Nucleic Acids Research, 2025, **53**, gkae1199
https://doi.org/10.1093/nar/gkae1199
Advance access publication date: 10 December 2024
Synthetic Biology and Bioengineering

High throughput variant libraries and machine learning yield design rules for retron gene editors

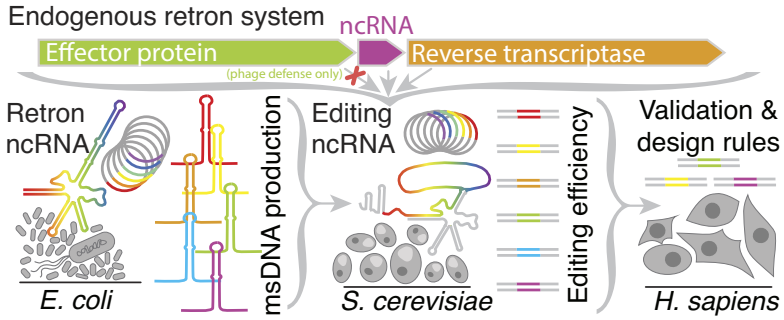
Kate D. Crawford[1,2], Asim G. Khan[1], Santiago C. Lopez[1,2], Hani Goodarzi[3,4,5] and Seth L. Shipman[1,6,7,*]

[1]Gladstone Institute of Data Science and Biotechnology, 1650 Owens St, San Francisco, CA 94158, USA
[2]Graduate Program in Bioengineering, University of California, San Francisco and Berkeley, 1700 Fourth St, San Francisco, CA 94158, USA
[3]Arc Institute, 3181 Porter Dr, Palo Alto, CA 94304, USA
[4]Department of Biochemistry and Biophysics, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158, USA
[5]Department of Urology, University of California, San Francisco, 400 Parnassus Ave, San Francisco, CA 94143, USA
[6]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158, USA
[7]Chan Zuckerberg Biohub San Francisco, 499 Illinois St, San Francisco, CA 94158, USA
[*]To whom correspondence should be addressed. Tel: +1 415 734 4058; Email: seth.shipman@gladstone.ucsf.edu

Abstract

The bacterial retron reverse transcriptase system has served as an intracellular factory for single-stranded DNA in many biotechnological applications. In these technologies, a natural retron non-coding RNA (ncRNA) is modified to encode a template for the production of custom DNA sequences by reverse transcription. The efficiency of reverse transcription is a major limiting step for retron technologies, but we lack systematic knowledge of how to improve or maintain reverse transcription efficiency while changing the retron sequence for custom DNA production. Here, we test thousands of different modifications to the Retron-Eco1 ncRNA and measure DNA production in pooled variant library experiments, identifying regions of the ncRNA that are tolerant and intolerant to modification. We apply this new information to a specific application: the use of the retron to produce a precise genome editing donor in combination with a CRISPR-Cas9 RNA-guided nuclease (an editron). We use high-throughput libraries in *Saccharomyces cerevisiae* to additionally define design rules for editrons. We extend our new knowledge of retron DNA production and editron design rules to human genome editing to achieve the highest efficiency Retron-Eco1 editrons to date.

Graphical abstract



Introduction

Retron components are increasingly being exploited for biotechnology due to their ability to produce DNA on demand in cells. In bacteria, retrons are a tripartite anti-phage system composed of a reverse transcriptase (RT), a non-coding RNA (ncRNA) that is reverse transcribed into DNA (multicopy single-stranded DNA; msDNA) and an effector protein (1,2). For Retron-Eco1 (used in this study), correct msDNA synthesis, initiated at a conserved guanosine via a 2′–5′ linkage, is crucial for phage defense (3,4) and results in filamentous sequestration of the toxic effector protein (5). A phage-encoded DNA cytosine methyltransferase triggers abortive infection by methylating the Retron-Eco1 reverse-transcribed DNA and results in nucleoside derivative depletion (6). The editron system uses only the RT and ncRNA from the retron as the effector protein is not necessary for reverse transcription.

In biotechnology, the retron RT is used to reverse transcribe modified forms of retron ncRNA into reverse-transcribed DNA (RT-DNA), or msDNA that has been used as: donor DNA for precise editing in bacteria (7–12), bacteriophage (13–15), plants (16,17) and eukaryotes (12,18–22); DNA barcodes to record molecular events (23,24); DNA containing transcription factor motifs for transcription factor activity

Received: July 2, 2024. **Revised:** November 14, 2024. **Editorial Decision:** November 17, 2024. **Accepted:** November 19, 2024

attenuation (25); DNA aptamers (26); and DNAzymes for messenger RNA cleavage (27).

Previous work has demonstrated that the abundance of retron reverse-transcribed DNA directly impacts the efficiency of downstream biotechnological applications. Specifically, modifications to the retron that generate more msDNA increase the efficiency of precise editing and the efficiency of event recording into a molecular ledger (19,23,25). These previous works used the same modification to the retron ncRNA for increased msDNA production—extension of the a1/a2 region. However, the retron ncRNA has not been systematically interrogated to determine which elements are necessary, which are tolerant to modifications, and where it may be possible to increase reverse transcription beyond the endogenous element.

In the context of precise genome editing technologies, there are additional parameters that have not been investigated systematically. An editron, which combines retron components with CRISPR-Cas9 components to generate both a programmed double-strand break and the reverse-transcribed donor to precisely repair it, has many degrees of freedom. These include among others, how to arrange the donor and guide RNA (gRNA) relative to each other, where to situate the edit within the donor, or how long of a donor to use. Without a set of clear design rules, users are left to either empirically test many designs for their desired edit or pick an arbitrary design which may not perform optimally.

To rectify this lack of systematic investigation, we comprehensively tested all parameters of the retron ncRNA for their effect on msDNA production in high throughput, used these findings to build a machine learning model of msDNA production, and used the output of the model to inform high-throughput tests of editing parameters in yeast. Finally, we extended these findings to human cells, resulting in a set of design rules for msDNA production and retron-based editing that apply broadly.

Materials and methods

Biological replicates were taken from distinct samples, not the same sample measured repeatedly. For *Escherichia coli* variant libraries, each biological replicate is an independent electroporation and expression of the libraries into the strain bSLS.114. For *Saccharomyces cerevisiae* variant libraries, each biological replicate is an independent transformation and expression of the variant libraries using a scaled-up version of the Zymo Frozen-EZ Yeast Transformation II Kit into the respective yeast strains containing the editing site. For human validation, each biological replicate is an independent transfection and expression of variants using Lipofectamine 3000 into a Cas9-containing HEK293T cell line.

All statistical tests and *P*-values are included in Supplementary Table S1.

Constructs and strains

A derivative of BL21-AI cells was used for all *E. coli* variant library experiments. This derivative, bSLS.114, has the endogenous Retron-Eco1 operon replaced by a chloramphenicol resistance cassette flanked by FRT recombinase sites using the method developed by Datsenko and Wanner (28). This knock-out cassette was amplified from pKD3, adding homology arms to the Retron-Eco1 locus with polymerase chain reaction (PCR) primers, and electroporated into BL21-AI cells with the Lambda Red recombination machinery (pKD46). After selecting clones on 10 μg/ml chloramphenicol plates, we genotyped to confirm the locus-specific knock-out and then excised the chloramphenicol resistance cassette using the FLP recombinase (pMS127).

All yeast variant libraries were cloned into pKDC.100, which contains, under control of a Gal7 promoter, the 5′ end of the *msr*/*msd* and PaqCI Golden Gate restriction enzyme sites at the 3′ end of the *msd* for insertion of variant parts. This plasmid contains a URA3 selection marker and an episomal origin of replication (CEN/ARS), and was constructed using Gibson assembly, with a Twist-synthesized gBlock containing the PaqCI sites and a PCR-amplified linear pSCL039[19]. Yeast plasmids containing the three editing sites in the HIS3 site were based off pZS.157[18]. These three variants (pSCL194: site 1; pSCL195: site 2; and pSCL368: site 3) contain galactose-inducible Retron-Eco1 RT and *Streptococcus pyogenes* Cas9 (Gal1-10 promoter) along with their respective sites. These plasmids were all constructed using Gibson assembly, using pZS.157 to create the backbone and Twist-synthesize gBlocks containing the editing sites. The strains containing these editing sites along with Cas9 and Retron-Eco1 RT were made using LiAc/SS carrier DNA/PEG transformation (29) of BY4742 (30). The respective plasmids were linearized using KpnI and transformed into BY4742 for homologous recombination into the *HIS3* locus. Clones were selected on SD-HIS media.

All human vectors are derivatives of pSCL.273[12], itself a derivative of pCAGGS. pCAGGS was modified by replacing the MCS and rb_glob_polyA sequence with an IDT gblock containing inverted BbsI restriction sites and a SpCas9 trans-activating CRISPR RNA (tracrRNA), using Gibson Assembly. The resulting plasmid, pSCL.273, contains an SV40 ori for plasmid maintenance in HEK293T cells. The strong CAG promoter is followed by the BbsI sites and SpCas9 tracrRNA. BbsI-mediated digestion of pSCL.273 yields a backbone for single or library cloning of plasmids by Gibson Assembly or Golden Gate cloning. Our backbone incorporated an EGFP-P2A and Eco1RT into pSCL.273. Twist-synthesized gBlocks encoding our various ncRNA donors were cloned into this backbone (pKDC.154) via Golden Gate Reaction with PaqCI. Plasmids were subsequently midiprepped according to manufacturer instruction (QIAGEN 12143). Human experiments were carried out in a HEK293T cell line which expresses Cas9 from a Piggybac-integrated, TRE3G-driven, doxycycline-inducible (1 μg/ml) cassette, which we have previously described (19).

All strains/lines are listed in Supplementary Table S3, and all plasmids in Supplementary Table S2.

Variant library cloning

Escherichia coli variant cloning was done as previously described (19) using BsaI Type IIS restriction sites and Golden Gate cloning. After high-efficiency cloning and electroporation, variant libraries were miniprepped for electroporation into the experimental strain (bSLS.114, described above). All *E. coli* variant parts were synthesized by Agilent.

All *S. cerevisiae* variant parts were synthesized by Twist. The variant part of the editron ncRNA was flanked by PaqCI Type IIS restriction sites and specific primers to amplify out sublibraries from a larger synthesis run. Each variant part

was padded by random nucleotides to 250 bp on the 3′ end, and sublibraries were segregated by original variant part length (gated to each sublibrary having <10% variance in the length) to avoid library bias with amplifying out sublibraries by PCR. Variant sublibraries were then combined with pKDC.100 in a Golden Gate reaction using PaqCI and the PaqCI activator (2:1 ratio), and T4 DNA ligase (NEB) to generate cloned sublibraries at high efficiency after electroporation into a cloning strain (ECloni Elite 10G, Biosearch Technologies). Sublibraries were then midiprepped and combined based on the number of variant parts in the sublibrary and the DNA concentration to create a final pooled library with equal distribution of variant parts (QIAGEN).

Variant library expression and sequencing

Escherichia coli variant libraries were grown overnight and diluted 1:500 into expression media (arabinose and isopropyl beta-D-thiogalactopyranoside, or IPTG, for the ncRNA, and erythromycin for the RT). At dilution, we also took a pre-expression sample. We then grew the cells for 5 h shaking at 37°C. After expression, we took two samples: one for variant plasmid quantification and the other for msDNA quantification.

The pre-expression and post-expression plasmid samples were mixed 1:1 with water and boiled at 95°C for 5 min, then plasmid variants were amplified using PCR primers Eco1_Variant_Plasmids_for_Sequencing_F and Eco1_Variant_Plasmids_for_Sequencing_R. *msd* variant plasmids were identified by their altered sequence without barcodes, while *msr* variant plasmids were identified by the matched barcode in the *msd* on the plasmid amplicon.

The msDNA expression sample was prepared as previously described (19). Briefly, DNA was purified using a modified miniprep protocol, treated with RNase A/T1 (New England Biolabs), and purified with single-stranded DNA (ssDNA)/RNA Clean & Concentrator kit from Zymo Research. After ssDNA isolation, we either amplified the DNA barcode with primers containing Illumina adapters (*msr* sublibraries msDNA samples; primers: Eco1_msdloop_for_Sequencing_F and Eco1_msdloop_for_Sequencing_R) or performed a non-sequence-biased sequencing preparation (*msd* msDNA sublibraries). To amplify msDNA without prior knowledge of the sequence, we treated the sample with DBR1 (Origene), extended the 3′ end with dCTP with TdT. We used Klenow fragment (3′→5′ exo-) to create the second complementary strand using a primer with six guanines and an Illumina adapter. After creating the second strand, we ligated an Illumina adapter to the 3′ end of the complementary strand using T4 ligase. All products were indexed and sequenced on the Illumina MiSeq. Sequencing primers are listed in Supplementary Table S4.

All yeast variant libraries were transformed into their matched strain using a 40× scaled-up version of the Zymo Frozen-EZ Yeast Transformation II Kit. After a recovery for 1 h in Yeast Extract Peptone Dextrose media (YPD) and an overnight growth shaking at 30°C in 2% raffinose SD-URA-HIS, a time = 0-h sample was taken and then yeast were passaged to 0.2 OD into 50 ml 2% galactose SD-URA-HIS. Cells were then grown for 24 h shaking at 30°C, a time = 24-h sample was taken. The yeast were then passaged again to 0.2 optical density at 600 nm (OD600) in 50 ml 2% galactose SD-URA-HIS and grown for another 24 h shaking at 30°C. After a total of 48 h of editing, the yeast optical densities were measured again and two aliquots of 500 million cells each were collected for the time = 48-h plasmid and genome sample.

Yeast genomic DNA (gDNA) was extracted as previously described (19). Briefly, cells were lysed in 120 µl lysis buffer (100 mM ethylenediaminetetraacetic acid pH 8, 50 mM Tris–HCl pH 8, 2% sodium dodecyl sulfate) and boiled for 15 min at 100°C. After cooling the lysate on ice, proteins were precipitated by adding 60 µl of ice-cold 7.5 M ammonium acetate and incubating at −20°C for 10 min. The samples were centrifuged at 17 000 × *g* for 15 min to pellet the protein, and the supernatant containing the gDNA was transferred to a new tube. The gDNA was precipitated in 1:1 ice-cold isopropanol at 4°C for 15 min, and then washed twice with 200 µl ice-cold 70% ethanol. The DNA pellet was dried at 65°C for 5–10 min to evaporate all ethanol, and resuspended in 40 µl water. Then, gDNA samples for deep-sequencing were amplified using primers around the editing site containing Illumina adapters. All products were indexed and sequenced on the Illumina MiSeq. Sequencing primers are listed in Supplementary Table S4.

Yeast plasmid DNA was extracted as previously described (31). The Zymo Yeast Miniprep Kit was scaled up to 500m cells. Briefly, we resuspended yeast in 1 ml digestion buffer and 30 µl zymolyase, and digested the cell wall for 3 h shaking at 900 r.p.m. at 37°C. We then added 1 ml of solution II (lysis buffer) to the tubes, split the sample across multiple microcentrifuge tubes and added 1:1 solution III (protein precipitation buffer). We then spun down the tubes and sequentially added the supernatant to the Zymo Yeast Miniprep spin column. After reconsolidating the sample, we washed the spin column with 550 µl wash buffer and eluted in 20 µl pre-warmed ultrapure nuclease-free H_2O at 37°C.

To prepare the plasmid samples for sequencing without the creation of hybrid products, we amplified the plasmid barcodes using 50 ng of plasmid DNA and 16 cycles of amplification, performing eight reactions in parallel per sample using primers containing the Illumina adapters. We then pooled the PCRs for each sample and removed primer-dimers through size-selective bead clean-up. We then use 5 µl of the cleaned-up plasmid DNA amplicons for indexing and sequencing on the Illumina MiSeq. Sequencing primers are listed in Supplementary Table S4.

Machine learning submethods

We split the Retron-Eco1 ncRNA variants and the associated msDNA production values into 2930 training sequences, 154 validation sequences and 342 test sequences. We then trained a convolutional neural network using one-hot-encoded retron ncRNA sequences as inputs and msDNA production as the output. The model parameters that were optimized using Ray Tune were number of layers, step size and number of dilations with a 3:1 train:validation scheme. The final model was made of two computational blocks and a residual dilated convolution block followed by a two-layer perceptron. All model code will be available on GitHub prior to peer-reviewed publication.

Human editing expression and analysis

All HEK cells were cultured in Dulbecco's modified Eagle's medium + GlutaMax supplement (Thermo Fisher

Scientific 10566016) + 10% heat-inactivated fetal bovine serum (HI-FBS). The six-well cultures were transiently transfected with 7.32 μg of plasmid per well using Lipofectamine 3000 (Thermo Fisher Scientific). Twenty-four hours after transfection, doxycycline was refreshed and cultures were passaged into T-25 flasks to be grown for an additional 48 h. Three days after transfection, cells were collected for fluorescence-activated cell sorting (FACS). DAPI dye was added to stain for live/dead and cells were gaited on DAPI and GFP with untransfected cells used as a negative control for background (BD FACSAria Fusion).

Human sample preparation

To prepare samples for sequencing, sorted cells were collected and gDNA was extracted using a QIAamp DNA Mini Kit according to the manufacturer's instructions. DNA was eluted in 50 μl of ultra-pure, nuclease-free water.

Two microliter of the gDNA was used as template in 25-μl PCR reactions with primer pairs to amplify the locus of interest which also contained adapters for Illumina sequencing preparation. Lastly, the amplicons were indexed, and sequenced on an Illumina MiSeq/NextSeq instrument.

msDNA production quantification

msDNA production was quantified as previously described (19). Briefly, custom Python software was used to extract the variant counts from the plasmid and msDNA samples. We then normalized raw counts to relative abundance (raw count over the total number of raw counts) and a variant's msDNA relative abundance to the same variant's plasmid relative abundance, using the average of the pre- and post-induction plasmid abundances to integrate the plasmid abundance over the 5-h expression window. Finally, these relative abundances were normalized to the Retron-Eco1 wild-type abundance, set at 100%.

Editing rate quantification

Custom software was built to quantify library-scale and individual validation editing rates in yeast and human cells. For yeast variant libraries, raw barcode counts were pulled from the 48-h genome (editing site) samples, and the 0-, 24- and 48-h plasmid samples. The read counts from the plasmids were summed across the three time samples to integrate the plasmid abundances over the editing window, and then each barcode read count was normalized against all barcode read counts in that sample. The relative abundance of an editor's barcode in the genome was then divided by the relative abundance of an editor's barcode in the integrated plasmid pool.

For human validation of individual variants, custom software was used to assess the number of reads with the precise edit divided by the number of reads with the wild-type sequence. All types of software used in the analysis of this paper are available on GitHub.

Results

msDNA production in *E. coli* from Retron-Eco1 ncRNA variant libraries

The Retron-Eco1 ncRNA is a highly structured RNA molecule with characteristic stem-loops and double-stranded regions that is partially reverse transcribed to generate abundant RT-DNA, or msDNA in cells (Figure 1A). As previous work found msDNA production was a limiting factor in using the retron as a template for precise editing in prokaryotes and eukaryotes and as a DNA barcode for molecular recording (12,19,23), we set out to systematically understand how variations in ncRNA sequence and structure impact msDNA production in *E. coli*. We constructed a 3443 member library of ncRNA variants, changing both the *msr* (non-reverse-transcribed region) and *msd* (reverse-transcribed region). This library contained all single-nucleotide substitutions, scanning deletions and insertions of varying sizes and variations on length and complementary of stem-loops and all permutation of the three-nucleotide RT recognition motif in the P3 loop. For variants with changes in the *msr*, we included a linked barcode in the P4 loop of the msDNA to allow amplification of the barcode via PCR. In all *msr* sublibraries, we also included a pseudo-wild-type control for normalization which had a linked barcode on the P4 loop to control for the effect of adding 10 nucleotides on msDNA production. The library was constructed using Golden Gate cloning, transformed into a B-strain *E. coli* bSLS.114 (BL21-AI ΔRetron-Eco1), and expressed along with the Retron-Eco1 RT for 5 h, after which we collected msDNA for quantification. All variant sequences are included in Supplementary Tables S9–S18.

To quantify the msDNA abundance of *msd* variants, we used a sequencing pipeline described previously that allows us to amplify msDNA without requiring prior knowledge of the msDNA sequence (4,12,19). Briefly, we (i) purified short ssDNA using a QIAGEN Midiprep Plasmid Plus Kit followed by a Zymo ssDNA Clean & Concentrator Kit, (ii) treated the resulting ssDNA with Dbr1 to remove the 2′–5′ linkage between the msDNA and ncRNA, (iii) extended the debranched ssDNA with a single polynucleotide using template-independent polymerase (TdT), (iv) generated a complementary strand using a primer consisting of the complementary single polynucleotide and an Illumina adaptor, (v) ligated an adaptor to the other end of the now double-stranded msDNA and lastly (vi) Illumina sequenced the now double-stranded msDNA with Illumina adaptors on both ends. msDNA barcodes linked to changes in the *msr* were quantified by amplifying the barcode for sequencing after purifying ssDNA. All variants were normalized against the production of the wild-type retron-derived msDNA and the abundance of the variant plasmid (Figure 1B). To quantify the relative abundance of each variant plasmid in the expression cells, we amplified the variable region of the ncRNA using plasmid-specific primers and sequenced the amplicons using Illumina sequencing.

Figure 1C shows single-nucleotide substitutions scanning across the Retron-Eco1 ncRNA, where we found substantial sequence flexibility on the single-nucleotide level except at two important positions: around the priming guanosine immediately after the a1 region, previously shown to be important for making the 2′–5′ linkage of ncRNA-to-msDNA (32); and around the previously known UUU putative recognition loop for the Retron-Eco1 RT (33) (Figure 1C; Supplementary Figure S1b–e).

We also analyzed deletions scanning across the Retron-Eco1 ncRNA that varied in length from one to five nucleotides. The Retron-Eco1 ncRNA is less tolerant to deletions than substitutions, particularly in the *msr* P2 and P3 stem loops, suggesting a greater influence of structure over sequence. In addition, deletions in the *msd* region directly flanking the a2 region were not tolerated. Larger deletions are less

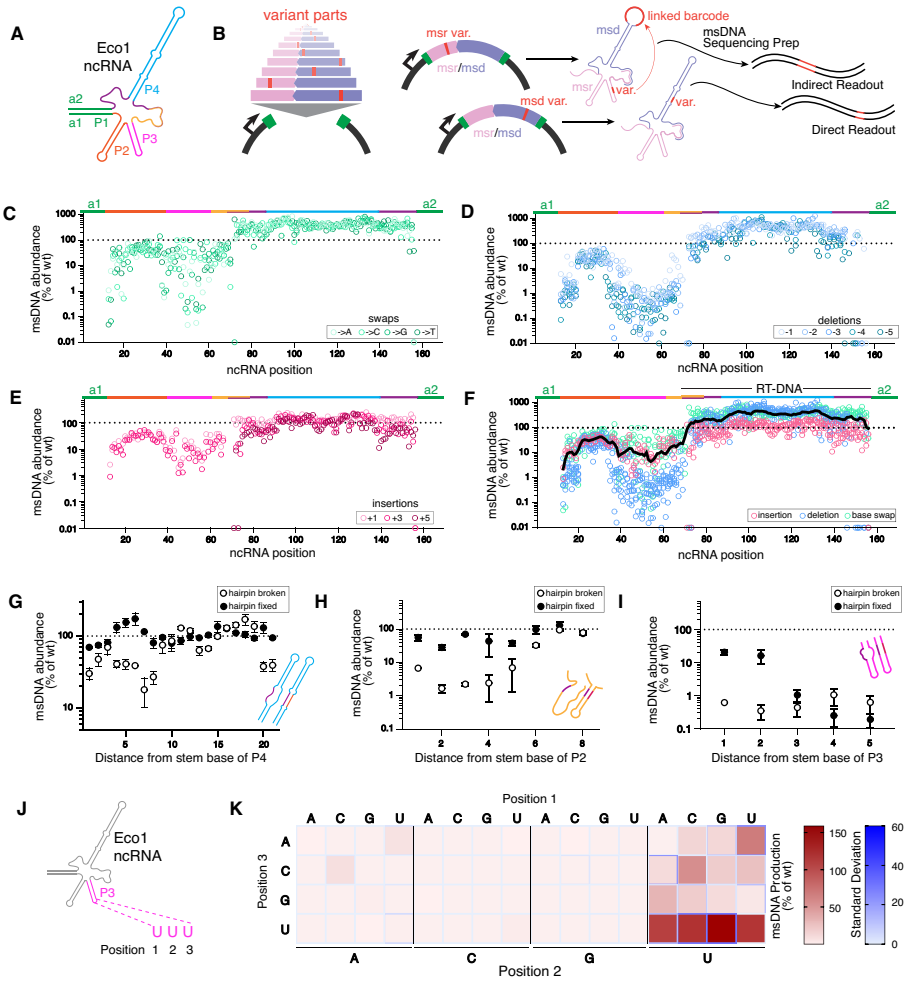


Figure 1. msDNA production of Retron-Eco1 variant libraries in *E. coli*. (**A**) Wild-type -Eco1 ncRNA structure. (**B**) Variant library schematic: variants were introduced on the *msr* (non-reverse-transcribed part of the ncRNA) or the *msd* (reverse-transcribed part of the ncRNA). After production of the msDNA libraries in *E. coli*, ssDNA was sequenced and variants quantified. *msd* variants were identified on the msDNA, while *msr* variants were identified through a barcode in the P4 loop. (**C**) msDNA production of all single-nucleotide substitutions relative to wild-type msDNA. Each open circle represents the mean of three biological replicates. (**D**) msDNA production of 1, 2, 3, 4 and 5 nucleotide deletions starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents the mean of three biological replicates. (**E**) msDNA production of 1, 3 and 5 nucleotide insertions starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents the mean of three biological replicates. (**F**) Summary of msDNA production relative to wild-type msDNA production of all single-nucleotide variants: insertions (pink), deletions (blue) and substitutions (green). msDNA production relative to wild-type msDNA is shown across the nucleotide positions in the ncRNA from 5′ to 3′. The black line on top is the mean of msDNA production of all the changes at that nucleotide position. Each open circle represents the mean of three biological replicates. (**G**) msDNA abundance of removing complementarity (black) and restoring complementarity (white) of stem P4 with different nucleotides along the distance from stem base relative to wild-type msDNA abundance. Each circle represents the mean of three biological replicates with error bars representing the standard error. The effect of breaking the stem is significant (one-way ANOVA using only broken stem and wild-type data, $P < 0.0001$) at positions 1, 4, 5, 6, 7, 8, 18, 20 and 21 compared with the wild-type stem (position 1, $P = 0.005$; position 4, $P = 0.0254$; position 5, $P = 0.0261$ position 6, $P = 0.0194$; position 7, $P = 0.0007$; position 8, $P = 0.003$; position 18, $P = 0.0045$; position 20, $P = 0.0164$; position 21, $P = 0.0208$) (Dunnett's corrected). Restoring the stem structure significantly increases msDNA production only at positions 7 and 21 (position 7, $P = 0.0023$; position 21, $P = 0.0285$) (Bonferroni corrected for multiple comparisons). (**H**) msDNA abundance of removing complementarity (black) and restoring complementarity (white) of stem P2 with different nucleotides along the distance from stem base relative to wild-type msDNA abundance. Each circle represents the mean of three biological replicates with error bars representing the standard error. The effect of breaking the stem is significant (one-way ANOVA using only broken stem and wild-type data, $P < 0.0001$) at all positions compared with the wild-type stem except position 7 compared with the wild-type stem (position 1, $P < 0.0001$; position 2, $P < 0.0001$; position 3, $P < 0.0001$; position 4, $P < 0.0001$; position 5, $P < 0.0001$; position 6, $P < 0.0001$; position 7, $P = 0.7977$; position 8, $P = 0.0029$) (Dunnett's corrected). Restoring the stem structure significantly increases msDNA production at positions 1, 2, 3 and 5 (position 1, $P = 0.01$; position 2, $P = 0.001$; position 3, $P < 0.0001$; position 5, $P = 0.03$) (Bonferroni corrected for multiple comparisons). (**I**) msDNA abundance of removing complementarity (black) and restoring complementarity (white) of stem P3 with different nucleotides along the distance from stem base relative to wild-type msDNA abundance. Each circle represents the mean of three biological replicates with error bars representing the standard error. The effect of breaking the stem is significant (one-way ANOVA using only broken stem data, $P < 0.0001$) at all positions compared with the wild-type stem (position 1, $P < 0.0001$; position 2, $P < 0.0001$; position 3, $P < 0.0001$; position 4, $P < 0.0001$; position 5, $P < 0.0001$) (Dunnett's corrected). Restoring the stem structure only significantly increases msDNA production in position 1 ($P = 0.0041$) (Bonferroni corrected for multiple comparisons). (**J**) Eco1 RT recognition motif UUU in the terminal loop of stem P3. (**K**) msDNA production of every permutation of Retron-Eco1 RT recognition motif relative to wild-type msDNA abundance. Position 1 is shown at the top of the heat map, position 3 on the left and position 2 on the bottom. msDNA production is scaled on the red–white color bar, while the standard deviation is represented by the blue around the squares of the heat map. Each square represents the mean of three biological replicates. There is a significant effect of the RT recognition motif (one-way ANOVA, $P < 0.0001$), with every permutation significantly different than the wild-type UUU ($P < 0.0001$) except UUA and AUU ($P = 0.8991$ and $P = 0.0551$, respectively) (Dunnett's corrected).

tolerated than smaller deletions in the critical region of the P3 stem-loop (Figure 1D; Supplementary Figure S1f–j).

We also assessed one-, three- and five-nucleotide insertions scanning across the Retron-Eco1 ncRNA (Figure 1E). While small insertions were slightly more tolerated than small deletions (Supplementary Figure S1g, k and l), larger insertions in the *msr* region resulted in undetectable levels of msDNA (Supplementary Figure S1m). Similarly to deletions, insertions directly adjacent to the a2 region in *msd* also greatly reduced msDNA production.

A summary of the effect of all nucleotide substitutions, deletions and insertions is shown in Figure 1F. Generally, the Retron-Eco1 ncRNA tolerates modifications in the P2 and P4 stem-loops, but is relatively intolerant to modifications around the priming guanosine and the stem-loop P3. The tolerance to mutations in the P4 stem is important for the use of Retron-Eco1 in biotechnology, as this is the position where editing donors and DNA barcodes have been encoded.

Next, we sought to assess the effect of structural variations. To do this, we quantified the effect of breaking complementarity in stem-loops P2, P3 and P4 by replacing one side of the stem with a non-complementary new sequence to create a nucleotide bubble of length 4 in stem-loops P2 and P3, and length 5 in stem-loop P4. To control for an effect of a sequence versus structural change, we also restored complementarity by changing the same position on the other side of the stem with the complement of the replaced nucleotides. Breaking P4 complementarity only affected msDNA production at the base and the tip of the stem, and fixing complementarity with different sequences restored wild-type levels of msDNA production (Figure 1G). Breaking stem-loop P2 complementarity closer to the base reduced msDNA production and restoring complementarity restored msDNA production (Figure 1H). Breaking stem-loop P3 complementarity reduces msDNA production, but restoring complementarity with an alternate sequence does not restore msDNA production (Figure 1I). Overall, there are clear structural requirements, most notably in P2/3: in P2, structure is important; and in P3, both sequence and structure are important for msDNA production.

We next sought to quantify how strictly required the UUU recognition motif in the loop of P3 is for RT recognition (Figure 1J). Testing every permutation of the UUU motif reveals low sequence flexibility in position 3 (vertical axis) and position 2 (bottom axis), requiring both of these to be uracils. However, there is significant flexibility in position 1, with every possible base approaching wild-type msDNA production levels (lower right square, UUU), with GUU having higher msDNA production than wild-type (Figure 1K; Supplementary Figure S2).

Machine learning on libraries reveals novel variables to increase msDNA production

Though we tested ∼3400 variants of the Retron-Eco1 ncRNA including all single-nucleotide substitutions, a variant library of all possible nucleotide combinations would number on the order of 10^{90} variants, without including insertions and deletions. Therefore, to explore more of the possible sequence space, we used the ncRNA variant library data to create a machine learning algorithm capable of predicting novel retron ncRNA sequences with enhanced msDNA production. The experimental values across ∼3400 measurements were inverse normal transformed and split into a train, validation and test sets. A convolutional neural network, named retDNN, was then used to learn the relationship between sequence and msDNA levels. The retDNN model comprises of two computational blocks and a residual dilated convolution block followed by a two-layer perceptron. The model was trained on 3084 measurements and tested on the held-out set, achieving an $R = 0.671$ performance ($R = 0.775$ on the training set) (Figure 2A and B). We then queried the retDNN model with *in silico* variants, including a P4 stem-loop of varying GC content. Interestingly, the model predicted that lowering GC content in the P4 stem-loop would increase msDNA production over wild-type, something untested in the original variant library. To validate this prediction, we synthesized and cloned the 500 queried variants of differing GC contents (25 variants per 10% GC content range) and experimentally validated msDNA production relative to wild-type through the same sequencing pipeline as above. As the algorithm predicted, lower GC percentages of the P4 stem-loop produced more msDNA (Figure 2C).

Editing performance in *S. cerevisiae* of Retron-Eco1 ncRNA variant libraries

Efficient msDNA production is critical for retron biotechnology, including the use of msDNA as the donor for precise genome editing. In this context, a Retron-Eco1 ncRNA is modified to encode a precise repair donor in the stem-loop of P4 and a gRNA for Cas9 double-strand DNA cleavage at the 3′ end of the ncRNA. This combination of CRISPR-Cas9 and retron immune systems has been called CRISPEY in yeast (18) or as an editron (12) to encompass its use in all eukaryotic cells. After determining the effect of ncRNA variations on msDNA production in *E. coli*, we sought to extend this understanding to editing and additionally investigate how donor, gRNA and ncRNA chassis variants all together affect precise editing rates in eukaryotes.

We designed a library to assess the contributions of structural, cut site and donor variables to precise genome editing by encoding unique donors in the P4 loop of the ncRNA, with each donor variant inserting a unique 10-bp barcode into the yeast genome at a designated site, along with changing the NGG *S. pyogenes* Cas9 protospacer adjacent motif (PAM) to NAT to prevent re-targeting of the edited site. We synthesized variant libraries for the same variables across three unique sites: two artificial, constructed sites with designed, symmetric PAMs around the edit site, and one site from the human genome (an intron in the *NPAS2* gene) with the same PAM locations as the constructed sites. These three sites were independently integrated into the *HIS* locus of *S. cerevisiae* to interrogate the local sequence effects on the editing efficiency, while ensuring the editing site remains active and open by also providing a copy of the *HIS* gene in *HIS* auxotrophic yeast, and maintaining strains in HIS media.

In these variant libraries, we assessed: five donor lengths (54, 64, 78, 94 and 112 nucleotides), five homology arm symmetries about the edit site per donor length, msDNA donors that are complementary to the target or non-target strand and five different cut sites (−16, −8, 0, +8 and +16 relative to barcode insertion point), leading to 175 donor/gRNA combinations per site (Figure 3A). We then combinatorially combined these donor/gRNA variants with 25 different ncRNA chassis: wild-type Eco1 ncRNA, CRISPEY ncRNA (18), the 13 best-performing structural variants from the *E. coli*

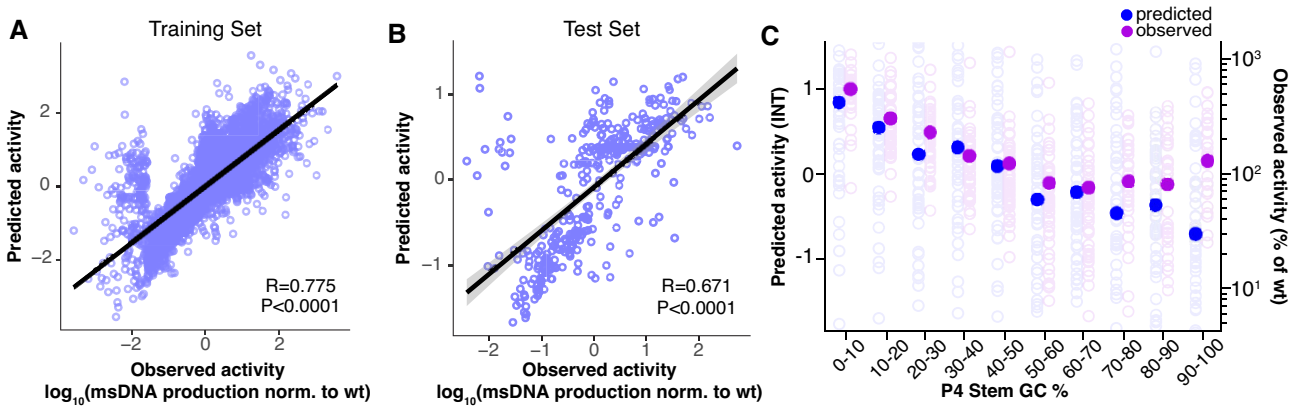


Figure 2. Machine learning on variant libraries guides novel predictors of msDNA production. (**A**) Machine learning algorithm performance on training set of ncRNA variants from *E. coli*. Input is ncRNA sequence and output is inverse-normalized variant msDNA production. Each open circle represents an individual ncRNA sequence. Linear regression *R* and *P*-values of ML predicted activity versus observed activity annotated on the plot. (**B**) Machine learning algorithm performance on held-out test data. Each open circle represents and individual ncRNA sequence. Linear regression *R* and *P*-values of ML predicted activity versus observed activity annotated on the plot. (**C**) Predicted (blue, left set of paired points) and experimentally determined (purple, right set of paired points) msDNA production of varying GC percentages in stem P4. Open circles represent means of two biological replicates of individual ncRNA variants and closed circles represent the mean of all ncRNA variants tested for that GC percentage. Linear regression slope of the predicted (blue) points has a slope of −0.0156 and a *P*-value of < 0.0001. Linear regression slope of the observed (purple) points has a slope of −3.7995 and a *P*-value = 0.0069.

variant libraries and 10 *de novo* predicted ncRNAs from the machine learning algorithm. In all, we tested 4275 variants per site. All variant sequences are included in Supplementary Tables S6–S8.

Three independent yeast lines were created, each with one of the three sites in the *HIS* locus of the yeast genome along with Cas9 and Retron-Eco1 RT under the control of a *GAL1/10* galactose-inducible divergent promoter (Figure 3B). These synthesized ncRNA variants for each site were encoded on a vector containing other necessary ncRNA components (ribozymes, tracrRNA) under the *GAL7* galactose-inducible promoter (Figure 3C). After transformation of the editing libraries into yeast, editing was performed for 48 h in galactose media.

To analyze the data, we sequenced the barcode distribution in the plasmid pool and the barcodes inserted into the correct site in the yeast genome after 48 h of editing. First, we calculated the proportion of each barcode's reads in the pool of reads (for barcodes edited into the genome: the reads at 48 h of editing; for barcodes in the plasmid pool: the reads as summed over samples taken at 0, 24 and 48 h of editing). This is to integrate the plasmid barcode pool over the entire editing period. Plasmid barcode read count was stable over the 48 h of editing (Supplementary Figure S3). Then, we normalized the individual barcode proportions as seen in the genome to the same barcode's proportion as seen in the plasmid pool (called barcode representation henceforth), and removed barcodes not seen at counts >10 in the plasmid pool or not seen at all in the genome pool (percent of working editors per library variable is shown in the Supplementary Figure S4). We then normalized along the axis of interest. For example, when assessing the effect of donor msDNA complementary to either the target or non-target strand (target strand: strand complementary to the gRNA/complementary to the PAM-containing strand; non-target strand: strand not complementary to gRNA/PAM-containing strand), we held all other variables constant (donor length, cut site, donor center and chassis) and normalized the target strand barcode representation to the non-target strand barcode representation of each specific group. This normalized barcode representation for every barcode for each biological replicate for each site is represented as a transparent circle in Figure 3D. We then took the median of each biological replicate of each site, based on the distribution on the right of Figure 3D, and averaged those across all sites to obtain the summary figure for that axis of interest. After performing this normalization, we found that, on average, target strand donors are worse editors than non-target strand donors because the barcode was inserted less often when holding all other variables constant, performing at about 50% efficiency as compared with the matched non-target strand donors (Figure 3E). Both target strand and non-target strand donors have about 50% functional editor variants, as other parameters also influence if an editor is functional (Supplementary Figure S4a). When examining whether cut position relative to edit affects strand polarity preferences, we find that donors complementary to the non-target strand perform worse or equal to donors complementary to the target strand, regardless of whether the cut is positioned on the 5′ or 3′ side of the edit (Supplementary Figure S5).

We analyzed the effect of cut site positioning relative to insertion point by using Cas9 spacer sequences eight nucleotides apart and analyzing as above, normalizing within-group to a cut position of 0, the site at which Cas9 cuts directly where the 10-bp barcode is then inserted. We noticed that the cut site of −8 for site 2 had an unusually low number of working donors (<20%), which was not observed when using other gRNAs at site 2 or with the −8 position at sites 1 and 3 (Supplementary Figure S4b). Given that we intend to quantify the effect of editron parameters and not local sequence around the gRNA, we excluded editrons with the −8 gRNA at site 2 from analysis. At site 1 and 3, we found that an edit on the PAM-proximal side of the cut site performed slightly better (∼130% efficiency at the cut site of −8 compared with a cut position of 0) and performed much better than an edit on the PAM-distal side of the cut site (∼65% efficiency at the cut site of +8), with consistency across sites, while cut sites far

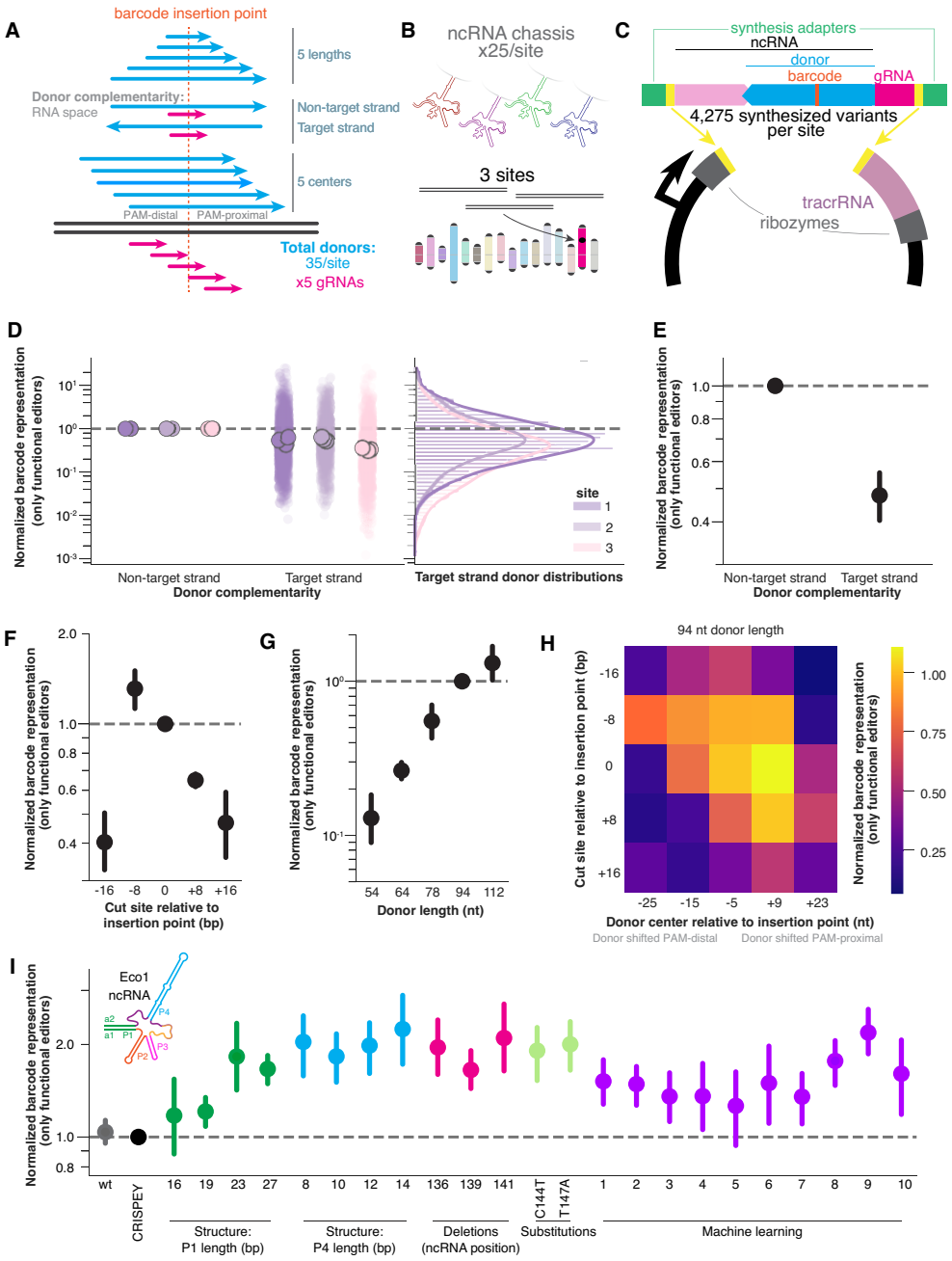


Figure 3. Precise editing of Retron-Eco1 editing variant libraries in *S. cerevisiae*. (**A**) HDR donor variant schematics and gRNA variants, with five donor lengths, two donor directions relative to the gRNA and five donor centers relative to edit and cut position for a total of 50 donors per editing site. There are five evenly spaced gRNAs per site relative to the edit position, for 250 donor/gRNA pairs per site. (**B**) There are 25 ncRNA chassis per donor/gRNA combination. Three sites integrated into the HIS locus of the yeast genome were tested: two synthesized and one from the human genome (NPAS2 locus). (**C**) Schematic for 4275 variant plasmids per site in the library. Each variant has a unique 10-bp barcode that can be read out from the plasmid or from the edit site in the genome. (**D**) All target-strand-homologous gRNA/donor variants' barcode representation normalized against its non-target strand homologous gRNA/donor variant, with all other variables held constant (chassis, donor length, center and gRNA). The variants for each site are broken apart from one another and plotted in different colors, and each biological replicate of a site is summarized by the median (left panel) of the distribution of variants (right panel). (**E**) Data in Figure 3E summarized as the mean of all sites and all biological replicates (closed circle) (±standard deviation), with target-strand-homologous donors editing at significantly lower frequencies (one-sample *t*-test; $P < 0.0001$). (**F**) Barcode representation of cut sites normalized to the cut site at the barcode insertion site (±standard deviation), with cut sites at −16, +8 and +16 editing at significantly lower frequencies (one-sample *t*-test, Bonferroni correction for multiple comparisons; $P < 0.0001$, $P < 0.0001$ and $P < 0.0001$, respectively, all other comparisons non-significant). (**G**) Barcode representation of donor lengths normalized to 94 nucleotide donor length (±standard deviation), with donor lengths <94 nucleotides editing at significantly lower frequencies (one-sample *t*-test, Bonferroni correction for multiple comparisons; $P < 0.0001$, $P < 0.0001$ and $P < 0.01$, respectively, all other comparisons non-significant). (**H**) Heat map of normalized barcode representation of cut site versus donor center (94 nucleotide donor length), normalized to the cut site at the barcode insertion site and donor center of 5 bp upstream the barcode insertion site. Cut site and donor center interact significantly (two-way ANOVA; *P*-value of interaction <0.0001). (**I**) Barcode representation of all chassis ncRNA normalized to the CRISPEY ncRNA (±standard deviation) chassis with a1/a2 27-bp length, 10-bp and 12-bp P4 length, deletion at position 139, substitutions at C144T and T147A and ML chassis 8 and 9 all edit at significantly higher frequencies (one-sample *t*-test, Bonferroni correction for multiple comparisons; $P = 0.004$, $P = 0.028$, $P = 0.036$, $P = 0.019$, $P = 0.049$, $P = 0.019$, $P = 0.024$ and $P = 0.009$, respectively).

from the insertion point resulted in lower frequency of precise editing (∼40–45% efficiency) (Figure 3F). However, it should be noted that only donors complementary to the non-target strand were included in this part of the editron library.

We examined the effect of donor length, normalizing within-group to a donor length of 94 nucleotides. In general, longer donors were more efficient editors than shorter donors, with a 54 nucleotide donor editing at ∼10% of the rate of to the 94 nucleotide donors, while 112 nucleotide had ∼130% efficiency compared with the 94 nucleotide donor (Figure 3G). The percentage of working donors per donor length also increased with donor length (Supplementary Figure S4c).

We assessed the effect of donor center and cut site together by first fixing the donor length (so each donor length is separately analyzed) and then normalizing within-group to the centered cut site (0) and centered donor (–5). The data for 94 nucleotide donor length is shown, as each different donor length has different donor center points. All other donor length results are shown in Supplementary Figure S6. As the higher normalized barcode representation goes from top left to bottom right for the 94 nucleotide donor, it was generally better to center the donor around the cut site than the insertion point, except for cases of cut sites very far from insertion point (top left and bottom right). In addition, for 94 nucleotide donors, when cut and insertion points were overlapping, a slightly PAM-proximal shifted donor performed slightly better than centered, at 110% efficiency compared with centered (Figure 3H). We observed similar but not identical results at other donor lengths (Supplementary Figure S6), potentially because symmetry requirements shift as donor length changes or due to outliers in those donor lengths. The percentage of working donors for donor center and cut site is included in Supplementary Figure S7.

Finally, we analyzed the effect of ncRNA chassis, normalizing within-group to the original CRISPEY chassis. In general, no structural variants performed worse than the CRISPEY chassis, and several variants performed significantly better (27 bp a1/a2 extension, 10 and 12 bp P4 stem length, deletion at position 139, C144T and T147A and machine learning (ML) chassis 8 and 9) (Figure 3I). Excitingly, we found that the machine learning-predicted chassis supported equally high rates of editing despite deviating from the natural sequence by 55–80% in the 20 nucleotide ML variable region, or up to 12% over the full Retron-Eco1 ncRNA including the 27-bp extended a1/a2 (logo map of ribonucleotide usage across the machine learning variable region in Supplementary Figure S9). Specific machine learning chassis structures and sequences can be found in Supplementary Figure S8. We found no evidence of a difference in the percentage of working donors across ncRNA chassis (Supplementary Figure S10).

Library-informed optimization of human editing

We next sought to understand if design rules learned in *E. coli* and *S. cerevisiae* extend to editing in human cells. Editrons contain the same constituent parts in human cells as in yeast, except for the editing ncRNA is driven by an H1 promoter for nuclear retention rather than being flanked by ribozymes. Our plasmids included an EGFP and Retron-Eco1 RT separated by a P2A driven by a CAG constitutive promoter and a ncRNA containing an editing donor fused to a single-guide RNA (sgRNA) driven by a Pol III H1 promoter. The addition of the EGFP enables selection of cells successfully transfected with at least one copy of the editing plasmid. The editing donors consist of a sequence homologous to the desired editing site in the genome but including a PAM recode (NGG > NAT), and a single nucleotide change. We chose to target an intron in the endogenous *NPAS2* site for human validation, using the exact ncRNA constructs used in the yeast libraries. All donors tested are included in Supplementary Table S5.

The editron plasmids were transfected into HEK293T cells containing an integrated doxycycline-inducible Cas9, whose expression was induced 24 h before transfection. Cells were collected 3 days after transfection and sorted via FACS to only include live and transfected cells, eliminating any variability to due transfection efficiency (Figure 4A). We used gRNA 5 for human validation, after an initial screen for gRNA efficacy showed it to have the highest rates of insertions/deletions (indels) of the three tested gRNAs, indicating highest cutting efficiency (Figure 4B). Consistent with our earlier findings in yeast, we demonstrated that a longer donor and a donor homologous to the non-target strand improve editing efficiency (Figure 4C and D). A 112 nucleotide donor increased precise editing from ∼5 to ∼12%, while a non-target strand homologous donor increased editing from ∼4 to ∼12%.

We chose to validate three chassis modifications in human cells. Longer a1/2 length increased editing compared with wild-type a1/a2 length. Excitingly, ML modifications enabled successful editing despite only 30% sequence similarity to wild-type, demonstrating the flexibility of the region (Figure 4E). Next, we sought to determine the ideal positioning of both the edit and the donor relative to a set cut site. We tested three edits: a middle edit at the cut site, an edit 20 bp upstream of the cut site, and an edit 20 bp downstream of the cut site. For each of these edits, we tested a donor which was non-symmetric about the edit with more homology on the 5′ side of the non-target strand, centered on the edit, or non-symmetric about the edit with more homology to the 3′ side of the edit site on the non-target strand (Figure 4F). All donors used were complementary to the non-target strand. We found that placing an edit at the cut site and on the PAM-proximal side both allowed successful editing, with a slight trend favoring the central cut. Additionally, the trend shows that a donor centered on the cut or with more homology on the PAM-proximal side donor both enable editing. None of the conditions with the edit on the PAM-distal side were edited successfully (Figure 4G).

Based on all our variant testing, we provide a set of generalizable design principles for creating future editrons for new targets. Testing several gRNAs to achieve optimal cutting efficiency is an important first step based on our findings showing the variability in indel rates among guides. Donors should be parallel to the guide and complementary to the non-target strand as msDNA, with a 112 nucleotide donor having the highest precise editing rate. Additionally, the cut should be centered or non-symmetrically shifted toward the PAM-proximal side of the non-target strand. When modifying the ncRNA, the a1/2 should be extended at least to 23bp. We also demonstrate flexibility in the 3′ region and the P4 length of the ncRNA, allowing for modifications as needed (Figure. 4H).

Discussion

In this work, we comprehensively evaluated the effect of ncRNA variations on msDNA production in bacteria from which we trained and validated a ML model. We then

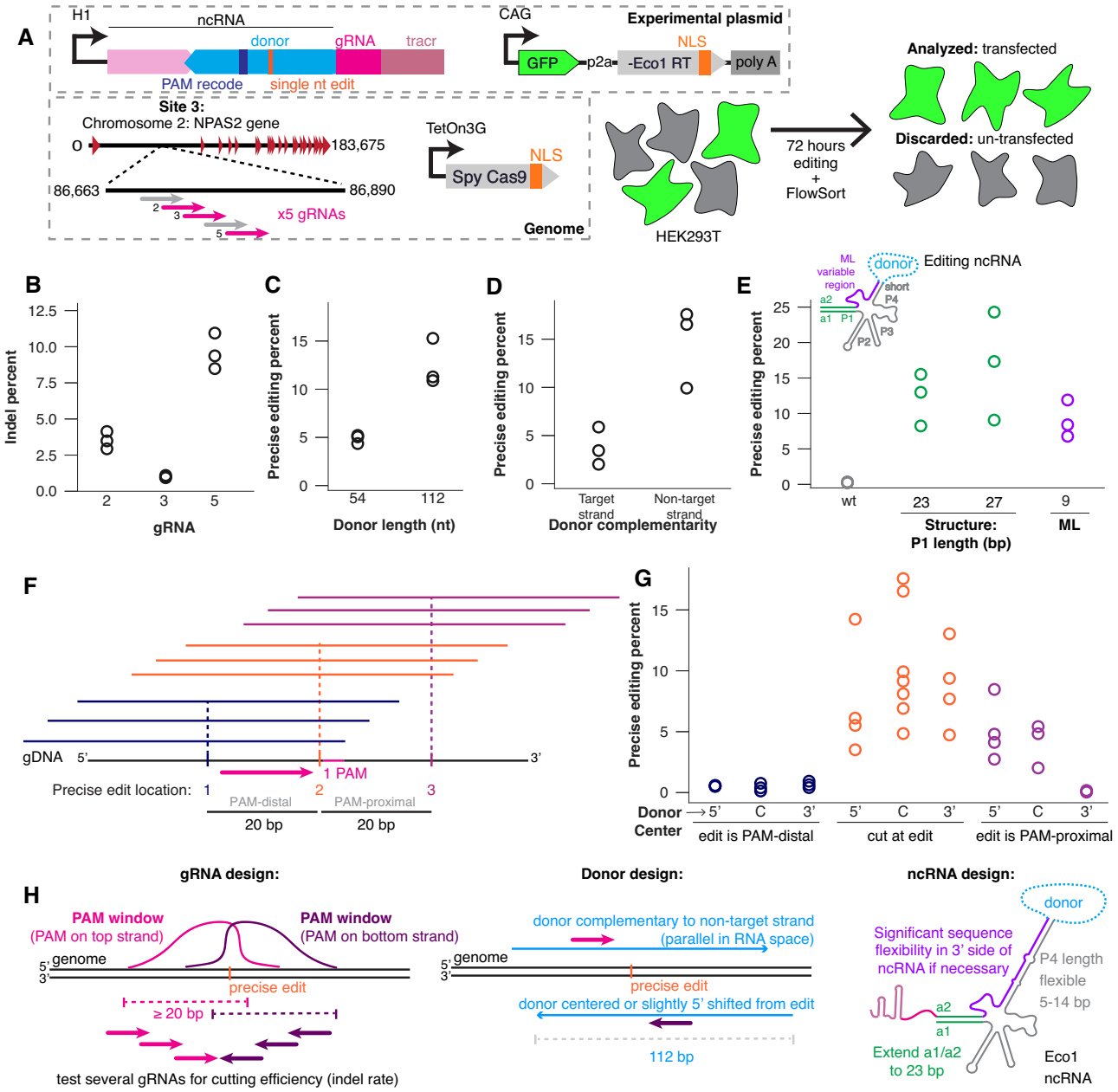


Figure 4. Validating yeast editing libraries with individual human variants. (**A**) Human editing schematic. HEK293T cells were transfected with a plasmid containing the editing ncRNA variant with a single nucleotide transversion as a precise edit, along with recoding the PAM NGG to NAT. The plasmid also contained a constitutively driven GFP-P2A-Eco1 RT. The editron targeted an intronic region of the NPAS2 gene on Chromosome 2 ('site 3' in the yeast data in Figure 3). The HEK293T line also had semi-randomly integrated *S. pyogenes* Cas9 by PiggyBac transposase under a dox-inducible promoter and a C-terminal NLS. Seventy-two hours after transfection, the HEK293T cells were sorted as GFP+/DAPI− (alive transfected cells) and their genomes were sequenced for precise edits. (**B**) Indel percent of the three tested gRNAs. Individual biological replicates are open circles. All gRNA indel rates are statistically different from one another (one-way ANOVA, $P < 0.0001$; Bonferroni *post hoc* test showed $P < 0.05$ for all comparisons). (**C**) Precise editing percentages of 52 nucleotide and 112 nucleotide long donors. Individual biological replicates are open circles. The 112 nucleotide donor is a significantly more efficient editor (paired *t*-test, $P = 0.025$). (**D**) Precise editing percentages of target and non-target strand homologous donors. Individual biological replicates are open circles. Non-target strand homologous donors are significantly more efficient editors (paired *t*-test, $P = 0.043$). (**E**) Precise editing percentages of four ncRNA chassis: wild-type Eco1 ncRNA, extended P1 (a1/a2) (23 and 27 bp) and machine learning chassis 9. Individual biological replicates are open circles. There is a significant effect of ncRNA chassis (one-way ANOVA, $P = 0.01$), with a1/a2 extensions of 23 ($P = 0.0267$) and 27 bp ($P = 0.0046$) performing significantly better than wild-type and ML chassis 9 not performing worse than wild-type ($P = 0.0993$) (Dunnett's corrected). (**F**) Schematic of donor center relative to precise edit site and cut site. Three precise edits were spaced 20-bp apart, with the cut site centered on the middle edit. Three different donor positions were used per edit: 5′-sided, centered and 3′-sided. (**G**) Precise editing percentages of the nine different donor center/edit combinations. Three datapoints in the central cut/centered donor are repeated from (D), as these replicates served as the controls for both the donor center/cut site experiment and the target strand experiment. There is a significant effect of edit site and donor symmetry (one-way ANOVA, $P = 0.0002$), with all edits on the PAM-distal side of the cut ($P = 0.0014$ for 5′ donor center, $P = 0.0012$ for centered donor and $P = 0.0016$ for 3′ centered donor) and the 3′ donor center on the PAM-proximal side ($P = 0.0009$) performing significantly worse than a central cut and edit (Dunnett's corrected). (**H**) Schematic illustrating final recommendations for editron design.

evaluated the effect of variations in donor and gRNA, along with ncRNA structure, on editing efficiency in yeast; and validated the major findings in human cells. From these variant libraries, we found that the *msd* region of the ncRNA is generally tolerant to alterations, specifically the stem-loop P4, in which programmable sequences for biotechnology can be inserted, like a donor sequence for precise editing or a transcription factor motif for attenuating transcription factor activity. We also characterized regions of the *msr* that are required for efficient reverse transcription, such as testing every permutation of the RT recognition motif in stem-loop P3 where the Retron-Eco1 RT initiates reverse transcription (33) and the UAGC sequence which includes the priming guanosine (33). In terms of editing parameters, we found higher rates of editing by increasing donor and a1/a2 length, and using a centered or slightly asymmetric donor with more homology on the PAM-proximal side of the non-target strand. We also demonstrated significant flexibility in the 3′ side of the *msd* sequence for editing, which we altered with targeted deletions, single-nucleotide changes, and stem length alterations. We also changed the 3′ side of the *msd* region to machine learning predicted *de novo* variants of 55–80% difference from the wild-type sequence in the 20 nucleotide ML variable region, or up to 12% over the full Retron-Eco1 ncRNA.

Editrons are conceptually similar to another precise editing approach called prime editing, which uses a nickase Cas9 fused to a promiscuous RT and a gRNA fused to a short donor. The RT extends from the nick using the donor to introduce a precise modification after flap excision and heteroduplex resolution (34). Editrons use prokaryotic, retron RTs, in contrast to the mammalian, viral, MMLV RT most typically used in prime editors. Retron RTs are smaller than MMLV RT, which can be advantageous for delivering parts to cells using plasmids or viruses, and are more processive than MMLV RT, which has enabled much longer insertions (18) than are possible without adding additional proteins, such as Bxb1 recombinase and a recombination donor (35) to prime editing. While prime editing has been extensively optimized, work like this study is necessary to realize the full potential of editrons.

Our variant libraries agree with previous optimizations with single-stranded oligonucleotides (ssODNs) in some aspects, and disagree in others. For example, previous work on ssODNs has found that ssODNs of 70–80 nucleotides have the highest rate of precise repair, and precise repair rates decline above 80 nucleotides (36,37). This is contrary to our finding that precise editing rates increase with increasing length of the msDNA past the previously found optimal length of ssODNs. This difference could be due to lower DNA transfection of longer oligonucleotides or due to the difficulty of synthesizing longer oligonucleotides (36). As our donor is created inside the nucleus of the cell by the Retron-Eco1 RT, our precise editing method will not be limited by synthesis or transfection limitations. We note that, eventually, the Retron-Eco1 RT processivity may hinder production of a longer donor, but that we do not believe we have reached that limit in this work, or that any processivity losses are offset by precise repair gains.

Prior optimization work of ssODN donors has also found that donors asymmetric about the cut site on the non-target strand have better precise editing outcomes, agreeing with our results (9,37,38). After cleavage, Cas9 releases the non-target strand, after which a 3′–to–5′ exonuclease, like Klenow, degrades the 3′ flap (39). Therefore, homology should be biased and asymmetric towards the PAM-proximal side of the non-target strand, as this strand is both free and non-degraded.

We only evaluated asymmetry in a donor homologous to the non-target strand in this study. This is because, in both yeast and human, across different cut sites, we find donors homologous to the non-target strand result in higher precise editing than the target strand, as fits with the mechanism of Cas9 above and to some ssODN studies (36). This is contrary to other ssODN studies, which find that strand polarity preference depends on cut position relative to the edit (40). However, because our editor is a ncRNA reverse-transcribed into a donor, we have the additional complexity of RNA:RNA hybridization. When the reverse-transcribed donor is homologous to the target strand, the gRNA would be homologous to the donor before reverse transcription and could cause the gRNA to be 'hidden' from Cas9 through base pairing with the ncRNA donor. This is an additional complexity not evaluated in optimizing ssODNs, and may increase the effect we observe, with non-target strand complementarity of the donor performing better than target strand complementarity and be the reason some ssODN studies find locus-dependence for strand preference, while we do not, though more loci will need to be tested before fully making this claim (37).

Our first variant library in *E. coli* was aimed at understanding parameters in the retron ncRNA that influence msDNA production. In contrast, our editron variant library used editing as the output, consistent with our goal of identifying parameters that influence editing. It is possible that some editing gains were due to increased msDNA production while others were due to the creation of more favorable donor–target-gRNA interactions. It is likely that the final optimized parts strike a balance between gains in msDNA production and gains from having ideal editing components. Ultimately, our high-throughput approach to testing thousands of variants enabled us to sample a wide space, including potential compromises between the multiple parameters influencing editing, which would have been impossible with traditional experiments testing one parameter at a time.

To our knowledge, this is the first demonstration of using variant libraries to train a ML library that we can query with *de novo* retron ncRNA sequences to assess their possible msDNA production. This high-throughput computational approach allowed us to screen many more sequences *in silico* than currently possible experimentally. Through this, we queried and validated new aspects of the ncRNA that can increase msDNA production, and thus editing. Importantly, we were able to use the output of the ML model to make semi-synthetic ncRNAs that are as functional as wild-type.

Data availability

All data supporting the findings of this study are available within the article and its supplementary information, or will be made available from the authors upon request. Sequencing data associated with this study are available in the NCBI SRA (PRNJNA1121319).

Code availability

Custom code to process or analyze data from this study is available on Github (https://github.com/Shipman-Lab/retron_ncRNA_ML_libraries/tree/master) and Zenodo (https://doi.org/10.5281/zenodo.14058431).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We would like to acknowledge the Gladstone Flow Cytometry Core, which performed the fluorescence-activated cell sorting of the human cells in Figure 4.

Author contributions: K.D.C. and S.L.S. conceived the study and, with the help of S.C.L., outlined the scope of the project and designed experiments. Experiments were performed and analyzed by S.L.S. (Figure 1; Supplementary Figure S1), S.L.S. and H.G. (Figure 2), K.D.C. (Figure 3; Supplementary Figure S2–S8) and A.G.K. and K.D.C. (Figure 4). K.D.C. and S.L.S. wrote the manuscript, with input from all authors.

Funding

National Science Foundation [MCB 2137692 and to K.D.C.]; National Institute of Biomedical Imaging and Bioengineering [R21EB031393]; W. M. Keck Foundation; Pew Biomedical Scholars Program [to S.L.S.]; Gary and Eileen Morgenthaler Fund [to S.L.S.]; UCSF Discovery Fellows Program [to K.D.C.]. Funding for open access charge: National Science Foundation.

Conflict of interest statement

S.L.S. is a co-founder of Retronix Bio and Sprint Synthesis.

References

1. Millman,A., Bernheim,A., Stokar-Avihail,A., Fedorenko,T., Voichek,M., Leavitt,A., Oppenheimer-Shaanan,Y. and Sorek,R. (2020) Bacterial retrons function in anti-phage defense. *Cell*, **183**, 1551–1561.
2. Bobonis,J., Mitosch,K., Mateus,A., Karcher,N., Kritikos,G., Selkrig,J., Zietek,M., Monzon,V., Pfalz,B., Garcia-Santamarina,S., *et al.* (2022) Bacterial retrons encode phage-defending tripartite toxin–antitoxin systems. *Nature*, **609**, 144–150.
3. Gao,L. (2020) Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, **369**, 1077–1084.
4. Palka,C., Fishman,C.B., Bhattarai-Kline,S., Myers,S.A. and Shipman,S.L. (2022) Retron reverse transcriptase termination and phage defense are dependent on host RNase H1. *Nucleic Acids Res.*, **50**, 3490–3504.
5. Carabias,A., Camara-Wilpert,S., Mestre,M.R., Lopéz-Méndez,B., Hendriks,I.A., Zhao,R., Pape,T., Fuglsang,A., Luk,S.H.-C., Nielsen,M.L., *et al.* (2024) Retron-Eco1 assembles NAD+-hydrolyzing filaments that provide immunity against bacteriophages. *Mol. Cell*, **84**, 2185–2202.
6. Wang,Y., Wang,C., Guan,Z., Cao,J., Xu,J., Wang,S., Cui,Y., Wang,Q., Chen,Y., Zhang,D., *et al.* (2023) Defense mechanism of a bacterial retron supramolecular assembly. bioRxiv doi: https://doi.org/10.1101/2023.08.16.553469, 16 August 2023, preprint: not peer reviewed.
7. González-Delgado,A., Lopez,S.C., Rojas-Montero,M., Fishman,C.B. and Shipman,S.L. (2024) Simultaneous multi-site editing of individual genomes using retron arrays. *Nat. Chem. Biol.*, **20**, 1482–1492.
8. Lim,H., Jun,S., Park,M., Lim,J., Jeong,J., Lee,J.H. and Bang,D. (2020) Multiplex generation, tracking, and functional screening of substitution mutants using a CRISPR/retron system. *ACS Synth. Biol.*, **9**, 1003–1009.
9. Schubert,M.G., Goodman,D.B., Wannier,T.M., Kaur,D., Farzadfard,F., Lu,T.K., Shipman,S.L. and Church,G.M. (2021) High-throughput functional variant screens via *in vivo* production of single-stranded DNA. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2018181118.
10. Ellington,A.J. and Reisch,C.R. (2022) Efficient and iterative retron-mediated *in vivo* recombineering in *Escherichia coli*. *Synth. Biol.*, **7**, ysac007.
11. Liu,W., Zuo,S., Shao,Y., Bi,K., Zhao,J., Huang,L., Xu,Z. and Lian,J.. (2023) Retron-mediated multiplex genome editing and continuous evolution in *Escherichia coli*. *Nucleic Acids Res.*, **51**, 8293–8307.
12. Khan,A.G., Rojas-Montero,M., González-Delgado,A., Lopez,S.C., Fang,R.F., Crawford,K.D. and Shipman,S.L. (2024) An experimental census of retrons for DNA production and genome editing. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-024-02384-z.
13. Goren,M.G., Mahata,T. and Qimron,U. (2023) An efficient, scarless, selection-free technology for phage engineering. *RNA Biol.*, **20**, 830–835.
14. Ramirez-Chamorro,L., Boulanger,P. and Rossier,O. (2021) Strategies for bacteriophage T5 mutagenesis: expanding the toolbox for phage genome engineering. *Front. Microbiol.*, **12**, 667332.
15. Fishman,C.B., Crawford,K.D., Bhattarai-Kline,S., Poola,D., Zhang,K., González-Delgado,A., Rojas-Montero,M. and Shipman,S.L. (2024) Continuous multiplexed phage genome editing using recombitrons. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-024-02370-5.
16. Molla,K.A., Shih,J., Wheatley,M.S. and Yang,Y. (2022) Predictable NHEJ insertion and assessment of HDR editing strategies in plants. *Front. Genome Ed.*, **4**, 825236.
17. Jiang,W., Sivakrishna Rao,G., Aman,R., Butt,H., Kamel,R., Sedeek,K. and Mahfouz,M.M. (2022) High-efficiency retron-mediated single-stranded DNA production in plants. *Synth. Biol.*, **7**, ysac025.
18. Sharon,E., Chen,S.-A.A., Khosla,N.M., Smith,J.D., Pritchard,J.K. and Fraser,H.B. (2018) Functional genetic variants revealed by massively parallel precise genome editing. *Cell*, **175**, 544–557.
19. Lopez,S.C., Crawford,K.D., Lear,S.K., Bhattarai-Kline,S. and Shipman,S.L. (2022) Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.*, **18**, 199–206.
20. Zhao,B., Chen,S.-A.A., Lee,J. and Fraser,H.B. (2022) Bacterial retrons enable precise gene editing in human cells. *CRISPR J.*, **5**, 31–39.
21. Kong,X., Wang,Z., Zhang,R., Wang,X., Zhou,Y., Shi,L. and Yang,H. (2021) Precise genome editing without exogenous donor DNA via retron editing system in human cells. *Protein Cell*, **12**, 899–902.
22. Doman,J.L., Pandey,S., Neugebauer,M.E., An,M., Davis,J.R., Randolph,P.B., McElroy,A., Gao,X.D., Raguram,A., Richter,M.F., *et al.* (2023) Phage-assisted evolution and protein engineering yield compact, efficient prime editors. *Cell*, **186**, 3983–4002.
23. Bhattarai-Kline,S., Lear,S.K., Fishman,C.B., Lopez,S.C., Lockshin,E.R., Schubert,M.G., Nivala,J., Church,G.M. and Shipman,S.L. (2022) Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature*, **608**, 217–225.
24. Farzadfard,F. and Lu,T.K. (2014) Genomically encoded analog memory with precise *in vivo* DNA writing in living cell populations. *Science*, **346**, 1256272.
25. Lee,G. and Kim,J. (2023) Engineered retrons generate genome-independent protein-binding DNA for cellular control. bioRxiv doi: https://doi.org/10.1101/2023.09.27.556556, 28 September 2023, preprint: not peer reviewed.
26. Vibhute,M.A., Machatzke,C., Bigler,K., Krümpel,S., Summerer,D. and Mutschler,H. (2024) Intracellular expression of a fluorogenic DNA aptamer using Retron Eco2. bioRxiv doi:

https://doi.org/10.1101/2024.05.21.595248, 24 May 2024, preprint: not peer reviewed.
27. Liu,J., Cui,L., Shi,X., Yan,J., Wang,Y., Ni,Y., He,J. and Wang,X. (2024) Generation of DNAzyme in bacterial cells by a bacterial retron system. *ACS Synth. Biol.*, **13**, 300–309.
28. Datsenko,K.A. and Wanner,B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 6640–6645.
29. Gietz,R.D. and Schiestl,R.H. (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.*, **2**, 31–34.
30. Baker Brachmann,C., Davies,A., Cost,G.J., Caputo,E., Li,J., Hieter,P. and Boeke,J.D. (1998) Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, **14**, 115–132.
31. Muller,R., Meacham,Z.A., Ferguson,L. and Ingolia,N.T. (2020) CiBER-seq dissects genetic networks by quantitative CRISPRi profiling of expression phenotypes. *Science*, **370**, eabb9662.
32. Inouye,S., Hsu,M.-Y., Xu,A. and Inouye,M. (1999) Highly specific recognition of primer RNA structures for 2′-OH priming reaction by bacterial reverse transcriptases. *J. Biol. Chem.*, **274**, 31236–31244.
33. Inouye,M., Ke,H., Yashio,A., Yamanaka,K., Nariya,H., Shimamoto,T. and Inouye,S. (2004) Complex formation between a putative 66-residue thumb domain of bacterial reverse transcriptase RT-Ec86 and the primer recognition RNA. *J. Biol. Chem.*, **279**, 50735–50742.
34. Chen,P.J. and Liu,D.R. (2023) Prime editing for precise and highly versatile genome manipulation. *Nat. Rev. Genet.*, **24**, 161–177.
35. Anzalone,A.V., Gao,X.D., Podracky,C.J., Nelson,A.T., Koblan,L.W., Raguram,A., Levy,J.M., Mercer,J.A.M. and Liu,D.R. (2022) Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat. Biotechnol.*, **40**, 731–740.
36. Yang,L., Guell,M., Byrne,S., Yang,J.L., De Los Angeles,A., Mali,P., Aach,J., Kim-Kiselak,C., Briggs,A.W., Rios,X., *et al.* (2013) Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.*, **41**, 9049–9061.
37. Liang,X., Potter,J., Kumar,S., Ravinder,N. and Chesnut,J.D. (2017) Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. *J. Biotechnol.*, **241**, 136–146.
38. Richardson,C.D., Ray,G.J., DeWitt,M.A., Curie,G.L. and Corn,J.E. (2016) Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.*, **34**, 339–344.
39. Wang,Y., Mallon,J., Wang,H., Singh,D., Hyun Jo,M., Hua,B., Bailey,S. and Ha,T.. (2021) Real-time observation of Cas9 postcatalytic domain motions. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2010650118.
40. Paix,A., Folkmann,A., Goldman,D.H., Kulaga,H., Grzelak,M.J., Rasoloson,D., Paidemarry,S., Green,R., Reed,R.R. and Seydoux,G. (2017) Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E10745–E10754.

Received: July 2, 2024. **Revised:** November 14, 2024. **Editorial Decision:** November 17, 2024. **Accepted:** November 19, 2024