Frequency and mechanisms of LINE-1 retrotransposon insertions at CRISPR/Cas9 sites

Jianli Tao, Qi Wang, Carlos Mendez-Dorantes, Kathleen H. Burns & Roberto Chiarle

CRISPR/Cas9-based genome editing has revolutionized experimental molecular biology and entered the clinical world for targeted gene therapy. Identifying DNA modifications occurring at CRISPR/Cas9 target sites is critical to determine efficiency and safety of editing tools. Here we show that insertions of LINE-1 (L1) retrotransposons can occur frequently at CRISPR/Cas9 editing sites. Together with PolyA-seq and an improved amplicon sequencing, we characterize more than 2500 de novo L1 insertions at multiple CRISPR/Cas9 editing sites in HEK293T, HeLa and U2OS cells. These L1 retrotransposition events exploit CRISPR/Cas9-induced DSB formation and require L1 RT activity. Importantly, de novo L1 insertions are rare during genome editing by prime editors (PE), cytidine or adenine base editors (CBE or ABE), consistent with their reduced DSB formation. These data demonstrate that insertions of retrotransposons might be a potential outcome of CRISPR/Cas9 genome editing and provide further evidence on the safety of different CRISPR-based editing tools.
Genome editing with CRISPR/Cas9 holds great promise for the treatment of human genetic diseases, resulting in a mix of intended and unintended genetic alterations. Canonical CRISPR/Cas9 DNA editing induces double-strand breaks (DSBs) that are resolved by the classical non-homologous end joining pathway (c-NHEJ) and typically result in insertions or deletions (indels) of relatively small DNA sequences at both on-target and off-target sites, leading to disruption of the target sequence.2–4 While c-NHEJ is the main cellular DNA repair pathway of CRISPR/Cas9-induced DSBs, homology-directed repair (HDR) can compete by inserting homologous DNA sequences that result in precise gene editing.5 More rarely, DSBs induced by CRISPR/Cas9 can lead to larger genomic rearrangements such as large chromosomal deletions, inversions, or translocations,6–10 or even more catastrophic events such as chromothripsis11 and chromosome loss12–14. Additional outcomes of the canonical CRISPR/Cas9 editing include integrations of exogenous sequences including lentivirus15,16, aden-associated virus (AAV)17, plasmids18,19, and small DNA fragments.20,21 These events exploit the availability of frequent DSBs generated by CRISPR/Cas9 to favor the insertion of exogenous DNA fragments using microhomology-mediated end joining (MMEJ).17,22

While it is known that mobile genetic elements can be captured at DSB sites,23–28 little is known on the frequency and mechanisms of retrotransposon integration into CRISPR/Cas9 editing sites because few such events have been described so far only in mouse zygotes.29 Retrotransposons are self-propagating sequences that generate de novo insertions through reverse transcription of RNA intermediates. In humans, the only autonomously active family is the long interspersed element-1 (LINE-1 or L1)29. L1 of RNA intermediates. In humans, the only autonomously active families that generate de novo insertions through reverse transcription of RNA intermediates are the long interspersed element-1 (LINE-1 or L1) and short interspersed element (SINE) families.30–32 LINE-1 and SINE retrotransposition events in cellular models33–35 can lead to large genomic rearrangements such as deletions, duplications, inversions, and insertions that disrupt the normal function of targeted genes.

Results

De novo L1-ORFeus insertions occur at DSBs induced by CRISPR/Cas9. To study the occurrence of de novo L1 retrotransposition at DSBs induced by canonical CRISPR/Cas9 activity, we examined repair outcomes of generated CRISPR/Cas9 DSBs in human cells expressing a L1 vector competent for retrotransposition. The L1-ORFeus construct contains a codon-optimized human L1 sequence46 coding for ORF1p, ORF2p, and a GFP reporter for retrotransposition3,41 (Fig. 1a, Supplementary Fig. 1a, Supplementary Sequence 1). First, we determined the efficiency of L1 retrotransposition in HEK293T and HeLa cells based on the percentage of induced GFP-expressing cells, and found that the proportion of cells containing the retrotransposed reporter increased over time up to 14 days after transfection and was much higher in HEK293T cells than in HeLa cells, as previously shown48 (Supplementary Fig. 1b). As previously reported29,37,38,43, L1-ORFeus mutations that disrupt either the endonuclease (EN) activity (L1-ORFeus-ENm (H230A), hereafter termed as L1-ENm) or reverse transcriptase activity (L1-ORFeus-RTm (D702Y), hereafter termed as L1-RTm) of ORF2p severely limited L1 integrations as compared to wild type (Supplementary Figs. 1a, c–e, 2a, b), despite comparable expression of ORF1 and ORF2 mRNA (Supplementary Fig. 1f) and ORF1p protein (Supplementary Fig. 1g). Given the higher efficiency, for most of the experiments, we decided to use HEK293T cells which have been widely used to test on-target and off-target activities of CRISPR/Cas9.40,41,42,43,49,50

We targeted one active gene (MYC) and one inactive gene (RAG1) for inducing DSBs by CRISPR/Cas9, two loci differing in their transcriptional activity and chromatin accessibility. Human RAG1 gene is also a proposed target for gene correction therapy.51,52 We transduced HEK293T cells with the CRISPR/Cas9-containing lentivirus at day 5 of L1 expression and retrotransposition, then we analyzed repair outcomes of CRISPR/Cas9-mediated DSBs at day 11 on the bulk cell population (Fig. 1b). CRISPR/Cas9 editing efficiency on MYC and RAG1 target sites was high and comparable in cells expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control (Supplementary Figs. 1h, 2c).

De novo insertions of L1-ORFeus were readily distinguishable from endogenous human L1 fragments (Supplementary Sequence 2). By improved amplicon sequencing of the fragments inserted into the CRISPR/Cas9-mediated DSBs (Supplementary Fig. 1i), we identified 546 de novo insertions of the ectopic L1-ORFeus at the MYC locus and 734 insertions at the RAG1 locus (Fig. 1d), consistent with the observation that L1 insertions are independent of gene expression or chromatin accessibility.43,44 The number of de novo L1-ORFeus insertions at the CRISPR/Cas9-mediated DSBs that were derived from the L1-EN mutant construct, were reduced relative to the L1-ORFeus reporter, and extremely low in the case of L1-RTm (Fig. 1c–e, Supplementary Fig. 2d, f), indicating that the EN activity of the ORF2p was dispensable while the RT activity was required for L1 integrations into CRISPR/Cas9-initiated DSBs. The rare L1-ORFeus insertions found in L1-RTm could be due to trans-complementation from
endogenous active L1s or more likely to the L1 plasmid integrations into CRISPR/Cas9-initiated DSB as in GFP control. De novo L1-ORFeus insertions showed a balanced 5′ to 3′ or 3′ to 5′ orientation, suggesting that L1 insertion can equally integrate at either side of the CRISPR/Cas9-mediated DSB (Fig. 2a–c, Supplementary Fig. 2i). L1-ORFeus insertions showed joining to the DSB either direct (no microhomology) or via short microhomologies, mostly 1 bp to 6 bp with a decreasing trend over length (Fig. 2d), consistent with the microhomologies reported for the 5′-junctions of human L1 insertions that occurred via TPRT53.

More importantly, the distribution of de novo L1 insertions derived from L1-ORFeus or L1-ENm reporters showed a strong enrichment for sequences mapping to the 3′ end of L1-GFP reporter and a sharp decrease in coverage spanning the intron (Fig. 1e, Supplementary Fig. 2f). These observations indicate that retrotransposition of the spliced L1-ORFeus at CRISPR/Cas9-mediated DSBs is RT-dependent and initiates at its 3′ end to generate 5′ truncated insertions, reminiscent of canonical L1 insertions mediated by the L1 protein machinery54. In contrast, insertions from the lentivirus encoding for Cas9 and the sgRNA were comparable in cells expressing L1-ORFeus, L1-ENm,
Fig. 1 De novo L1-ORFeus insertions at CRISPR/Cas9 target sites in a reverse transcriptase (RT) dependent manner. a De novo L1-ORFeus retrotransposition reporter. The L1-ORFeus expression plasmid contains an eGFP cassette in the L1 3′UTR which is in the opposite orientation of the L1 and is interrupted by an intron. Cells express eGFP only after the L1-ORFeus transcript undergoes splicing and intron removal, reverse transcription, and integration into chromosomal DNA. b Schematic of the experimental strategy employed to examine de novo L1-ORFeus retrotransposition events at the MYC CRISPR/Cas9 target site in HEK293T cells. c Pie charts show the relative abundance of three main types of obtained by amplicon sequencing at the MYC locus targeted by MYC CRISPR/Cas9 in HEK293T cells expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. d Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC locus targeted by MYC CRISPR/Cas9 (d, left) or at the RAG1 locus targeted by RAG1 CRISPR/Cas9 (d, right) in HEK293T cells expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. Number of L1-ORFeus insertions was pooled from 4 independent experiments with similar results. e Numbers and fragment lengths of L1-ORFeus insertions obtained by amplicon sequencing at the MYC locus targeted by MYC CRISPR/Cas9 in HEK293T cells expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. The X-axis indicates the nucleotide position of the full-length L1 (ORFeus) reporter cassette. Top: each histogram represents the number of L1 insertions mapping to a 100 bp interval of the L1 (ORFeus) sequence represented on the X-axis. In orange are insertions oriented from 5′ to 3′ end of the L1 (ORFeus) sequence (+), in blue are insertions oriented from 3′ to 5′ (−); bottom: Length of the L1 insertions mapping to the L1 (ORFeus) sequence represented on the X-axis. Each dot indicates the center of the insertion fragment oriented either 5′ to 3′ (+) or 3′ to 5′ (−). Each bar represents the projection of the dot to indicate the nucleotide position in the full-length L1 (ORFeus) reporter cassette. Example L1-ORFeus insertions bridging the intron of GFP in cells expressing L1-ORFeus are shown in Supplementary Sequence 6.

L1-RTm, or GFP control, and evenly distributed with peaks at the flanking LTRs as previously described37 (Supplementary Figs. 1j, k, 2e, g). Likewise, insertions of genomic fragments were similar in cells expressing each distinct L1 construct, and originated from regions (approx. ±5 kb) flanking the 5′ and 3′ ends of the CRISPR/Cas9-induced DSB, consistent with previous reports9 (Supplementary Figs. 1j, 1k, 2e, h).

Furthermore, we found de novo L1-ORFeus insertions at induced DSBs at a third locus (i.e., CCR5) where CRISPR/Cas9 has been applied for therapeutic editing to prevent HIV entry into cells55,56 (Supplementary Fig. 3a, b, d, e), though we detected a lower number of L1-ORFeus insertions compared to MYC and RAG1 loci, likely due to the lower efficiency of the CCR5 CRISPR/Cas9 (Supplementary Fig. 3c) as well as only one copy of the CCR5 gene can be edited since HEK293T cells are heterozygous for CCR5 delta 32 mutation57. Again, retrotransposition of the spliced L1-ORFeus transcript were observed at CCR5 CRISPR/Cas9 editing site only in L1-ORFeus, but not in L1-RTm (Supplementary Fig. 3g), in contrast to insertions of lentiviral plasmid (Supplementary Fig. 3f, h) and genomic fragments (Supplementary Fig. 3f) that were comparable in cells expressing L1-ORFeus or L1-RTm. The relative frequency of L1-ORFeus, genomic fragments, and lentiviral insertions varied between MYC, RAG1, and CCR5 editing sites, indicating a competition between these insertional events for the same DSB (Fig. 1c, Supplementary Figs. 2d, 3d).

De novo L1-ORFeus insertions at DSBs induced by CRISPR/Cas9 are insertions independent of the EN consensus motif. We next sought to evaluate features of de novo L1-ORFeus reporter insertions at CRISPR/Cas9-initiated DSBs versus de novo L1-ORFeus reporter insertions genome-wide. By adapting the high-throughput genome-wide translocation sequencing (HTGTS) technique58,59, we implemented one-sided, nested amplification approach to enrich de novo L1 insertion sites for sequencing by capturing the 3′ ends of newly inserted L1 and their flanking genomic DNA. This technique, termed Polya-seq, originates from SV40 Polya and imposes a requirement for L1 insertions to start from the intact 3′ ends and have a minimum 15 bp Poly(A) tail typical of canonical TPRT products, similar to a previously described approach43 (Supplementary Fig. 4a, b). Using Polya-seq, we detected >36,000 of de novo L1-ORFeus insertions from L1-ORFeus expressing cells in the entire genome (Supplementary Fig. 4c) that displayed enrichment of the 7mer consensus motif 5′-TT/AAAAA-3′ at L1 pre-integration sites (Supplementary Fig. 4d), consistent with the previously described L1 integration preference mediated by the ORF2p EN activity37,43,44. As expected, these L1-ORFeus insertions ended with 3′ poly(A) tracts (range 15–75 bp) (Fig. 3a, Supplementary Fig. 4e). In contrast, L1-ORFeus insertions obtained by Polya-seq were markedly diminished with L1-ENm, indicating that Polya-seq preferentially captured endonuclease-dependent L1 retrotransposition events.

By Polya-seq, we identified de novo L1-ORFeus insertions from L1-ORFeus expressing cells that peaked in occurrence at the CRISPR/Cas9 target sites and, strikingly, extended for about 5 kilobases 5′ or 3′ of the DSB (Fig. 3b, c, Supplementary Fig. 4g). These L1-ORFeus insertions were markedly reduced with L1-ENm and undetectable with L1-RTm (Fig. 3b). Interestingly, the consensus motif of L1 pre-integration site was lost within ±5 kb of the CRISPR/Cas9-induced DSB whereas it was readily identified in insertions distant from the DSB (Fig. 3d, Supplementary Fig. 4f, h), indicating that L1 retrotransposition in the region flanking the CRISPR/Cas9-induced DSB does not rely on the EN activity of ORF2p.

PolyA-seq also revealed that de novo L1-ORFeus integrations occurred at CRISPR/Cas9 off-target sites, although less frequently than at on-target sites (Fig. 4a, b). Amplicon sequencing at MYC CRISPR off-target #1 site (MYC OT1)39 further validated these findings, in which RT-dependent retrotransposition of the spliced L1-ORFeus transcript were observed (Fig. 4c, d, Supplementary Fig. 5c), although about 50 times lesser frequency than MYC on-target. In contrast, insertions of lentiviral plasmid or genomic fragments were comparable in cells expressing L1-ORFeus or L1-RTm (Supplementary Fig. 5b, d). L1 insertions also competed with integrations of genomic DNA and lentiviral plasmid at the off-target sites (Fig. 4c, Supplementary Fig. 5a).

Key experiments were reproduced with CRISPR/Cas9 editing the MYC locus and expressing L1-ORFeus or L1-RTm in HeLa (Supplementary Fig. 6a–c) and U2OS cells (Supplementary Fig. 7a–c). Similar to HEK293T cells, amplicon sequencing detected RT-dependent de novo L1-ORFeus retrotransposition events mediated by short microhomologies (Fig. 5a, Supplementary Fig. 6d, e, h, i), in contrast to the comparable lentiviral plasmid and genomic integrations in cells expressing L1-ORFeus or L1-RTm (Supplementary Fig. 6d, f, g). Polya-seq revealed 307 (HeLa) or 370 (U2OS) RT-dependent de novo L1-ORFeus insertions at the MYC locus, respectively (Fig. 5b, c, Supplementary Figs. 6j, k, 7d–g). Like in HEK293T cells, the consensus motif of L1 pre-integration site was lost within ±5 kb of the CRISPR/Cas9-induced DSB (Fig. 5d, Supplementary Fig. 7h).

RT-dependent de novo L1RIP insertions at CRISPR/Cas9 target site. While the L1-ORFeus reporter is advantageous to distinguish
**Fig. 2 Detailed characterization of de novo L1-ORFeus insertions at CRISPR/Cas9 target sites.**

**a** A schematic model of L1-ORFeus insertions at either DNA end of the CRISPR/Cas9-mediated DSB. CRISPR/Cas9 cutting and subsequent end-resection provide a 3′ overhang (green) that can be primed by 3′ end of L1-ORFeus mRNA (red) for retrotranscription to synthesize the L1 cDNA. The L1 insertions (+) in orange represents the orientation of sequenced L1-ORFeus sequence is from 5′ to 3′ (a, left), while (−) in blue represents the orientation of sequenced L1-ORFeus sequence is from 3′ to 5′ (a, right).

**b** L1-ORFeus junction analysis obtained by amplicon sequencing at the MYC locus targeted by MYC CRISPR/Cas9 in HEK293T cells expressing L1-ORFeus. The sequence alignment was centered on the left end of CRISPR-Cas9-mediated DSBs. Other sources include genomic fragments or plasmid insertions.

**c** Density and distribution of L1-ORFeus insertion junctions pooled from MYC (b) and RAG1 (Supplementary Fig. 2i). The bar in the bottom represents the position of the insertion junction that join to the left or right end of the CRISPR/Cas9-mediated DSB in HEK293T cells expressing L1-ORFeus. The orientation of L1-ORFeus insertion junction is shown in orange when the fragment is oriented 5′ to 3′ (+) or blue when 3′ to 5′ (−).

**d** Histogram plots of microhomology lengths in junctions (from c) joining L1-ORFeus to the left or right end of the CRISPR/Cas9-mediated DSB in HEK293T cells expressing L1-ORFeus. The orientation of L1-ORFeus insertions is shown in orange when the fragment is oriented 5′ to 3′ (+) or blue when 3′ to 5′ (−).
de novo L1 insertions from insertions of endogenous L1 genomic fragments, we sought to reproduce key findings with a more physiologic wild-type L1Hs retrotransposon reporter. For these assays, we used a construct containing a full-length endogenous L1Hs sequence competent of retrotransposition (L1RP). This L1RP construct uses native L1RP sequences, together with the native L1 5′ UTR promoter to more closely mimic endogenous transcription. It also contains a 3′ GFP-artificial intron cassette. The L1RP sequence was further mutated to generate the EN mut (H230A) (L1RP-ENm) or the RT mut (D702Y) (L1RP-RTm) reporters (Supplementary Fig. 8a, Supplementary Sequence 13).

Immunoblot analysis showed an expression of the ORF1p protein consistent across L1RP, L1RP-ENm, and L1RP-RTm reporters, and slightly lower than ORF1p protein expression observed with the
Fig. 3 EN consensus motif-independent de novo L1-ORFeus insertions at CRISPR/Cas9 target sites. a Distribution of 3' Poly(A) tract lengths of L1-ORFeus insertions obtained by PolyA-seq in HEK293T cells expressing L1-ORFeus and targeted by MYC CRISPR/Cas9. b Numbers of L1-ORFeus insertions obtained by PolyA-seq at the MYC locus targeted by MYC CRISPR/Cas9 (b, left) or at the RAG1 locus targeted by RAG1 CRISPR/Cas9 (b, right) in HEK293T cells expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. c Detailed view of the distribution of L1-ORFeus insertions at the MYC locus obtained by PolyA-seq in HEK293T cells targeted by MYC CRISPR/Cas9, L1-ORFeus, L1-ENm, L1-RTm, or GFP control. Each line corresponds to an independent L1-ORFeus insertion; number of L1-ORFeus insertions is pooled from two independent PolyA-seq. Corresponding GRO-seq, ChIP-seq (including Pol II, H3K4me3, and H3K27ac), and ATAC-seq profiles are shown (c, bottom). Red arrow shows the MYC CRISPR/Cas9 target site. d Top: Sequence logo representing the consensus motif detected at L1-ORFeus pre-integration sites in example regions proximal or distant to the MYC CRISPR/Cas9 DSB. Each dot represents a chromosomal region encompassing 30 independent L1-ORFeus insertions. Bottom: Spline interpolation curve of consensus motif position-weighted matrix (PWM) score in chromosome 8 obtained by PolyA-seq in HEK293T cells targeted by MYC CRISPR/Cas9 that express L1-ORFeus. The mean interval size covered by the chromosomal region encompassing 30 independent L1-ORFeus insertions was 1353.5 bp (range 198–8440 bp) within approx. ±5 kb of the MYC CRISPR/Cas9 target site, and 2,824,123 bp (range 801,848–6,764,412 bp) outside of the CRISPR/Cas9 target site.

Fig. 4 RT-dependent de novo L1-ORFeus insertions at CRISPR/Cas9 off-target sites. a Detailed view of the distribution of L1-ORFeus insertions at two representative MYC CRISPR/Cas9 off-targets (OT), MYC OT1 (ChrX:153,624,558–153,624,577) (a, left) and MYC OT2 (MYO1SA) (a, right), obtained by PolyA-seq in HEK293T cells targeted by MYC CRISPR/Cas9 expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. b Detailed view of the distribution of L1-ORFeus insertions at the RAG1 CRISPR off-target #1 (RAG1 OT1, DAZAP1) obtained by PolyA-seq in HEK293T cells targeted by RAG1 CRISPR/Cas9 expressing L1-ORFeus, L1-ENm, L1-RTm, or GFP control. c Pie charts show the relative abundance of indicated three main types of insertions (i.e., L1-ORFeus, lentiCRISPR v2 plasmid, and genomic fragments) obtained by amplicon sequencing at the MYC OT1 locus in HEK293T cells targeted by MYC CRISPR/Cas9 that express L1-ORFeus or L1-RTm. d Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC OT1 locus in HEK293T cells targeted by MYC CRISPR/Cas9 that express L1-ORFeus, L1-ENm, L1-RTm, or GFP control.
**Fig. 5** RT-dependent de novo L1-ORFeus insertions at CRISPR/Cas9 target site in HeLa cells.  

- **a** Detailed view of the distribution of L1-ORFeus insertions at the left or right end of the MYC locus obtained by amplicon sequencing at the MYC locus targeted by MYC CRISPR/Cas9 in HeLa cells expressing L1-ORFeus. 
- **b** Numbers of L1-ORFeus insertions at the MYC locus obtained by PolyA-seq targeted by MYC CRISPR/Cas9 in HeLa cells that express L1-ORFeus or L1-RTm. 
- **c** Detailed view of the distribution of L1-ORFeus insertions at the MYC locus obtained by PolyA-seq in HeLa cells targeted by MYC CRISPR/Cas9 expressing L1-ORFeus or L1-RTm. 
- **d** Top: Sequence logo representing the consensus motif detected at L1-ORFeus pre-integration sites in example regions proximal or distant to the MYC CRISPR/Cas9 DSB. Each dot represents a chromosomal region encompassing 30 independent L1 insertions. Bottom: Spline interpolation curve of consensus motif position-weighted matrix (PWM) score in chromosome 8 obtained by PolyA-seq in HeLa cells targeted by MYC CRISPR/Cas9 that express L1-ORFeus. The mean interval size covered by the chromosomal region encompassing 30 independent L1-ORFeus insertions was: 235 bp (range 28–442 bp) within approx. ±5 kb of the MYC CRISPR/Cas9 target site, and 8,025,664 bp (range 863,222–16,390,320 bp) outside of the CRISPR/Cas9 target site.
L1-ORFeus reporter (Supplementary Fig. 8c). The percentage of GFP+ cells was lower with the L1RP reporter compared to L1-ORFeus reporter (Fig. 6a) and the PCR assay across the intron of the GFP reporter showed that the removal of the intron in the L1RP reporter was less efficient than in the L1-ORFeus reporter, and not detected in the L1RP-ENm and L1RP-RTm mutants (Fig. 6b). Overall, these data are consistent with the lower retrotransposition efficiency of a wild-type L1Hs retrotransposon reporter construct compared to the L1-ORFeus reporter.

Next, we repeated key experiments in HEK293T cells with these L1RP reporters (Supplementary Fig. 8b). We detected de novo L1RP retrotransposition events at CRISPR/Cas9 target site by PolyA-seq, indicating that also a more physiologic L1 reporter integrates at CRISPR/Cas9 genome editing site (Fig. 6c, d). Remarkably, no insertions were observed with the L1RP-RTm reporter and only few insertions with the L1RP-ENm reporter (Fig. 6c). The consensus motif of L1RP pre-integration site was lost within ±5 kb of the CRISPR/Cas9-induced DSB whereas it...
De novo L1-ORFeus insertions occur less frequently at editing sites targeted by prime editors. Canonical CRISPR/Cas9-mediated editing has shown some clinical success\(^6\), yet novel CRISPR-based genome editing tools have been developed to expand editing flexibility and enhance safety. Prime editors (PE) enable precise genome editing, including targeted insertions, deletions, all possible base-to-base conversions and their combinations, using Cas9 nickase (H840A) fused to an engineered Moloney murine leukemia virus (M-MLV) reverse transcriptase (RT) that is programmed with a pegRNA to edit DNA sites\(^6\). PE2 nick the target site to expose a 3′-hydroxyl group that could be used to prime the reverse transcription of DNA-containing pegRNA directly into the target site, while PE3 makes a second nick on the non-edited strand to induce its replacement and increase editing efficiency. Importantly, PE3b was then developed as an improved strategy that nick the non-edited strand only after edited strand flap resolution to minimize the presence of concurrent nicks and potential DSB in PE3\(^4\) (Fig. 7a). PE do not require DSB intermediates for their editing activity, raising the possibility that L1 insertions might not occur frequently in their editing sites. Since HEK293T cells have been widely used to validate PE efficiency\(^6\),\(^6\),\(^5\), we used this cell line to assay for de novo L1-ORFeus insertions at editing sites targeted by various prime editors: PE2, PE3, and PE3b, as well as nick only for PE3 served as a nickase-Cas9 generated single-strand break (SSB) control, or nick only for PE3b served as a negative control since this nicking sgRNA binds only to the edited sequence\(^6\) (Fig. 7a).

First, we examined editing sites at the same position in the MYC locus targeted by canonical CRISPR/Cas9 using two different PE editing approaches (+2-4AAAdel and +5GtoC, Supplementary Figs. 9a–c, 10a). Prime editing efficiency at the targeted sites varied with different PE systems ranging from 4 to 14% with PE2, 10 to 19% with PE3b, and 23 to 31% with PE3 (Supplementary Figs. 9d, 10c). Amplicon sequencing at editing sites targeted by prime editors showed the expected integrations of the plasmid expressing the prime editing system (Supplementary Figs. 9f, 11c). Similar L1-ORFeus insertion frequency and characteristics were observed at one additional editing site (FANCF +5GtoT)\(^6\) (Supplementary Figs. 10b, d, f, 11b, d), as well as comparable pegRNA, plasmid, and genomic fragments insertions in cells expressing L1-ORFeus or L1-RTm (Supplementary Figs. 10f, 11f, d, h). Mechanistically, the outcomes of de novo L1-ORFeus insertions both at MYC and FANCF PE3-editing sites were either L1-ORFeus joined to the pegRNA-edited sequence or to a non-edited or editing site deleted sequence (Supplementary Fig. 12e–g). These L1-ORFeus insertions showed joining to the nicked sites via short microhomologies, mostly 1 to 6 bp with a decreasing trend over length (Supplementary Fig. 12h).

Moreover, PolyA-seq validated de novo L1-ORFeus insertions at PE3 editing sites in a RT-dependent manner, and again were at a lower level compared to canonical CRISPR/Cas9 (Fig. 7f, Supplementary Fig. 12i). Accordingly, de novo L1-ORFeus insertions were undetectable at PE3b editing sites by PolyA-seq (Fig. 7f), demonstrating that PE3b significantly improves the safety of prime editing, consistent with its reduced indels formation by minimizing the presence of DSB\(^6\). Overall, we concluded that RT-dependent L1-ORFeus insertions can exploit DNA lesions associated with PE3 editing though were at a lower level compared to canonical CRISPR/Cas9, and are rare during PE2 or PE3b editing, shedding light on the safety concerns of research and possible clinical applications of these prime editors.

De novo L1-ORFeus insertions are rare during base editors mediated DNA editing. Two classes of base editors have been developed to install or correct these pathogenic point mutations\(^6\). In cytosine base editors (CBEs), the APOBEC1 enzyme is fused with a nickase Cas9 (D10A) for targeting via sgRNA and with UGI to increase the accuracy and efficacy of base editing, converting G:C base pair to T:A base pair. In adenosine base editors (ABEs), a E. coli-derived tRNA adenine deaminase (Tada) is fused with nickase Cas9 to convert an A:T base pair to a G:C base pair. Briefly, in contrast to dCas9-fused BE2, BE3 uses nickase Cas9 to specifically nick the non-edited strand to increase the editing efficiency\(^6\),\(^6\),\(^9\). Fusing BE4 to Gam, a bacteriophage Mu protein that binds DSBs greatly reduces indel formation during base editing\(^6\), while optimization of codon usage and nuclear location sequences (NLS) resulted in the most advanced and efficient AncBE4max or ABE8e\(^7\),\(^1\). Similar to prime editors, BE systems do not require the formation of a DSB intermediate for editing\(^6\). Thus, we also assayed for de novo L1-ORFeus insertions in multiple editing sites targeted by several CBEs (BE2, BE3, BE4-Gam, and AncBE4max) and one ABE (ABE8e) (Fig. 8a, Supplementary Fig. 13a, b).

In both MYC and FANCF loci, the base editing efficiency was high in all five BEs, with highest in AncBE4max and ABE8e, and comparable in cells expressing L1-ORFeus or L1-RTm (Supplementary Fig. 13d, e). Amplicon sequencing analysis revealed that de novo L1-ORFeus insertions were rare at the MYC site (Fig. 8b) and not significantly frequent at FANCF site (Fig. 8c, Supplementary Fig. 13l) edited with all five BE systems, which also validated by PolyA-seq (Fig. 8e, Supplementary Fig. 12j). We could find few plasmid and genomic fragments integrations among these BEs, though they were at much lower levels than in cells edited with canonical CRISPR/Cas9 or PE3 (Supplementary Fig. 13g, h). Notably, BE2 and BE4-Gam showed the lowest insertions (including L1-ORFeus, plasmid or genomic fragments) compared to the other BE systems (Fig. 8b, c, Supplementary Fig. 13g, h), consistent with their lowest rate of indels formation\(^6\),\(^7\).
Moreover, we mapped de novo L1-ORFeus insertions in cells undergoing ABE8e editing at two disease-relevant loci (BCL11A enhancer and HBG1/2 promoter) currently being evaluated for gene therapy of sickle cell disease and β-thalassemia. Not surprisingly, de novo L1-ORFeus insertions were almost undetectable both by amplicon sequencing (Fig. 8d) and PolyA-seq (Fig. 8f, Supplementary Fig. 13k, n) at all editing sites despite a high base editing efficiency (Supplementary Fig. 13f). The few insertions of plasmid and genomic fragments at BCL11A enhancer and HBG1/2 promoter editing sites were consistent with their low levels of indels formation as reported (Supplementary Fig. 13i). Overall, these data suggest that de novo L1-ORFeus insertions are rare with BE systems that avoid DSB formation, supporting their promising efficacy and safety for potential gene therapy even in cell types known to support retrotransposition of endogenous L1.

**Discussion**

In this work, we describe LINE-1 retrotransposon insertions into CRISPR/Cas9-mediated DSBs as an outcome of canonical CRISPR/Cas9 genome editing as well as PE3 prime editing. Retrotransposed sequences are thus added to a growing list of insertions known to occur at DSBs generated by CRISPR/Cas9.
**Fig. 7 Safety evaluation of prime editors in HEK293T cells.** a Example diagram of the PE systems editing MYC + 5Gtoc. While nCas9 (H840A) and pegRNA are required for all prime editing strategies, PE3 contains a nicking sgRNA at +42 bp from the pegRNA-mediated nick to increase the editing efficiency. PE3b contains a sgRNA (−5 bp from the pegRNA-mediated nick) with spacer that match the edited strand to minimize the presence of concurrent nicks. Nick only for PE3 introduces a single-strand break in the non-edited strand by a nicking sgRNA. b Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC locus targeted by PE2 (MYC + 2-4AAA del), PE3 (MYC + 2-4AAA del), or nick only for PE3 in HEK293T cells expressing L1-ORFeus, L1-Enm, L1-RTm, or GFP control. Number of L1-ORFeus insertions was pooled from three independent experiments with similar results. c Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC locus targeted by PE2 (MYC + 5Gtoc), PE3 (MYC + 5Gtoc), PE3b (MYC + 5Gtoc), or nick only for PE3/PE3b in HEK293T cells expressing L1-ORFeus or L1-RTm. d Numbers and fragment lengths of L1-ORFeus insertions obtained by amplicon sequencing at the MYC locus targeted by PE3 (MYC + 2-4AAA del) in HEK293T cells expressing L1-ORFeus or L1-RTm. Example L1-ORFeus insertional sequences in cells expressing L1-ORFeus are shown in Supplementary Sequence 4. Example L1-ORFeus insertions bridging the intron of GFP in cells expressing L1-ORFeus are shown in Supplementary Sequence 5. e Pie charts show the relative abundance of four main types of insertions obtained by amplicon sequencing at the MYC locus targeted by PE3 (MYC + 2-4AAA del) in HEK293T cells expressing L1-ORFeus, L1-Enm, L1-RTm, or GFP control. f Detailed view of the distribution of L1-ORFeus insertions at the MYC locus obtained by PolyA-seq in HEK293T cells targeted by PE3 (MYC + 2-4AAA del), PE3 (MYC + 5Gtoc), or PE3b (MYC + 5Gtoc) that express L1-ORFeus or L1-RTm.

**Fig. 8 De novo L1-ORFeus insertions are rare in base editing.** a Diagram of the base editors. BE2, BE3, BE4-Gam, and AncBE4max are developed as cytosine base editors (CBEs), while ABE8e is an adenine base editor (ABE). b, c Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC (b) or FANC (c) base editing site targeted by BE2, BE3, BE4-Gam, AncBE4max, or ABE8e in HEK293T cells expressing L1-ORFeus or L1-RTm. d Numbers of L1-ORFeus insertions obtained by amplicon sequencing at the MYC base editing site targeted by ABE8e in HEK293T cells expressing L1-ORFeus or L1-RTm. e Detailed view of the distribution of L1-ORFeus insertions at the MYC base editing site obtained by PolyA-seq in HEK293T cells targeted by AncBE4max or ABE8e that express L1-ORFeus or L1-RTm. f Detailed view of the distribution of L1-ORFeus insertions at the BCL11A enhancer base editing site obtained by PolyA-seq in HEK293T cells targeted by ABE8e that express L1-ORFeus or L1-RTm.

including various vectors such as plasmids used for HR-mediated editing\(^{18,19}\), lentiviral\(^{20}\), or adeno-associated virus (AAV)\(^{21}\) vectors encoding CRISPR/Cas9, as well as DNA fragments from freely available DNA\(^{22}\). In contrast to these known events that rely on double-stranded DNA (dsDNA) sequences that are inserted during DSB repair, the mechanism of de novo L1 insertions is quite distinct, and appears to rely on the ability of RNA to template break repair in an RT-dependent manner at CRISPR/Cas9-mediated DSBs. Since large insertions are difficult to detect and can be easily missed by PCR and DNA sequencing that normally used to evaluate the safety and outcomes of CRISPR/Cas9 editing\(^{18,19}\), we developed PolyA-seq together with improved amplicon sequencing to fully appreciate these L1 insertions. Together, these approaches captured more than 2500 de novo L1 retrotransposition events at CRISPR/Cas9 editing sites, as well as 141,490 insertions of the lentiviral sequences that encode for the CRISPR/Cas9 and 179,264 integrations of genomic DNA sequences that mostly originated from regions flanking the DSB.

Mechanistically, de novo L1 insertions at sites of genome editing are similar to canonical retrotransposition events in some...
ways, but distinct in others. Like canonical retrotransposition, de novo L1 insertions at genome editing sites are dependent on RT activity but they are distinct in that they are largely independent of EN activity and show a random sequence motif pattern instead of the typical preference for TT/AAA/AA target sites (Figs. 3d, 5d, 6e, Supplementary Figs. 4h and 7h) with junctions that are either direct or mediated by short microhomologies (Fig. 2d, Supplementary Fig. 6i). This suggests that the initiation of reverse transcription depends on a DNA break generated by CRISPR/Cas9 and that its resolution requires initiation of reverse transcription depends on a DNA break resection of the DSB before L1 integration. In contrast, canonical retrotransposition events, which were EN-dependent and showed TT/AAA insertion site preference predominated in more distant regions from the CRISPR/Cas9-mediated DSBs, including on the same chromosome or in the rest of the genome. Of note, plasmids are known to integrate into DSBs initiated by CRISPR/Cas9 cutting18,19. Although we induced CRISPR/Cas9 DSBs 5 days after L1 plasmid transfection, still few copies of plasmid are present in cells at the time of CRISPR/Cas9 cutting. Consistently, the rare sequences found in GFP control samples aligned to the GFP portion of the vector and are likely plasmid integrations (Fig. 1e, Supplementary Fig. 2f). The few insertions mapping to the intron of the GFP cassette observed in L1-ORFeus and L1-ENm, but not in the GFP control and L1-RtM (Fig. 1e, Supplementary Fig. 2f) might be also plasmid insertions or due to an inefficient intron removal during the pre-mRNA splicing, that can occur in experiments conducted with a transfected plasmid reporter75.

Unwanted outcomes of DSBR repair after canonical CRISPR/Cas9 targeting are a potential threat even if occurring in a small fraction of the edited cell population. These unwanted outcomes can occur both at the on-target or off-target sites of CRISPR/Cas9 editing. Notably, Polya-seq allows for genome-wide surveys of de novo L1 integrations, and not only identified L1 insertions at on-target but also at multiple off-targets generated by CRISPR/Cas9 (Fig. 4, Supplementary Fig. 5). Thus, detection of recurrent de novo L1 insertion sites could represent an additional method to detect off-target activities of various nuclease-based genome editing tools, parallel to the existing GUIDE-seq20, SITE-seq76, LAM-HTGTS10, PEM-seq9, IDLV assay16, CHANGE-seq49, CIRCLE-seq72, and DISCOVER-seq78.

Increasing the safety of CRISPR/Cas9 editing is essential for gene therapy.9 To this end, several CRISPR/Cas9-based genome editing tools that do not require the formation of DSB intermediates or donor DNA have been recently developed, including PE and BE systems. PE2 and PE3b systems as well as CBE and ABE systems rarely allowed de novo L1 insertions, similar to the low genomic or plasmid integrations at the editing site, consistent with their low propensity to permit indels.4,80. We also compared canonical CRISPR/Cas9 with nickase-Cas9 using nick only for PE3 (Fig. 7a). In nick only for PE3, when SSBs are generated by nickase-Cas9 only rare L1-ORFeus, plasmid or genomic insertions were detected (Fig. 7b, c, Supplementary Figs. 9e and 10e, f). These results demonstrate that L1 insertions are much more frequent in DSBs than in SSBs generated by CRISPR/Cas9 (Supplementary Figs. 12a–g) through a lower level compared to canonical CRISPR/Cas9. Interestingly, sequences with de novo L1-ORFeus insertions in the PE3 system from ampiclon sequencing showed an increased editing efficiency compared to the bulk sequences (Supplementary Fig. 12a–d). A similar increase of editing efficiency was also observed in sequences with inserted PE plasmid, pegRNA, or genomic fragments (Supplementary Fig. 12a–c).

While our experimental models are based on ectopically-expressed L1-ORFeus or L1RP reporters to show de novo L1 insertions, insertions generated by active endogenous L1 sequences are expected to be lower, and may vary according to the expression level and retrotransposition activity of the endogenous L1 in the cell type that is being edited. L1 retrotransposition is commonly suppressed in human somatic cells, but it appears to be reactivated not only in human cancers11,83, but also in human embryonic stem cells (hESCs) and induced pluripotent stem cells (iPSCs) that are frequently used for CRISPR/Cas9 editing12,84,85. Although HEK293T express lower L1 ORF1p than iPSCs (Supplementary Fig. 1g), suggestive of a low endogenous activity, we analyzed whether endogenous L1 insertions could be detected at the CRISPR/Cas9-induced DSBs by aligning to L1Hs sequences. Currently, Homo sapiens specific L1 (L1Hs) is the most active autonomous L1 element in the genome34,35,86. We detected L1Hs sequences with bias toward the 3′ of the L1Hs both in MYC and RAG1 editing sites (Supplementary Fig. 14a–c, Supplementary Sequence 12). However, it is difficult to conclude whether these L1Hs sequences represent retrotransposition events of active endogenous L1Hs retrotransposon occurring at CRISPR/Cas9-mediated DSBs or rather genomic insertions of non-active L1 sequences.

Overall, our results show that retrotransposition of L1 elements is a frequent outcome of canonical CRISPR/Cas9 editing. Although these findings will need to be further confirmed in more physiological or clinical settings, they should be carefully evaluated and searched for when editing is intended for a therapeutic approach given the potential pathogenetic effect of L1 insertions34.

Methods

Plasmids, oligonucleotides, and cloning. The human codon-optimized, synthetic L1-ORFeus was modified from the human codon-optimized, synthetic L1-ORFeus constructs published by the Boeke Lab86. The detailed information of pCEP4 GFP control, pCEP4 L1-ORFeus, pCEP4 L1-ORFeus-ENm (H230A) (L1-ENm), and pCEP4 L1-ORFeus-RtM (D702Y) (L1-RtM) can be found in Supplementary Table 1. Briefly, CMV promoter controls expression of the L1-ORFeus reporter cassette in the pCEP4-derived constructs. The L1-ORFeus mutant constructs contain the following modifications: the EN mutant construct contains the H230A residue in ORF2p (L1-ENm) and the RT mutant construct contains the D702Y residue in ORF2p (L1-RtM), which are disrupted residues critical for retrotransposition, as previously shown87,88. The GFP control vector was generated by cloning an intron-less GFP reporter in the same pCEP4 vector backbone. pCEP4 L1RP (L1RP) is modified from the Boeke Lab and published previously89. (Addgene plasmid #131392), pCEP4 L1RP uses native L1RP sequences and the native L1 STUB promoter, sharing the same pCEP4 vector backbone and GFP. Artificial Intron cassette with pCEP4 L1RP vector; likewise, two mutants pCEP4 L1RP-ENm (H230A) and pCEP4 L1RP-RtM (D702Y) were generated (see Supplementary Fig. 8a and Supplementary Sequence 13), lentiCRISPR v2 (Addgene plasmid #52961) was a gift from Feng Zhang. Plasmids expressing sgRNAs were constructed by ligation of annealed oligonucleotides into BsmBI-digested LentCRISPR v2. pCMV-PGK-P2 (Addgene plasmid #132775) and pU6-pegRNA-GG acceptor (Addgene plasmid #132777) were gifts from David Liu. Plasmids expressing pegRNAs or nicking sgRNAs for (PE3 and PE3b) were constructed by Golden Gate assembly using Bsal-digested u6-pegRNA-GG acceptor vector as reported previously44. pCMV-BE2 (Addgene plasmid #73020), pCMV-BE3 (Addgene plasmid #73021), pCMV-BE4-Gam (Addgene plasmid #100806), pCMV-ANC-Be4max (Addgene plasmid #112094), and pCMV-AB-E8e (Addgene plasmid #138489) were gifts from David Liu. Oligonucleotides were purchased from Integrated DNA Technology (IDT), and sequences of sgRNAs and pegRNAs native to the work were listed in Supplementary Table 1. CRISPR off-targets (OT) were identified as previously reported10,19, sequences were listed in Supplementary Table 2. All constructs were sequence-verified and purified by NucleoBond Plasmid DNA purification PC100/500 (Takara).
Cell culture. HEK293T, HeLa, and U2OS cells were maintained in DMEM supplemented with 10% fetal bovine serum (FBS), penicillin-streptomycin (100 units per ml), and L-glutamine (2 mM). Cells were cultured at 37°C with 5% CO₂ and tested negative for mycoplasma. HEK293T and HeLa cells were authenticated as they were purchased from ATCC; U2OS cells were provided by Jeremy Stark, and tested negative for mycoplasma. HEK293T and HeLa cells were authenticated as they were purchased from ATCC; U2OS cells were provided by Jeremy Stark, and tested negative for mycoplasma. HEK293T, HeLa, and U2OS cells were maintained in 10% FBS-containing DMEM, supplemented with 1% penicillin/streptomycin. U2OS cells were authenticated as they were purchased from ATCC; U2OS cells were provided by Jeremy Stark, and tested negative for mycoplasma. HEK293T, HeLa, and U2OS cells were maintained in 10% FBS-containing DMEM, supplemented with 1% penicillin/streptomycin. U2OS cells were authenticated as they were purchased from ATCC; U2OS cells were provided by Jeremy Stark, and tested negative for mycoplasma. HEK293T and HeLa cells were authenticated as they were purchased from ATCC; U2OS cells were provided by Jeremy Stark, and tested negative for mycoplasma.

Amplicon sequencing data analysis. Sequences in library were processed as previously described92. Briefly, reads for each experimental condition were demultiplexed by designed forward and reverse barcodes (followed portion of primer totally plus barcode up to 8 bp was appended in case barcode length is <8 bp). To enhance specificity of target sequences, reads were further filtered by the presence of primer plus additional 10 downstream bases.

Insertion sources analysis. For focusing on insertion portion of PCR fragments obtained by amplicon sequencing, the barcode, primer, and target reference portion of the sequences were trimmed. Based on trimmed sequences, the exact duplicates treated as PCR repeats and sequences with length <30 bp were further eliminated. To analyze insertion sources, we next employed BOWTIE2 v2.1 to align the processed sequences to different references like human genome (GRCh38/ hg38), exogenous LINE1 (ORFeus), and youngest active L1Hs93,94. LIRP(L1PA1), and relatively younger inactive L1PA2, L1PA3, L1PA8, to identify endogenous L1 insertions we particularly mapped sequences against the youngest reference elements annotated as L1Hs in RepeatMasker v4.0.8 (http://www.repeatmasker.org). Therefore, we specifically excluded the alignments where contain the common sequences with lentCRISPR v2 plasmid and attributed those reads to lentCRISPR v2 plasmid; excluded the alignments where contain the common sequences with endogenous L1Hs and attributed those reads to L1Hs reference. As for plasmids also, the blastn-indicated plasmid was chosen for correct reads alignment, i.e., lentCRISPR v2 for canonical CRISPR/Cas9 editing analysis, pCMV-PE2 and pU6-ppegRNA for PE editing analysis, pCMV-BE2 for BE editing analysis, pCMV-BE3 for BE3 editing analysis, pCMV-BE4-Gam for BE4-Gam editing analysis, pCMV-AnC-Be4max for AncB-Be4max editing analysis, pCMV-ABE8e for ABE8e editing analysis, respectively (see Supplementary Table 3). The aligned sequences were further cleaned by removing duplicates by using PICKARD MarkDuplicates tool and those with low mapping quality (mapq < 2). Next, we used BEDTOOLS v2.29 bamtobed to convert paired-end sequence alignments in BAM format into BEDPE records, and further excluded redundant sequences having same sequence alignment length within 1 bp difference, reads for each experimental condition were aligned to L1-ORFeus insertions and their alignments from each reference. Insertion dot and fragment maps were plotted by using R ggplot to depict the distribution of insertion sources in reference.

L1-ORFeus insertion junction analysis. To understand how L1-ORFeus mechanically joins to break point of CRISPR/Cas9-mediated DSBs to complete its insertion, we analyzed the position distribution and microhomology of insertion junction. First, the raw paired-end sequences that contain L1-ORFeus insertions identified by above insertion sources analysis obtained by amplicon sequencing were particularly selected and aligned to amplicon target reference sequence and LINE1 (ORFeus) by using BLAST v2.11, respectively. The forward sequences alignments records against target reference and LINE1 (ORFeus) were used to determine the left end of DSBs. The PE mapping of the reads containing paired-end pairs between break end and insertion junction were calculated as microhomology, while the gap between them was considered as “other sources” of insertions (include genomic fragments or plasmid insertions). Likewise, the reverse alignments records were calculated for right end of DSB and insertion junction. Both forward and reverse records aligning insertions against LINE1 (ORFeus), and indicated plasmids, respectively. Because the PE alignment of CRISPR/Cas9-mediated cut site (three base pairs before PAM) were plotted by using R ggplot to illustrate the mechanisms of insertions at DSBs. BLASTN parameters ‘-reward’ and ‘penalty’ were set as 1 and −2 for more conserved result. Only alignments record with alignment start in PCR target reference no more than 8 bp from 5’ end for forward and 3’ end for reverse sequence were considered as valid portion of PCR target reference. Furthermore, sequences with PCR target reference alignment length <18 bp or LINE1 (ORFeus) alignment length <30 bp were determined as other sequence alignment. The reads containing pegRNA nick strand were also aligned to p6-ppegRNA-GG-acceptor plasmid to identify pegRNA portion required to have edited base, minus alignment orientation.
and at least 10 bp alignment length. Then by taking into account the LINE1 (ORFeus) alignment records, pegRNA portion and the barcode and primer portion of the library were aligned to target genomic DNA. Subsequently, sequences were masked for alignment analysis. Next, the processed sequences were aligned to target genomic DNA (GRCh38/hg38) using BLAT, and finally aligned sequences were further cleaned by removing PCR repeats (reads with same nucleotide in position in alignment to the reference genome and a start position in the read <3 bp apart), invalid alignments (including alignment scores <30, reads with multiple alignments having a score difference <4 and alignments having 10- bp gaps) and alignment artifacts (for example, random ligation with HaeIII restriction sites). L1 insertion junction position was determined based on the genomic position of the 5' end of the aligned read.

Consensus motif analysis at L1 insertion sites. The genomic DNA sequence flanking L1 insertion positions −3 (upstream) and +8 (downstream) was retrieved to analyze the consensus sequence motif at the L1 insertion sites obtained by PolyA-seq. R package ‘Biostings’ and ‘gseqlogo’ were performed to generate sequence logos. The corresponding position-weighted matrix (PWM) was calculated by R package ‘seqLogo’. The motif score is defined as the sum of information content at each position.

L1 3' Poly(A) tract analysis. To identify the 3' Poly(A) tract additionally added to L1 intact 3' end, we generated 200 bp Poly(A) reference sequence and BLASTN all reads with L1 insertion to such Poly(A) reference, restricting query start in read mapped to Poly(A) reference less than query start in read to Human genome to ensure that the identified Poly(A) was inserted and followed by genomic sequence at insertion junction. The default parameter ‘word_size’ of BLAST is 11, so Poly(A) size <11 was ignored accordingly. The above approach was applied to PolyA-seq.

Editing efficiency analysis. Nucleotide mutations were calculated from raw sequences as previously described. Briefly, sequences with mean quality score <20 and length <50 were excluded. The remaining sequences were aligned to target reference sequence using BLASTN. Only sequences with alignment length >80 were considered for calculation of editing efficiency (mutation rate). The real mutations were determined by passing a Neighborhood Quality Standard (NQS) filter criterion that requires a minimum Phred score of 30 for the mutation itself, and 20 for the five adjacent bases on either side. Finally, the percentage of reads that mutated at target base site represents editing efficiency.

Data reporting. No statistical methods were used to predetermine sample size.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The source data used to generate figures in this study are provided in the Supplementary Information/Source data file. Information for amplicon sequencing libraries can be found online in Supplementary Data 1: Information of Amplicon sequencing libraries; information for PolyA-seq libraries can be found online in Supplementary Data 2: Information of PolyA-seq libraries. All sequencing data have been deposited in the Gene Expression Omnibus database under accession number GSE178440. For GRO-seq analysis, publicly available HeK293T GRO-seq data were acquired from NCBI Gene Expression Omnibus (GEO) database (accession: GSM1249889). The downloaded hg18-based bigwig files were converted to hg38 by using CrossMap v0.5. For ChiP-seq and ATAC-seq analysis, publicly available HeK293T ChiP-seq and ATAC-seq data were acquired from GEO, and the accession numbers are Pol II: GSM1249891; H3K4me3: GSM1249885, H3K27ac: GSM1249889, ATAC-seq: GSM3271043. Raw reads are aligned against to Human genome (hg38) using BWA. Aligned BAM files are filtered by removing low quality and unpaired reads, and de-duplicated. Source data are provided with this paper.

Code availability Source code for genome editing analysis tools (GEAT) developed in our laboratory to perform the analysis is available at https://github.com/geatools/geat97.

Received: 24 September 2021; Accepted: 14 June 2022; Published online: 27 June 2022

References
1. Malek, H. L. Treatment by CRISPR-Cas9 gene editing—a proof of principle. N. Engl. J. Med. 384, 286–287 (2021).
2. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821 (2012).
3. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013).
4. Mali, P. et al. RNA-guided human genome engineering via Cas9. Science 339, 823–826 (2013).
5. Chu, V. T. et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat. Biotechnol. 33, 543–548 (2015).
6. Adikusuma, F. et al. Large deletions induced by Cas9 cleavage. Nature 560, E8–E9 (2018).
7. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat. Biotechnol. 36, 765–771 (2018).
8. Cullot, G. et al. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. Nat. Commun. 10, 1136 (2019).
9. Liu, M. et al. Global detection of DNA repair outcomes induced by CRISPR-Cas9. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab686 (2018).
10. Trock, R. L. et al. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat. Biotechnol. 33, 179–186 (2015).
11. Leibowitz, M. L. et al. Chromothripsis as an on-target consequence of CRISPR-Cas9 genome editing. Nat. Genet. 53, 895–905 (2021).
12. Zuccaro, M. V. et al. Allele-specific chromosome removal after Cas9 cleavage in human embryos. Cell 183, 1650–1664.e1615 (2020).
13. Zuo, E. et al. CRISPR-Cas9-mediated targeted chromosome elimination. Genome Biol. 18, 224 (2017).
14. Papathanasou, S. et al. Whole chromosome loss and genomic instability in mouse embryos after CRISPR-Cas9 genome editing. Nat. Commun. 12, 5855 (2021).
15. Manoj, A. G. et al. Unexpected gene activation following CRISPR-Cas9-mediated genome editing. Embo Rep. 23, e53902 (2022).
16. Wang, X. et al. Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nat. Biotechnol. 33, 175–178 (2015).
17. Harlon, K. S. et al. High levels of AAV vector integration into CRISPR-induced DNA breaks. Nat. Commun. 10, 4439 (2019).
18. Norris, A. L. et al. Template plasmid integration in germline genome-edited cattle. Nat. Biotechnol. 38, 163–164 (2020).
84. Klawitter, S. et al. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat. Commun. 7, 10286 (2016).
85. Alanis-Lobato, G. et al. Frequent loss-of-heterozygosity in CRISPR-Cas9 edited early human embryos. Proc. Natl Acad. Sci. USA 202004832, https://doi.org/10.1073/pnas.2004832117 (2021).
86. De Masi, C., Spitalieri, P., Murdocca, M., Novelli, G. & Sangiuliano, F. Application of CRISPR/Cas9 to human-induced pluripotent stem cells: from gene editing to drug discovery. Hum. Genomics 14, 25 (2020).
87. Smit, A. F., Tóth, G., Riggs, A. D. & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. 246, 401–417 (1995).
88. Adney, E. M. et al. Comprehensive scanning mutagenesis of human retrotransposon LINE-1 identifies motifs essential for function. Genetics 213, 1401–1414 (2019).
89. Kimberland, M. L. et al. Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum. Mol. Genet. 8, 1557–1560 (1999).
90. Bhargava, R., Lopezcolorado, F. W., Tsai, L. J. & Stark, J. M. The canonical non-homologous end joining factor XLF promotes chromosomal deletion rearrangements in human cells. J. Biol. Chem. 295, 125–137 (2020).
91. Zhang, H. et al. Rat and human STINGs profile similarly towards anticancer/antiviral compounds. Sci. Rep. 5, 18035 (2015).
92. Tao, J. et al. Nonspecific DNA binding of cGAS N terminus promotes cGAS activation. J. Immunol. 198, 3627–3636 (2017).
93. Gao, J. et al. Identification and characterization of phosphodiesterases that specifically degrade 3’-cyclic GMP-AMP. Cell Res. 25, 539–550 (2015).
94. Deng, M. J. et al. Novel mechanism for cyclic dinucleotide degradation revealed by structural studies of vibrio phosphodiesterase V-cGAP3. J. Mol. Biol. 430, 5080–5093 (2018).
95. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic Genome Res. 110, 462–467 (2005).
96. Liu, M. et al. Two levels of protection for the B cell genome during somatic hypermutation. Nature 451, 841–845 (2008).
97. Wang, Q. Frequency and mechanisms of LINE-1 retrotransposon insertions at CRISPR/Cas9 sites. qwang-flyinggene/GEAT: GEAT v0.1 (v0.1). Zenodo https://doi.org/10.5281/zenodo.6592772 (2022).

Acknowledgements
We thank John V. Moran for critical reading the manuscript; William Pu and Maksymilian Prondzynski for providing iPSC (WTC-11) protein extract, Thorsten Schlaeger and Yang Tang for providing hESC protein extract. We thank Boxun Zhao, Daniel Ardeljan, and Jared P. Steranka for helpful discussion. This work was supported by 1R01-CA222598 to R.C.

Author contributions
J.T. and R.C. conceived the project, designed experiments, and wrote the manuscript with the help of Q.W.; J.T. performed all of the experiments; Q.W. designed and performed the bioinformatics analysis and tools development; C.M.-D. and K.B. provided key reagents, contributed to design and interpretation of the experiments, and contribute to writing the manuscript; J.T., Q.W., and R.C. analyzed and interpreted the data, designed the figures; R.C. supervised the study.

Competing interests
The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31322-3.

Correspondence and requests for materials should be addressed to Jianli Tao or Roberto Chiarle.

Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.