Title: CRISPR-Enabled Autonomous Transposable Element (CREATE) for RNA-based gene editing and delivery

Authors: Yuxiao Wang[1]*†, Ruei-Zeng Lin[1]†, Meghan Harris[1], Bianca Lavayen[1], Bruce McCreedy, Robert Hofmeister[1] and Daniel Getts[1]*

Affiliations:

[1] Myeloid Therapeutics Inc., Cambridge, MA 02139, USA

* Corresponding authors. Email: daniel@myeloidtx.com (D.G.); ywang@myeloidtx.com (Y.W.);

† These authors contributed equally

Abstract:

To address a wide range of genetic diseases, genome editing tools that can achieve targeted delivery of large genes without causing double-stand breaks (DSBs) or requiring DNA templates are necessary. Here, we introduce the CRISPR-Enabled Autonomous Transposable Element (CREATE), a genome editing system that combines the programmability and precision of CRISPR/Cas9 with the RNA-mediated gene insertion capabilities of the human LINE1 (L1) element. CREATE employs a modified L1 mRNA to carry a payload gene, and a Cas9 nickase to facilitate targeted editing by L1-mediated reverse transcription and integration without relying on DSBs or DNA templates. Using the system, a 1.1 kb gene cassette comprising an EF1α promoter and green fluorescent protein (GFP) gene is inserted into several genomic loci of multiple human cell lines. Mechanistic studies reveal that the CREATE system is highly specific with no observed off-target events. Together, these findings establish CREATE as a programmable gene delivery technology solely based on RNA components, enabling large-scale in vivo genome engineering with broad therapeutic potential.

Introduction

The recent FDA approval of exagamglogene autotemcel (Casgevy®) for sickle cell disease has marked a significant milestone in CRISPR/Cas9-based gene therapies[1]. Traditional CRISPR/Cas9 systems, which introduce genomic double-strand breaks (DSBs), are best suited to disrupt defective genes and while CRISPR/Cas9 can also be used to insert new DNA templates via the homology-directed repair (HDR) pathway, the integration efficiency is low[2]. To avoid DSB-associated chromosomal deletions or translocations[3–6], Base Editors (BEs) and Prime Editors (PEs), which perform gene edits by nicking a single DNA strand rather than creating DSBs, have been introduced[7,8]. BEs enable targeted single nucleotide conversions, while PEs can insert or replace up to 40 base pairs without a donor DNA template. Despite these improvements, these editors face limitations in payload capacity, which restricts their application in treating most genetic disorders involving a spectrum of mutations, insertions, or deletions across extensive genomic regions.

Expanding on the capabilities of PEs, twin-prime and template-jumping PEs have been developed to incorporate larger payloads of a few hundred base pairs (bps)[9]. Furthermore, combining PEs with sequence-specific serine recombinases has extended this capacity up to 36 kB for the delivery of large DNA fragments encoded in a donor plasmid[10,11]. Despite their potential, the technical complexity of these systems, which require the co-delivery of multiple components such as the PE editor with pegRNA, the serine recombinase, and a DNA donor plasmid, presents substantial challenges for clinical applications[12].

The LINE1 retrotransposon (referred to as L1) is a mobile genetic element that comprises 17% of the human genome[13]. It naturally replicates via an RNA intermediate, capable of efficiently reverse-transcribing large mRNA sequences into cDNA and integrating into human genome[14,15]. L1 replication involves a bicistronic mRNA encoding two proteins, ORF1p and

48 ORF2p[14]. The ORF1p is an RNA-binding protein that interacts with L1 mRNA transcripts[16].
49 ORF2p encodes an endonuclease (EN) and a reverse transcriptase (RT) domain[17,18]. A distinctive
50 feature of L1 is its 'cis preference,' where ORF1p and ORF2p associate with their own mRNA
51 transcript to form a ribonuclear protein complex (RNP)[19], facilitating the reverse transcription
52 and insertion of the L1 mRNA[20,21]. The endonuclease domain of ORF2p cleaves at a redundant
53 5'TTTT/AA3' consensus sequence, initiating a target-primed reverse transcription (TPRT)
54 process mediated by the RT domain of ORF2p that converts the L1 mRNA into cDNA and
55 integrates into the genome[22,23]. The ORF2p RT domain was shown to be highly processive,
56 capable of reverse transcription of long RNA sequences[23,24]. However, the consensus motif
57 recognized by the EN domain of L1 ORF2p is found at numerous sites in the genome, precluding
58 its application for targeted, programmable gene editing. A recent study attempting to combine
59 Cas9 with another retroelement, the R2 retroelement via a direct fusion approach was unable to
60 achieve complete retrotransposition and integration[25].

61 Intrigued by the properties of L1, we developed the CRISPR-Enabled Autonomous
62 Transposable Element (CREATE) system co-opting the transposable element for site-specific
63 gene delivery. By combining the precision and programmability of CRISPR/Cas9 with the large-
64 scale genome re-writing capability of L1, we inserted a 1.1 kb nucleotide sequence comprising a
65 promoter and green fluorescent protein (GFP) into several sites in the human genome guided by
66 Cas9. Together our findings show that the CREATE system is a programmable and highly
67 specific gene delivery platform capable of delivering a payload of over 1 kb with only RNA
68 components, thus offering considerable prospects for diverse therapeutic applications.

69

Results

Design of the CREATE editing system

The CREATE system encodes, on a single mRNA, the codon optimized L1 components ORF1p and ORF2p, separated by the native inter-ORF sequence which mediates an unconventional bicistronic expression of the two proteins[26] (Fig. 1a). To facilitate nuclear import, the ORF2p is flanked with an N-terminal SV40 nuclear localization signal (NLS) and a C-terminal nucleoplasmin NLS, respectively. The inactivation of the catalytic residue within the EN domain of ORF2p (mutation D205A) is a critical modification to prevent DNA cleavage of the degenerate consensus motif recognized by natural L1. The payload cassette consisting of a promoter and the gene of interest is placed at the 3'UTR of the L1 mRNA to form the CREATE mRNA (Fig. 1a). To ensure that the payload cannot be expressed by direct translation prior to genome integration, the payload is encoded in the antisense orientation (Fig. 1a). Critically, the payload is flanked by sequences designed to hybridize with single-guide RNA (sgRNA) target sites in the genome to prime reverse transcription. These sequences are referred to as primer binding site 1 (PBS1) and reverse complement (RC) primer binding site 2 (RC-PBS2).

Upon cellular uptake, the CREATE mRNA is translated into ORF1p and ORF2p^{ENdead} proteins, which then co-assemble with the mRNA to form CREATE RNP and enter cell nucleus (Fig. 1a, step 1). Natural L1 retrotransposition relies on nicking of the non-targeting DNA strand by the EN domain of ORF2p[23,27]. To harness this mechanism, a Cas9^{H840A} non-target strand nickase is employed to introduce a single strand nick guided by sgRNA1 (Fig 1a, step 2). The liberated DNA 3'-flap then hybridizes with the PBS1 sequence on the CREATE mRNA, serving as the primer for the RT domain of ORF2p for first strand cDNA synthesis (Fig. 1a, step 3). The newly synthesized cDNA strand includes both the payload and the PBS2 sequences in the sense orientation (Fig. 1a, step 4). Subsequently, PBS2 hybridizes with the 3' flap released from the

94 second nicked site generated by sgRNA2, invoking the L1 template jump mechanism to
95 complete the second strand synthesis and integration[23] (Fig. 1a, step 5-6). The outcome of the
96 editing cycle is the successful integration of the payload cassette between the PBS sites,
97 discarding the L1 components, resulting in safe non-replicative gene insertion.

98

99 **CREATE-mediated large payload gene delivery in mammalian cells**

100 To demonstrate that CREATE can be used to deliver functional genes, we designed a reporter
101 payload cassette consisting of an EF1α core promoter-driven GFP with an SV40 poly(A) signal
102 (totaling 1.1 kb). The AAVS1 site, a well-characterized genomic safe harbor site, was selected to
103 show programmable targeted insertion. We chose two sgRNAs that have been validated
104 previously, targeting sites within AAVS1 that are 90 bp apart[28]. Successful editing would replace
105 the 90 bp DNA segment with the 1.1 kb payload cassette (Fig. 1b). Baldwin et al. showed that
106 ORF2p can efficiently initiate TPRT with a 7-20 bp primer length[23]. Therefore, we designed
107 PBS1 and RC-PBS2 in this range to anneal with the DNA flaps released from two sgRNA cut
108 sites, and positioned them on either side of the payload.

109 Both Cas9 and ORF2p are large multidomain proteins, each exceeding 100 kDa. We
110 were concerned that direct Cas9-ORF2p fusions would compromise activity and the ability to
111 enter the nucleus. Since ORF2p could scan genomic DNA for entry sites, we hypothesized that it
112 could independently identify sites nicked by Cas9, eliminating the need for a direct fusion. To
113 test this concept, we developed a HEK293T cell line engineered to stably express a SpCas9 non-
114 targeting strand nickase (293T-nCas9^{H840A}). Fluorescence-activated cell sorting (FACS) was
115 used to quantify the percentage of cells expressing GFP as a measurement of successful payload
116 integration. Transfection of cells with CREATE mRNA and the two sgRNAs resulted in 0.5%

117 and 0.4% GFP expression on Day 3 and Day 8, respectively (Fig. 1b). No GFP expression was
118 observed in 293T-nCas9^{H840A} cells transfected with non-targeting control sgRNAs or in wildtype
119 HEK293T cells (Fig. 1b). These results confirmed sgRNA-guided integration and ruled out the
120 possibility of Cas9-independent, nonspecific insertion. Furthermore, we demonstrated that
121 payload expression required co-transfection of both sgRNAs, as delivery of either sgRNA alone
122 did not result in GFP positive cells (Fig. 1c). This result indicated that nicking on opposite DNA
123 strands was necessary for TPRT and successful integration. Such requirement inherently
124 enhances specificity, as a single independent sgRNA off-target binding would be insufficient to
125 cause payload integration.

126 FACS based cell sorting was employed to enrich the CREATE edited cells to
127 approximately 70% GFP positive, after which cells were maintained for an additional two weeks
128 (Fig. 1c). During this time GFP expression remained stable (>70%), indicative of successful and
129 stable genomic integration of the payload. To validate the site-specificity of the insertion and
130 search for potential off-target editing, we performed next generation sequencing with target
131 enrichment by using hybridization probes to capture EF1α-GFP payload sequences from
132 genomic DNA extracted from edited cells. Insertion sites analysis across the whole genome
133 showed highly specific editing at the intended AAVS1 locus at chromosome 19 (Fig. 1d).
134 Detailed examination of the sequencing reads showed no off-target edits at other genomic
135 regions (Fig. 1d and Supplementary Table 3). These results highlighted the target specificity of
136 CREATE system.

137 To further appreciate the temporal requirements for delivery of the CREATE
138 components, sequential transfection of RNA for the individual CREATE elements was
139 performed. The specific protocols included (1) Simultaneous transfection of CREATE mRNA
140 and sgRNAs (One-shot Protocol); (2) Transfecting CREATE mRNA four hours prior to sgRNAs

141 (Two-step, mRNA-first Protocol); (3) Transfecting sgRNAs four hours before CREATE mRNA
142 (Two-step, sgRNA-first Protocol). In Protocols (2) and (3), culture media was replaced
143 immediately prior to the second transfection. Observations supported superior performance of
144 the two-step protocols when compared to the one-shot approach (Supplementary Fig. 1a).
145 Notably, the mRNA-first strategy yielded the highest CREATE editing efficiency, reaching
146 ~1.2% GFP expression on day 3 and remaining stable on day 8. The improvement is likely due
147 to the necessary time for the ORF1p and ORF2p mRNA to be translated, co-assemble with
148 CREATE mRNA and subsequently enter the cell nucleus. Delivery of sgRNAs before the
149 CREATE RNP complex forms might result in the single strand nicks being repaired by the cell
150 before the CREATE editing process can occur. A time interval of 4 to 8 hr between mRNA and
151 sgRNA transfections yielded the highest editing rates of approximately 1.5%. Likewise, an
152 increase of sgRNA above 0.242 μg per each well improved the editing efficiency
153 (Supplementary Fig. 1b and 1c).

154

155 **CREATE editing is predicated on L1 retrotransposition mechanism**

156 Mutational analysis of the key components of CREATE was performed to investigate the
157 proposed editing mechanisms. We first examined whether CREATE editing required non-target
158 strand ($Cas9^{H840A}$) rather than target strand nickase ($Cas9^{D10A}$) activity. When we transfected
159 CREATE mRNA and the two AAVS1-targeting sgRNAs into HEK293T cells stably expressing
160 either $Cas9^{D10A}$ or $Cas9^{H840A}$ nickase, we observed GFP expression and integration exclusively in
161 the presence of $Cas9^{H840A}$ nickase (Fig. 2a), confirming the natural L1 replication mechanism
162 that starts with non-target strand cleavage. Of note, the activity of $Cas9^{H840A}$ also destroys the
163 PAM sequence after successful insertion, ensuring a single allele may only be edited once,

164 mitigating concerns about the potential for continuous nicking of the edited genome. Next, we
165 introduced a D702A mutation to inactivate the RT domain of L1 ORF2p (CREATE^{RTdead}). This
166 modification entirely abolished the integration and expression of the GFP payload (Fig. 2a)
167 confirming L1 RT activity was required for cDNA synthesis. Similarly, the removal of the entire
168 EN domain (CREATE mRNA$^{\Delta EN}$) resulted in a marked decrease in editing efficiency, indicating
169 that while the EN domain must be rendered inactive to avert cleavage of the genome at
170 redundant consensus sites recognized by L1 naturally, its complete removal was detrimental to
171 the editing process. This is consistent with the recent reports showing that the EN domain and
172 RT domain of L1 ORF2p are highly integrated structurally and functionally[23,27].

173

174 **sgRNA target sites and PBS sequences influence editing efficiency**

175 To determine if PBS length impacts insertion efficiency, we designed a series of CREATE
176 mRNA constructs with different PBS1 and RC-PBS2 hybridization sequence lengths (17 bp, 30
177 bp, and 50 bp) and measured insertion efficiency at the AAVS1 sites targeted with the same
178 sgRNAs. In these experiments 17-30 bp appeared to support the most efficient editing, while 50
179 bp was associated with a reduction of GFP integration efficiency (Fig. 2b). These results are with
180 a recent study investigating the poly-A length required for efficient TPRT by L1 ORF2[27].

181 As a search-and-replace tool, CREATE swaps out the genomic sequence between
182 sgRNA1 and sgRNA2 for the payload (Fig. 1b). Next, we investigated the impact of the distance
183 between the two sgRNA nicking sites on editing efficiency. Three sgRNA1s with binding sites
184 located 90 bp, 481 bp or 976 bp from a fixed site targeted by sgRNA2 were designed. While
185 CREATE achieved successful integration of a GFP cassette and replacement of the segment

186 between the sgRNAs in all cases, the editing efficiency dropped slightly with increasing distance
187 between the sgRNAs (Fig. 2b).

188

189 **A requirement for paired sgRNAs and matched PBS sequences impart high fidelity and**
190 **precision on CREATE**

191 To demonstrate the importance of two PBS sites flanking the payload, we designed CREATE
192 mRNA with either PBS1 or a RC-PBS2 sequence. In the presence of AAVS1 targeting sgRNAs,
193 the removal of PBS1 or RC-PBS2 resulted in a lack GFP expression on day 3 or day 8 (Fig. 2c).
194 Additionally, exchanging the position of PBS1 and RC-PBS2 while simultaneously encoding
195 them in a reverse-complementary sequence ("RC-PBS1" and "PBS2) restored the GFP
196 expression. The requirement for the two PBS sequences that matched the two sgRNA cut sites
197 were critical for a highly specific integration. When we assessed the top 5 predicted off-target
198 sites for sgRNA1 and sgRNA2 by PCR, no integration of GFP cassette into theses potential sites
199 was observed (Supplementary Fig. 2). Together, these data validate the proposed mechanism of
200 CREATE and show that the editing process has several inherent checkpoints that prevent off-
201 target activity. These include: 1) correct base-pairing of PBS1 with sgRNA1 cut sites; 2) correct
202 base-pairing of reverse transcribed RC-PBS2 with sgRNA2 cut site and 3) adjacent sgRNA1 and
203 sgRNA2 nicking.

204

205 **CREATE-mediated gene integration into different genomic loci**

206 To determine the ability of the CREATE system to edit and deliver genes specifically into sites
207 in the genome beyond AAVS1, we engineered CREATE to deliver GFP to other genomic loci
208 including HEK3, PRNP, and IDS. The HEK3 locus, situated on chromosome 1 is frequently

209 chosen to benchmark gene editing tools[29–31]. The PRNP locus on chromosome 20 encodes the
210 prion protein implicated in multiple neurodegenerative diseases[32,33], and IDS gene on
211 chromosome X is associated with Hunter's syndrome[34], a lysosomal storage disease. For each
212 targeted site, a pair of sgRNAs 70-90 bp apart were selected. CREATE mRNA, which carried
213 the corresponding 30 bp PBS1 and RC-PBS2 flanking the GFP expression cassette, was
214 deployed to each target site. Flow cytometric analysis on day 3 demonstrated all three sites
215 exhibited comparable editing efficiency, as determined by GFP expression (Fig. 3a).

216 To characterize the editing accuracy at the junction sites between payload sequences and
217 genomic sequence, we performed amplicon sequencing by designing PCR primers to cover the
218 expected junctions (Fig. 3b). Interestingly, for AAVS1 edited cells we observed varying levels of
219 indel formation at the PBS1 junction, while PBS2 junction was >95% without indels (Fig. 3c).
220 Changing of the length of PBS from 17 bp to 30 bp reduced indel formation at PBS1 (Fig. 3c).
221 We also performed the same analysis on the junction regions for cells with targeted insertion at
222 HEK3, PRNP and IDS sites. The results showed that for HEK3 and IDS, the insertion were more
223 than 95% accurate; while the PRNP site showed indel formation only at the PBS1 junction but
224 not PBS2 (Fig. 3d). Taken together, these results showed that the formation of indels occurs only
225 at the junction site and are likely dependent on the sgRNA sequence for the corresponding target
226 locus. Optimization of target sgRNAs for each particular target site is necessary to reduce
227 junction infidelity and improve editing efficiency.

228

229 **CREATE as a fully RNA-based gene delivery tool for different mammalian cell types**

230 Finally, an important attribute of CREATE gene editing is the ability to deliver all components
231 as RNA into mammalian cells (Fig. 4a). To demonstrate this, we synthesized mRNA encoding

232 nCas9^{H840A} with N- and C-terminal NLS. Electroporation was used to deliver the nCas9 mRNA,
233 CREATE mRNA and two sgRNAs designed to target AAVS1 locus into HEK293T cell and the
234 immortalized liver cell line Huh7. We found that one-time electroporation led to ~1% and 0.2%
235 GFP payload re-expression in Huh7 and HEK293T cells, respectively (Fig. 4b). Taken together,
236 these results highlight the potential to develop CREATE into a fully RNA based gene delivery
237 tool for in vitro and in vivo applications.

238

239 **Discussion**

240 Here we have engineered CREATE, a programmable, site-specific tool for integration of
241 full-length genes into the human genome, based on the human L1 retrotransposon. Key design
242 elements for the site specificity of CREATE include the modification of the natural L1 ORF2p
243 endonuclease and the introduction of a Cas9 nickase. Guided by two sgRNAs, Cas9 nickase
244 targets the desired integration sites without DSBs, thereby mitigating the risk of chromosomal
245 instabilities. Two primer binding sites flanking the payload are critical for the initiation of
246 reverse transcription and converting the payload RNA into cDNA for integration into the
247 genome. This is the first evidence of using L1 or any other retrotransposon system to deliver
248 genes with specificity and programmability. An earlier study explored the direct fusion of a Cas9
249 or Cas12a endonuclease with the reverse transcriptase domain of L1 ORF2p for site-specific
250 integration[35]. While Cas-directed cleavage and subsequent RNA payload integration was
251 achieved, the study was limited to payload insertion into high-copy plasmids in *E. coli*. The use
252 of an active Cas9 endonuclease exhibited significant toxicity. In contrast, CREATE employs a
253 Cas9 nickase in combination with the full-length, endonuclease inactivated ORF2p for reverse
254 transcription. Recent high-resolution cryo-EM structures of ORF2p complexed with DNA and
255 RNA substrates shed light on the process of target site recognition and initiation of reverse

256 transcription[22,23]. Multiple domains in ORF2p fold into an integrated structure that wrap around
257 the substrates to effectively couple genomic DNA binding, target site cleavage and target-primed
258 reverse transcription. Our findings support the notion that preserving all essential domains of
259 ORF2p can facilitate productive integration in mammalian genomes.

260 Non-LTR retroelements have gained increasing interest for the delivery of larger
261 payloads. R2 retroelements, found in metazoans and birds, have been used to insert large RNA
262 payloads (up to 4kb) into eukaryotic genomes. However, R2 elements are naturally restricted to
263 the ribosomal DNA (rDNA) loci, which limits their utility for editing in other genomic locations.
264 In addition, insertion at rDNA site may result in reduced expression over time due to rDNA array
265 instability[36]. An attempt to integrate R2 elements with Cas9 for RNA-guided reverse
266 transcription and integration was unsuccessful to achieve complete cDNA synthesis and
267 integration[25]. Exploiting the natural replication mechanisms of an engineered L1 element,
268 CREATE achieved an integration efficiency of approximately 1.5% in multiple mammalian cell
269 lines, inserting a 1.1 kb GFP cassette without detectable off-target integration in multiple
270 genomic sites. Given L1's natural capacity to mediate insertion of approximately 6 kb of RNA,
271 we anticipate that future iterations of CREATE could deliver even larger sequences.

272 The L1 element has been associated with increased risk of cancer and cellular
273 senescence[37,38]. A gradual loss of epigenetic suppression of L1 transcription, can lead to
274 reactivation of L1 retrotransposition which can cause insertional mutagenesis or chronic
275 inflammation triggered a type I interferon response[39–41]. In the present study, no off-target
276 integrations were observed suggesting that CREATE does not cause genomic instability. Several
277 properties and modification ensure that the CREATE system integrates the payload gene without
278 propagating natural L1 components or perturbing their repressed state. CREATE mRNA is only
279 transiently expressed and once on-target integration has occurred, the sgRNA binding sites are

280 altered which prevents any further re-editing event. Finally, the EN domain of ORF2p in the
281 CREATE system is inactivated to prevent non-specific cleavage of genomic DNA.

282 Varying degrees of indel formation at the transgene/genomic locus junction were noted.
283 This phenomenon has also been described for PE-related technologies that utilize double-
284 nicking. 10-20% indel by-products were observed for the PE3 system[8]. GRAND editing, a PE-
285 derived technology, yielded 5.8% to 63.0% accurate editing with substantial indel formation[42].
286 Systematic optimization of pegRNA design and manipulation of cellular DNA repair pathways
287 can reduce indel by-products[43,44]. We anticipate that a more in-depth understanding of the
288 CREATE integration mechanism along with sgRNA and target site optimization will enhance
289 junction fidelity and overall editing efficiency. Furthermore, the combination of the modified L1
290 element with alternative programmable RNA-guided DNA endonucleases such as Cas12, TnpB
291 and Fanzor can be explored to expand the capability of the CREATE system[45–47]

292 In summary, the development of CREATE combines the gene integration capacity of a
293 human L1 retroelement with the CRISPR/Cas9 precision. As a fully RNA-based genome
294 engineering tool that does not require DNA templates, it complements existing approaches and
295 has the potential to address a large variety of genetic diseases.

296

297 **Methods**

298 **Cell culture**

299 HEK293T (ATCC, cat. no. CRL-3216) and Huh7 cells (JBRC, Cat. No. JCRB0403) were
300 maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% (v/v)
301 fetal bovine serum (FBS, Gibco) and 1% (v/v) Penicillin/Streptomycin/Glutamine (Gibco). Cells
302 were cultured at 37°C with 5% CO_2. TrypLE Select (Thermo Fisher Scientific) was used for
303 passaging and harvesting cells.

304

305 **sgRNAs**

306 sgRNAs were acquired from Synthego or Thermo Fisher Scientific and reconstituted with TE
307 buffer into 100 pmol/µL stock solution. Sequences of sgRNAs are listed in Supplementary
308 Table 1**.**

309

310 **Generation of Cas9-expressing HEK293T cell lines**

311 PiggyBac vectors expressing Cas9 nickase H840A or D10A variants with an N-terminal SV40
312 NLS and a C-terminal Nucleoplasmin NLS under the CMV promoter were constructed by
313 Vector Builder. PiggyBac vectors (1.25 µg) and hyPBase mRNA (0.7 µg) were co-transfected
314 into 2.5×10^6 HEK293T cells using MaxCyte ATx electroporator by following the manufacturer's
315 protocol. Three days after transfection, cells were selected with hygromycin (400 µg/mL) for 10
316 days, to acquire stable Cas9-expressing HEK293T cell lines.

317

318 **Cloning of CREATE constructs**

319 DNA sequences encoding human L1 ORF1p (GenBank: AAB59367.1) and ORF2p (GenBank:
320 AAA51622.1) with N-terminal SV40 NLS and C-terminal nucleoplasmin NLS were codon

321 optimized and synthesized at Twist Bio. Endogenous L1 interORF region sequences flanked by
322 restriction sites PshAI (GACAGCCGTC) and AccI (GTATAC) was synthesized at Twist Bio.
323 The full constructs containing all L1 components including ORF1p, interORF and ORF2p were
324 then assembled using Hifi Assembly mix (NEB) into pCMV vector with the CMV promoter
325 region removed. A T7 promoter was introduced upstream of ORF1p for in vitro transcription.
326 The payload cassette including an EF1α core promoter, EGFP and SV40 poly(A) signal was
327 synthesized by Integrated DNA Technologies and inserted at the 3'UTR of the L1 mRNA using
328 Hifi Assembly mix (NEB). gBlocks gene fragments containing different PBS site sequences
329 were synthesized by Integrated DNA Technologies, and subsequently cloned into the vector to
330 flank the payload gene. Sequences of PBS sites and payloads are listed in Supplementary
331 Table 2. Mutation of ORF2 RT domain was introduced using Q5 site-directed mutagenesis kit
332 (NEB). Deletion of ORF2 EN domain was performed by PCR and ligation cloning.
333

334 **mRNA synthesis via in vitro transcription**

335 Plasmid templates for *in vitro* transcription were linearized by restriction enzyme and templates
336 were purified by column clean up (GeneJET PCR Purification Kit, Thermofisher) and eluted in
337 40µl of nuclease free water. All in house generated mRNA was transcribed via T7 polymerase
338 and co-transcriptionally capped utilizing New England Biolabs HiScribe T7 mRNA Kit with
339 CleanCap Reagent AG according to the manufacturer's protocols. The in vitro transcription
340 reaction was allowed to proceed for 2 hours at 37 ºC. Reaction mixtures were treated with DNase
341 I and incubated at 37 ºC for 30 minutes before immediate purification (Monarch® RNA Cleanup
342 Kit, NEB). A sample of each mRNA was reserved for quality analysis and transcripts were tailed
343 with *E. coli* poly(A) polymerase (NEB) at 37 ºC for 30 minutes, and subsequently purified
344 (Monarch® RNA Cleanup Kit, NEB). All tailed and untailed mRNA samples were normalized

345 to a concentration of 50ng/µl and quality was assessed via TapeStation RNA ScreenTape
346 Analysis (Agilent, RNA ScreenTape, 5067-5576) according to the manufacturer's protocol.
347
348 **Lipofectamine transfection**
349 HEK293T cells were seeded on 24-well plates at 80,000 cells the day prior to transfection in
350 DMEM with 2% FBS. Transfection was performed by either one of these protocol:
351 - One-shot Protocol (#1): L1-mRNA (3 ug per well) and sgRNAs (0.121 ug each
352 sgRNA per well) were transfected using Lipofectamine MessengerMAX (Invitrogen).
353 - Two-step, mRNA first Protocol (#2): L1-mRNA (3 ug per well) was transfected using
354 Lipofectamine MessengerMAX. Transfected cells were incubated at 37°C with 5%
355 CO_2 for 4 hour. After refreshing the media, cells were then transfected with sgRNAs
356 (0.121 ug each sgRNA per well) using Lipofectamine RNAiMax (Invitrogen).
357 - Two-step, sgRNA first Protocol (#3): sgRNAs were transfected using Lipofectamine
358 RNAiMax. Transfected cells were incubated at 37°C with 5% CO_2 for 4 hour. After
359 refreshing the media, cells were then transfected with L1-mRNA (3 ug per well)
360 using Lipofectamine MessengerMAX.
361
362 **Electroporation**
363 Huh7 cells were harvested by TrypLE Select and resuspended in MaxCyte buffer at a density of
364 6.25 x 10^7 cells/mL. For MaxCyte electroporation, 15 µg of L1-mRNA, 7.5 µg of nCas9 mRNA,
365 and 2.5 µg/each dual sgRNAs were gently mixed with 1.25 x 10^6 Huh7 cells in a total volume of
366 27.5 µL MaxCyte buffer. 25 µL of cell-RNA mixture was transferred into an OC-25x3 cuvette.
367 Electroporation was carried out on the MaxCyte ATX at the indicated electroporation energy.
368 Immediately after electroporation, the OC-25x3 cuvettes with cells were incubated at 37ºC for 15

369 min to allow for cell membrane recovery. Finally, cells were added to a 6-well plate feeded with
370 DMEM with 5% FBS for the first 3 days after electroporation. The medium was removed and
371 switched back to DMEM with 10% FBS after day 3. Cells were harvested on day 3 and day 8
372 post-transfection for FACS analysis.
373

374 **Flow cytometry analysis and cell sorting**

375 Flow cytometry analysis was performed on day 3 and day 8 after transfection. Transfected cells
376 were collected after PBS washing and TrypLE Select digestion and resuspended in PBS with 1%
377 FBS and 2mM EDTA for flow cytometry analysis (CYTEK). Data were analyzed by FlowJo
378 10.9.0 software.
379 For experiments that require enriched GFP-positive cells, cells were collected three to
380 five days after transfection, and sorted by flow cytometry (BD FACSMelody sorter) to obtain the
381 GFP-positive cells. Sorted cells were cultured in DMEM with 20% FBS for the initial 7 days to
382 facilitate cell recovery.
383

384 **Genomic DNA extraction**

385 One million collected cells were centrifuged at *5,000 g* for 5 min at 4°C and the pellet was
386 resuspended in 100-µL cold PBS. Genomic DNA was extracted using Monarch Genomic DNA
387 Purification Kit (NEB) according to manufacturer's instructions and eluted in 20 µL water.
388

389 **Target hybridization sequencing**

390 Genomic DNA (gDNA) extracted was used to prepare Insertion Site Analysis (ISA) libraries
391 using the Twist Target Enrichment procedure. Briefly, gDNA was enzymatically fragmented,
392 followed by end repair, dA-tailing and adapter ligation, and library amplification using indexed

393 primers. The resulting pre-hybridization libraries underwent QC followed by hybridization-
394 capture with baits corresponding to the transgene/insert, and a final library amplification using
395 universal primers. The final library was sequenced by Illumina platform using 2x150bp
396 sequencing configuration in paired end mode.
397
398 For data analysis, raw BCL files generated by the sequencer were converted to raw fastq files for
399 each sample using bcl2fastq v.2.20. Fastq reads were then trimmed with adapter sequences.
400 Homology analysis was performed for vector genome against human reference genome (build
401 hg38) and high homology regions identified in reference human genome were hard masked. A
402 combined reference genome consisting of vector genome and hard masked human reference was
403 generated and reads were aligned to the combined genome using bwa v0.7.4. The bam file was
404 subsequently processed with Samtools v0.1.16. Finally, integration sites were identified by
405 taking into account pairs of reads that support insertions and alignment to both vector and human
406 reference genomes. Identified potential insertion sites were clustered and annotated by Azenta
407 proprietary ISA pipeline. Potential insertion sites with >5 confirmed reads per location were
408 analyzed and plotted. Detailed analysis of identified insertions sites with >5 confirmed reads are
409 shown in Supplementary Table 3.
410
411 **Amplicon sequencing**
412 To validate the site-specific integration of CREATE payload, target regions were PCR amplified
413 and analyzed by sequencing methods. Genomic DNA samples were PCR amplified with Q5
414 High-Fidelity 2X PCR Master Mix (NEB) based on the manufacturer's protocol. PCR amplicon
415 primers were designed specifically to integration locus-CREATE junctions and are listed in

416 Supplementary Table 4. Amplicons were analyzed by 1% agarose gel electrophoresis and
417 purified according to the sizes using the DNA Gel Extraction Kit (NEB).
418 To quantify integration of CREATE payload by targeted deep sequencing, amplicons
419 from PCR amplification of genomic-CREATE insertion junctions were prepared for sequencing
420 on a MiSeq (Illumina). Alignment of amplicon reads to a reference sequence was performed
421 using CRISPResso2. Reads with a mean quality score <30 were discarded.
422

423 **Determination of CREATE editing activity at potential off-target sites**

424 To evaluate the off-target activity of CREATE editing, the potential off-target sites of a single
425 sgRNA were detected by targeted PCR amplification. Potentially top-ranking predicted off-target
426 sites of spacers were selected using the CRISPR-Cas9 guide RNA design checker (Integrated
427 DNA Technologies). PCR primers and predicted off-target insertion are listed in Supplementary
428 Table 5. The PCR reactions were performed using Q5 High-Fidelity 2X PCR Master Mix (NEB)
429 with optimized Tm and the PCR products were detected by 1% agarose gel.
430

431 **Statistics and reproducibility**

432 At least three replicates were analyzed in each independent experiment to ensure the
433 experimental data were reliable. Data are expressed as mean ± s.d. Statistical analyses were
434 performed with GraphPad Prism (version 9.0, GraphPad Software).

435 **Data availability**

436 Sequences including primers, sgRNAs are available in the supplementary materials.
437

438 **Reference**

439

440 1. Frangoul, H. *et al.* Exagamglogene Autotemcel for Severe Sickle Cell Disease. *Blood* **142**,
441 1052 (2023).

442 2. Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the
443 beginning. *Science* **379**, eadd8643 (2023).

444 3. Leibowitz, M. L. *et al.* Chromothripsis as an on-target consequence of CRISPR–Cas9 genome
445 editing. *Nat. Genet.* **53**, 895–905 (2021).

446 4. Brunet, E. & Jasin, M. Chromosome Translocation. *Adv. Exp. Med. Biol.* **1044**, 15–25 (2018).

447 5. Song, Y. *et al.* Large-Fragment Deletions Induced by Cas9 Cleavage while Not in the BEs
448 System. *Mol. Ther. - Nucleic Acids* **21**, 523–526 (2020).

449 6. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–
450 Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).

451 7. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without
452 DNA cleavage. *Nature* **551**, 464–471 (2017).

453 8. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or
454 donor DNA. *Nature* **576**, 149–157 (2019).

455 9. Chen, P. J. & Liu, D. R. Prime editing for precise and highly versatile genome manipulation.
456 *Nat. Rev. Genet.* **24**, 161–177 (2023).

457 10. Yarnall, M. T. N. *et al.* Drag-and-drop genome insertion of large sequences without double-
458 strand DNA cleavage using CRISPR-directed integrases. *Nat. Biotechnol.* **41**, 500–512 (2023).

459 11. Anzalone, A. V. *et al.* Programmable deletion, replacement, integration and inversion of
460 large DNA sequences with twin prime editing. *Nat. Biotechnol.* **40**, 731–740 (2022).

461 12. Tang, S. & Sternberg, S. H. Genome editing with retroelements. *Science* **382**, 370–371
462 (2023).

463 13. Consortium, I. H. G. S. *et al.* Initial sequencing and analysis of the human genome. *Nature*
464 **409**, 860–921 (2001).

465 14. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 Elements in Structural
466 Variation and Disease. *Annu. Rev. Genom. Hum. Genet.* **12**, 187–215 (2011).

467 15. Han, J. S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent
468 developments, and unanswered questions. *Mob. DNA* **1**, 15 (2010).

469 16. Khazina, E. *et al.* Trimeric structure and flexibility of the L1ORF1 protein in human L1
470 retrotransposition. *Nat. Struct. Mol. Biol.* **18**, 1006–1014 (2011).

471 17. Mathias, S. L., Scott, A. F., Jr., H. H. K., Boeke, J. D. & Gabriel, A. Reverse Transcriptase
472 Encoded by a Human Transposable Element. *Science* **254**, 1808–1810 (1991).

473 18. Feng, Q., Moran, J. V., Kazazian, H. H. & Boeke, J. D. Human L1 Retrotransposon Encodes
474 a Conserved Endonuclease Required for Retrotransposition. *Cell* **87**, 905–916 (1996).

475 19. Doucet, A. J. *et al.* Characterization of LINE-1 Ribonucleoprotein Particles. *PLoS Genet.* **6**,
476 e1001150 (2010).

477 20. Kojima, K. K. Different integration site structures between L1 protein-mediated
478 retrotransposition in cis and retrotransposition in trans. *Mob. DNA* **1**, 17 (2010).

479 21. Kulpa, D. A. & Moran, J. V. Cis-preferential LINE-1 reverse transcriptase activity in
480 ribonucleoprotein particles. *Nat. Struct. Mol. Biol.* **13**, 655–660 (2006).

481 22. Thawani, A., Ariza, A. J. F., Nogales, E. & Collins, K. Template and target site recognition
482 by human LINE-1 in retrotransposition. *Nature* 1–3 (2023) doi:10.1038/s41586-023-06933-5.

483 23. Baldwin, E. T. *et al.* Structures, functions and adaptations of the human LINE-1 ORF2
484 protein. *Nature* 1–13 (2023) doi:10.1038/s41586-023-06947-z.

485 24. Piskareva, O. & Schmatchenko, V. DNA polymerization by the reverse transcriptase of the
486 human L1 retrotransposon on its own template in vitro. *FEBS Lett.* **580**, 661–668 (2006).

487 25. Wilkinson, M. E., Frangieh, C. J., Macrae, R. K. & Zhang, F. Structure of the R2 non-LTR
488 retrotransposon initiating target-primed reverse transcription. *Science* **380**, 301–308 (2023).

489 26. Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V. Unconventional
490 translation of mammalian LINE-1 retrotransposons. *Genes Dev.* **20**, 210–224 (2006).

491 27. Thawani, A., Ariza, A. J. F., Nogales, E. & Collins, K. Template and target-site recognition
492 by human LINE-1 in retrotransposition. *Nature* 1–8 (2023) doi:10.1038/s41586-023-06933-5.

493 28. Zheng, C. *et al.* Template-jumping prime editing enables large insertion and exon rewriting
494 in vivo. *Nat. Commun.* **14**, 3369 (2023).

495 29. Zheng, C. *et al.* A flexible split prime editor using truncated reverse transcriptase improves
496 dual-AAV delivery in mouse liver. *Mol. Ther.* **30**, 1343–1351 (2022).

497 30. Kong, X. *et al.* Precise genome editing without exogenous donor DNA via retron editing
498 system in human cells. *Protein Cell* **12**, 899–902 (2021).

499 31. Kweon, J. *et al.* Targeted genomic translocations and inversions generated using a paired
500 prime editing strategy. *Mol. Ther.* **31**, 249–259 (2023).

501 32. Zhang, W. *et al.* Mutational analysis of PRNP in Alzheimer's disease and frontotemporal
502 dementia in China. *Sci. Rep.* **6**, 38435 (2016).

503 33. Palmer, M. S., Dryden, A. J., Hughes, J. T. & Collinge, J. Homozygous prion protein
504 genotype predisposes to sporadic Creutzfeldt–Jakob disease. *Nature* **352**, 340–342 (1991).

505 34. Wilson, P. J. *et al.* Hunter syndrome: isolation of an iduronate-2-sulfatase cDNA clone and
506 analysis of patient DNA. *Proc. Natl. Acad. Sci.* **87**, 8531–8535 (1990).

507 35. Manoj, F., Tai, L. W., Wang, K. S. M. & Kuhlman, T. E. Targeted insertion of large genetic
508 payloads using cas directed LINE-1 reverse transcriptase. *Sci. Rep.* **11**, 23625 (2021).

509 36. Zhang, X. *et al.* Harnessing eukaryotic retroelement proteins for transgene insertion into
510 human safe-harbor loci. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-024-02137-y.

511 37. Xiao-Jie, L., Hui-Ying, X., Qi, X., Jiang, X. & Shi-Jie, M. LINE-1 in cancer: multifaceted
512 functions and potential clinical implications. *Genet. Med.* **18**, 431–439 (2016).

513 38. Gorbunova, V. *et al.* The role of retrotransposable elements in ageing and age-associated
514 diseases. *Nature* **596**, 43–53 (2021).

515 39. Cecco, M. D. *et al.* L1 drives IFN in senescent cells and promotes age-associated
516 inflammation. *Nature* **566**, 73–78 (2019).

517 40. Mendez-Dorantes, C. & Burns, K. H. LINE-1 retrotransposition and its deregulation in
518 cancers: implications for therapeutic opportunities. *Genes Dev.* **37**, 948–967 (2023).

519 41. Grundy, E. E., Diab, N. & Chiappinelli, K. B. Transposable element regulation and
520 expression in cancer. *FEBS J.* **289**, 1160–1179 (2022).

521 42. Wang, J. *et al.* Efficient targeted insertion of large DNA fragments without DNA donors.
522 *Nat. Methods* **19**, 331–340 (2022).

523 43. Nelson, J. W. *et al.* Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol.*
524 **40**, 402–410 (2022).

525 44. Chen, P. J. *et al.* Enhanced prime editing systems by manipulating cellular determinants of
526 editing outcomes. *Cell* **184**, 5635-5652.e29 (2021).

527 45. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas
528 System. *Cell* **163**, 759–771 (2015).

529 46. Saito, M. *et al.* Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**,
530 660–668 (2023).

531 47. Karvelis, T. *et al.* Transposon-associated TnpB is a programmable RNA-guided DNA
532 endonuclease. *Nature* **599**, 692–696 (2021).

533 48. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence
534 analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).

535

536

537 **Acknowledgments:**

538 We thank R. Vale (HHMI Janelia) & S. Mukherjee (Columbia U) for critical reading of the

539 manuscript and the entire Myeloid Therapeutics team for support.

540 **Author contributions:**

541 Conceptualization: YW, DG, RL, BM

542 Methodology: YW, DG, RL, MH

543 Experiments and Data Analysis: RL, MH, BL, ND, YW

544 Writing: YW, RL, DG, RH

545 **Competing interests:** Y.W., R.L., M.H., B.L., R.H. and D.G. are current employees of Myeloid

546 Therapeutics. B.M. was a past employee of Myeloid Therapeutics. All authors hold

547 equity interest in Myeloid Therapeutics.

548

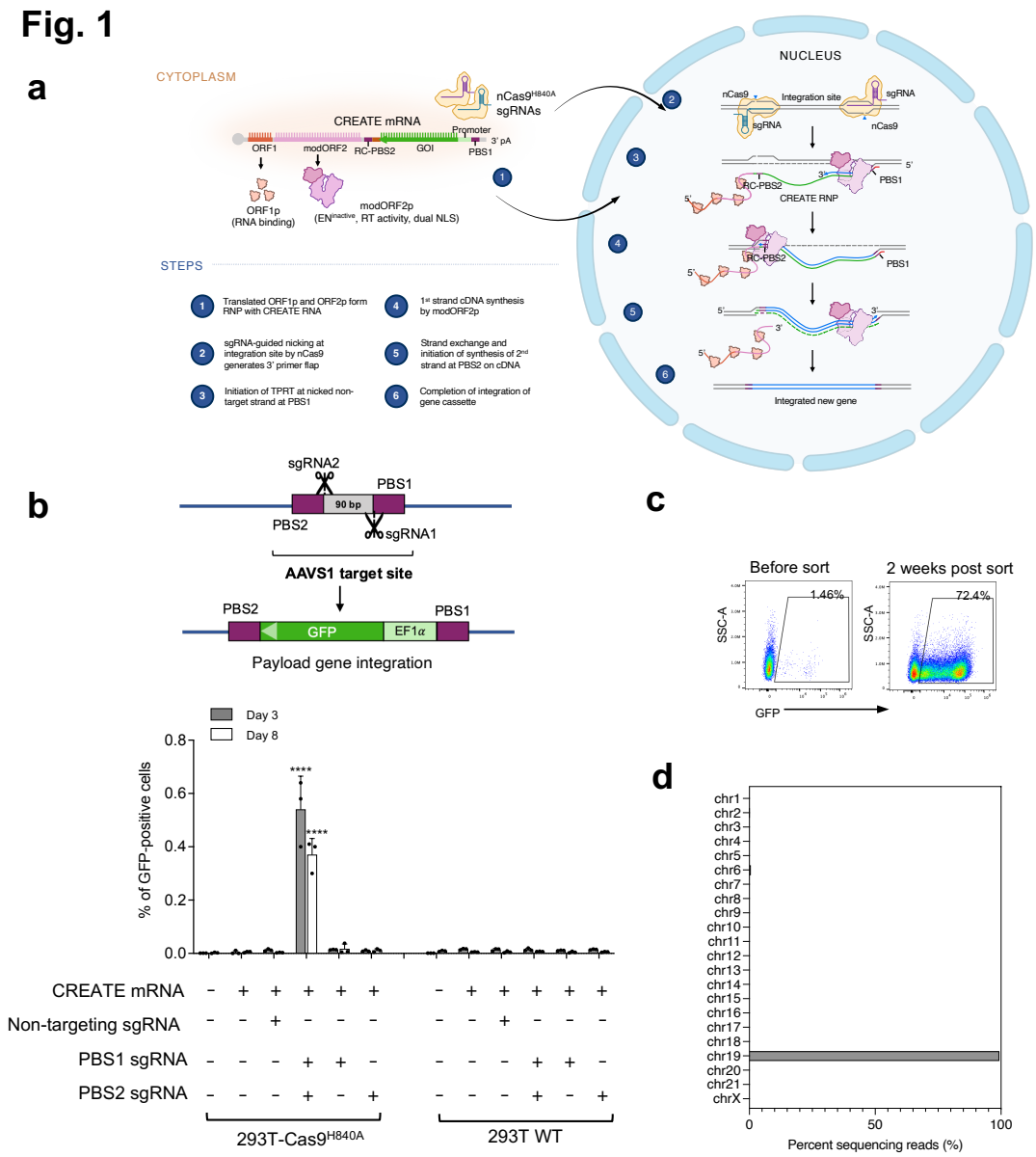
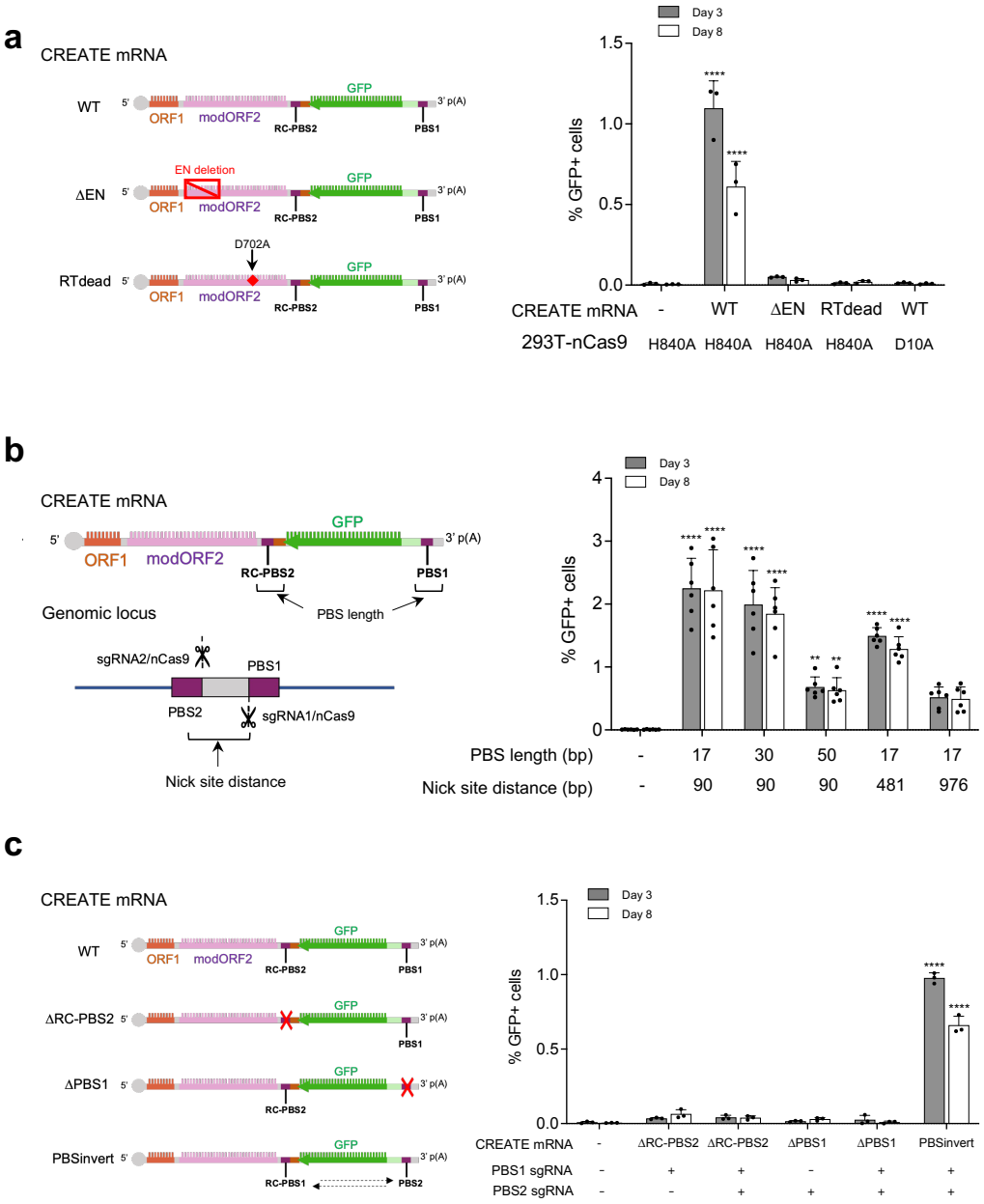


Fig. 1. CREATE editing design and proof-of-concept in mammalian cells. (a) Design of the CREATE system and the mechanism for gene integration. (**b**) Proof-of-concept study demonstrating replacement of a 90 bp segment at AAVS1 locus with a GFP expression cassette using CREATE. Quantification of payload integration efficiencies shows successful editing requires Cas9 nickase activity and dual sgRNAs. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample. **** indicates $p < 0.0001$. Statistical significance is labeled for samples with $p < 0.01$. Data are mean ± SD. (**c**) AAVS1 edited cells before and after enrichment for GFP positive cells. (**d**) Genome-wide target hybridization sequencing confirms specific integration of GFP payload gene at AAVS1 locus on chromosomal 19. Identified potential sites with >5 sequencing reads were plotted across all chromosomes (see Supplementary Table 3 for details).

562

Fig. 2



563
564 **Fig. 2. Mechanistic analysis of CREATE editing**. (**a**) Mutational analysis of the key enzymatic
565 activities and protein domains involved in editing. (**b**) Influence of the length of PBS sequences
566 in CREATE mRNA and the distance between sgRNA nick sites in the genome on editing
567 efficiency. (**c**) CREATE editing requires two PBS sequences that matches the two sgRNA target
568 sites. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple
569 comparisons test comparing against the first sample. **** $p < 0.0001$.** $p<0.01$. Statistical

570 significance is labeled for samples with $p < 0.01$. Data shown are representative of at least two
571 independent experiments. All data are mean ± SD.
572

Fig. 3

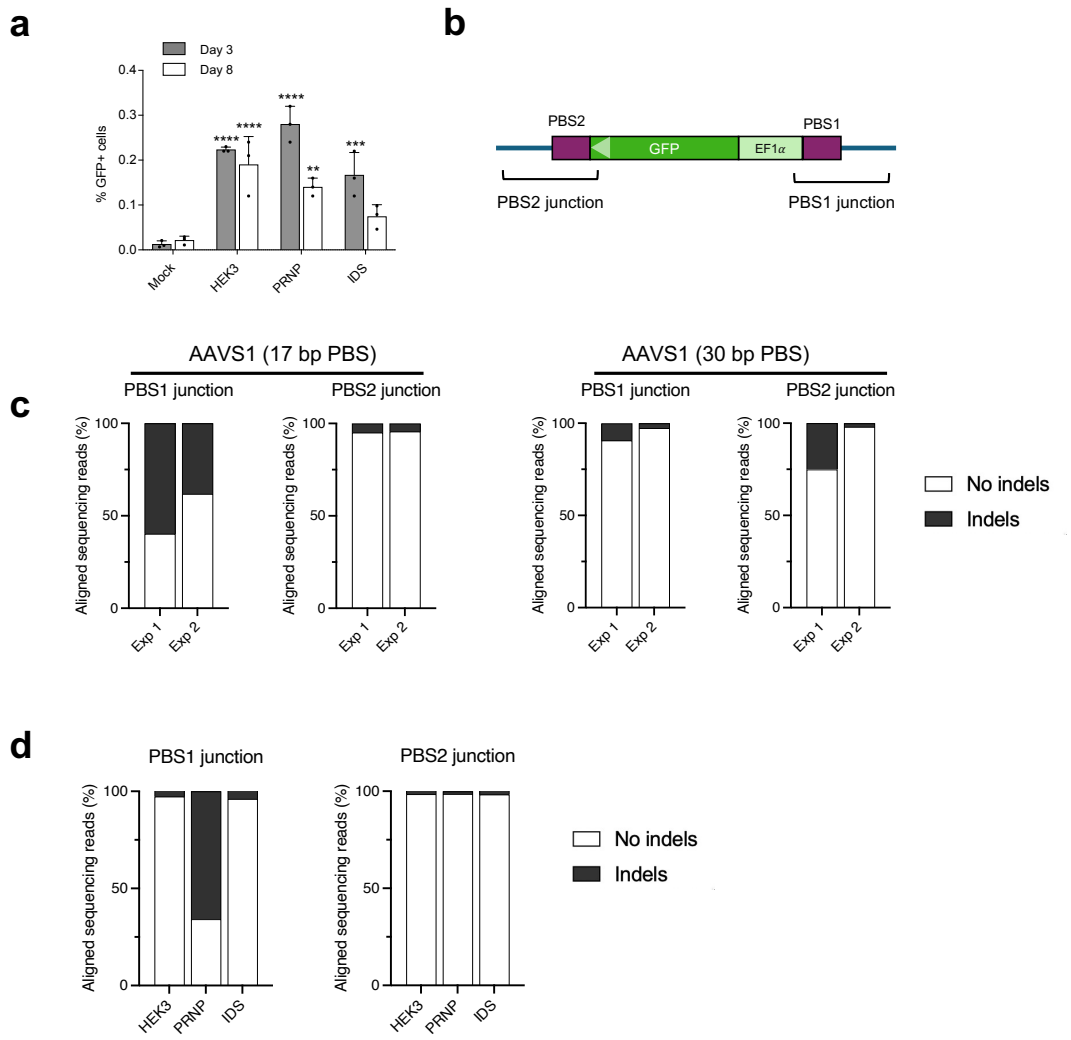
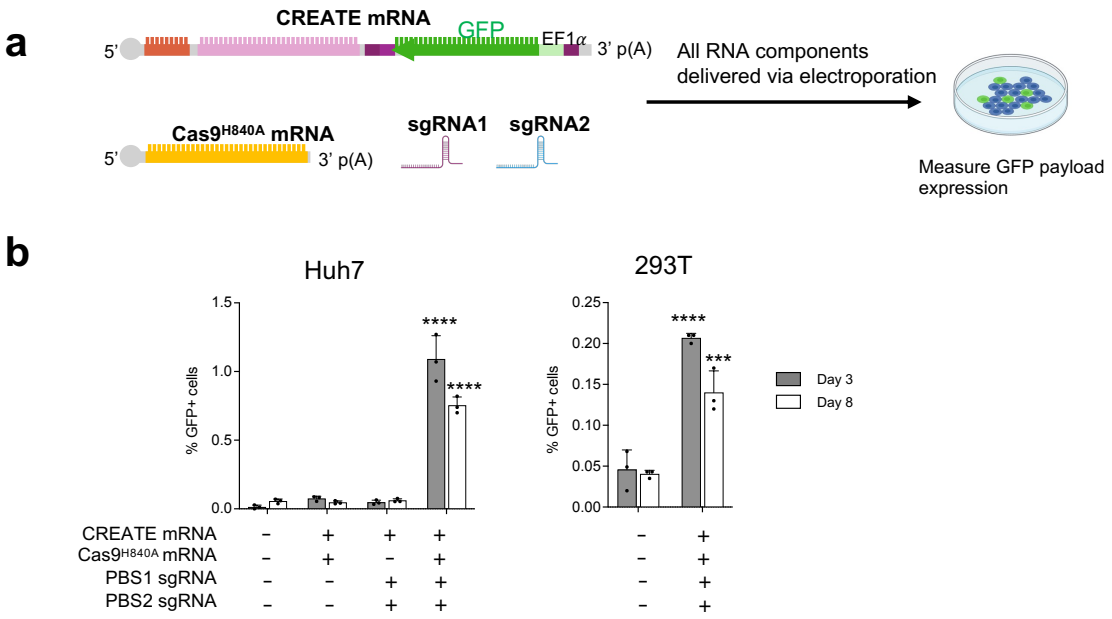


Fig. 3. CREATE as a fully programmable gene delivery tool in mammalian cells. (**a**) Editing of three additional genomic loci by designing unique sgRNAs and matching PBSs. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample. **** $p < 0.0001$. *** $p<0.001$.** $p<0.01$. Statistical significance is labeled for samples with $p <0.01$. Data shown are representative of at least two independent experiments. All data are plotted as mean ± SD. (**b**) Amplicon sequencing to characterize the junction regions of integrated payload genes. (**c**) The percentage of indels at the integration junction at AAVS1 locus. Two different CREATE mRNAs were designed with 17 bp or 30 bp PBSs, both targeting AAVS1 locus. Exp 1 and Exp 2 represent two independent experiments. (**d**) The percentage of indels at integration junctions at HEK3, PRNP and IDS loci. Edited cells were from experiment shown in Fig. 3a. Data from (**c**) and (**d**) were analyzed and plotted using CRISPResso2[48].

587

Fig. 4



588

589 **Fig. 4. CREATE as a fully RNA-based gene delivery system to multiple cell types** (**a**) Co-
590 delivery of CREATE editing components as mRNA/sgRNA. (**b**) RNA-based GFP payload
591 insertion at AAVS1 locus in Huh7 and HEK293T cells. Statistical analysis was performed using
592 two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample.
593 **** $p < 0.0001$. *** $p<0.001$. Statistical significance is labeled for samples with $p <0.01$. Data
594 shown are representative of at least two independent experiments. All data are plotted as mean ±
595 SD.