion, regional fragmentation, and sloping copy-number variation on 2q in **GFP+ clone C1**

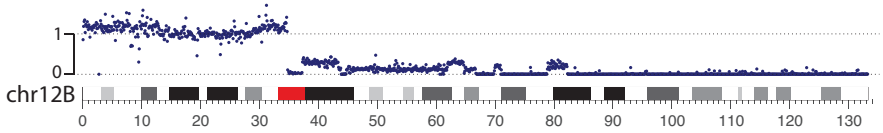


B Complex rearrangements on chr4A in **GFP+ clone H7**

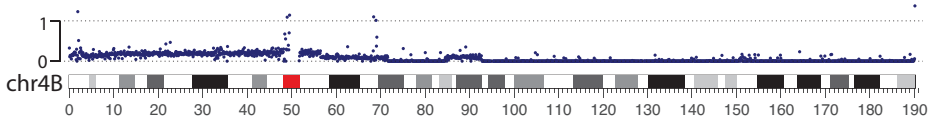
1. q-terminal deletion (149Mb-qter.); 2. regional chromothripsis;
3. RT at DNA ends on chr4 & insertion of chr4 fragments to other RT junctions
4. subclonal DNA losses



C q-terminal deletion and q-arm chromothripsis on chr12B in the **GFP+ clone A8**



q-terminal deletion and q-arm chromothripsis on chr4B in the **GFP+ clone B6**



chromothripsis of chr8B in **GFP+ clone H5**

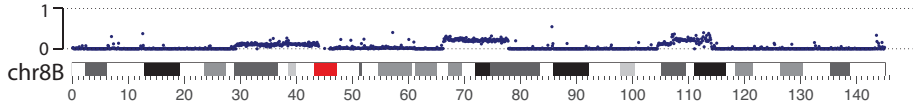


Figure S18 | Spectra of single-base substitutions (SBS) in clones with L1 expression (Tet-On L1 with Dox treatment and GFP+) and control clones (Tet-On L1 with DMSO treatment and GFP-) suggest no apparent SBS signature associated with L1 expression.