

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Optimizing retron-based genome engineering across the kingdoms of life

Permalink

<https://escholarship.org/uc/item/9hj700zj>

Author

Crawford, Katherine

Publication Date

2024

Supplemental Material

<https://escholarship.org/uc/item/9hj700zj#supplemental>

Peer reviewed|Thesis/dissertation

Optimizing retron-based genome engineering across the kingdoms of life

by

Kate Crawford

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

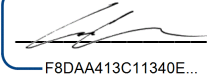
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:



F8DAA413C11340E...

Seth Shipman

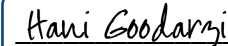
Chair

Signed by:



Liana Lareau

DocuSigned by:



FDD44359FCC6487...

Hani Goodarzi

Committee Members

Copyright 2024

by

Kate Crawford

Dedication:

To my family: my mom, Barb; my brother, Matt; my husband,
Topher; and, of course, my dad, Greg

ACKNOWLEDGEMENTS

This Ph.D. took a village to complete and is immeasurably better for it. I want to acknowledge the following people, and apologize to the people that I've surely missed here.

First, to Seth, my Ph.D. mentor: you took in someone who had never done a PCR before and somehow taught me molecular biology, which is an accomplishment in and of itself. You've also provided kind, supportive mentorship, gently guided me through all the attendant crises of a Ph.D., and awakened in me a love for understanding the more fundamental rules of how the world works. Thank you so much for everything you've done over the past five years.

To current and past committee members, Liana, Leor, and Hani; and past mentors, Shannon, Dino, Henry, Hamdi, and Prof Orwin: you've helped me navigate failed projects and made successful projects possible. You helped mould me into a more thoughtful scientist. Thank you so much for dedicating not-inconsiderable time to helping me grow.

To past and present members of the Shipman lab: your wide-ranging expertise, willingness to always teach and help, and joyful attitude about science inspires me every day. I will miss working with all of you.

To Gladstone: I couldn't imagine a better, more supportive place to do a Ph.D. We are so lucky to have the fourth floor lab aides, especially Lyle, who work tirelessly to make our experiments go as smoothly as they possibly can. I also want to thank our cores, especially Jane at the Flow Core, the Facilities staff, for fulfilling so many work requests to make the lab easier to use, and our entire administrative staff, but particularly Sue, who supports the lab in many ways than I can count, and Sudha, Alicia, and Emma, who spend

so much time and care making sure that graduate students feel supported and specifically have supported me through some difficult times in my Ph.D.

To my lovely friends: Sakshi, Amanda, and Jenn for braving the PhD program with me, and always providing a sounding board and a hug when things are not going well; Izzy and Helen, for sticking it out in San Francisco when all of our other friends were leaving and always being interested in my research even when it sounds like Goobledegook; Sakshi (again) and Joanna, for always being up for any adventure, even though Joanna was one of the aforementioned friends who moved out of San Francisco to New York; and all of our dinner co-op, which has been one of the highlights of my time in Oakland.

To our stinky, stupid, and wonderful cat, Florence, who makes me laugh every day.

To my family, Mom and Matt, for always cheering me on and shaping me into the person I am today, as well as to my extended family, for always being a safe and loving place to return to.

To Topher: you supported me in uncountable ways over these past six years. I am so lucky to have you and it's hard to imagine how I would have done it without you. Here's to our next adventure.

To my dad, Greg: I know you would be so proud of me. I miss you every day.

CONTRIBUTIONS

Some of the material in this dissertation is adapted from or has been published in the references below. The co-authors listed in these publications assisted, directed, or supervised the research that forms the basis for this dissertation.

Chapter 2:

Lopez SC, **Crawford KD**, Lear SK, et al. (2022), Precise genome editing across kingdoms of life using retron-derived DNA. *Nat Chem Biol* **18**, <https://doi.org/10.1038/s41589-021-00927-y>

Chapter 3:

Fishman CB*, **Crawford KD***, Bhattarai-Kline S*, Poola D*, et al. (2024), Continuous multiplexed phage genome editing using recombitrons. *Nat Biotech* **xx**, <https://doi.org/10.1038/s41587-024-02370-5>

Chapter 4:

Crawford KD, Khan AG, Lopez SC, Goodarzi H, and Shipman SL. (2024), High throughput variant libraries and machine learning yield design rules for retron gene editors. *bioRxiv*.

<https://doi.org/10.1101/2024.07.08.602561>

Messieurs, ce sont les microbes qui auront le dernier mot.

Gentlemen, it is the microbes who will have the last word.

— Louis Pasteur

Optimizing retron-based genome engineering across the kingdoms of life

Katherine Doherty Crawford

ABSTRACT

Since the discovery that CRISPR-Cas9 is an RNA-guided DNA-endonuclease and can perform programmable cutting, the genome engineering field has moved from making a simple double-stranded break towards performing a precise edit, where you change the identity of one or many nucleotides to another. However, making a precise repair requires a template for that repair, often made out of single-stranded DNA. In this dissertation, I will detail optimization of one such tool for intracellular DNA production, the bacterial retron, and its utilization as a template for precise repair in: eukaryotes (Chapter 2), bacteriophage (Chapter 3), and then high-throughput libraries for higher rates of precise editing in human cells (Chapter 4).

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION	1
1.1 CRISPR-Cas9	1
A brief primer on CRISPR-Cas9	1
Cas9 as a cutter of bacteriophage DNA	3
Cas9 as a cutter of human DNA	4
1.2 Precise editing tools using Cas9	5
Bacteriophage precise editing	5
Human precise editing	6
1.3 Retrons as intracellular factories for DNA	7
Retron biology.....	7
Retrons as biotechnology	8
1.4 Retrons for precise editing	8
CHAPTER 2 PRECISE GENOME EDITING ACROSS KINGDOMS OF LIFE	
USING RETRON-DERIVED DNA	10
2.1 Abstract	10
2.2 Introduction	11
2.3 Results	13
Modifications to the retron ncRNA affect RT-DNA production.....	13
Modifications to the retron ncRNA affect RT-DNA production.....	18
Improvements extend to applications in genome editing	20
Precise editing by retons extends to human cells	26
2.4 Discussion	29

2.5 Methods	32
Constructs and strains	32
qPCR	38
RT-DNA purification and PAGE analysis.....	39
Variant library cloning.....	40
Variant library expression and analysis.....	41
Recombineering expression and analysis	42
Yeast editing expression and analysis	43
Human editing expression and analysis	45
Reporting Summary	46
Data availability.....	46
Code availability.....	46
2.6 Supplementary Information	47
2.7 Supplemental Files	51
Supplementary_Information_Chapter2.pdf	51
Supplementary_Dataset_Chapter2.xlsx	51
CHAPTER 3 CONTINUOUS MULTIPLEXED PHAGE GENOME EDITING USING RECOMBITRONS	52
3.1 Abstract	52
3.3 Results	56
Recombitrons target phage genomes for editing	56
Continuous editing	62
Optimizing recombitron parameters.....	63

Optimizing host strain	66
Insertions and deletions using recombitrons	68
Multiplexed phage engineering using recombitrons	72
Combinatorial mutations of the T7 tail fiber	76
3.5 Methods	84
Bacterial strains and growth conditions	84
Plasmid construction.....	85
Phage strains and propagation.....	85
Plaque assays	86
Recombineering and sequencing	87
Enrichment of phage in nonediting hosts.....	89
Editing rate quantification	89
Data availability.....	90
Code availability.....	90
Acknowledgements.....	90
Contributions.....	91
3.6 Supplementary Information	92
Supplementary_Information_Chapter3.pdf.....	96
Supplementary_Table2_Chapter3.xlsx	96
CHAPTER 4 CHAPTER 4 HIGH THROUGHPUT VARIANT LIBRARIES AND	
MACHINE LEARNING YIELD DESIGN RULES FOR RETRON GENE EDITORS	97
4.1 Abstract	97
4.3 Results	99

msDNA Production in <i>E. coli</i> from Retron-Eco1 ncRNA Variant Libraries	99
Machine Learning on Libraries Reveals Novel Variables to Increase mDNA Production	105
Editing Performance in <i>S. cerevisiae</i> of Retron-Eco1 ncRNA Variant Libraries.....	107
Library-Informed Optimization of Human Editing	114
4.4 Discussion	119
4.5 Methods	123
Constructs and strains	123
Variant library cloning.....	125
Variant library expression and sequencing	126
Machine learning submethods	129
Human editing expression and analysis	129
Human sample preparation	129
msDNA production quantification.....	130
Editing rate quantification	130
Data availability.....	131
Code availability.....	131
Acknowledgements.....	131
Contributions.....	132
4.6 Supplementary Information	133
Supplementary_Tables_Chapter4_1-5.xlsx	140
Supplementary_Tables_Chapter4_6-8.xlsx	140
Supplementary_Tables_Chapter4_9-18.xlsx	140
CHAPTER 5 REFERENCES	142

LIST OF FIGURES

Figure 2-1: Bacterial retrons enable RT-DNA production.	13
Figure 2-2: Modifications to retron ncRNA affect RT-DNA production	16
Figure 2-3: RT-DNA production in eukaryotic cells	19
Figure 2-4: Improvements extend to applications in genome editing	22
Figure 2-5: Precise editing by retrons extends to human cells	27
Figure 3-1: Recombitrons target phage genomes for continuous editing.	57
Figure 3-2: Optimizing recombitron parameters.	64
Figure 3-3: Insertions and deletions using recombitrons.	69
Figure 3-4: Multiplexed phage engineering using recombitrons.	74
Figure 3-5: Combinatorial mutations of the T7 tail fiber.	78
Figure 4-1: msDNA production of Retron-Eco1 variant libraries in E. coli.	104
Figure 4-2: Machine learning on variant libraries guides novel predictors of msDNA production.	107
Figure 4-3: Precise editing of retron Eco1 editing variant libraries in S. cerevisiae.	113
Figure 4-4: Validating yeast editing libraries with individual human variants.	117
Extended Data Fig 2-1: RT-DNA sequencing prep	47
Extended Data Fig 2-2: RT-DNA production in eukaryotic cells	47
Extended Data Fig 2-3: Precise genome editing rates across additional genomic loci in E. coli	48

Extended Data Fig 2-4: Imprecise editing profile of the yeast ADE2 locus	49
Extended Data Fig 2-5: Genome editing rates across additional genomic loci in yeast	50
Extended Data Fig 2-6: Imprecise editing rates across genomic loci in human cells	50
Extended Data Fig 3-1: Accompaniment to Fig 3-1	92
Extended Data Fig 3-2: Accompaniment to Fig 3-2	93
Extended Data Fig 3-3: Accompaniment to Fig 3-3	94
Extended Data Fig 3-4: Accompaniment to Fig 3-4	95
Extended Data Fig 4-1: Substitution, deletion, and insertion sub-library msDNA production in E. coli.	133
Extended Data Fig 4-2: msDNA production of every permutation of Retron-Eco1 reverse transcriptase recognition motif as a barplot.	134
Extended Data Fig 4-3: Correlation in plasmid read counts over an example 48-hr editing window in S. cerevisiae.	135
Extended Data Fig 4-4: Fraction working editors for different editing variables.	136
Extended Data Fig 4-5: Normalized barcode representation of target and non-target strand donors broken out by cut site.	136
<i>Extended Data Fig 4-6: Normalized barcode representation of donors of varying cut sites vs. donor centers in S. cerevisiae.</i>	137
<i>Extended Data Fig 4-7: Standard deviation of normalized barcode representation of donors of varying cut sites vs. donor centers in S. cerevisiae.</i>	137
<i>Extended Data Fig 4-9: ncRNA structure and sequence of top machine learning ncRNA chassis.</i>	138

Extended Data Fig 4-10: Usage of ribonucleotides in ML ncRNA chassis across variable region. Ribonucleotide height scaled with usage, created by the Python logomaker package. 139

Extended Data Fig 4-11: Fraction working editors across ncRNA chassis. 139

LIST OF ABBREVIATIONS

Abi	Abortive infection
Cas	CRISPR-associated
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
dsDNA	Double-stranded deoxyribonucleic acid
DSB	Double-stranded break
gRNA	Guide RNA
HDR	Homology directed repair
HR	Homologous recombination
Indel	Insertion / deletion
IPTG	Isopropyl β -d1-thiogalactopyranoside
KO	Knock-out
msDNA	Multicopy single-stranded DNA
ncRNA	Non-coding RNA
PAGE	Polyacrylamide gel electrophoresis
PAM	Protospacer adjacent motif
(q)PCR	(Quantitative) polymerase chain reaction
PE	Prime editing
pegRNA	Prime editing guide RNA
Phage	Bacteriophage
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
RBS	Ribosome binding site
RNA	Ribonucleic acid
RT	Reverse transcriptase
RT-DNA	Reverse-transcribed DNA
SSAP	Single-stranded Annealing Protein
SSB	Single-stranded Binding Protein
ssDNA	Single-stranded deoxyribonucleic acid
WT	Wild-type

Chapter 1 Introduction

As every high schooler learns, the central dogma of molecular biology is that: DNA is transcribed into RNA which is translated into proteins. Painting with a broad stroke, proteins are the doers of cells: the things that perform the functions of metabolism, division, contraction, oxygen transport, digestion, and so much more¹. However, proteins can be fleeting, with 100 proteins measured in cancer cells having half-lives of 45 min to 22.5 hr². Therefore, if we truly want to cure diseases where proteins are misbehaving or understand how changes in the protein alter its function, we want to look back to the original cookbook where the recipes for all the proteins are written: DNA.

However, making changes to DNA is not trivial. While many technologies have been developed to make changes to our DNA (gene therapies), we will specifically focus on technologies utilizing a protein called CRISPR-Cas9 in this thesis; discuss some improvements that can be made to Cas9-based gene editing technologies; and, in one case, move beyond technologies requiring Cas9 to allow for more flexible edits to the genome.

1.1 CRISPR-Cas9

A brief primer on CRISPR-Cas9

In 1987, scientists noticed an unusual feature of one of the most ubiquitously-studied organisms in lab, *E. coli*³. Upon sequencing, it was a set of repetitive DNA sequences with a semi-palindromic structure, all of the same length, and separated by unique sequences, also all of a characteristic length. These repetitive sequences were highly-conserved: of the 21 repeats, 15 had the exact same sequence. The same

scientists also found these repeats in closely related species. *Salmonella typhimurium* and *Shigella dysenteriae*⁴. Several years later, tantalizingly, another group found a very similar repeat structure in the halophilic archaea *Haloferax mediterranei* and *Haloferax volcanii*^{5,6}. While the sequence of the repeats were different, the structure was eerily similar: repeats with a semi-palindromic structure separated by unique 'spacers'.

In 2002, these repetitive genomic elements of this structure were named CRISPR (Clustered Regularly Interspersed Short Palindromic Repeats), and they were discovered all over the archaeal and bacterial domains, more than 40 in total. In addition, multiple different strains of the same species were identified with these CRISPR arrays: within a species, the repeat sequence remained constant, while the spacer sequences remained variable, though always within a characteristic length. The same group also identified four genes commonly associated with these repeats, that almost always occur directly upstream or downstream of the CRISPR array, termed *CRISPR-associated* or *cas* genes⁷.

One of these *cas* genes eventually discovered was termed *Cas9*. After much work showing that these CRISPR arrays were transcribed and processed into individual repeats, termed crRNAs, the Doudna lab showed, in 2012, that, outside of cells, the *Cas9* protein could bind to a spacer-based crRNA (guide crRNA), a trans-activating crRNA, and induce a double-stranded break in a piece of DNA that matched the sequence of the guide crRNA⁸. The Doudna lab also created a chimeric RNA that combined both the guide crRNA and the trans-activating crRNA, and termed it a single guide RNA, or simply gRNA. Thus, *Cas9* was shown to be a programmable cutter of double-stranded DNA, and the race to use it to engineer genomes began.

Cas9 as a cutter of bacteriophage DNA

But how did the Doudna lab begin to think Cas9 would cut DNA? The original mechanisms suggested for the CRISPR system were as part of the system for partitioning DNA between two daughter cells during replication. The answer came from the slow unveiling of the CRISPR systems natural function in cells, as an adaptive immune system.

Bacteria have been predated by bacterial viruses, called bacteriophage (phage), for billions of years. In 2005, three groups identified the origin of some of these hypervariable spacers that occur between the repeats: bacteriophage⁹⁻¹¹. Because of the paucity of full genome sequences at that time, the origin of only 20-35% of spacers were identifiable, but the majority of identified spacers came from genetic invaders of bacteria: bacteriophage and extrachromosomal plasmids. Importantly, one group⁹ mapped these spacer sequences back to the genomes from which they originated (protospacers), and found a common motif “purine-pyrimidine-A-A-a sequences downstream from spacer-matching stretch.” This became known as the protospacer adjacent motif, or PAM, and is of utmost importance in the immune and genome engineering function of Cas9.

While these papers provided *in silico* evidence that the CRISPR system may act as an adaptive immune system, where the bacteria keeps a ledger of viruses it has encountered (similarly to how human memory T- and B-cells ‘remember’ a previously-encountered pathogen), they provided no experimental evidence. To do this, Barrangou et al., 2007, challenged *S. thermophilus* with two different phages and showed, after infection, that phage-resistant bacteria that emerged had added one to four additional spacers to their CRISPR array, and that those spacers matched the a protospacer in the

challenging phages genome¹². They then isolated mutant escapee phages, that were able to evade CRISPR system surveillance, and found that those phage had mutated the protospacer sequences within their genomes! A year later, Marraffini and Sontheimer showed similar results with conjugative plasmids, but went one step further to provide evidence that the CRISPR system was directly targeting the plasmid DNA, not inhibiting plasmid RNA production¹³.

Cas9 as a cutter of human DNA

After Cas9, one specific protein in the CRISPR system, was shown cut DNA in a test tube by Jennifer Doudna's group, the remaining question is: could Cas9 cut human genomic DNA inside the nucleus of human cells? Because Cas9 is easily programmable simply by changing the gRNA without needing to re-engineer the protein (as was required for earlier DNA cutters like TALENs or zinc-finger nucleases), if it could be ported into human cells, it could make both large-scale experiments and human therapeutic gene editing more feasible than previously possible.

In early 2013, just a few months after the in vitro cutting results were published, three labs reported site-specific, programmable cutting of human DNA¹⁴⁻¹⁶. All three labs showed Cas9 cutting of the human genome inside the cell and the random repair of the genome initiated by this double-strand break (DSB), leading to small insertions and deletions (indels) at the cut site in the genome. While just creating programmable indels was a step towards programmable gene editing, two lab showed templated repairs, where they used Cas9 and a template piece of DNA to create a precise change in the genome, and targeted deletions by producing two gRNAs within the cell and having the cell remove

the sequence between the two gRNAs^{14,16}. In addition, one lab showed that templated repair was possible beyond immortalized cancer cell lines by showing a Cas9 cut and templated repair in human induced pluripotent stem cells¹⁴.

1.2 Precise editing tools using Cas9

Bacteriophage precise editing

Bacteriophage, as natural predators of bacteria, have emerged in recent years as an alternative to broad-spectrum antibiotics, to which bacteria are aggressively evolving resistance¹⁷. However, current methods for finding a bacteriophage that can treat a multi-drug resistant microbial infection are low-throughput, requiring screening and characterization of many phages before finding one that can treat that specific strain of bacteria^{18,19}. In order to scale and speed up phage therapy development, we require genome engineering tools that can precisely edit phage genome. These tools will allow us to both perform basic science experiments to understand how phage recognize, inject, replicate, and lyse their host, and, in the future, engineer phage to specifically kill a strain of bacteria²⁰.

Classically, phage engineering has been achieved through homologous recombination (HR), where the phage infects a bacterial cell containing a double-stranded piece of DNA that matches the phage genome, except for the desired edit. During replication, the phage will incorporate this piece of DNA into its genome at low rates ($\sim 10^{-10}$ - 10^{-4})²¹. To boost the efficiency of recombination, researchers have turned to CRISPR systems which can cut DNA or RNA as a way to select against the wild-type (WT) phage without the edit, and enrich the phage that contains the edit, called counterselection²²⁻²⁶.

However, there are several issues with CRISPR counterselection. First, phages have been battling against CRISPR systems as a bacterial line of defense against phage infection for billions of years. Thus, phages have evolved diverse mechanisms to escape CRISPR counterselection, including anti-CRISPR proteins²⁷, genome modifications to make gRNA binding more difficult²⁶, and, as a last line of defence, simply mutating their genomes at the protospacer sequence or the PAM to escape targeting^{12,22,23,28}. Secondly, the use of CRISPR counterselection limits the type of edits you can make to the phage genome. Most most single point mutations will not disrupt gRNA binding enough to allow escaping CRISPR targeting, and thus, larger genomic changes are required for efficient sorting of edited phage from WT²⁶.

Human precise editing

Human precise editing also uses the pieces of DNA homologous to the genome to initiate homology directed repair after a Cas9 DSB. In initial Cas9 studies, this piece of DNA was supplied on a plasmid^{14,16} or introduced as a linear piece of double-stranded DNA²⁹. However, it was soon discovered that single-stranded DNA (ssDNA) donors are less toxic than dsDNA³⁰. However, co-delivering a piece of DNA, as well as Cas9 and gRNA (sometimes as a ribonucleoprotein) necessitates all parts successfully being delivered to the cell for the precise repair to happen, and is not easily trackable, as the donor DNA disappears from cell over time due to exonucleases.

Thus, next-generation editors use additional proteins beyond Cas9 to either produce donor DNA within the cell or act directly on the DNA to change a base identity. To produce donor DNA within the cell, researchers use a reverse transcriptase (RT),

which can produce DNA from RNA. Prime editing (PE) uses the mammalian viral RT, M-MLV RT, to produce a fused gRNA and donor template termed a pegRNA, as well as a nickase Cas9, which only cuts one strand of DNA³¹, and has gone through multiple rounds of optimization to reach high editing levels in primary human cells³², as well as fusion with other proteins, such as recombinases, to create larger insertions than possible with just the M-MLV RT³³. Our lab, along with others, also use a bacterial RT called the retron, detailed in the next section and the focus of this dissertation. Base editing is another next-generation technology which fuses a Cas9 nickase to a ssDNA deaminase enzymewhich can change the identity of a base (C-to-T or A-to-G, or the inverse) within the genomic DNA exposed after Cas9 binding and unwinding of dsDNA³⁴. This suite of technologies, along with others, have allowed trackable, library-scale precise editing^{35,36}, and editing in cells which cannot repair DSBs³⁴.

1.3 Retrons as intracellular factories for DNA

Now, like CRISPR-Cas9, we will look at another bacterial immune system that help bacteria avoid phage predation, and has been repurposed for biotechnology.

Retron biology

Similarly to the CRISPR array, retons were originally discovered when researchers observed an unexpected band of DNA from *Myxococcus xanthus* of approximately 180 bp. The researchers cut the band out of a gel, sequenced it, and found something even more peculiar: it was ssDNA primed by an RNA segment on the 5' end³⁷. At this time, reverse transcription had only been identified in mammalian viruses and

eukaryotes, not prokaryotes³⁸. Through pioneering years of work, this system, called the retron, was shown to be composed of a reverse transcriptase, a non-coding RNA (ncRNA) which is often, but perhaps not always, reverse transcribed into a specific piece of ssDNA, and an accessory protein³⁹. This system also confers defense against bacteriophage^{40,41}, often through an abortive infection (Abi) mechanism, utilizing the accessory protein as a toxin.

Retrons as biotechnology

However, the retron system does not strictly require the accessory protein that serves as a toxin; if the retron RT and ncRNA are co-expressed in a cell, that cell will produce many copies of the retron ssDNA, called multi-copy single-stranded DNA (msDNA), at a copy number of ~500-1000 copies^{37,42}. This retron ncRNA can be altered to include sequences of interest, including donor templates for genome engineering⁴²⁻⁵⁶; DNA barcodes^{57,58}; transcription factor decoys⁵⁹; DNA aptamers⁶⁰; and DNazymes⁶¹.

1.4 Retrons for precise editing

In my PhD, I have spent considerable time developing and optimizing retron-based tools for precise genome engineering. In Chapter 2, I will demonstrate of retron msDNA in human cells and use of engineered retron msDNA as a donor for human editing, work done by Santiago López, also in the lab. In Chapter 3, I will describe a multi-pronged, multi-teammate effort to optimize the retron for use as a donor for continuous, multiplexed phage editing. Finally, in Chapter 4, I will delve into work done with Asim Khan to engineer the retron ncRNA architecture, along with editing components such as the gRNA and

donor DNA, to all work in harmony to achieve much higher editing rates than the wild-type version of that retron.

Chapter 2 Precise genome editing across kingdoms of life using retron-derived DNA

2.1 Abstract

Exogenous DNA can be a template to precisely edit a cell's genome. However, the delivery of in vitro-produced DNA to target cells can be inefficient, and low abundance of template DNA may underlie the low rate of precise editing. One potential tool to produce template DNA inside cells is a retron, a bacterial retroelement involved in phage defense. However, little effort has been directed at optimizing retrons to produce designed sequences. Here, we identify modifications to the retron non-coding RNA (ncRNA) that result in more abundant reverse-transcribed DNA (RT-DNA). By testing architectures of the retron operon that enable efficient reverse transcription, we find that gains in DNA production are portable from prokaryotic to eukaryotic cells and result in more efficient genome editing. Finally, we show that retron RT-DNA can be used to precisely edit cultured human cells. These experiments provide a general framework to produce DNA using retrons for genome modification.

2.2 Introduction

Exogenous DNA, which does not match the genome of the cell where it is harbored, is a fundamental tool of modern cell and molecular biology. This DNA can serve as a template to modify a cell's genome, subtly alter existing genes or even insert wholly new genetic material that adds function or marks a cellular event, such as lineage. Exogenous DNA for these uses is typically synthesized or assembled in a tube and then physically delivered to the cells that will be altered. However, it remains an incredible challenge to deliver exogenous DNA to cells in universally high abundance and without substantial variation between recipients⁶². These technical challenges likely contribute to low rates of precise editing and unintended editing that occurs in the absence of template DNA^{63–65}. Effort has been made to bias cells toward template-based editing by manipulating the proteins involved in DNA repair or tethering DNA templates to other editing materials to increase their local concentration⁶⁶. However, a simpler approach may be to eliminate DNA delivery problems by producing the DNA inside the cell.

In recent years, it has been shown that retroelements can be used to produce DNA for genome editing within cells by reverse transcription^{54,55,58,67}. This RT-DNA is produced in cells from plasmids, transgenes or viruses and benefits from transcriptional amplification to create high cellular concentrations that overcome inefficiencies in genome editing. One retroelement class that has been useful in this regard is bacterial retrons^{54,55,58}, which are elements involved in phage defense^{40,41,68,69}. Retrons are attractive as tools for biotechnology due to their compact size, tightly defined sites of reverse transcription initiation and termination, lack of known host factor requirements

and lack of transposable elements. Indeed, retron-generated RT-DNA has demonstrated utility in bacterial^{54,58} and eukaryotic⁵⁵ genome editing.

Despite the potential of the retron as a component of molecular biotechnology, so far, it has been modified as little as is necessary to produce an editing template. Given that the advantage of the retroelement approach is the increased cellular abundance of RT-DNA, we asked whether we could identify retron modifications that would yield even more abundant RT-DNA and increase editing efficiency. Further, most work with retrons has been performed in bacteria, with only one functional demonstration of RT-DNA production in yeast⁵⁵ and only a brief description of reverse transcription in mammalian cells (NIH3T3 mouse cells)⁷⁰. Therefore, we wanted to engineer a more flexible architecture for retron expression across kingdoms of life to serve as a universal framework for RT-DNA production.

Here, we used variant libraries in *Escherichia coli* to show that extension of complementarity in the a1/a2 region of the retron ncRNA increases production of RT-DNA. This effect was generalized across different retrons and kingdoms, from bacteria to yeast. Moreover, retron DNA production across kingdoms was possible using a universal architecture. We found that increasing the abundance of RT-DNA in the context of genome engineering increased the rate of editing in both prokaryotic and eukaryotic cells, simultaneously showing that the template abundance is limiting for these editing applications and demonstrating a simple means of increasing genome-editing efficiency. Finally, we show that the retron RT-DNA can be used as a template for editing human cells to enable further gains in both future research and therapeutic ventures.

2.3 Results

Modifications to the retron ncRNA affect RT-DNA production

A typical retron operon consists of a reverse transcriptase (RT), an ncRNA that is both the primer and template for the RT and one or more accessory proteins⁷¹ (**Fig 2-1a**). The RT partially reverse transcribes the ncRNA to produce a single-stranded RT-DNA with a characteristic hairpin structure, which varies in length from 48 to 163 base pairs (bp)⁷². The ncRNA can be subdivided into a region that is reverse transcribed (*msd*) and a region that remains RNA in the final molecule (*msr*), which are partially overlapping^{37,73–75}.

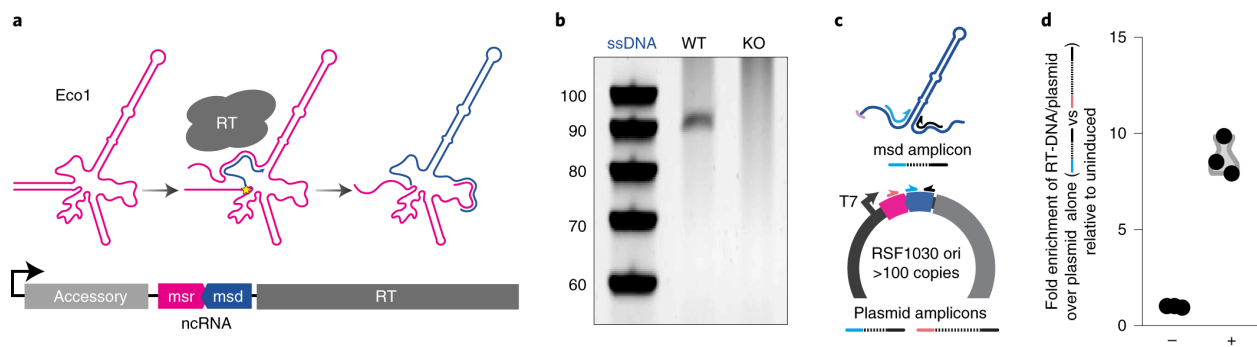


Figure 2-1: Bacterial retrons enable RT-DNA production.

a. Top, conversion of the ncRNA (pink) to RT-DNA (blue); bottom, schematic of the Eco1 retron operon. **b.** Representative image from $n > 3$ PAGE analyses of endogenous RT-DNA produced from Eco1 in BL21-AI wild-type (WT) cells and a knockout (KO) of the retron operon; ssDNA, single-stranded DNA. **c.** Quantitative PCR (qPCR) analysis schematic for RT-DNA. The blue/black primer pair will amplify using both the RT-DNA and the *msd* portion of the plasmid as a template. The red/black primer pair will only amplify using the plasmid as a template; ori, origin of replication. **d.** Enrichment of the RT-DNA/plasmid template over the plasmid alone relative to the uninduced condition, as measured by qPCR; induced versus uninduced: $P = 0.0002$, unpaired t -test; $n = 3$ biological replicates. Circles represent each of the three biological replicates.

One of the first described retrons was found in *E. coli*, Eco1 (previously ec86)⁷⁵. In BL21 cells, this retron is both present and active and produces RT-DNA that can be detected at the population level, which is eliminated by removing the retron operon from the genome (**Fig 2-1b**). In the absence of this native operon, the ncRNA and RT can be

expressed from a plasmid lacking the accessory protein, which is a minimal system for RT-DNA production. We quantified this RT-DNA using qPCR. Specifically, we compared amplification from primers that anneal to the *msd* region, which can use both the RT-DNA and plasmid as a template, to amplification from primers that only amplify the plasmid (**Fig 2-1c,d**). In *E. coli* lacking an endogenous retron, overexpression of the ncRNA and RT from a plasmid yielded an ~eight to tenfold enrichment of the RT-DNA/plasmid region over the plasmid alone, which is evidence of robust reverse transcription (**Fig 2-1d**).

Given that retron utility in biotechnology relies on increasing the RT-DNA abundance in cells above what can be achieved with delivery of a synthetic template, we sought to identify aspects of ncRNA that could be modified to produce more abundant RT-DNA. To do this, we synthesized variants of the Eco1 ncRNA and cloned them into a vector for expression, with the RT expressed from a separate vector. Our initial library contained variants that extended or reduced the length of the hairpin stem of the RT-DNA. This variant cloning took place in single-pot, Golden Gate reactions, and the resulting libraries were purified and then cloned into an expression strain for analysis of RT-DNA production (**Fig 2-2a**). Cells harboring these library vector sets were grown overnight and then diluted, and ncRNA expression was induced during growth for 5 h.

We quantified the relative abundance of each variant plasmid in the expression strain by multiplexed Illumina sequencing before and after expression. After expression, we additionally purified RT-DNA from pools of cells harboring different retron variants by isolating cellular nucleic acids, treating that population with an RNase mixture (A/T1) and isolating single-stranded DNA from double-stranded DNA using a commercial column-based kit. We then sequenced the RT-DNAs and compared their relative abundance to

that of their plasmid of origin to quantify the influence of different ncRNA parameters on RT-DNA production. To sequence the RT-DNA variants in this library, we used a custom sequencing pipeline to prepare each RT-DNA without biasing toward any variant. This involved tailing purified RT-DNA with a string of polynucleotides using a template-independent polymerase (terminal deoxynucleotidyl transferase (TdT)) and generating a complementary strand via an adapter-containing, inverse anchored primer. Finally, we ligated a second adapter to this double-stranded DNA and proceeded to indexing and multiplexed sequencing (**Extended Data Fig 2-1a,b**).

In this first library, we modified the *msd* stem length from 0 to 31 bp and found that stem length can have a large impact on RT-DNA production (**Fig 2-2b**). The RT tolerated modifications of the *msd* stem length that deviate by a small amount from the WT length of 25 bp. However, variants with stem lengths of <12 and >30 bp produced less than half as much RT-DNA than the WT. Therefore, we used a stem length of between 12 and 30 bp going forward.

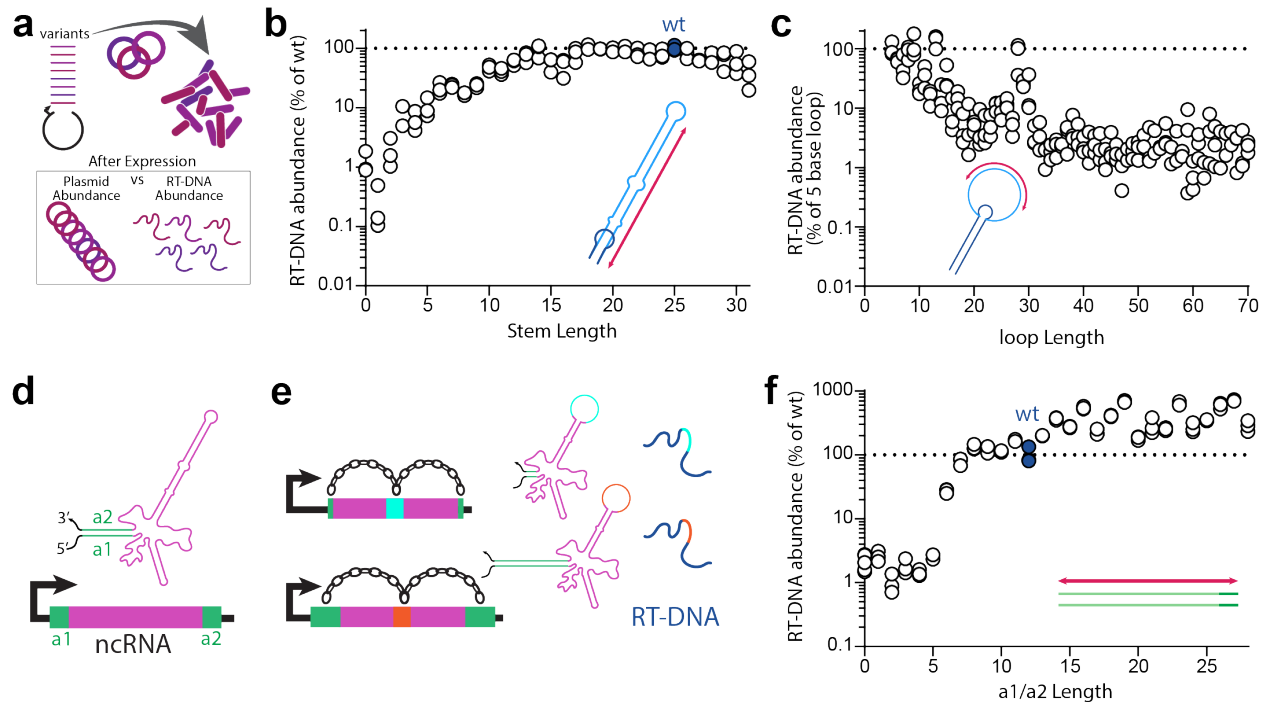


Figure 2-2: Modifications to retron ncRNA affect RT-DNA production

a. Schematic of variant library construction and analysis. **b.** Relative RT-DNA abundance of each stem length variant represented as percentage of WT. Circles represent each of the three biological replicates. WT length is shown in blue along with a dashed line at 100%; effect of stem length: $P < 0.0001$, one-way analysis of variance (ANOVA); $n = 3$ biological replicates. **c.** Relative RT-DNA abundance of each loop length variant represented as a percentage of the value of 5-bp loops. Circles represent each of the three biological replicates, each of which is the average of five loops at that length with differing base content. A dashed line is shown at 100%; effect of loop length: $P < 0.0001$, one-way ANOVA; $n = 3$ biological replicates. **d.** Schematic illustrating the a1 and a2 regions of the retron ncRNA. **e.** Variants of the a1/a2 region are linked to a barcode in the msd loop for sequencing. **f.** Relative RT-DNA abundance of each a1/a2 length variant as a percentage of WT. Circles represent each of the three biological replicates. WT length is shown in blue along with a dashed line at 100%; effect of a1/a2 length: $P < 0.0001$, one-way ANOVA; $n = 3$ biological replicates.

In a second library, we investigated the effect of increasing the loop length at the top of what becomes the RT-DNA stem (**Fig 2-2b**). To do this, we created five random sequences of 70 bp each. We then synthesized variant ncRNAs incorporating 5–70 of these bases into the msd top loop. Thus, we tested five versions of each loop length, each with different base content, and then averaged each variant's RT-DNA production at every loop length. We did not include the WT loop in this library, so we normalized RT-DNA

production to the 5-bp loops, which are closest in size to the WT length of 4 bp. We found a substantial decline in RT-DNA production as loop length increased from 5 to ~14 bp, but we observed almost no continued decline beyond that point other than a single point at 28 bp, which inexplicably produced more RT-DNA than its neighboring loops. While we were limited by our synthesis and sequencing parameters to 70 bp, our conclusion is that loops shorter than 14 bp are ideal for RT-DNA production; however, loops that extend beyond 14 bp do not additionally reduce RT-DNA production.

The other parameter we investigated was the length of a1/a2 complementarity, a region of the ncRNA structure where the 5' and 3' ends of the ncRNA fold back on themselves that we hypothesized plays a role in initiating reverse transcription (**Fig 2-2d**). Because this region of the ncRNA is not reverse transcribed, we could not sequence the variants in the RT-DNA population directly. Instead, we introduced a 9-bp barcode in an extended loop of the *msd* that we could sequence as a proxy for the modification (**Fig 2-2d**). We amplified these barcodes directly from the purified RT-DNA for sequencing (**Fig 2-2e**) or prepared the RT-DNA using the TdT extension method described above (**Extended Data Fig 2-1c**). In both cases, we found a similar effect; reducing the length of complementarity in this region below 7 bp substantially impaired RT-DNA production, consistent with a critical role in reverse transcription (**Fig 2-2f**). However, extending the a1/a2 length resulted in increased production of RT-DNA relative to the WT length. Importantly, this is the first modification to a retron ncRNA that has been shown to increase RT-DNA production.

Modifications to the retron ncRNA affect RT-DNA production

We next wondered whether increased RT-DNA production by the extended a1/a2 region would be a portable modification to other retrons and to eukaryotic systems. To facilitate expression of Eco1 in eukaryotic cells, we inverted the operon from its native arrangement⁷⁶. In the endogenous arrangement, the ncRNA is in the 5'-untranslated region (UTR) of the RT transcript, requiring internal ribosome entry for the RT from a ribosome-binding site (RBS) that is contained in or near the a2 region of the ncRNA. In eukaryotic cells, this arrangement puts the entire ncRNA between the 5' mRNA cap and the initiation codon for the RT. This increased distance between the cap and initiation codon, and the ncRNA structure and out-of-frame ATG codons, is expected to negatively affect RT translation^{76,77}. Moreover, altering the a1/a2 region in the native arrangement could have unintended effects on RT translation. In the inverted architecture, the RT is driven by an RNA polymerase II (Pol II) promoter directly with its initiation codon near the 5' end of the transcript and the ncRNA in the 3'-UTR, where variations are unlikely to influence RT translation (**Fig 2-3a**).

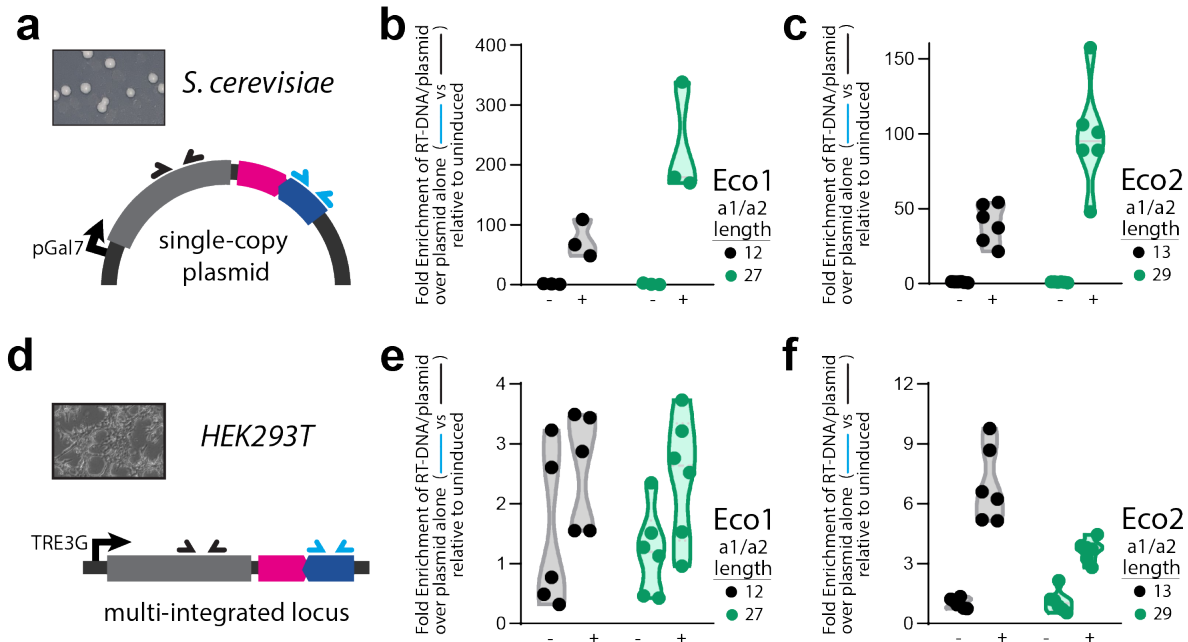


Figure 2-3: RT-DNA production in eukaryotic cells

a. Schematic of the retron cassette for expression in yeast with qPCR primers indicated. **b.** Enrichment of the Eco1 RT-DNA/plasmid template over the plasmid alone by qPCR in yeast, with each construct shown relative to uninduced. Circles show each of the three biological replicates, with black for the WT a1/a2 length and green for the extended a1/a2 length; one-way ANOVA with Sidak's multiple comparisons test (corrected): a1/a2 length 12, induced versus uninduced: $P = 0.2898$; a1/a2 length 27, induced versus uninduced: $P = 0.0015$; a1/a2 length 12 versus 27, induced: $P = 0.0155$; $n = 3$ biological replicates. **c.** qPCR of Eco2 in yeast, otherwise identical to **b**; one-way ANOVA, Sidak's multiple comparisons test (corrected): a1/a2 length 13, induced versus uninduced: $P = 0.006$; a1/a2 length 29, induced versus uninduced: $P < 0.0001$; a1/a2 length 13 versus 29, induced: $P < 0.0001$; $n = 6$ biological replicates. **d.** Schematic of the retron for expression in mammalian cells with qPCR primers indicated. **e.** qPCR of Eco1 in HEK293T cells, otherwise identical to **b**; one-way ANOVA with Sidak's multiple comparisons test (corrected): a1/a2 length 12, induced versus uninduced: $P = 0.2897$; a1/a2 length 27, induced versus uninduced: $P = 0.1358$; a1/a2 length 12 versus 27, induced: $P = 0.9957$; $n = 5$ biological replicates. **f.** qPCR of Eco2 in HEK293T cells, otherwise identical to **b**; one-way ANOVA with Sidak's multiple comparisons test (corrected): a1/a2 length 13, induced versus uninduced: $P < 0.0001$; a1/a2 length 29, induced versus uninduced: $P = 0.0012$; a1/a2 length 13 versus 29, induced: $P < 0.0001$; $n = 6$ biological replicates.

We first tested this arrangement for Eco1 in *Saccharomyces cerevisiae* by placing the RT ncRNA cassette under the expression of a galactose-inducible promoter on a single-copy plasmid. We detected RT-DNA production using a qPCR assay analogous to that described for *E. coli* above and compared amplification from primers that could use the plasmid or RT-DNA as a template to amplification from primers that could anneal only

to the plasmid. Here, we found that increasing the length of the Eco1 a1/a2 region from 12 to 27 bp resulted in more abundant RT-DNA production (**Fig 2-3b** and **Extended Data Fig 2-2a**). We then extended this analysis to another retron, Eco2¹. We found a similar effect; although the WT ncRNA produced detectable RT-DNA, a version extending the a1/a2 region from 13 to 29 bp produced significantly more RT-DNA (**Fig 2-3c** and **Extended Data Fig 2-2a**). In each case, we compared induced to uninduced cells, which likely underreports the total RT-DNA abundance if there is any transcriptional 'leak' from the plasmid in the absence of inducers. Indeed, we detected RT-DNA production in the uninduced condition relative to a control expressing a catalytically dead RT, indicating some transcriptional 'leak' (**Extended Data Fig 2-2b**).

We then moved from yeast to cultured human HEK293T cells. Using a similar gene architecture to yeast, but with a genome-integrating cassette (**Fig 2-3d**), we found that Eco1 does not produce significant abundance of RT-DNA in human cells that we could detect by qPCR, regardless of a1/a2 length (**Fig 2-3e**), from a tightly regulated promoter (**Extended Data Fig 2-2c**). By contrast, Eco2 produces detectable RT-DNA, with both a WT and extended a1/a2 region (**Fig 2-3f**). In human cells, however, the introduction of an extended a1/a2 region diminished, rather than enhanced, production of RT-DNA. Nevertheless, this demonstrates RT-DNA production by a retron in human cells.

Improvements extend to applications in genome editing

In prokaryotes, retron-derived RT-DNA can be used as a template for recombineering^{54,58}. The retron ncRNA is modified to include a long loop in the *msd* that contains homology to a genomic locus along with one or more nucleotide modifications

(Fig 2-4a). When RT-DNA from this modified ncRNA is produced along with a single-stranded annealing protein (for example, λ Red β), the RT-DNA is incorporated into the lagging strand during genome replication, thereby editing the genome of half of the cell progeny. This process is typically performed in modified bacterial strains with numerous nucleases and repair proteins knocked out, because editing occurs at a low rate in WT cells⁵⁴. Therefore, we asked whether increasing RT-DNA abundance using retransposons with extended a1/a2 regions could increase the rate of editing in relatively unmodified strains.

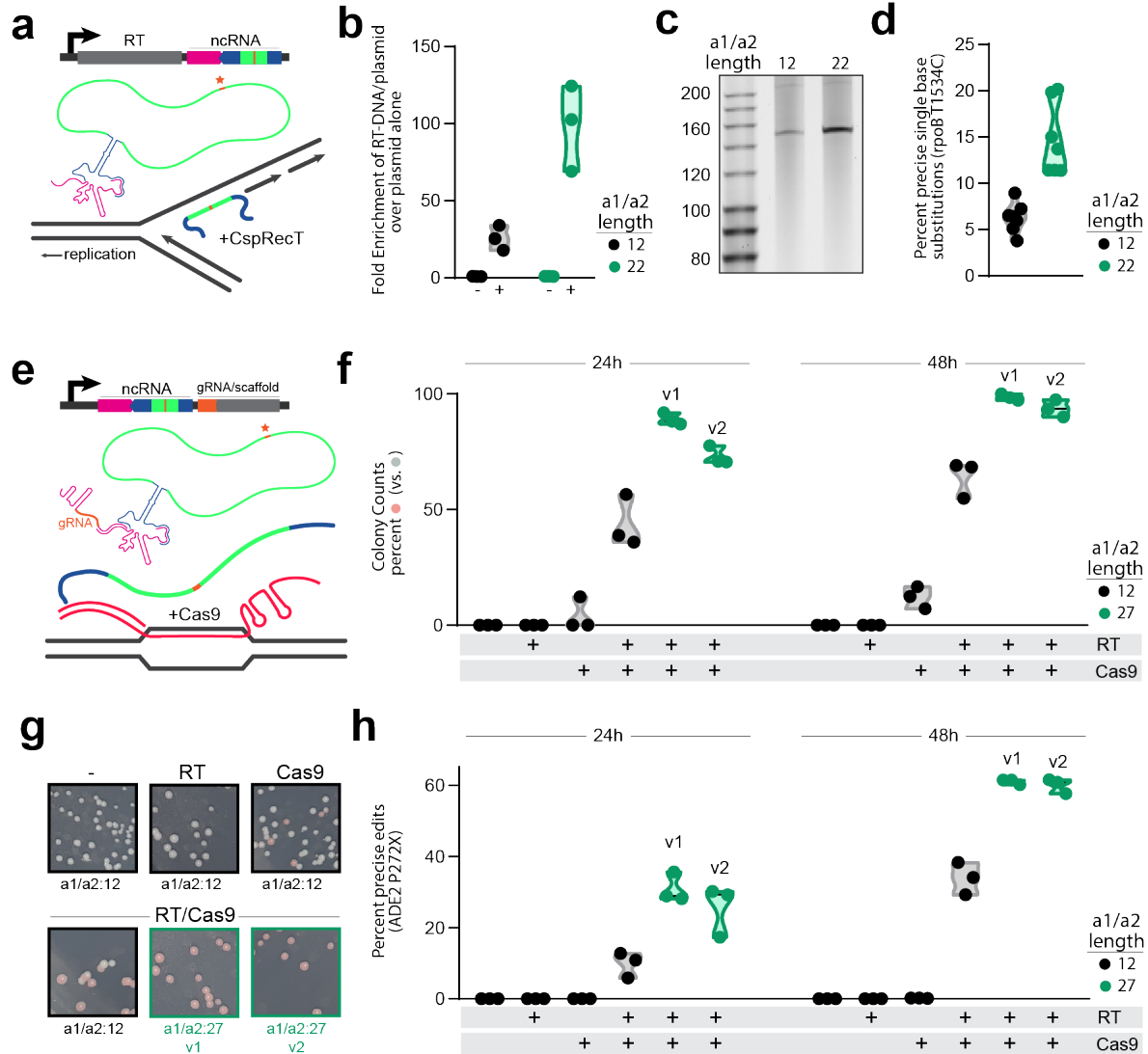


Figure 2-4: Improvements extend to applications in genome editing

a. Schematic of an RT-DNA template for recombineering. **b.** Fold enrichment of the *Eco1*-based recombineering RT-DNA/plasmid template over the plasmid alone by qPCR in *E. coli*, with each construct shown relative to uninduced. Circles show each of the three biological replicates, with black for the WT a1/a2 length and green for the extended a1/a2 length; one-way ANOVA with Sidak's multiple comparisons test (corrected): a1/a2 length 12, induced versus uninduced: $P = 0.1953$; a1/a2 length 22, induced versus uninduced: $P = 0.0001$; a1/a2 length 12 versus 22, induced: $P = 0.0008$; $n = 3$ biological replicates. **c.** PAGE gel showing purified RT-DNA for the WT (a1/a2 length: 12 bp) and extended (a1/a2 length: 22 bp) recombineering constructs to support qPCR; $n = 1$. **d.** Percent of cells precisely edited, quantified by multiplexed sequencing, for the WT (black) and extended (green) recombineering constructs; unpaired *t*-test: a1/a2 length 12 versus 22: $P = 0.1953$; a1/a2 length 22, induced versus uninduced: $P = 0.0001$; a1/a2 length 12 versus 22, induced: $P = 0.0002$; $n = 6$ biological replicates. **e.** Schematic of an RT-DNA/gRNA hybrid for genome editing in yeast. **f.** Percentage of colonies edited based on phenotype (pink colonies) at 24 and 48 h. Circles show each of the three biological replicates, with black for the WT (a1/a2 length: 12 bp) and green for the extended a1/a2 (two extended versions, v1 and v2: a1/a2 length, 27 bp). Induction conditions are shown below the graph for the RT and Cas9; two-way ANOVA: effect of condition (construct/induction), $P < 0.0001$; effect of time: $P < 0.0001$; $n = 3$ biological replicates. **g.** Representative (Figure caption continued on the next page)

(Figure caption continued from the previous page)

images from each condition plotted in **f** at 24 h. Induction conditions are above each image. **h**. Quantification of precise editing of the ADE2 locus in yeast by Illumina sequencing plotted as in **f**; two-way ANOVA: effect of condition (construct/induction), $P < 0.0001$; effect of time: $P < 0.0001$; $n = 3$ biological replicates.

We produced RT-DNA to edit a single nucleotide in the *rpoB* gene. We designed the retron using the same flexible architecture that we used for both yeast and mammalian expression, with the ncRNA in the 3'-UTR of the RT. We used a 12-bp stem for the msd, which retains near-WT RT-DNA production. We constructed two versions of the editing retron, one with the WT 12-bp a1/a2 region and another with an extended 22-bp a1/a2 length. Using qPCR and PAGE analysis, we confirmed that the extended a1/a2 version produced more abundant RT-DNA (**Fig 2-4b,c**). Finally, we expressed each version of the ncRNA along with CspRecT, a high-efficiency single-stranded annealing protein⁷⁸, and mutL E32K, a dominant-negative mutL that eliminates mismatch repair at sites of single-base mismatch^{79,80}, in BL21-AI cells that were unmodified other than the removal of the endogenous Eco1 retron operon. Both ncRNAs resulted in appreciable editing after a single 16-h overnight expression, but the extended version was significantly more effective (**Fig 2-4d**). To test whether the effect of the a1/a2 extension was locus-specific or generalized across genomic sites, we tested an additional three loci⁸¹ for precise editing. We found that the engineered retron mediated editing at each additional loci and that the efficiency of editing was improved by the a1/a2 extension at all three additional sites (**Extended Data Fig 2-3**). This shows that the abundance of the RT-DNA template for recombineering is a limiting factor for editing and that modified ncRNA can be used to introduce edits at a higher rate.

Retron-derived RT-DNA can also be used to edit eukaryotic cells⁵⁵. Specifically, in yeast, the ncRNA is modified to contain homology to a genomic locus and to add one or more nucleotide modifications in the loop of the *msd*, similar to the prokaryotic template. However, in this version, the ncRNA is on a transcript that also includes a *Streptococcus pyogenes* Cas9 (SpCas9) guide RNA (gRNA) and scaffold. When these components are expressed along with RT and SpCas9, the genomic site is cut and repaired precisely using the RT-DNA as a template (**Fig 2-4e**). We tested our modified ncRNAs using an architecture that was otherwise unchanged from a previously described version⁵⁵. The ncRNA/gRNA transcript was expressed from a galactose-inducible promoter on a single-copy plasmid flanked by ribozymes. Along with the plasmid-encoded ncRNA/gRNA, we expressed either Eco1RT, Cas9, both the RT and Cas9 or neither from galactose-inducible cassettes integrated into the genome. The ncRNA/gRNA was designed to target and edit the *ADE2* locus, resulting in both a two-nucleotide modification and a cellular phenotype (pink colonies).

Using the ncRNA with a 12-bp a1/a2 length, we found that the expression of both the RT and Cas9 was necessary for editing based on pink colony counts, with only a small amount of background editing when we expressed Cas9 alone (**Fig 2-4f,g**). This is consistent with the reverse transcription of the ncRNA being required rather than having the edit arise from the plasmid as a donor. To test the effect of extending the a1/a2 region on genome-editing efficiency, we designed two versions of the a1/a2 extended forms, both of which had a length of 27 bp but differed in their a1/a2 sequence. We found that both versions outperformed the standard 12-bp form for precise genome editing (**Fig 2-4f,g**). Consistent with our results in *E. coli*, this indicates that RT-DNA production is a

limiting factor for precise genome editing and that extended a1/a2 length is a generalizable modification that enhances retron-based genome engineering. We further confirmed these phenotypic results by sequencing the *ADE2* locus from batch cultures of cells (**Fig 2-4h**). Precise modifications of the site, resulting from edits that use the RT-DNA as a template, follow the same pattern as the phenotypic results, showing editing that depends on both the Cas9 nuclease and RT, and are increased by extension of the a1/a2 region.

We also found that the rates of precise editing determined by sequencing from batch cultures were consistently lower than those estimated from counting colonies. This is likely due to additional editing that continues to occur on the plate before counting and our method of counting colonies as pink even if they were only partially pink. Another source of pink colonies could be any imprecise edits to the site that result in a non-functional *ADE2* gene. Indeed, we observed some *ADE2* loci that matched neither the WT nor precisely edited sequence. These occurred at a low rate (~1–3%) in all conditions, which was slightly elevated by Cas9 expression but unaffected by RT expression/RT-DNA production (**Extended Data Fig 2-4a**). This, as well as the pattern of insertions, deletions, transitions and transversions, is consistent with a combination of sequencing errors and Cas9-produced insertion–deletions (indels) (**Extended Data Fig 2-4b,c**).

As in the bacterial experiments, we tested whether the extended a1/a2 modification was a generalizable improvement by targeting additional loci across the genome. To this end, we generated WT and extended a1/a2 retons to edit four additional loci¹ in yeast (*TRP2*, *FAA1*, *CAN1* and *LYP1*). We found that for three of the four additional loci, the extended a1/a2 retons yielded higher rates of precise editing, whereas

one site showed lower, but still substantial, rates of editing with the extended version (**Extended Data Fig 2-5**). Overall, across the nine sites tested in bacteria and yeast, the a1/a2 extension improved editing rates at eight sites.

Precise editing by retrons extends to human cells

Finally, we sought to test whether retron-produced RT-DNA could be used for precise editing of human cells as a step toward future therapeutic applications and research applications seeking to unravel the mechanisms of genetic disease. Porting the editing machinery to cultured human cells required some additional modifications. In yeast, we produced both Cas9 and the retron RT from separate promoters. In human cells, expressing both of these proteins from a single promoter would greatly simplify the system and increase its portability. To identify an optimal single-promoter architecture, we tested six arrangements in yeast: four fusion proteins using two different linker sequences with both orientations of Cas9 and Eco1RT, and two versions where Cas9 and Eco1RT were separated by a P2A⁸² sequence in both possible orientations. These constructs were coexpressed with the best-performing *ADE2*-editing ncRNA/gRNA construct described above (extended v1, a1/a2 length of 27 bp). We found that expression of these constructs resulted in a range of precise editing rates, with the Cas9–P2A–RT version yielding editing rates comparable to our previous versions based on two promoters (**Fig 2-5a**).

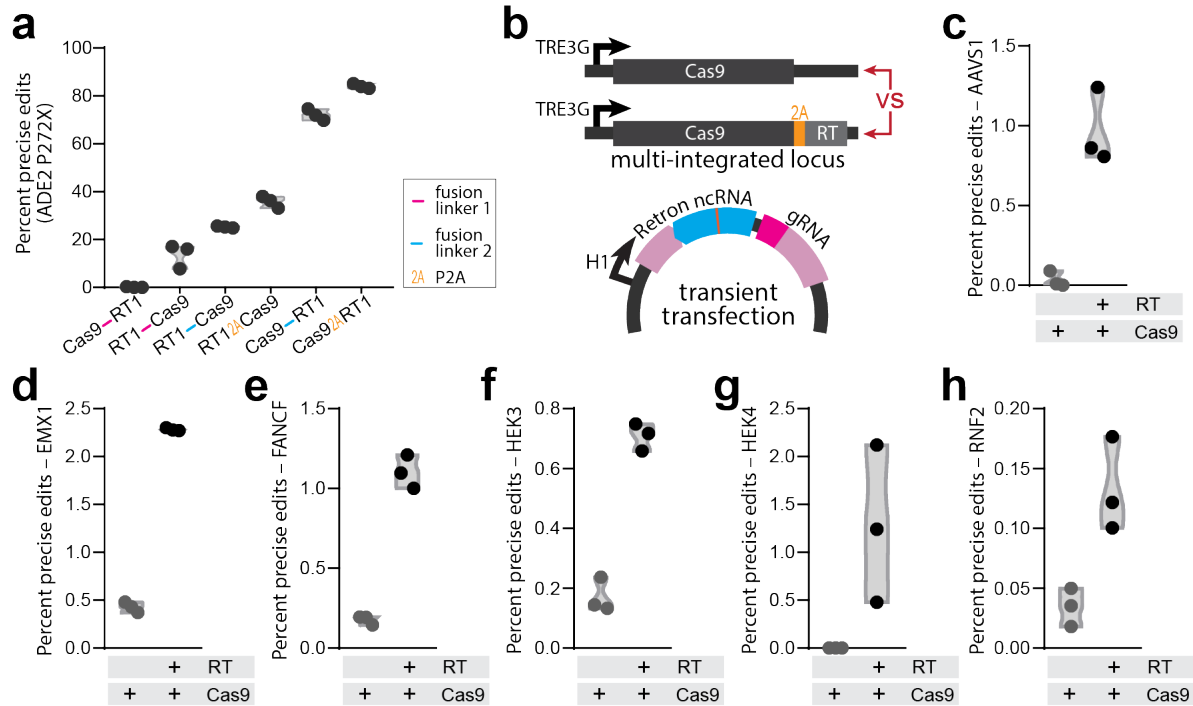


Figure 2-5: Precise editing by retrons extends to human cells

a. Testing different single-promoter architectures for editing the *ADE2* locus in *S. cerevisiae*. The arrangement of proteins is indicated below, and the fusion linkers are listed in the **Methods**. Circles show each of the three biological replicates; one-way ANOVA, effect of construct: $P < 0.0001$; $n = 3$ biological replicates. **b.** Schematic showing the elements for editing in human cells. Top, integrated protein cassettes that are compared in **c–h**. Bottom, plasmid for transient transfection of the site-specific ncRNA/gRNA. **c.** Quantification of precise editing of the *AAVS1* locus in HEK293T cells by Illumina sequencing. Proteins present are shown below. Circles represent each of the three biological replicates; unpaired *t*-test: effect of Cas9 alone versus Cas9 and RT: $P = 0.0026$; $n = 3$ biological replicates. **d–h.** Experiments and plots identical to **c**, but for *EMX1* (**d**), *FANCF* (**e**), *HEK3* (**f**), *HEK4* (**g**) and *RNF2* (**h**) loci, respectively; for **d–h**, unpaired *t*-test: effect of Cas9 alone versus Cas9 and RT: $P < 0.0001$, $P = 0.0001$, $P = 0.0002$, $P = 0.0543$ and $P = 0.0158$, respectively; $n = 3$ biological replicates.

We then created two HEK293T cell lines that each harbored one of two integrating cassettes: Cas9 alone or Cas9-P2A-Eco1RT (**Fig 2-5b**). We initially tested precise genome editing using a Pol II-driven ncRNA/gRNA flanked by ribozymes, as we had in yeast. However, we found no evidence of either precise editing or indels, consistent with previous reports of inefficient ribozyme-mediated gRNA release in human cells⁸³. Therefore, we changed the expression of our retron ncRNA/gRNA to be driven by a Pol III H1 promoter, which was carried on a transiently transfected plasmid (**Fig 2-5b**). Six

genomic loci (*HEK3*, *RNF2*, *EMX1*, *FANCF*, *HEK4* (ref.⁶⁷) and *AAVS1* (ref.¹⁴)) were selected for editing, and an ncRNA/gRNA plasmid aiming to target and edit the site was generated.

The repair template was designed to introduce two distinct mutations separated by at least 2 bp: the first introduced a single-nucleotide change near the cut site, and the second recoded the PAM nucleotides (NGG → NHH, H: non-G nucleotide). The reasoning for this was twofold. First, the multiple changes should both eliminate Cas9 cutting of the ncRNA/RT plasmid and recutting of the precisely recoded site. Second, these multiple, separated changes make it much less likely to mistakenly assign a Cas9-induced indel as a precise edit. As a technical aside, we would recommend against using single-base modifications to benchmark Cas9-induced precise editing applications, as they are a common outcome of imprecise repair and can easily lead to inaccurate estimates of editing rate. We induced expression of the protein(s) for 24 h, transfected the ncRNA/gRNA plasmids and collected cells 3 d after transfection. Using targeted Illumina sequencing, we found precise editing of each site in the presence of the RT, well above the background rate of editing in the absence of the RT (**Fig 2-5c–h**). We believe that the small percentage of precise edits in the absence of the RT likely represents use of the plasmid as a repair template, and the gain in the editing rate in the presence of the RT indicates edits using RT-DNA as the template. Interestingly, we see that the rates of imprecise edits (indels) decline in the presence of the RT by roughly the same magnitude as the precise edits themselves, suggesting that the RT-DNA is being used to precisely edit sites that would have otherwise been edited imprecisely (**Extended Data Fig 2-6**).

2.4 Discussion

The bacterial retron is a molecular component that can be exploited to produce designer DNA sequences in vivo. Our results yield a generalizable framework for retron RT-DNA production. Specifically, we show that a minimal stem length must be maintained in the msd to yield abundant RT-DNA and that the msd loop length affects RT-DNA production. We also show that there is a minimum length for the a1/a2 complementary region. Perhaps most importantly, we demonstrate that the a1/a2 region can be extended beyond its WT length to produce more abundant RT-DNA and that increasing template abundance in both bacteria and yeast increases editing efficiency.

Importantly, these modifications are portable, both across retrons and across species. The extended a1/a2 region produces more RT-DNA using Eco1 in bacteria and both Eco1 and Eco2 in yeast. Oddly, the extended a1/a2 region did not increase RT-DNA production in cultured human cells. Further work will be necessary to optimize RT-DNA production in human cells specifically. Nonetheless, we provide a clear demonstration of retron-produced RT-DNA in human cells.

Retrons have been used to produce DNA templates for genome engineering^{54,55,58}, driven by the rationale that an intracellularly produced template eliminates the issues related to exogenous template delivery and availability. However, there have been no investigations of whether RT-DNA templates are abundant enough to saturate the editing or if even more template would lead to higher rates of editing. Our results establish that editing template abundance is limiting for genome editing in both bacteria and yeast because extension of the a1/a2 region, which increases the abundance of the RT-DNA, also increases editing efficiency.

Additionally, the inverted arrangement of the retron operon, with the ncRNA in the 3'-UTR of the RT transcript, was found to produce RT-DNA in bacteria, yeast and mammalian cells. Here, we show that a single, unifying retron architecture is compatible with all of these host systems, simplifying comparisons and portability across kingdoms.

We also show, consistent with contemporaneous studies⁵³, that the retron RT-DNA can be used as a template to precisely edit human cells. Further, our repair template design allows us to confidently call the precise editing rates. Importantly, we have also applied the same analysis to the Cas9-only conditions and reported the precise editing rates therein and recommend that this approach be applied in future work. We believe that this will allow for estimations of the proportion of precise editing attributable to nuclease-only activity and will ultimately help in obtaining more realistic estimates of the precise editing rates attributable to the genome-engineering tool of interest.

One major difference between the two eukaryotic systems (yeast/humans) is the ratio of precise to imprecise editing. Yeast RT-DNA-based editing occurs at a ratio of ~74:1 precise edits:imprecise edits, while human editing inverted at a ratio of ~1:15 precise edits to imprecise edits. Whether this is a result of differences in repair pathways or the substantial difference in the abundance of retron-produced RT-DNA between yeast and human cells that we report here, it represents a clear direction for future research and technological advances in this area. In summary, this work represents an important advance in the versatile use of retron in vivo DNA synthesis and RT-DNA for genome editing across kingdoms.

2.5 Methods

All biological replicates were collected from distinct samples and not from the same sample measured repeatedly. Full statistics can be found in **Supplementary Table 2-4**.

Constructs and strains

For bacterial expression, a plasmid encoding the Eco1 ncRNA and RT in that order from a T7 promoter (pSLS.436) was constructed by amplifying the retron elements from the BL21-AI genome and using Gibson assembly for integration into a backbone based on pRSFDUET1. The Eco1RT was cloned separately into the erythromycin-inducible vector pJKR-O-mphR⁸⁴ to generate pSLS.402. Eco1 ncRNA variants were cloned behind a T7/lac promoter in a vector based on pRSFDUET1 with BsaI sites removed to facilitate Golden Gate cloning (pSLS.601) and is described further below. Eco1 RTs along with recombineering ncRNAs driven by T7/lac promoters (pSLS.491 and pSLS.492) were synthesized by Twist in pET-21(+).

Bacterial experiments were performed in BL21-AI cells or a derivative of BL21-AI cells. These cells harbor a T7 polymerase driven by a ParaB arabinose-inducible promoter. A KO strain for the Eco1 operon (bSLS.114) was constructed from BL21-AI cells using a strategy based on Datsenko and Wanner⁸⁵ to replace the retron operon with an FRT-flanked chloramphenicol resistance cassette. The replacement cassette was amplified from pKD3, adding homology arms to the Eco1 locus. This amplicon was electroporated into BL21-AI cells expressing lambda Red genes from pKD46, and clones were isolated by selection on 10 $\mu\text{g ml}^{-1}$ chloramphenicol plates. After genotyping to

confirm locus-specific insertion, the chloramphenicol cassette was excised by transient expression of FLP recombinase to leave only an FRT scar.

For yeast expression, four sets of plasmids were generated. The first set of plasmids, designed to express the protein components for yeast genome editing, were based off of pZS.157 (ref.⁵⁵), an HIS3 yeast integrating plasmid for galactose-inducible Eco1RT and Cas9 expression (Gal1-10 promoter). A first set of variants of pZS.157, designed to compare the effect of WT versus extended a1/a2 region lengths on genome editing, were generated by PCR and expressed either an empty cassette (pSCL.004), only Cas9 (pSCL.005), only the Eco1RT (pSCL.006) or both (pZS.157). A second set of variants was generated to test single-promoter expression of Cas9–Eco1RT variants. We designed six such plasmids: Eco1RT–linker 1–Cas9 (pSCL.71), Cas9–linker 1–Eco1RT (pSCL.72), Eco1RT–linker 2–Cas9 (pSCL.94), Cas9–linker 2–Eco1RT (pSCL.95), Eco1RT–P2A–Cas9 (pSCL.102) and Cas9–P2A–Eco1RT (pSCL.103). The intervening sequences used were linker 1 (GGTSSGGSGTAGSSGATSGG), linker 2 (SGGSSGGSSGSETPGTSESATPESSGGSSGGSS)⁶⁷ and P2A (ATNFSLLKQAGDVEENPGP)⁸².

The second set of plasmids built for the genome-editing experiments were based off of pZS.165 (ref.⁵⁵), a URA3⁺ centromere plasmid for galactose (Gal7)-inducible expression of a modified Eco1 retron ncRNA, which consists of an Eco1 *msr-ADE2*-targeting gRNA chimera flanked by HH-HDV ribozymes. An initial variant of pZS.165 was generated by cloning an IDT-synthesized gBlock consisting of an Eco1 ncRNA (a1/a2 length: 12 bp), which, when reverse transcribed, encodes a 200-bp *ADE2* repair template to introduce a stop codon (P272X) into the *ADE2* gene (pSCL.002). Two additional

plasmids were generated to extend the a1/a2 region of the Eco1 ncRNA to 27 bp, with variations in the a1/a2 sequence (pSCL.039 and pSCL.040).

The third set of plasmids was built to assess the generalizability of the extended a1/a2 modification. The plasmids carrying WT-length a1/a2 retrans are based off of pSCL.002, where the *ADE2*-targeting gRNA and *ADE2*-editing msd were replaced with analogous sequences to target and insert the following mutations: *Can1* G444X (pSCL.106), *Lyp1* E27X (pSCL.108), *Trp2* E64X (pSCL.110) and *Faa1* P233X (pSCL.112). The plasmids carrying extended-length a1/a2 retrans are based off of pSCL.039 and were generated similar to the WT-length a1/a2 retrans-encoding plasmids (*Can1* G444X (pSCL.107), *Lyp1* E27X (pSCL.109), *Trp2* E64X (pSCL.111) and *Faa1* P233X (pSCL.113)).

The last set of plasmids, designed to compare the levels of RT-DNA production by the different retrans systems, were derived from pSCL.002. IDT-synthesized gBlocks encoding a mammalian codon-optimized Eco1RT and ncRNA (WT), a dead Eco1RT and ncRNA (WT) and a human codon-optimized Eco2RT and ncRNA (WT) were cloned into pSCL.002 by Gibson Assembly, generating pSCL.027, pSCL.031 and pSCL.017, respectively. pSCL.027 was used to generate pSCL.028 by PCR, which carries a mammalian codon-optimized Eco1RT and ncRNA (extended a1/a2: 27 bp). Similarly, pSCL.017 was used to generate pSCL.034 by PCR, which carries a mammalian codon-optimized Eco2RT and ncRNA (extended a1/a2: 29 bp).

All yeast strains were created by LiAc/SS carrier DNA/PEG transformation⁸⁶ of BY4742 (ref.⁸⁷). Strains for evaluating the genome-editing efficiency of various retrans ncRNAs were created by BY4742 integration of plasmids pZS.157, pSCL.004, pSCL.005

or pSCL.006 using KpnI-linearized plasmids for homologous recombination into the *HIS3* locus. Transformants were isolated on SC-HIS plates. To evaluate the effect of the length of the Eco1 ncRNA a1/a2 region on genome-editing efficiency, these parental strains were transformed with episomal plasmids carrying the different retron ncRNA cassettes (pSCL.002, pSCL.039 or pSCL.040), and double transformants were isolated on SC-HIS-URA plates. The result was a set of control strains that lacked one or both components of the genome-editing machinery (that is, Eco1RT and Cas9) and three strains that had all components necessary for retron-mediated genome editing but differed in the length of the Eco1 ncRNA a1/a2 region (12 bp versus 27 bp).

Strains designed to assess the generalizability of the extended a1/a2 modification were created by transformation of a HIS3:pZS.157 yeast strain with plasmids carrying either WT or extended a1/a2 retrons for editing of the four additional loci. Transformants were isolated on SC-HIS-URA plates. Strains to test single-promoter expression of Cas9–Eco1RT variants were created by BY4742 integration of plasmids pSCL.71, pSCL.72, pSCL.94, pSCL.95, pSCL.102 or pSCL.103 using KpnI-linearized plasmids for homologous recombination into the *HIS3* locus. Transformants were isolated on SC-HIS plates. These strains were then transformed with pSCL.39, and transformants were isolated on SC-HIS-URA plates.

Strains designed to compare the levels of RT-DNA production by the different retron constructs were created by transformation of plasmids pSCL.027, pSCL.037 and pSCL.028 for Eco1 (WT, WT dead RT and extended a1/a2, respectively) into BY4742 and pSCL.017 and pSCL.031 for Eco2 (WT and extended a1/a2, respectively) into BY4742. Transformants were isolated by plating on SC-URA agar plates. Expression of proteins

and ncRNAs from all yeast strains was performed in liquid SC-Ura 2% galactose medium for 24 h unless otherwise specified.

For mammalian retron expression and quantification of RT-DNA production, synthesized gBlocks encoding human codon-optimized Eco1 and Eco2 were cloned into a PiggyBac integrating plasmid for doxycycline-inducible human protein expression (TetOn-3G promoter). Eco1 variants were WT retron-Eco1RT and ncRNA (pKDC.018 with an a1/a2 length of 12 bp), extended a1/a2 length ncRNA (pKDC.019 with an a1/a2 length of 27 bp) and a dead Eco1RT control (pKDC.020 with an a1/a2 length of 27 bp). Eco2 variants were WT retron-Eco2RT and ncRNA (pKDC.015 with an a1/a2 length of 13 bp) and extended a1/a2 length ncRNA (pKDC.031 with an a1/a2 length of 29 bp).

Stable mammalian cell lines for assessing RT-DNA production by WT and extended a1/a2 regions were created using the Lipofectamine 3000 transfection protocol (Invitrogen) and a PiggyBac transposase system. T25 flasks of 50–70% confluent HEK293T cells were transfected using 8.3 μ g of retron expression plasmids (pKDC.015, pKDC.018, pKDC.019, pKDC.020 or pKDC.031) and 4.2 μ g PiggyBac transposase plasmid (pCMV-hyPBase). Stable cell lines were selected with puromycin.

For assessment of retron-mediated precise genome editing in mammalian cells, two sets of plasmids were generated. The first set of plasmids, carrying either the SpCas9 gene or the SpCas9-P2A-Eco1RT construct, was built by restriction cloning of the respective genes (PCR amplified off of the aforementioned yeast vectors) into a PiggyBac integrating plasmid for doxycycline-inducible human protein expression (TetOn-3G promoter). The second set of plasmids carried the ncRNA/gRNA targeting one of six loci in the human

genome: *HEK3* (pSCL.175), *RNF2* (pSCL.176), *EMX1* (pSCL.177), *FANCF* (pSCL.178), *HEK4* (pSCL.179) and *AAVS1* (pSCL.180). These were generated by restriction cloning of the ncRNA/gRNA cassette (built by primer assembly¹) into an H1 expression plasmid (FHUGW).

The ncRNA/gRNA cassette was designed as follows. The msd contained a repair template-encoding, 120-bp sequence in its loop. The plasmid-encoded repair template was slightly asymmetric (49 bp of genome site homology upstream of the Cas9 cut site; 71 bp of genome site homology downstream of the cut site) and was complementary to the target strand; in practice, this means that after reverse transcription, the repair template RT-DNA is complementary to the non-target strand, as recommended in previous studies⁸⁸. The repair template carried two distinct mutations: the first introduces a 1-bp single-nucleotide polymorphism (SNP) at the Cas9 cut site, and the second (designed to be at least 2 bp away from the first mutation) recodes the Cas9 PAM (NGG → NHH, where H is any nucleotide beside G). The gRNA is 20 bp.

Stable mammalian cell lines for assessing retron-mediated precise genome editing were created using the Lipofectamine 3000 transfection protocol (Invitrogen) and a PiggyBac transposase system. T25 flasks of 50–70% confluent HEK293T cells were transfected using 8.3 μ g of protein expression plasmids (pSCL.139 and pSCL.140) and 4.2 μ g of PiggyBac transposase plasmid (pCMV-hyPBBase). Stable cell lines were selected with puromycin.

Plasmids and strains are listed in **Supplementary Tables 2-1 and 2-2**. Primers used to generate and verify strains are listed in **Supplementary Table 2-3**. All plasmids will be made available on Addgene at the time of peer-reviewed publication.

qPCR

qPCR analysis of RT-DNA was performed by comparing amplification from samples using two sets of primers. One set could only use the plasmid as a template because they bound outside the msd region (outside), and the other set could use either the plasmid or RT-DNA as a template because they bound inside the msd region (inside). Results were analyzed by first taking the difference in cycle threshold (C_t) between the inside and outside primer sets for each biological replicate. Next, each biological replicate's ΔC_t value was subtracted from the average ΔC_t of the control condition (for example, uninduced). Fold change was calculated as $2^{-\Delta\Delta C_t}$ for each biological replicate. This fold change represents the difference in abundance of the inside versus outside template, where the presence of RT-DNA leads to fold change values of >1 .

For the initial analysis of Eco1 RT-DNA when overexpressed in *E. coli*, the qPCR analysis used just three primers, two of which bound inside the msd and one which bound outside. The inside PCR was generated using both inside primers, while the outside PCR used one inside and one outside primer. For all other experiments, four primers were used. Two bound inside the msd and two bound outside the msd in the RT. qPCR primers are all listed in **Supplementary Table 2-3**.

For bacterial experiments, constructs were expressed in liquid culture maintained with shaking at 37 °C for 6–16 h, after which a volume of 25 μ l of culture was collected, mixed with 25 μ l of water and incubated at 95 °C for 5 min. A volume of 0.3 μ l of this boiled culture was used as a template in 30- μ l reactions using KAPA SYBR FAST qPCR mix.

For yeast experiments, single colonies were inoculated into SC-URA 2% glucose and grown with shaking overnight at 30 °C. To express the constructs, the overnight cultures were centrifuged, washed and resuspended in 1 ml of water, passaged at a 1:30 dilution into SC-URA 2% galactose and grown with shaking for 24 h at 30 °C. Aliquots (250 μ l) of the uninduced and induced cultures were collected for qPCR analysis. For qPCR sample preparation, the aliquots were centrifuged, resuspended in 50 μ l of water and incubated at 100 °C for 15 min. The samples were then briefly centrifuged and placed on ice to cool, and 50 μ l of the supernatant was treated with Proteinase K by combining with 29 μ l of water, 9 μ l of CutSmart buffer and 2 μ l of Proteinase K (New England Biolabs) followed by incubation at 56 °C for 30 min. The Proteinase K was inactivated by incubation at 95 °C for 10 min, followed by a 1.5-min centrifugation at maximum speed (~21,000g). The supernatant was collected and used as a template for qPCR reactions consisting of 2.5 μ l of template in 10- μ l KAPA SYBR FAST qPCR reactions.

For mammalian experiments, retron expression in stable HEK293T cell lines was induced using 1 μ g ml⁻¹ doxycycline for 24 h at 37 °C in six-well plates. Aliquots (1 ml) of induced and uninduced cell lines were collected for qPCR analysis. qPCR sample preparation and reaction mix followed the yeast experimental protocol.

RT-DNA purification and PAGE analysis

To analyze RT-DNA on a PAGE gel after expression in *E. coli*, 2 ml of culture was pelleted, and nucleotides were prepared using a Qiagen mini prep protocol, substituting Epoch mini spin columns and buffers MX2 and MX3 for Qiagen components. Purified DNA was then treated with additional RNase A/T1 mix (New England Biolabs) for 30 min

at 37 °C, and single-stranded DNA was isolated from the preparation using an ssDNA/RNA Clean & Concentrator kit from Zymo Research. The purified RT-DNA was then analyzed on 10% Novex TBE-Urea gels (Invitrogen) with 1× TBE running buffer that was heated to >80 °C before loading. Gels were stained with SYBR Gold (Thermo Fisher) and imaged on a Gel Doc imager (Bio-Rad).

To analyze RT-DNA on a PAGE gel after expression in *S. cerevisiae*, 5 ml of overnight culture in SC-URA 2% galactose was pelleted, and RT-DNA was isolated by RNase A/T1 treatment of the aqueous (RNA) phase after TRIzol extraction (Invitrogen), following the manufacturer's recommendations with few modifications, as noted here. Cell pellets were resuspended in 500 μ l of RNA lysis buffer (100 mM EDTA pH 8, 50 mM Tris-HCl pH 8, 2% SDS) and incubated for 20 min at 85 °C before the addition of the TRIzol reagent. The aqueous phase was chloroform extracted twice. Following isopropanol precipitation, the RNA + RT-DNA pellet was resuspended in 265 μ l of TE and treated with 5 μ l of RNase A/T1 + 30 μ l of NEB2 buffer. The mixture was incubated for 25 min at 37 °C, after which the RT-DNA was reprecipitated by addition of equal volumes of isopropanol. The resulting RT-DNA was analyzed on Novex 10% TBE-Urea gels as described above.

Variant library cloning

Eco1 ncRNA variant parts were synthesized by Agilent. Variant parts were flanked by Bsal type IIS cut sites and specific primers that allowed for amplification of the sublibraries from a larger synthesis run. Random nucleotides were appended to the 3' end of synthesized parts so that all sequences were the same length (150 bp). The vector to accept these parts (pSLS.601) was amplified with primers that also added Bsal sites

so that the ncRNA variant amplicons and amplified vector backbone could be combined into a Golden Gate reaction using BsaI-HFv2 and T4 ligase to generate a pool of variant plasmids at high efficiency when electroporated into a cloning strain. Variant libraries were minipreped from the cloning strain and electroporated into the expression strain. Primers for library construction are listed in **Supplementary Table 2-3**. Variant parts are listed in **Supplementary Data 2-2**.

Variant library expression and analysis

Eco1 ncRNA variant libraries were grown overnight and then diluted 1:500 for expression. A sample of the culture preexpression was taken to quantify the variant plasmid library, mixed 1:1 with water, incubated at 95 °C for 5 min and frozen at -20 °C. Constructs were expressed (arabinose and IPTG for the ncRNA, erythromycin for the RT) as the cells grew with shaking at 37 °C for 5 h, after which two samples were collected. One was collected to quantify the variant plasmid library. That sample was mixed 1:1 with water, incubated at 95 °C for 5 min and frozen at -20 °C, identical to the preexpression sample. The other sample was collected to sequence the RT-DNA. That sample was prepared as described above for RT-DNA purification.

The two variant plasmid library samples (boiled cultures) taken before and after expression were amplified by PCR using primers flanking the ncRNA region that also contained adapters for Illumina sequencing preparation. The purified RT-DNA was prepared for sequencing by first treating with DBR1 (OriGene) to remove the branched RNA and then extending the 3' end with a single nucleotide, dCTP, in a reaction with TdT. This reaction was performed in the absence of cobalt for 120 s at room temperature with

the aim of adding only five to ten cytosines before inactivating the TdT at 70 °C. A second complementary strand was then created from that extended product using Klenow Fragment (3' → 5' exo-) with a primer containing an Illumina adapter sequence, six guanines and a non-guanine (H) anchor. Finally, Illumina adapters were ligated on at the 3' end of the complementary strand using T4 ligase. In one variation, the loop of the RT-DNA for the a1/a2 library was amplified using Illumina adapter-containing primers in the RT-DNA but outside the variable region from the purified RT-DNA directly. All products were indexed and sequenced on an Illumina MiSeq. Primers used for sequencing are listed in **Supplementary Table 2-3**.

Python software was custom written to extract variant counts from each plasmid and RT-DNA sample. In each case, these counts were then converted to a percentage of each library or relative abundance (for example, raw count for a variant over total counts for all variants). The relative abundance of a given variant in the RT-DNA sample was then divided by the relative abundance of that same variant in the plasmid library using the average of the pre- and postinduction values to control for differences in the abundance of each variant plasmid in the expression strain. Finally, these corrected abundance values were normalized to the average corrected abundance of the WT variant (set to 100%) or the loop length of five (set to 100%).

Recombineering expression and analysis

In experiments using the retron ncRNA to edit bacterial genomes, the retron cassette was coexpressed with CspRecT and mutL E32K from the plasmid pORTMAGE-Ec1 (ref.¹) for 16 h with shaking at 37 °C. After expression, a volume of 25 μ l of culture

was collected, mixed with 25 μ l of water and incubated at 95 °C for 5 min. A volume of 0.3 μ l of this boiled culture was used as a template in 30- μ l reactions with primers flanking the edit site, which additionally contained adapters for Illumina sequencing preparation. These amplicons were indexed and sequenced on an Illumina MiSeq instrument and processed with custom Python software to quantify the percentage of precisely edited genomes.

Yeast editing expression and analysis

For yeast genome-editing experiments, single colonies from strains containing variants of the Eco1 ncRNA–gRNA cassette (WT or extended a1/a2 length for WT versus extended a1/a2 region experiments; extended a1/a2 length v1 to test single-promoter expression of Cas9–Eco1RT variants) and editing machinery (–/+ Cas9, –/+ Eco1RT for WT versus extended a1/a2 region experiments; Eco1RT–linker 1–Cas9, Cas9–linker 1–Eco1RT, Eco1RT–linker 2–Cas9, Cas9–linker 2–Eco1RT, Eco1RT–P2A–Cas9, Cas9–P2A–Eco1RT to test single-promoter expression of Cas9–Eco1RT variants) were grown in SC-HIS-URA 2% raffinose for 24 h with shaking at 30 °C. Cultures were passaged twice into SC-URA 2% galactose (1:30 dilutions) for 24 h for a total of 48 h of editing. At each timepoint (after 24 h of raffinose, 24 h of galactose, 48 h of galactose), an aliquot of the cultures was collected, diluted and plated on SC-URA low-ADE plates. Plates were incubated at 30 °C for 2–3 d until visible and countable pink (*ADE2* KO) and white (*ADE2* WT) colonies grew. Editing efficiency was calculated in two ways. The first was by calculating the ratio of pink colonies to total colonies on each plate for each timepoint. This counting was performed by an experimenter blinded to the condition. The second

was by deep sequencing of the target *ADE2* locus. For this, we collected cells from 250- μ l aliquots of the culture for each timepoint in PCR strips and performed a genomic preparation as follows. The pellets were resuspended in 120 μ l of lysis buffer (see above), heated at 100 °C for 15 min and cooled on ice. Protein precipitation buffer (60 μ l; 7.5 M ammonium acetate) was added, and the samples were gently inverted and placed at -20 °C for 10 min. The samples were then centrifuged at maximum speed for 2 min, and the supernatant was collected in new Eppendorf tubes. Nucleic acids were precipitated by adding equal parts ice-cold isopropanol and incubating the samples at -20 °C for 10 min followed by pelleting by centrifugation at maximum speed for 2 min. The pellets were washed twice with 200 μ l of ice-cold 70% ethanol and dissolved in 40 μ l of water. gDNA (0.5 μ l) was used as template in 10- μ l reactions with primers flanking the edit site in *ADE2*, which additionally contained adapters for Illumina sequencing preparation (see **Supplementary Table 2-4** for oligonucleotide sequences). Importantly, the primers do not bind to the ncRNA/gRNA plasmids. These amplicons were indexed and sequenced on an Illumina MiSeq instrument and processed with custom Python software to quantify the percentage of P272X edits caused by Cas9 cleavage of the target site on the *ADE2* locus and repair using the Eco1 ncRNA-derived RT-DNA template.

The editing experiments at additional loci were performed as described above, with the difference that editing was quantified by amplifying 0.5 μ l of the gDNA with locus-specific primers, adapters for Illumina sequencing preparation. These primers are listed in **Supplementary Table 2-3**. Custom Python software was used to quantify the percentage of precise edits caused by Cas9 cleavage of the target site on the *ADE2* locus and repair using the Eco1 ncRNA-derived RT-DNA template.

Human editing expression and analysis

For human genome-editing experiments, Cas9 or Cas9–P2A–Eco1RT expression in stable HEK293T cell lines was induced using $1 \mu\text{g ml}^{-1}$ doxycycline for 24 h at 37°C in T12.5 flasks. Cultures were transiently transfected with a plasmid constitutively expressing ncRNA/gRNA at a concentration of $5 \mu\text{g}$ of plasmid per T12.5 using Lipofectamine 3000 (see plasmid list described above and **Supplementary Table 2-1**). Cultures were passaged, and doxycycline was refreshed the following day for an additional 48 h. Three days after transfection, cells were collected for sequencing analysis.

To prepare samples for sequencing, cell pellets were processed, and gDNA was extracted using a QIAamp DNA mini kit according to the manufacturer's instructions. DNA was eluted in $200 \mu\text{l}$ of ultra-pure, nuclease-free water. Then, $0.5 \mu\text{l}$ of the gDNA was used as template in $12.5\text{-}\mu\text{l}$ PCR reactions with primer pairs to amplify the locus of interest, which also contained adapters for Illumina sequencing preparation (see **Supplementary Table 2-4** for oligonucleotide sequences). Importantly, the primers do not bind to the ncRNA/gRNA plasmids. The amplicons were purified using a QIAquick PCR purification kit according to the manufacturer's instructions, and the amplicons were eluted in $12 \mu\text{l}$ of ultra-pure, nuclease-free water. Lastly, the amplicons were indexed and sequenced on an Illumina MiSeq instrument and processed with custom Python software to quantify the percentage of on-target precise and imprecise genomic edits.

Reporting Summary

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

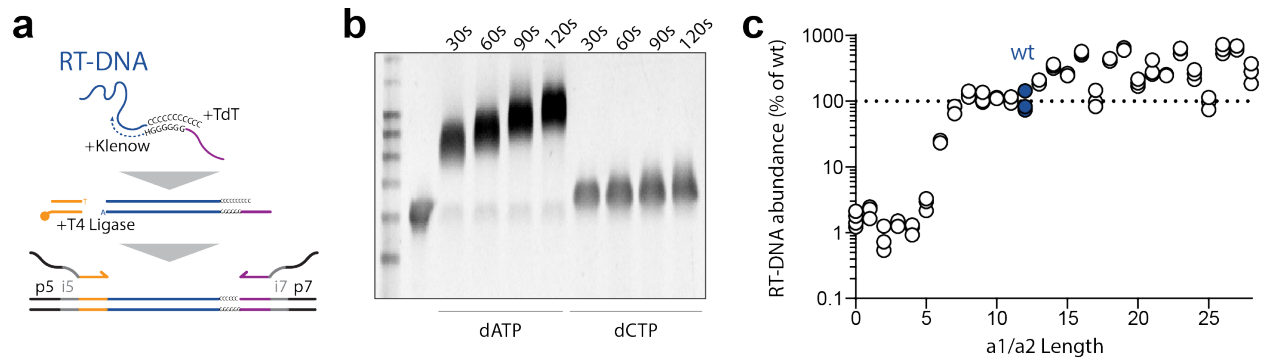
Data availability

All data supporting the findings of this study are available within the article and its **Supplementary Information**. Sequencing data associated with this study are available through the NCBI BioProject database under accession number PRJNA770365. Source data are provided with this paper.

Code availability

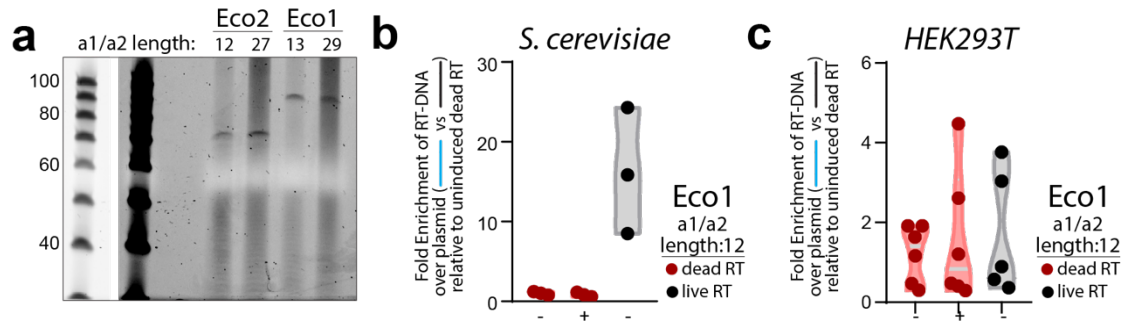
Custom code to process or analyze data from this study is available on GitHub at https://github.com/Shipman-Lab/retron_architectures.

2.6 Supplementary Information



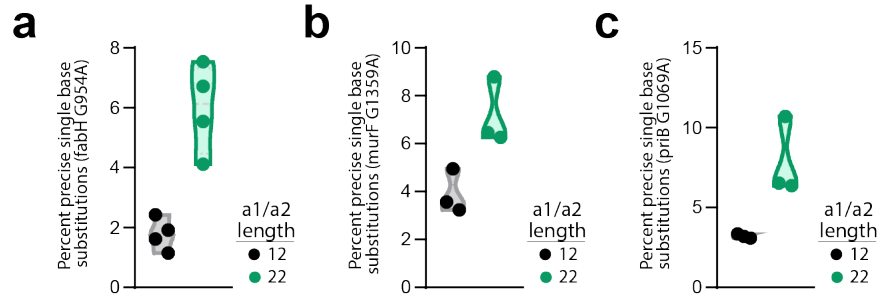
Extended Data Fig 2-1: RT-DNA sequencing prep

a. Schematic of the sequencing prep pipeline for RT-DNA. **b.** Representative image of a PAGE analysis showing the addition of nucleotides to the 3' end of a single-stranded DNA, controlled by reaction time. The experiment was repeated twice with similar results. **c.** Alternate analysis of the RT-DNA for the a1/a2 length library, using a TdT-based sequencing preparation. **Related to Fig 2-2.**



Extended Data Fig 2-2: RT-DNA production in eukaryotic cells

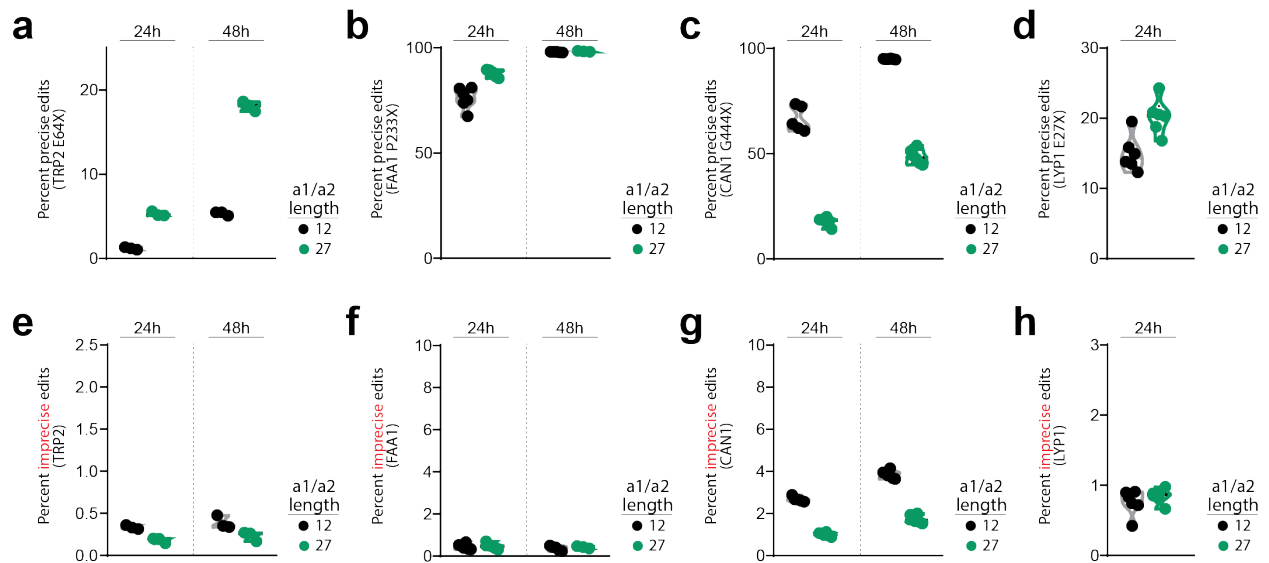
a. Representative image of a PAGE analysis of Eco1 and Eco2 RT-DNA isolated from yeast. The ladder is shown at a different exposure to the left of the gel image. The experiment was repeated twice with similar results. **b.** Enrichment of the Eco1 RT-DNA/plasmid template when uninduced compared to a dead RT construct. Closed circles show each of three biological replicates, with red for the dead RT version and black for the live RT. **c.** Identical analysis as in **b**, but for Eco1 in HEK293T cells. **Related to Fig 2-3.**



Extended Data Fig 2-3: Precise genome editing rates across additional genomic loci in *E. coli*
a-c. Percent of cells precisely edited, quantified by multiplexed sequencing, for the wt (black) and extended (green) recombineering constructs for three additional loci in *E. coli*. **Related to Fig 2-4a-d.**

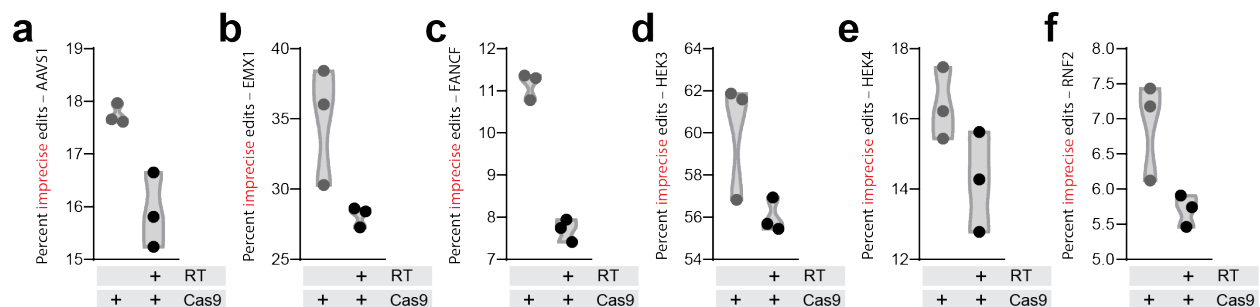
(Figure caption continued from the previous page)

a. Percent of ADE2 loci with imprecise edits or sequencing errors at 24 and 48 hours. Closed circles show each of three biological replicates, with black for the wt a1/a2 length and green for the extended a1/a2 (two extended versions, v1 and v2). Induction conditions are shown below the graph for the RT and Cas9. **b.** Breakdown of the data in a. by type of edit/error. **c.** Imprecise edits and sequencing errors found in all data sets, ranked by frequency. Above the graph are the wt ADE2 locus and intended precise edit. On the Y axis are the imprecise edits and sequencing errors found. X axis represents count of each sequence in all data sets. **Related to Fig 2-4h.**



Extended Data Fig 2-5: Genome editing rates across additional genomic loci in yeast

a-d. Percent of cells precisely edited, quantified by multiplexed sequencing, for the wt (black) and extended (green) recombineering constructs for four additional loci in *S. cerevisiae* at 24 and 48 hours. Cultures edited at the LYP1 E27X site were not viable beyond 24 hours. **e-h.** Percent of imprecise edits or sequencing errors for the loci in a-d. **Related to Fig 2-4e-h.**



Extended Data Fig 2-6: Imprecise editing rates across genomic loci in human cells

a-f. Percent of cells imprecisely edited (indels), quantified by multiplexed sequencing, in the presence of the ncRNA/gRNA plasmid and either Cas9 alone or Cas9 and Eco1 RT (as indicated below). Individual circles represent each of three biological replicates. **Related to Fig 2-5.**

2.7 Supplemental Files

Supplementary_Information_Chapter2.pdf

This PDF file contains:

- Supplementary Table 2-1: Plasmids used in this study
- Supplementary Table 2-2: Strains used in this study
- Supplementary Table 2-3: Primers used in this study
- Supplementary Table 2-4: Per-figure statistics

Supplementary_Dataset_Chapter2.xlsx

This excel file contains the Eco1 ncRNA variant library parts.

Chapter 3 Continuous multiplexed phage genome editing using recombitrans

3.1 Abstract

Bacteriophage genome editing can enhance the efficacy of phages to eliminate pathogenic bacteria in patients and in the environment. However, current methods for editing phage genomes require laborious screening, counterselection or in vitro construction of modified genomes. Here, we present a scalable approach that uses modified bacterial retrons called recombitrans to generate recombineering donor DNA paired with single-stranded binding and annealing proteins for integration into phage genomes. This system can efficiently create genome modifications in multiple phages without the need for counterselection. The approach also supports larger insertions and deletions, which can be combined with simultaneous counterselection for >99% efficiency. Moreover, we show that the process is continuous, with more edits accumulating the longer the phage is cultured with the host, and multiplexable. We install up to five distinct mutations on a single lambda phage genome without counterselection in only a few hours of hands-on time and identify a residue-level epistatic interaction in the T7 gp17 tail fiber.

3.2 Introduction

Bacteriophages (phages) naturally influence the composition of microbial ecosystems through selective infection of bacterial species. Humans have long sought to harness this power of phages to make targeted interventions to the microbial world, such as delivering phages to a patient suffering from an infection to eliminate a bacterial pathogen. This approach to mitigate pathogenic bacterial infections, known as phage therapy, predates the discovery of penicillin, with 100 years of evidence for efficacy and safety¹. However, the success of small-molecule antibiotics over the same period of time has overshadowed and blunted innovation in phage therapy.

Unfortunately, it is now clear that reliance on small-molecule antibiotics is not a permanent solution. Antimicrobial resistance in bacteria was associated with 1–5 million deaths worldwide in 2019¹, a figure that is projected to rise in the coming decades¹. As such, there is a pressing need for alternatives or adjuvants to small-molecule antibiotics, such as phage therapy, to avoid returning to the rampant morbidity of bacterial infections of the preantibiotic era. This work has already begun, with researchers and clinicians using phage-screening pipelines to identify natural phages for use in patients to overcome antimicrobial-resistant infections^{1,2}.

While these efforts demonstrate the potential of phage therapeutics, they do not scale well. Screening natural phages for each patient is time and effort intensive, requires massive repositories of natural phages and results in the clinical use of biological materials that are not fully characterized¹. To functionally replace or supplement small-molecule antibiotics, phage therapy needs to be capable of industrialization and more rapid iteration. This will likely include modifying known phages to create engineered

therapeutic tools that target specific pathogens and evade natural bacterial immunity, rather than just the opportunistic isolation of natural phages. There are a few recent examples of therapies using engineered phages^{1,2}. However, such approaches are limited by the relative difficulty in modifying phage genomes¹.

The various approaches to modify phage genomes and the limitations imposed by each were recently reviewed¹. One approach is to modify phage genomes by recombination within their bacterial host, which is inefficient and requires laborious screening of phage plaques. That screening effort can be reduced by imposing a counterselection on the unedited phage, such as clustered regularly interspaced short palindromic repeats (CRISPR)-based depletion of the wild-type phage¹⁻⁴. However, this negative selection is not universally applicable to all edit types because it requires functional disruption of a protospacer sequence¹. Phages also frequently escape CRISPR targeting¹, which can result in most selected phages containing escape mutations outside of the intended edit¹. Another approach is to ‘reboot’ a phage by assembling a modified phage genome in vitro and repackaging it in a host¹. Such rebooting can require extensive work to enable in vitro assembly of a phage genome, although this is an area where technical advances are emerging¹. However, phage genome size limits the efficiency of transformation for rebooting in natural hosts¹. A fully cell-free packaging system eliminates the issue of inefficient transformation but instead requires substantial upfront technical development for each additional host¹.

Because of the urgent need for innovation in phage therapeutics and the clear technical hurdle in modifying phage genomes, we developed an alternative system for phage engineering. Our system edits phage genomes as they replicate within their

bacterial hosts by integrating a single-stranded DNA (ssDNA) donor encoding the edit into the replicating phage genome using a single-stranded annealing protein (SSAP) and single-stranded binding protein (SSB), a process known as recombineering. We use a modified bacterial retron¹⁻⁴ to continuously produce the editing ssDNA donor by reverse transcription within the host. Thus, propagating a population of phage through this host strain enables continuous accumulation of the intended edit over generations within a single culture.

Moreover, this system enables a more complex form of editing in which the bacterial culture is composed of multiple, distinct editing hosts, each producing donors that edit different parts of the phage genome. Propagation of phages through such a complex culture leads to the accumulation of multiple distinct edits at distant locations in individual phage genomes. Here, we demonstrate this approach, showing that the editing is a continuous process in which edits accumulate over time; moreover, it can be applied to multiple types of phage and used to introduce different edit types, it can be optimized to reach efficiencies that do not require counterselection and it can be used to make multiplexed edits across a phage genome and enables rapid biological investigation of formerly resource-intensive and time-intensive problems. For disambiguation with other techniques, we call this approach 'phage retron recombineering' and term the molecular components that include a modified retron a 'recombitron'.

3.3 Results

Recombitrons target phage genomes for editing

There are four core molecular components of a recombitron: a retron noncoding RNA (ncRNA) that is modified to encode an editing donor, a retron reverse transcriptase (RT), an SSB and an SSAP (**Fig 3-1a**). Endogenous retrons partially reverse-transcribe a short (~200 nt) ncRNA into a single-stranded reverse-transcribed DNA (RT-DNA) of ~90 nt. In bacteria, this RT-DNA is used in conjunction with retron accessory proteins to detect and respond to phage infection. The retron accessory proteins are necessary for the phage defense phenotype and are not included in the recombitron to avoid reconstituting an antiphage system¹⁻⁴. We modified the retron ncRNA by adding nucleotides to the reverse-transcribed region that are homologous to a locus in the phage genome and carry the edit we aim to incorporate.

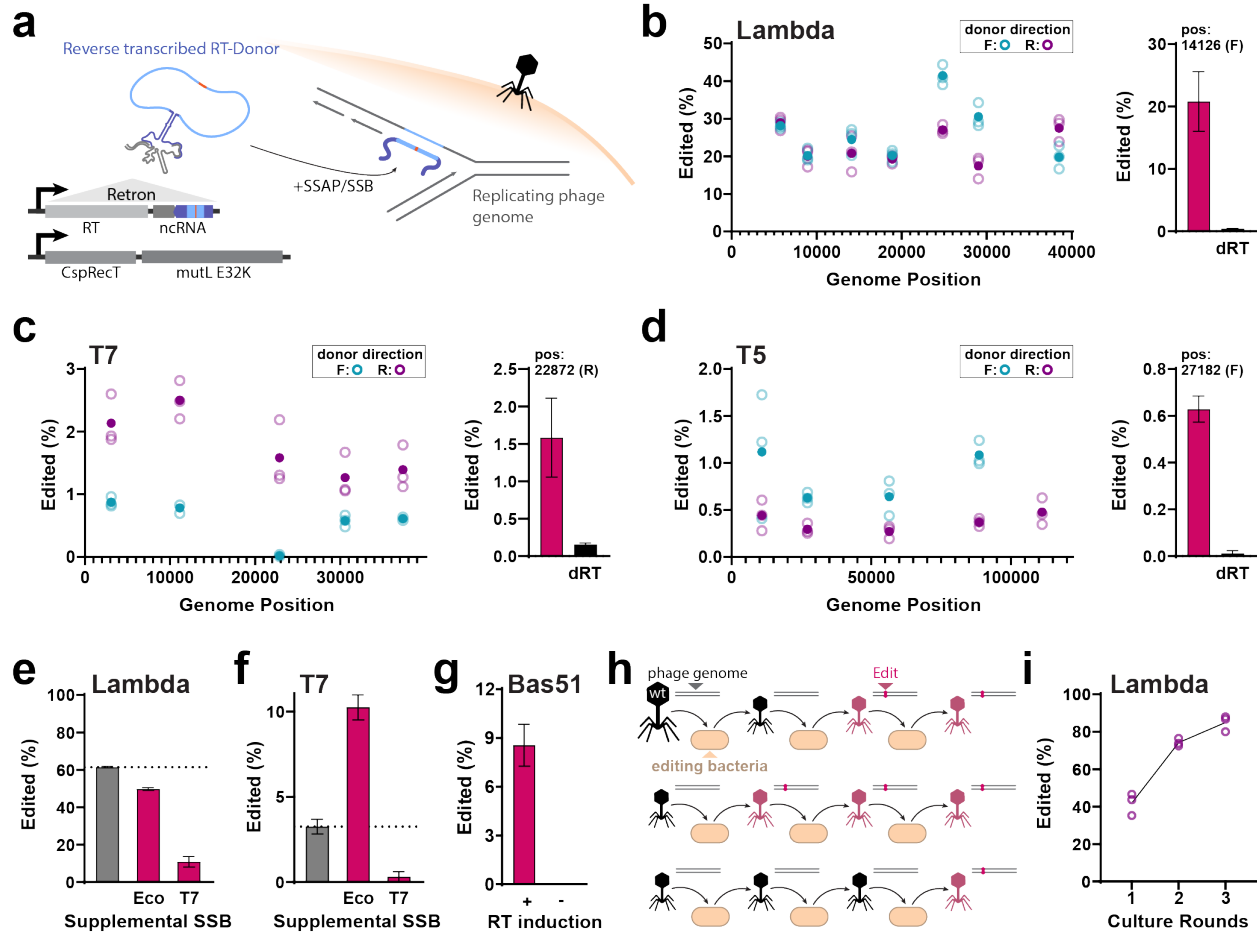


Figure 3-1: Recombitrons target phage genomes for continuous editing.

a. A modified retron generates an ssDNA donor that is integrated during replication by SSB and SSAP. The reverse-transcribed region of the ncRNA is shown in blue, the donor sequence is shown in light blue and the edit site is shown in orange. A separate plasmid expresses CspRecT and mutL E32K. **b.** Left, edited phage genomes (as a percentage of all genomes) across lambda phage. Editing with a forward RT-DNA is shown in blue, while editing with a reverse RT-DNA is shown in purple. Three biological replicates are shown as open circles; closed circles are the means. Right, recombiron editing at site 14,126 is significantly greater than a dRT control (two-sided *t*-test, $P = 0.0018$). **c.** Left, edited phage T7 genomes for three biological replicates at each position, displayed as in **b.** Right, recombiron editing at site 22,872 is significantly greater than a dRT control (two-sided *t*-test, $P = 0.0094$). **d.** Left, edited phage T5 genomes for three replicates at each position, as in **b.** Right, recombiron editing at site 27,182 is significantly greater than a dRT control (two-sided *t*-test, $P < 0.0001$). **e.** Editing of lambda at site 30,840 (F) compared to editing with supplemental expression of *E. coli* SSB or T7 SSB. Three biological replicates are shown as open circles; closed circles are the means. The effect of SSB expression is significant (one-way ANOVA, $P < 0.0001$, $n = 3$), with both *E. coli* ($P = 0.005$) and T7 ($P \leq 0.0001$) different from the no-SSB condition (Dunnnett's, corrected). **f.** Editing of T7 at site 11,160 (R) compared to editing with supplemental expression of *E. coli* SSB or T7 SSB. Three biological replicates are shown as open circles; closed circles are the means. The effect of SSB expression is significant (one-way ANOVA, $P < 0.0001$, $n = 3$), with both *E. coli* ($P = 0.0002$) and T7 ($P = 0.0127$) different from the no-SSB condition (Dunnnett's, corrected). **g.** Editing of phage Bas51 (T28791A) showing induced RT versus uninduced RT (two-sided *t*-test, $P = 0.0003$). Three biological replicates are shown as open circles; closed circles are the means. **h.** Schematic illustrating the accumulation of edited phages with multiple rounds of editing. **i.** The proportion of edited lambda phage increases over three rounds of editing at site 30,840 (F). Three replicates per round are shown in open circles. Additional statistical details are provided in **Supplementary Table 3-1**.

The recombitron we began with specifically contains a modified retron-Eco1 ncRNA expressed on the same transcript as a retron-Eco1 RT, which produces a 90-nt-long reverse-transcribed editing donor (RT-Donor) nested inside the natural RT-DNA sequence. Once produced, the SSB binds the editing RT-Donor to destabilize internal helices and promote interaction with an SSAP. In this initial recombitron, we leveraged the endogenous *Escherichia coli* SSB. Next, an SSAP promotes annealing of RT-Donor to the lagging strand of a replication fork, where the sequence is incorporated into the newly replicated genome¹. From a separate promoter, we expressed the SSAP CspRecT along with an optional recombitron element mutL E32K, a dominant-negative version of *E. coli* mutL that suppresses mismatch repair¹⁻³. Such suppression helps when creating single-base mutations but is not required for larger insertions or deletions.

This system benefits from stacking several previously discovered beneficial modifications to the retron and recombineering machinery. First, the retron ncRNA, in addition to being modified to encode an RT-Donor, was also modified to extend the length of its a1/a2 region, which we previously found to increase the amount of RT-DNA produced^{1,2}. Second, the SSAP, CspRecT, is more efficient than the previous standard, lambda β , and is known to be compatible with *E. coli* SSB¹. Third, the dominant-negative mutL E32K eliminates the requirement to pre-engineer the host strain to remove mutS^{1,2}. A previous study attempted a single edit using homologous recombination with a retron-produced donor for T5 phage but only reported editing after counterselection¹. By using these stacked innovations, we aimed to produce a system that does not require counterselection.

To edit phages, we designed recombitrons targeting 5–7 sites across the genome of four *E. coli* phages: lambda, T7, T5 and T2. Each recombitron was designed to make synonymous single-base substitutions to a stop codon, which we assumed to be fitness neutral, although this assumption was not directly tested. We constructed separate recombitrons to produce the RT-Donor in either of the two possible orientations, given that the mechanism of retron recombineering requires integration into the lagging strand¹, and created a recombitron with a catalytically dead RT at one position per phage as a control. We pre-expressed the recombitron for 2 h in BL21-AI *E. coli* that lacks the endogenous retron-Eco1, then added phage at a multiplicity of infection (MOI) of 0.1 and grew the cultures for 16 h overnight. The next day, we pelleted the cultures to collect phage in the supernatant. We used PCR to amplify the regions of interest in the phage genome (~300 nt) surrounding the edit sites. We sequenced these amplicons on an Illumina MiSeq and quantified editing with custom software.

We observed RT-dependent, counterselection-free editing in lambda (~20%) (**Fig 3-1b**), T7 (~1.5%) (**Fig 3-1c**) and T5 (~0.6%) (**Fig 3-1d**). We did not observe editing in T2 above the dead RT background level (**Extended Data Fig 3-1a**). This could be because of the substitution of cytosines with modified 5-hydroxymethylcytosines in T2 DNA¹. In support of that hypothesis, we found that T4, which similarly contains modified cytosines¹, was also poorly edited relative to the other phages tested, and a T4 variant (T4 GT7) that cannot modify its cytosines¹ was edited at a higher rate than the wild-type version (**Extended Data Fig 3-1b**). We interpret this result cautiously as the unmodified variant T4 has severely impaired fitness and we do not know how that could affect editing¹.

In lambda, T7 and T5, we obtained similar maximum editing efficiencies at each site across the genome. Editing efficiency was affected by RT-Donor direction in a manner generally consistent with individual phage replication mechanisms. Lambda has an early bidirectional phase of replication and a later unidirectional phase¹. Thus, either strand may be lagging at some point and, accordingly, we observed similar efficiencies for the forward and reverse recombitron (**Fig 3-1b**). T7 has a single origin of replication at one end of its linear genome and we observed directional editing favoring the predicted lagging (reverse) strand (**Fig 3-1c**)¹. Interestingly, T5 replication is less well studied, with one report describing bidirectional replication from multiple origins¹. However, we found a clear strand preference indicating dominant unidirectional replication from one end of the phage genome (**Fig 3-1d**).

Editing efficiency differed among the phages tested, with lambda being the most efficiently edited. One possible explanation for this difference is that lambda is the only temperate phage tested; thus, hypothetically, the editing could have occurred while lambda was integrated into the *E. coli* genome. However, we obtained two results inconsistent with this hypothesis. First, we generated a lambda strain with an inactivated *cl* gene (Δcl) that is required for prophage maintenance using a recombitron to introduce two premature stop codons and found similar levels of editing in that lambda strain as compared with our standard lambda strain (**Extended Data Fig 3-1c**)¹. Next, we modified the bacterial host to remove the lambda prophage integration site (*attB*) and again found similarly high levels of editing of the Δcl strain in the $\Delta attB$ bacteria (**Extended Data Fig 3-1c**)¹. Perhaps the recombitron differentially affects phage replication, which in turn affects editing efficiency. However, we found no effect on

replication of any phage (as measured by phage titer) with recombitron expression (**Extended Data Fig 3-1e-f**). Another possibility is that the recombination system carried by lambda, including *beta*, an SSAP encoded by lambda, assists the recombitron¹. We found no editing of lambda in the absence of CspRecT expression, which is inconsistent with that possibility (**Extended Data Fig 3-1d**). We further tested overexpression of lambda genes *gam* (a RecBCD nuclease inhibitor) and *beta* in addition to the full recombitron system. We found a very slight increase in editing T7 while expressing lambda *beta* but no increase in lambda editing and no effect of *gam* on either phage (**Extended Data Fig 3-1g,h**). However, the increase in T7 editing with *beta* expression was minor relative to the difference in editing rates between the phages.

Another alternative explanation is SSB compatibility¹. Phage T7, unlike lambda, encodes a separate SSB in its genome (gp2.5)¹. Perhaps T7 SSB competes with the endogenous *E. coli* SSB for the recombitron RT-DNA, which could inhibit interactions with CspRecT. To test this possibility, we repeated the lambda and T7 editing experiments at one locus each, while overexpressing either the *E. coli* or the T7 SSB. Consistent with this explanation, we found that overexpression of the T7 SSB had a large negative impact on editing of lambda, while the *E. coli* SSB had a much smaller negative impact (**Fig 3-1e**). We similarly found that overexpression of the T7 SSB strongly reduced editing of T7, whereas overexpression of the *E. coli* SSB had a large positive effect on T7 editing, more than doubling the efficiency compared to a condition without *E. coli* SSB overexpression, to a rate of 10% (**Fig 3-1f**). Thus, the T7 SSB appears to inhibit the retron recombineering approach but can be counteracted by overexpressing a compatible SSB. Lambda *gam* or *beta* added no additional benefit to T7 or lambda editing in the presence

of *E. coli* SSB (**Extended Data Fig 3-1i,j**). We tried a similar SSB approach with T5, which has a much less characterized SSB (PC4-like, by homology to phage T4)¹. While the T7 SSB similarly inhibited T5 editing, the T5 PC4-like protein had a small negative effect on editing and the *E. coli* SSB did not improve T5 editing (**Extended Data Fig 3-1k-m**).

Another approach to edit phage genomes involves in vitro assembly and rebooting by electroporation of the modified genome into a bacterial host. However, rebooting can suffer from limitations with phages that have large genomes because of the size dependence of electroporation. We found that we could efficiently edit Bas51, an *E. coli* phage with a 140,659-bp genome (**Fig 3-1g**)¹. Recombitrons, therefore, present a particularly attractive approach for modifying phages with large genomes. In contrast, two related large phages with modified nucleotides were poorly edited (Bas46 and Bas47)¹, consistent with results in other modified phages T2 and T4 (**Extended Data Fig 3-1n**).

Continuous editing

One clear advantage of using recombitorons for phage editing is the fact that they edit continuously while the phage is propagating through the culture, increasing the proportion of edited phages with every generation (**Fig 3-1h,i**). This is quite distinct from other methods of recombineering in phages where editing donors are delivered by transformation of the host at a single point in time¹. To illustrate the continuous nature of this approach, we edited lambda over three rounds of overnight culture. Editing and analysis were performed as described above; however, at the end of each round of editing, we propagated the resulting phage population through a fresh culture of the

editing host. The percentage of edited genomes increased with each additional round, reaching >80% of phages edited after three rounds (**Fig 3-1i**).

Optimizing recombitron parameters

Given the initial success of recombitrons, we next tested the parameters of the system to achieve optimal editing. The first parameter we tested was length of the RT-Donor. We designed a set of recombitrons with different RT-Donor lengths, each encoding a lambda edit (C14070T) in the center of the homologous donor (**Fig 3-2a**). We found no editing with a 30-nt RT-Donor but all other lengths tested from 50 to 150 nt produced successful editing (**Fig 3-2b**). The highest overall editing rates occurred with a 70-nt RT-Donor. However, for donors between 50 and 150 nt, an analysis corrected for multiple comparisons only found a significant difference between 70-nt and 150-nt donors, indicating that long RT-Donor length is detrimental to editing efficiency (**Fig 3-2b**).

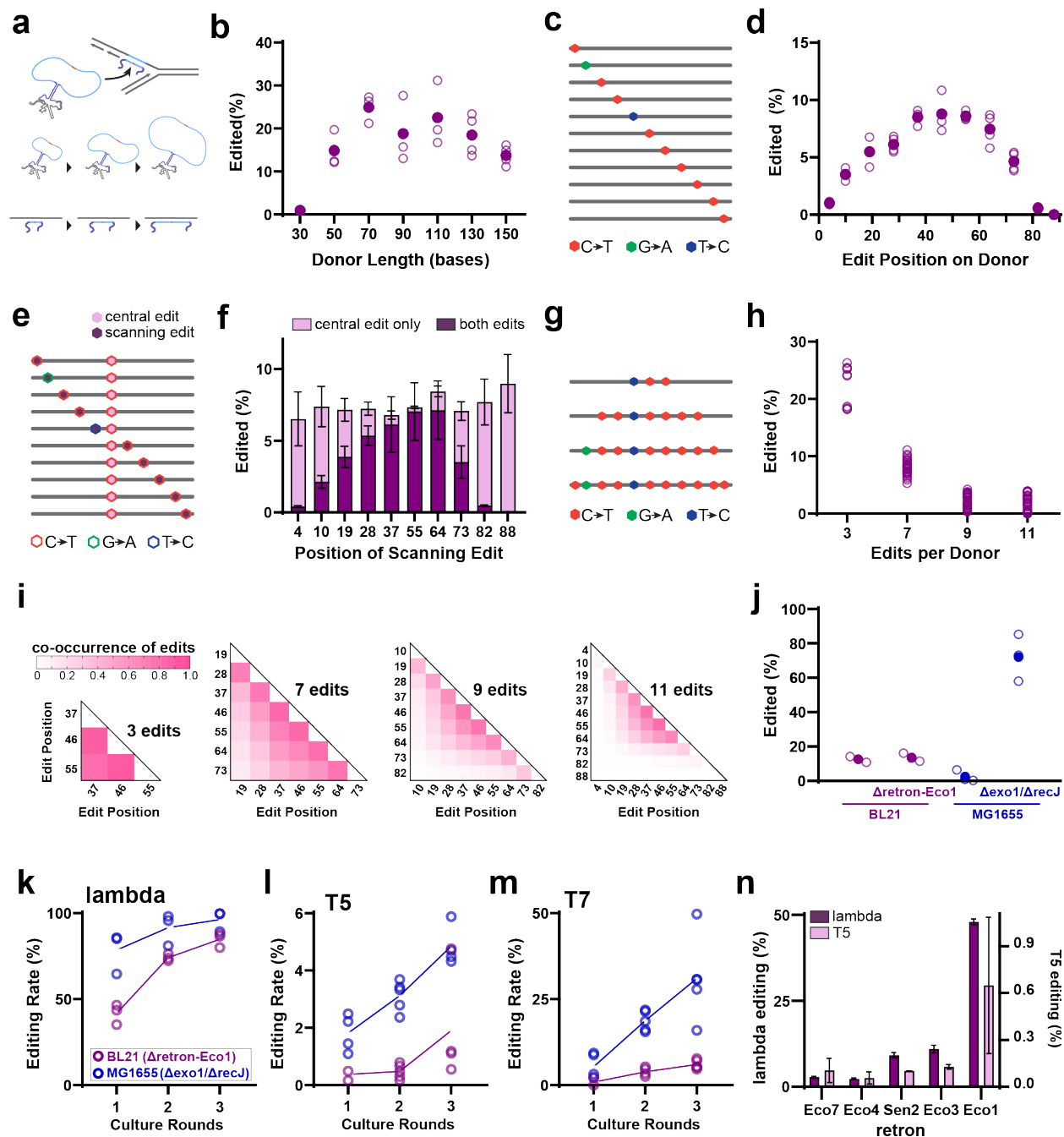


Figure 3-2: Optimizing recombitoron parameters.

a. Schematic of RT-donor length. **b.** Replicates of donor length are shown as open circles ($n=3$ for 90,110; $n=4$ for others); closed circles show the means. The effect of length (one-way ANOVA, $P < 0.0001$) is shown; length >50 versus 30 (Šidák's corrected $P < 0.05$) and length 150 versus 70 (Šidák's corrected $P < 0.05$). **c.** Schematic of RT-donors containing scanning edits. **d.** Scanning edits, shown as in **b** ($n=4$). The effect of placement (one-way ANOVA, $P < 0.0001$) is shown; position ≤ 28 or ≥ 73 was significantly worse than position 46 (Dunnett's corrected $P < 0.001$). **e.** Schematic of RT-donors containing a scanning and central edit. **f.** Bottom, the percentage of genomes with only central edit; top, the percentage of genomes with both edits. Plots are as in **b**. The position of scanning edit does not affect total editing (Figure caption continued on the next page)

(Figure caption continued from the previous page)

(one-way ANOVA, $P=0.7868$) but does affect whether both edits are installed (one-way ANOVA, $P<0.0001$). **g.** Multiedit recombitrons containing 3, 7, 9 and 11 edits per RT-Donor. **h.** Multiediting replicates at each site are shown as open circles ($n=3$ for 11 edits; $n=4$ for others). The number of edits per RT-Donor affects the average editing rate across sites (one-way ANOVA, $P<0.0001$), with 3 edits performing better than >3 edits (Dunnett's corrected $P<0.0001$). **i.** The co-occurrence of edits from multiedit donors is shown as the correlation coefficient r^2 normalized to the overall editing rate for each donor at each site. **j.** Editing in different host strains at position 14,126 (R), shown as in **b** ($n=3$). The effect of strain (one-way ANOVA, $P<0.0001$) is shown, where the MG1655 ($\Delta\text{exo1}/\Delta\text{recJ}$) strain yielded more editing than BL21 (Dunnett's corrected $P<0.0001$). **k.** Editing increases over three rounds of editing, where the engineered K-strain outperformed the B-strain (two-way ANOVA, strain, $P<0.001$; round, $P<0.0001$; interaction, $P=0.0282$). Three biological replicates are shown as open circles; lines connect the means. **l.** Editing T5 over three rounds, shown as in **k**. The engineered K-strain outperformed the B-strain (two-way ANOVA, strain, $P<0.0001$; round, $P=0.003$). **m.** Editing T7 over three rounds, shown as in **k**. The engineered K-strain outperformed the B-strain (two-way ANOVA, strain, $P<0.0001$; round, $P<0.0001$; interaction, $P=0.0011$). **n.** Editing of lambda (left axis) and T5 (right axis) using recombitrons based on different retons, shown as in **b** ($n=3$) (two-way ANOVA, retron, $P<0.0001$; phage, $P<0.0001$; interaction, $P<0.0001$). Additional statistical details are provided in **Supplementary Table 3-1**.

The next parameter we tested was positioning of the edit within the RT-Donor. We tested a set of recombitrons where we held a 90-nt region of homology to lambda constant and encoded an edit at different positions along the donor, each of which introduces a distinct synonymous single-base substitution (**Fig 3-2c**). Here we found that the editing rate increased as the edit approached the center of the donor (**Fig 3-2d**). After correcting for multiple comparisons, we found no difference between edit placement between a central 27 nt from position 37 to 64 but a significant decline in editing outside the central region.

We also tested the effect of position when encoding multiple edits on a single RT-Donor. In this case, we tested a set of recombitrons with edits at different positions along the RT-Donor as above but additionally included a central edit on every RT-Donor (**Fig 3-2e**). Similar to the effect above, we found that the rate of incorporating both edits on a single phage genome declined as the scanning edit approached either edge of the RT-Donor (**Fig 3-2f**). However, there was a consistent editing rate for all recombitrons in phage genomes that were only edited at the central site. When both edits were positioned

toward the center of the donor, most genomes contained both edits, whereas, when the scanning edit was positioned toward the edge of the RT-Donor, there was a significant increase in likelihood of incorporating only the central edit (**Fig 3-2f**). We interpret this result as evidence that the RT-Donor can be partially used, with a bias toward using the central part of the RT-Donor. We also found only a very low rate of the scanning edit being incorporated without the central edit (**Extended Data Fig 3-2a**).

Next, we tested the effect of increasing the number of edits per recombitron. We constructed a set of recombitrons with 3, 7, 9 or 11 of the scanning edits used above (**Fig 3-2g**). We found that editing with the three-edit recombitron was comparable to the single edits we previously tested but editing rates declined across all edit sites when additional edits were encoded on a single recombitron (**Fig 3-2h**). We interpret this as an effect of decreasing homology across the donor, making it more difficult to anneal to the target site. To further assess which edits on the same RT-Donor were likely to be acquired together, we analyzed the co-occurrence of edits. Because of the effect of decreasing homology on the editing rate, we calculated the correlation coefficient r^2 and normalized it to the overall editing rate for each donor at each site. We found that there was a bias toward acquiring the more central edits overall and a tendency to coedit nearby sites (**Fig 3-2i**), again supporting the potential for partial use of the RT-Donor.

Optimizing host strain

We also tested the effect of the editing host. Specifically, we compared editing in B-strain *E. coli* (BL21-AI), a derivative of BL21-AI lacking the endogenous retron-Eco1, K-strain *E. coli* (MG1655) and derivative of MG1655 with premature stop codons

in *exo1* and *recJ*—nucleases whose removal was previously shown to increase recombination rates using synthetic oligonucleotides^{1,2}. We found no difference between the BL21-AI and retron-deletion derivative, indicating that the endogenous retron does not interfere with the recombitron system (**Fig 3-2j**). We found decreased editing in the wild-type K-strain that was not statistically significant as compared to the B-strain but significantly improved editing in the $\Delta\textit{exo1};\Delta\textit{recJ}$ K-strain versus the B-strain (**Fig 3-2j**). This increase in editing observed in the modified K-strain was not directly accounted for by a gross increase in the amount of RT-DNA produced (**Extended Data Fig 3-2b**).

We next tested adding additional rounds of culture for multiple phages in the original B-strain and modified K-strain. For lambda, we found that three rounds of culture in the modified K-strain resulted in editing rates >95% (**Fig 3-2k**). This strain improvement also extended to T5, which reached 5% editing, and T7, which reached >30% editing, after three rounds of culture (**Fig 3-2l,m**). We also tested whether different retrons could be used to generate phage-editing donors in the recombitron format. We tested five different retrons, all of which supported editing of lambda and T5 (**Fig 3-2n**). Retron-Eco1-derived recombitrons resulted in the highest editing rates for both phages. Lastly, we tested the effect of culture conditions for editing of lambda, finding that a lower initial MOI and a lower culture temperature both resulted in higher rates of editing (**Extended Data Fig 3-2c,d**), pointing to potential gains from widening the editing time window within a given round of culture.

Insertions and deletions using recombitorons

Genomic deletions and insertions are also useful for engineered phage applications. Deletions can be used to remove potential virulence factors or toxins from phage genomes or to optimize phages by minimization. Insertions can be used to deliver cargo, such as nucleases that can help kill target cells or anti-CRISPR proteins to escape bacterial immunity. Therefore, we tested the efficiency of engineering deletions and insertions into the lambda genome.

We began by comparing the efficiency of deleting 2, 4, 8, 16 or 32 nt to one of the single-base synonymous substitutions we tested previously (**Fig 3-3a**). These experiments were performed in the engineered K-strain that we found to yield higher editing efficiencies using inactivated cl lambda as the phage. We deleted bases preceding position 37,673 using recombitorons where a 90-nt RT-Donor contained flanking homology around the deletion site but omitted the bases to be deleted. We found consistent editing of ~45% for each of the deletions (**Fig 3-3b**) using Illumina amplicon sequencing. The deletion size did not significantly affect the efficiency of editing, although all the deletions exhibited significantly lower editing than the synonymous single-base substitution.

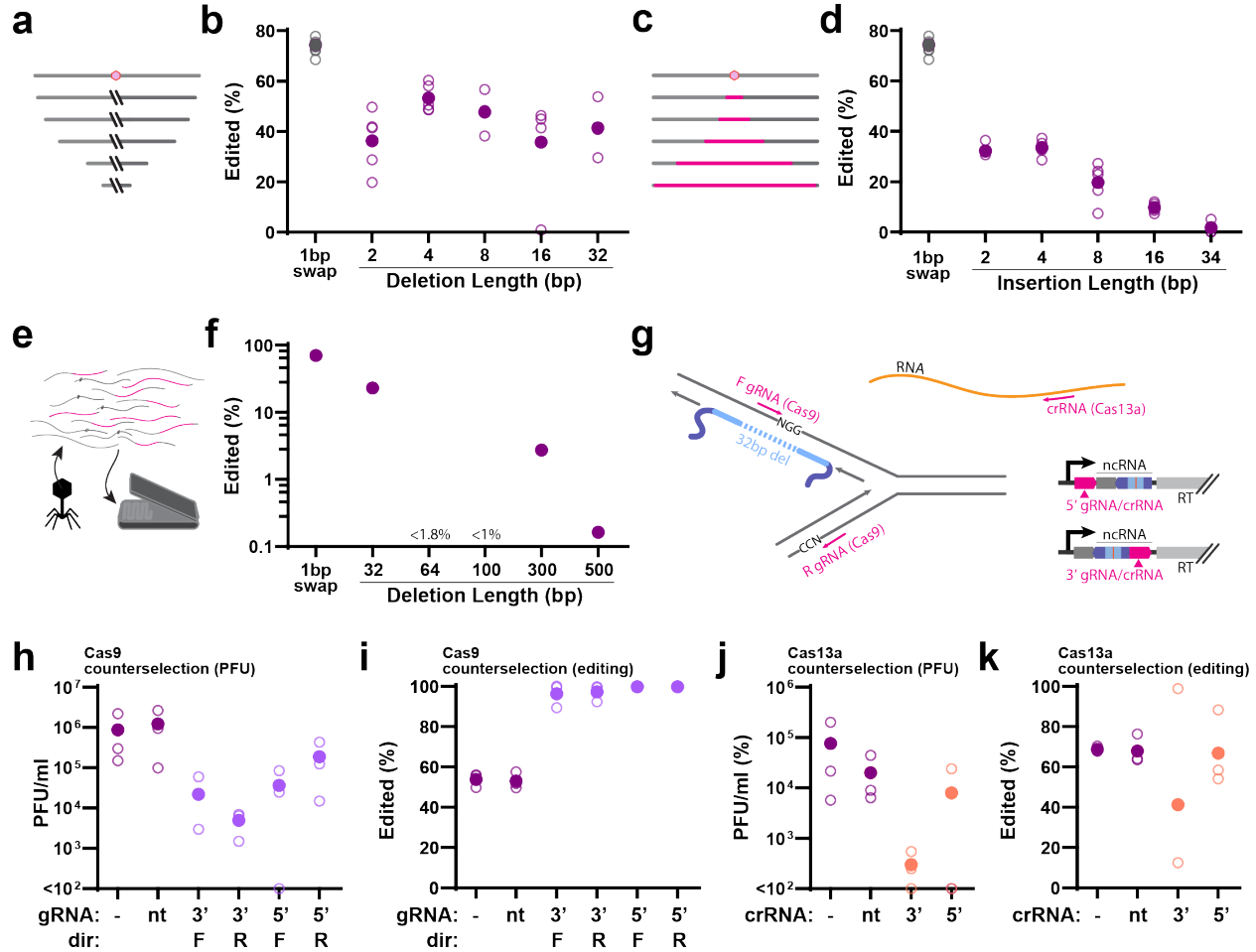


Figure 3-3: Insertions and deletions using recombitrons.

a. Phage genome deletions of increasing size around a common central site (pink hexagon). **b.** Quantification of editing efficiency for deletions preceding position 37,673 as compared to a single-base synonymous substitution. Five biological replicates are shown as open circles; closed circles are the means. There was no effect of deletion length (one-way ANOVA, $P = 0.1564$) but all deletions occurred at a lower frequency than the single-base substitution (one-way ANOVA, $P < 0.0001$; Dunnett's corrected versus 2, $P < 0.0001$; versus 4, $P = 0.0083$; versus 8, $P = 0.0008$; versus 16, $P < 0.0001$; versus 32, $P = 0.0005$). **c.** Phage genome insertions of increasing size ending at a common position (pink hexagon). **d.** Quantification of editing for insertions at position 37,673 as compared to a single-base synonymous substitution, shown as in **b** ($n = 5$). The same single-base substitution was used as a parallel control for both deletions and insertions. There was a significant effect of insertion length (one-way ANOVA, $P < 0.0001$) and all insertions occurred at a significantly lower frequency than the single-base substitution (one-way ANOVA, $P < 0.0002$; Dunnett's corrected $P < 0.0001$ for all insertions). **e.** Schematic illustrating long-read quantification. **f.** Editing efficiency using Nanopore. Points shown represent the fraction of edited reads pooled across three biological replicates. The 64-nt and 100-nt deletions were not detected in 57 and 102 reads aligned to the edit region, respectively. **g.** Schematic of simultaneous phage editing and counterselection. **h,i.** Number of viable phages (**h**) and percentage of edited phages (**i**) after simultaneous editing and counterselection by Cas9 (editing: one-way ANOVA, $P < 0.0001$; Dunnett's corrected for all on-target gRNA conditions versus no gRNA, $P < 0.0001$). A dash indicates no gRNA; NT, nontargeting; 3' and 5' indicate the fusion position of gRNA to editing ncRNA; F and R indicate the phage strand targeted by the guide. Three biological replicates are shown as in **b**. **j,k.** Number of viable phages (**j**) and percentage of edited phages (**k**) after simultaneous editing and counterselection by Cas13a (editing: one-way ANOVA, $P = 0.5594$), as in **h** ($n = 3$). Additional statistical details are provided in **Supplementary Table 3-1**.

We next tested the insertion of 2, 4, 8, 16 or 32 nt into the same lambda site (Fig 3-3c). Here, we found a range of editing efficiencies (Fig 3-3d). Like the deletions, all insertions were significantly less efficient than a synonymous single-base substitution. Unlike the deletions, insertion size significantly affected efficiency, favoring smaller insertions.

Whereas the synonymous single-base substitution is assumed to be neutral for phage fitness, we cannot assume the same for the deletions and insertions, which could affect transcription or phage packaging. This may underlie the lower, although still substantial, rate of deletions and insertions overall. Additionally, the PCR required for analysis of editing by amplicon sequencing is known to be biased toward smaller amplicons, which could inflate the measured rates of the larger deletions and decrease the measured rates of larger insertions.

To test yet larger insertions in a manner that is not subject to the size bias of PCR, we edited phages and then sequenced their genomes without amplification using long-read Nanopore sequencing (Fig 3-3e). After editing, we isolated phage genomes using the Norgen Phage DNA Isolation kit, attached barcodes and Nanopore adaptors and sequenced molecules for 24–48 h on a MinION (Oxford Nanopore Technologies). We quantified the resulting data using custom analysis software that binned reads by alignments to three possible genomes: a wild-type lambda genome, a lambda genome containing the relevant edit or the BL21 *E. coli* genome as a negative control. Basic local alignment search tool (BLAST) alignment scores (percentage identity, *E* value and alignment length) were compared for any read aligning to the lambda genome (wild-type

or edited) in the region of the intended edit to quantify wild-type versus edited genomes. Consistent with a PCR bias for size, we found that amplification-free sequencing resulted in comparable editing rates to amplification-based sequencing for quantifying a single-base substitution but slightly lower rates of editing when measuring a 32-nt deletion as compared to amplification-based sequencing (**Extended Data Fig 3-3a**).

This amplification-free sequencing approach enabled us to extend our deletion and insertion size ranges. For deletions, we tested 32, 64, 100, 300 or 500 nt. We found deletions of 32 nt at a frequency of ~23%, 300 nt at a frequency of ~2.7% and 500 nt at a frequency of 0.16% (**Fig 3-3f**). We did not observe deletions of 64 or 100 nt in our data. However, our long-read coverage of the editing region in these samples was limited; hence, we can only conclude that, if these edits occur, they are present at <1.8% and <1%, respectively, across the three replicates of our sequencing data (**Extended Data Fig 3-3b,c**). In cases where we observed deletions, they were reliably found at the intended location and of the intended size (**Extended Data Fig 3-3d**).

We also tested the insertion of larger sequences. We built recombitrons to insert a 34-nt flippase recognition target (FRT) recombination site, a 264-nt anti-CRISPR protein (AcrIIA4), a 393-nt anti-CRISPR protein (AcrIIA13) and a 714-nt sfGFP. As anticipated from the low insertion rates of insertions >8 nt, we did not observe any of these larger insertions in our long-read sequencing data, which were read-limited to a detection level of ~1% (**Extended Data Fig 3-3e**). Thus, the recombitrons enable high-efficiency deletions of up to 32 nt and insertions of up to 8 nt but are not practical for counterselection-free isolation of larger deletions or insertions.

To extend the utility of recombitrans for larger insertions and deletions, we performed simultaneous retron recombination and counterselection (**Fig 3-3g**). We tested *Streptococcus pyogenes* Cas9 (DNA-targeting) and *Leptotrichia buccalis* Cas13a (RNA-targeting) as counterselection modalities, along with different guide RNA (gRNA) and CRISPR RNA (crRNA) architectures. For Cas9, we tested gRNAs targeting both the top and the bottom strand of the genome (F, forward; R, reverse). For both Cas9 and Cas13a, we tested fusing the gRNA along with required ncRNA scaffolding at the 5' end and 3' end of the recombित्रon ncRNA. Simultaneous counterselection with Cas9 increased the editing frequency of a 32-nt deletion from ~50% (no gRNA) to >99% (on-target gRNA), while only decreasing the output number of phage 10–100× (**Fig 3-3h,i**). Cas13a counterselection did not increase the frequency of a 32-nt deletion when an on-target gRNA was present, although there was much more variability in the editing frequency and a substantial reduction in phage output of 100–1,000× (**Fig 3-3j,k**).

Multiplexed phage engineering using recombitrans

One shortcoming of current phage-editing approaches is the time and labor required to introduce multiple nonadjacent edits on the same phage. The ability to introduce parallel modifications simultaneously would be of great benefit to efforts aimed at engineering phages for targeted killing of pathogenic bacteria and directed interactome analyses, but current approaches require cycles of editing, isolation and re-editing that are impractical in an academic or industrial setting. We reasoned that a slight modification of our recombित्रon approach would enable such parallel, multiplexed editing of distant sites across a genome. In this modification, multiple bacterial strains that each harbor

distinct recombitrons targeting different parts of the phage genome are mixed in a single culture. The phage to be edited is then propagated through that mixed culture. Every new infection event is an opportunity to acquire an edit from one of the recombitrons and, over time, these distinct, distant edits accumulate on individual phage genomes (**Fig 3-4a,b**).

We ran these experiments in the engineered K-strain.

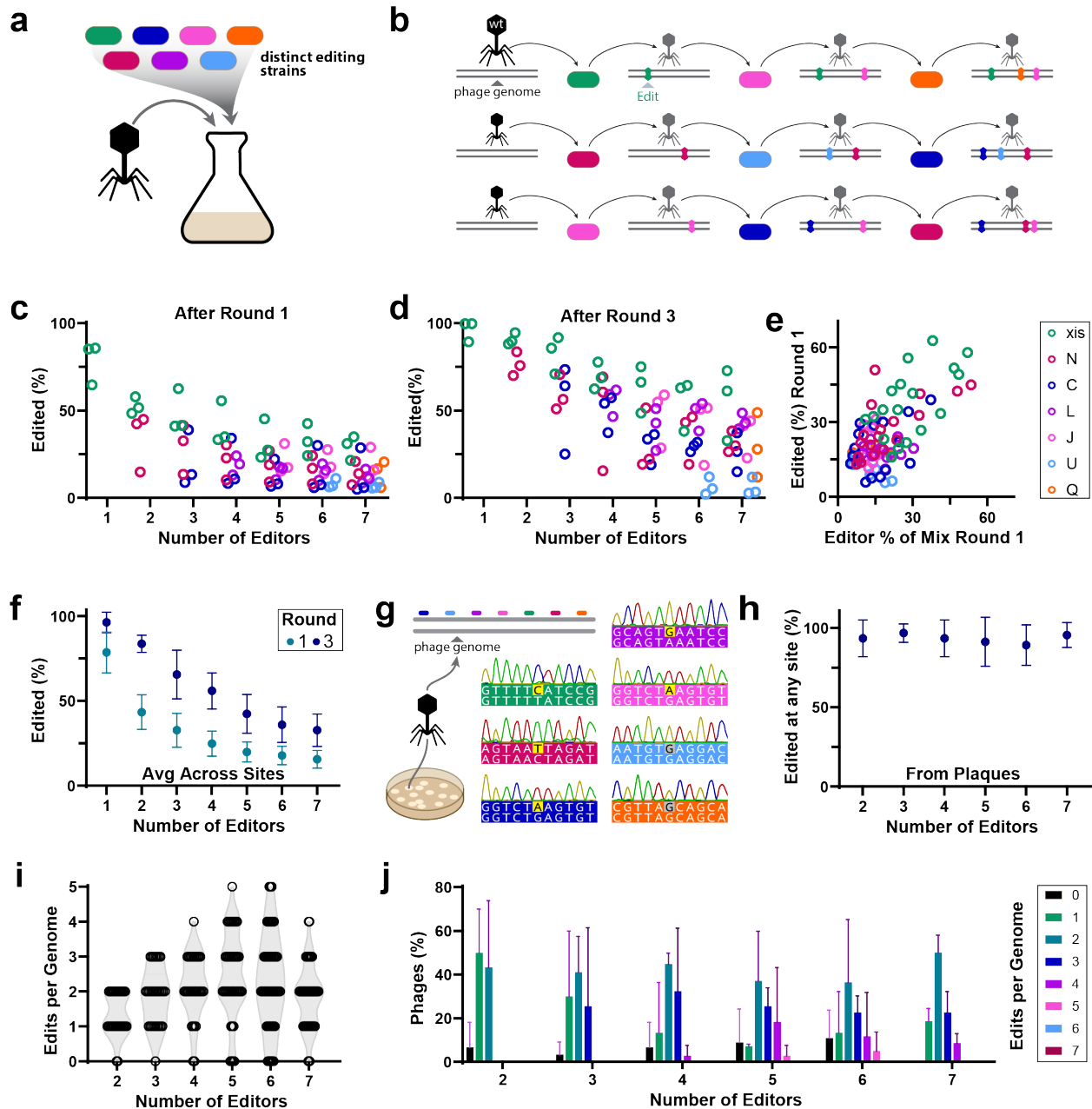


Figure 3-4: Multiplexed phage engineering using recombitrons.

a. Schematic illustrating the propagation of phages through cells expressing distinct recombitrons in a coculture. **b.** Over generations of phage propagation, phage genomes accumulate multiple edits from different RT-Donors. **c.** Editing (%) of each site from mixed recombitron cultures after one round of editing. Three biological replicates are shown in open circles for each site, clustered over the number of recombitrons used. **d.** Editing (%) of each site from mixed recombitron cultures after three rounds of editing, shown as in **c.** **e.** Editing (%) of each site after one round of editing in relation to the proportion of that recombitron strain in the culture (%) (two-tailed Pearson correlation, $P < 0.0001$). **f.** Average editing (%) of all sites in each mixed culture. Points represent the mean of three biological replicates \pm s.d. (two-way ANOVA, number of editors, $P < 0.0001$; round, $P < 0.0001$; interaction, $P = 0.234$) **g.** Schematic illustrating (Figure caption continued on the next page)

(Figure caption continued from the previous page)

*Sanger sequencing of plaques, which enables quantification of each editing site from clonal phages. Representative data from one plaque with five edits (yellow) and two wild-type sites (gray). Data in **h–j** are from three biological replicates. Total number of plaques: two editors, 30 plaques; three editors, 29 plaques; four editors, 31 plaques; five editors, 42 plaques; six editors, 53 plaques; seven editors, 55 plaques. **h**. Overall editing (%) of any site on plaques isolated from mixed cultures. Points represent the mean of three biological replicates \pm s.d. **i**. Number of edits made on a single phage genome from mixed cultures as a function of the number of recombitrons. Individual plaques from three biological replicates are shown as open circles on a violin plot of distribution. **j**. Number of edits made on a single phage genome from mixed cultures as a function of the number of recombitrons, shown as a histogram (mean \pm s.d. for three biological replicates). Additional statistical details are provided in **Supplementary Table 3-1**.*

We created seven bacterial editing strains using the lambda recombitrons tested in **Fig 3-1b**. We then made mixed cultures of one, two, three, four, five, six or all seven bacterial strains and performed editing of lambda phage in these mixtures. These strains were grown individually in liquid culture overnight, then separately preinduced for 2 h, before being diluted to OD 0.25 and mixed together in equal proportions. We infected these cultures with lambda (ΔcI) at an MOI of 0.1, grew the cultures for 16 h overnight and collected the phage lysate the next morning. We then used that phage stock to perform two additional rounds of editing in mixed-host cultures prepared identically to those of the first round. On the basis of our results that the amount of phage put into the culture is roughly the amount of phage extracted (**Extended Data Fig 3-1d,e**), we added the same volume of phage in each round as a function of the titer determination before round one. We quantified editing across phage genomes and genomic loci for each of the mixtures using Illumina sequencing. We found editing across all sites from these mixed cultures after one round and increased editing in all cultures and sites after three rounds (**Fig 3-4c,d**). We found that the editing rate at any given site across all the mixtures was well correlated with the percentage of that editor in the mixture after round one (**Fig 3-4e**) and that the overall editing rate across all sites declined with the number of recombitron strains used (**Fig 3-4f**). This is consistent with a dilution effect of the strains on each other,

which suggests that the overall editing rate is limited by the number of cells expressing each recombitron available for the phage to propagate through.

While amplicon sequencing showed substantial editing across many sites in a population of phages, we cannot use it to determine whether multiple edits are accumulating on individual phage genomes. Therefore, we plated phages after three rounds of editing from each condition on nonediting bacterial lawns and Sanger-sequenced individual plaques at each edit site (**Fig 3-4g**). This showed that 93.2% of phages were edited at one site or more, without using any form of counterselection (**Fig 3-4h**). The plaque-based analysis strongly mirrored the Illumina sequencing analysis of editing rate per site by mixture when plaque data for a condition were pooled (**Extended Data Fig 3-4**). With the plaque sequencing, however, we can also look at edits per phage genome. We found that multiple sites were edited on a majority of the phages across all conditions, with individual phages edited at up to five distinct locations (**Fig 3-4i,j**). This represents a milestone in the continuous, multiplexed, selection-free engineering of phage genomes.

Combinatorial mutations of the T7 tail fiber

We next used phage retron recombineering as a tool to understand the fundamental biology of phage–host interaction by making a library of combinatorial mutations within the T7 receptor-binding protein. Specifically, we focused on the lipopolysaccharide (LPS)-recognizing tip domain of the T7 tail fiber protein gp17. Mutations in *E. coli* genes *rfaG* and *rfaD* lead to LPS truncation at the outer and inner

core, respectively, which disrupts gp17 binding and T7 absorption, thus facilitating phage escape¹.

Previously performed deep mutational scanning of gp17 quantified the effect of individual amino acid substitutions on phage fitness across *rfaG* and *rfaD* mutants, identifying phage mutants that were both host tolerant, enabling replication on both wild-type and mutant hosts, and host switching, enabling enhanced replication on mutant hosts¹. However, this approach was restricted to point mutants because of the technical complexity of exploring large combinatorial space with synthesized, barcoded libraries—a problem that recombitrons are well suited to address. We selected four highly consequential residues on the four exterior loops of the gp17 protein, which are close to one another in the folded protein, but nonadjacent in the phage genome. For each, we chose a host-switching and a host-tolerant substitution (**Fig 3-5a,b**). We constructed eight recombitron plasmids to create these mutations, carried in eight parallel strains. As in our previous experiment editing lambda, we mixed these strains together and propagated T7 through this editing mixture three times. We then took the resulting library of phages, which could have acquired any individual mutation or any combination of mutations, and ran a round of enrichment through either wild-type or one of two LPS mutants ($\Delta rfaG$ and $\Delta rfaD$) obtained from the Keio collection (**Fig 3-5c**). Pre-enrichment and postenrichment phages were sequenced through the gp17 region to quantify enrichment or depletion in each particular strain.

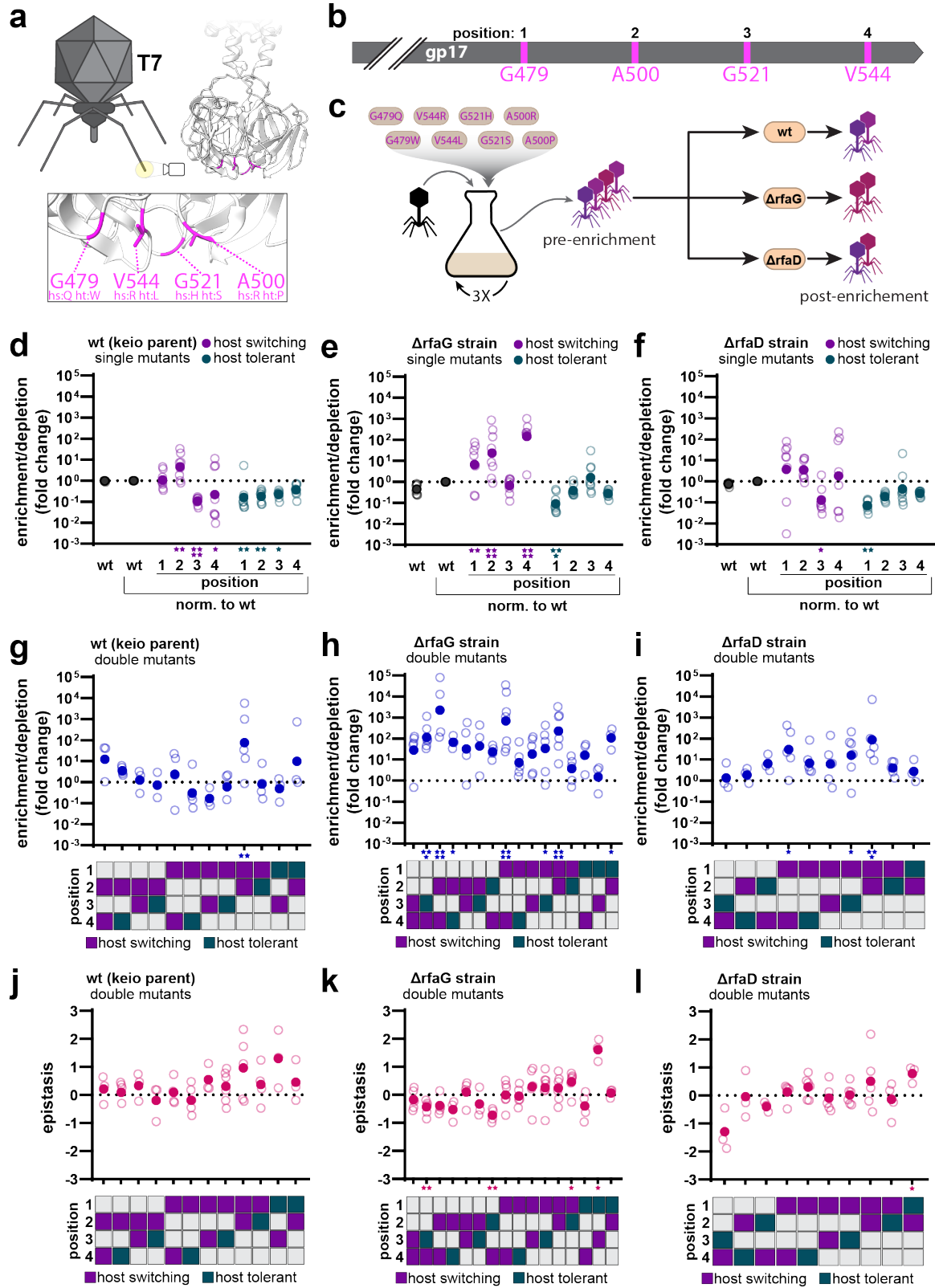


Figure 3-5: **Combinatorial mutations of the T7 tail fiber.**
 (Figure caption continued on the next page)

(Figure caption continued from the previous page)

a. Structural location of the targeted residues on the T7 gp17 tail fiber protein and their corresponding edit to host switching (*hs*) or host tolerant (*ht*). **b.** Genomic location of edits. **c.** Schematic illustrating the workflow to study combinatorial mutations. **d–f.** Quantification of the fold enrichment or depletion of single-residue mutants, both *hs* and *ht*, relative to the gp17 wild-type sequence of the postselection mixtures from the KEIO parent (**d**), $\Delta rfaG$ (**e**) and $\Delta rfaD$ (**f**) strains, respectively. Open circles are biological replicates; closed circles are the means. **g–i.** Quantification of the fold enrichment/depletion of combinatorial double-residue mutants of the postselection mixture from the KEIO parent (**g**), $\Delta rfaG$ (**h**) and $\Delta rfaD$ (**i**) strains, respectively. Open circles are biological replicates; closed circles are the means. **j–l.** Plot epistasis values calculated by subtracting the log-transformed enrichment or depletion of the individual single-site mutants from that of the simultaneous double-site mutant from the KEIO parent (**j**), $\Delta rfaG$ (**k**) and $\Delta rfaD$ (**l**) strains, respectively. Open circles are biological replicates; closed circles are the geometric means. Asterisks below a particular condition summarize *P* values (**P* < 0.05, ***P* < 0.01, ****P* < 0.001 and *****P* < 0.0001) for either a Dunnett's corrected multiple comparisons test versus wild type (**d–i**) or a Holm–Šidák adjusted two-sided, one sample *t*-test (**j–l**). Full statistical details are provided in **Supplemental Table 3-1**.

After enrichment through the wild-type (Keio parent) strain, we found that phages carrying one of the eight single substitutions were enriched (A500R), while phages carrying any one of the other substitutions in isolation were either mildly depleted or showed no effect (**Fig 3-5d**). In the $\Delta rfaG$ strain, single-mutant phages carrying three of the host-switching substitutions were enriched (G479Q, A500R and V544R) and the only depleted single mutant carried a putative host-tolerant substitution (G479W) (**Fig 3-5e**). The only significant effect of single substitutions in the $\Delta rfaD$ strain was mild depletion (**Fig 3-5f**).

We next examined combinations of substitutions. In general, as anticipated, double mutants displayed larger magnitudes of effects. For instance, a double mutant (G479Q;A500R) was ~100-fold enriched relative to wild-type T7 after passage through the Keio parent strain (**Fig 3-5g**). This effect was most pronounced in $\Delta rfaG$, where many double mutants were 100–5,000-fold enriched relative to the wild type (**Fig 3-5h**). The most enriched mutants were combinations of host-switching substitutions (A500R;V544R, G479Q;V544R and G479Q;A500R). Similarly, three double mutants were significantly enriched in $\Delta rfaD$ even though the single substitutions contained in

those double mutants only trended toward enrichment when present individually (**Fig 3-5f,i**).

On the basis of the single-mutant and double-mutant enrichment and depletion data, we next looked for evidence of epistatic interactions among the substitutions at these sites. Epistasis was calculated by taking the log-transformed enrichment or depletion of a given double-site mutant and subtracting the log-transformed enrichment or depletion of each individual single-site mutant contained in the double. We found no significant epistasis among the double mutants after enrichment through the wild-type strain (**Fig 3-5j**). Likewise, the double mutants that were strongly enriched by selection in the $\Delta rfaG$ strain (for example, A500R;V544R, >1,000-fold enriched) displayed no epistasis but rather were enriched exactly as expected by the additive effects of the single mutants they contained (**Fig 3-5k**). With such large effect sizes, it is a testament to the approach that such additive effects were able to be accurately measured. We did find evidence of negative epistasis when one of the most strongly positive single mutants (V544R) was combined with a host-tolerant substitution that had no effect on its own (G521S) (**Fig 3-5k**). After enrichment in $\Delta rfaD$ strain, the fitness of all double mutants was as expected by the additive effects of the single mutants (**Fig 3-5l**). This work identifies particular combinations of T7 tail fiber substitutions that lead to improved fitness in LPS mutant strains and finds, surprisingly, that nearly all substitutions were additive in effect without significant epistatic effects. This work also demonstrates the utility of the multiplexed recombitron approach, where eight simple plasmids could be used to quickly generate 24 distinct, precise phage mutants for direct phenotyping, without a step of counterselection or isolation.

3.4 Discussion

Here, we present an approach to phage editing using recombitrans that leverages modified retrons to continuously provide RT-Donors for recombineering in phage hosts. Recombitrans enable counterselection-free generation of phage mutants across multiple phages, with optimized forms yielding up to 100% editing efficiency in lambda. Moreover, recombitrans can be multiplexed to generate multiple distant mutations on individual phage genomes. Critically, this approach is easy to perform. Recombitrans are generated from simple, standard cloning methods using inexpensive, short oligonucleotides. The process of editing requires propagating the bacteria–phage culture, with no intervening transformations or special reagents. Once recombitrans are cloned and phage stocks are prepared, the generation of lambda phage edited at up to five distinct positions required hands-on time of less than 2 h.

Much previous work modifying phage genomes uses recombination followed by CRISPR counterselection to achieve acceptable rates of editing. However, such counterselection is not always possible when creating defined mutations, particularly in the case of small mutations (1–3 nt). In the case of Cas9 counterselection, the mutation must eliminate or modify a protospacer-adjacent motif (PAM) or the seed bases of the gRNA target to protect the edited phage from cleavage. When using a PAM-less CRISPR nuclease such as Cas13a, it has been shown that multiple contiguous mutations in the seed region are required to confer resistance; thus, a precise single-base mutation or small noncontiguous amino acid modification cannot be created without recoding a more substantial region¹. Therefore, a counterselection-free approach is preferable to create particular precisely defined modifications to phage genomes.

While we were able to edit a number of distinct phages, we found substantial variation in the editing rates among different phages. In some cases, we were able to explain these differences (for example, the T7 SSB that interfered with recombineering and could be fixed with the overexpression of a compatible SSB or the modified cytosine that appeared to inhibit recombineering in T4). However, we were not able to explain all the differences. One possible outstanding explanation is the replication speed and burst size of different phages, which may make the editing window in a single bacterium or single culture longer or shorter. Indeed, the fact that a lower initial MOI and lower culture temperature lead to higher levels of editing in lambda suggests that extending the number of cycles per culture is useful and a large burst size, for instance, would do the exact opposite. Another variable is the extent and speed of host genome degradation during phage infection^{1,2}, which could affect the additional production of editing materials or critical host factors. These variables will need to be further explored as this approach is scaled up.

This technical advance is poised to change the way we approach innovation in phage biology and phage therapeutics. For instance, studying epistasis in phage genomes becomes a feasible experiment that merely requires mixing recombinon strains in different combinations, as supported by our intraprotein epistasis studies on the T7 gp17 tail fiber. The technical hurdle of phage library generation is also dramatically reduced. In our multiplexed phage lambda experiment, 93.2% of phages were edited. Host range is a major determinant of phage therapy efficacy^{1,2}, which could be addressed by rapidly screening a large library targeting the tail fiber or other critical host range genes for a fraction of the effort of current approaches.

Some outstanding questions and further engineering challenges remain. For instance, overexpression of a compatible SSB provided a large benefit to the editing of T7, which demonstrates that recombitrons can be optimized for generality across phages. However, that effect was not immediately transferable to T5. Presumably, T5 may be limited by another phage-specific requirement and T2 is presumably limited by the use of modified nucleotides. Further work will seek to engineer around these idiosyncrasies and increase the extensibility of this approach to other phages. Lastly, we achieved our most efficient editing in a host *E. coli* lacking *exo1* and *recJ*. Moreover, we overexpressed both the *E. coli* SSB to increase efficiency and a dominant-negative *mutL* to suppress mismatch repair. These modifications may limit the ease with which the phage recombineering approach can be ported to other bacterial species. Testing and overcoming such limitations will be a major objective of future work.

3.5 Methods

Biological replicates were taken from distinct samples, not the same sample measured repeatedly.

Bacterial strains and growth conditions

This work used the following *E. coli* strains: NEB 5-alpha (NEB, C2987; not authenticated), BL21-AI (Thermo Fisher, C607003; not authenticated), bMS.346 and bSLS.114. The bMS.346 (used previously¹) strain was generated from *E. coli* MG1655 by inactivating the *exoI* and *recJ* genes with early stop codons. The bSLS.114 (used previously¹) strain was generated from BL21-AI by deleting the retron-Eco1 locus by lambda Red recombinase-mediated insertion of an FRT-flanked chloramphenicol resistance cassette, which was subsequently excised using FLP recombinase¹. The bCF.5 strain was generated from bSLS.114, also using the lambda Red system. A 12.1-kb region was deleted that contains a partial lambda*B prophage that is native to BL21-AI cells within the *attB* site, where temperate lambda integrates into the bacterial genome¹. Keio collection *E. coli* strains were used for the selection of T7 mutants (Keio parent, $\Delta rfaG$ and $\Delta rfaD$ ¹).

Phage retron recombineering cultures were grown in Luria–Bertani (LB) medium, shaking at 37 °C with appropriate inducers and antibiotics. Inducers and antibiotics were used at the following working concentrations: 2 mg ml⁻¹ l-arabinose generally or 10 mg ml⁻¹ when performing counterselection (GoldBio, A-300), 1 mM IPTG (GoldBio, I2481C), 1 mM *m*-toluic acid (Sigma-Aldrich, 202-723-9), 35 µg ml⁻¹ kanamycin (GoldBio, K-120), 100 µg ml⁻¹ carbenicillin (GoldBio, C-103) and

25 $\mu\text{g ml}^{-1}$ chloramphenicol (GoldBio, C-105; used at 10 $\mu\text{g ml}^{-1}$ for selection during bacterial recombineering for strain generation).

Plasmid construction

All cloning steps were performed in *E. coli* NEB 5-alpha. pORTMAGE-Ec1 was generated previously (Addgene, plasmid no. 138474)¹. Derivatives of pORTMAGE-Ec1 (pCF.109, pCF.110 and pCF.111) were cloned to contain an additional SSB protein, amplified with PCR from its host genome, using Gibson assembly. Plasmids for RT-Donor production, containing the retron-Eco1 RT and ncRNA with extended a1/a2 regions, were cloned from pSLS.492. pSLS.492 was generated previously (Addgene, plasmid no. 184957)¹. Specific donor sequences for small edits were encoded in primers and substituted into the RT-DNA-encoding region of the ncRNA with a PCR and KLD (kinase, ligase and DpnI) reaction (NEB M0554). Donor sequences for larger insertions were cloned through Gibson assembly using synthesized gene fragments (Twist Biosciences). Recombitron ncRNAs encoding the editing donors are listed in **Supplementary Table 3-2**.

Phage strains and propagation

Phages were propagated from American Type Culture Collection stocks (Lambda 97538, Lambda WT 23724-B2, T7 BAA-1025-B2, T5 11303-B5 and T2 11303-B2) into a 2-ml culture of *E. coli* (BL21 Δ Eco1) at 37 °C at an OD600 0.25 in LB medium supplemented with 0.1 mM MnCl₂ and 5 mM MgCl₂ (MMB) until culture collapse, according to established techniques^{1,2}. The culture was then centrifuged for 10 min at

3,434g and the supernatant was filtered through a 0.2- μ m filter to remove bacterial remnants. Lysate titer was determined using the full-plate plaque assay method as described by Kropinski et al¹. We used recombitrons to edit this lambda strain to encode two early stop codons in the *cI* gene, responsible for lysogeny control, to ensure the phage was strictly lytic (lambda ΔcI). After recombineering, we Sanger-sequenced plaques to check the edit sites. We isolated an edited plaque and used Illumina Miseq of its lysate to ensure purity of the edited phage. We used this strictly lytic version for all experiments involving lambda phage, unless otherwise noted.

Genomic locations used to label edits are from wild-type reference sequences of phages available through the National Center for Biotechnology Information (NCBI) GenBank: lambda (J02459.1), T5 (AY587007.1), T7 (V01146.1) and T2 (AP018813.1). We found that the strain of phage lambda we used naturally contains a large genomic deletion between 21,738 and 27,723. This region encodes genes that are not well characterized but may be involved in lysogeny control^{1,2}.

Plaque assays

Small-drop and full-plate plaque assays were performed similarly to Mazzocco et al.¹, starting from bacteria grown overnight at 37 °C. For small-drop plaque assays, 200 μ l of the bacterial culture was mixed with 2 ml of melted MMB agar (LB + 0.1 mM MnCl₂ + 5 mM MgCl₂ + 0.75% agar) and plated on MMB agar plates. Tenfold serial dilutions in MMB were performed for each of the phages and 2- μ l drops were placed on the bacterial layer. The plates were incubated overnight at 37 °C. Full-plate plaque assays were set up by mixing 200 μ l of the bacterial culture with 20 μ l of phage lysate, using

tenfold serial dilutions of the lysate to achieve 200–10 plaques. After incubating at room temperature without shaking for 5 min, the mixture was added to 2 ml of melted MMB agar and poured onto MMB agar plates. The plates were incubated overnight at 37 °C. Plaque-forming units were counted to calculate the titer.

Recombineering and sequencing

The retron cassette, with modified ncRNA to contain a donor, was coexpressed with CspRecT and mutL E32K from the plasmid pORTMAGE-Ec1. All experiments, except multiplexed lambda editing and large insertions or deletions (>30 nt), were conducted in 500- μ L cultures in a deep 96-well plate. Multiplexed lambda editing and large insertions or deletions (>30 nt) were conducted in 3-ml cultures in 15-ml tubes. For amplification-free sequencing by Nanopore, larger culture volumes (25-ml cultures in 250-ml flasks) were used to enable collection of a higher quantity of phage DNA. Cultures were induced for 2 h at 37 °C with shaking. The OD₆₀₀ of each culture was measured to approximate cell density and cultures were diluted to an OD₆₀₀ of 0.25. Phages were originally propagated through the corresponding host that would be used for editing (B-strain or K-strain *E. coli*). A volume of pretitered phage was added to the culture to reach an MOI of 0.1. The infected culture was grown overnight for 16 h before being centrifuged for 10 min at 3,434g to remove host cells. The supernatant was filtered through a 0.2- μ m filter to isolate phage.

For amplicon-based sequencing, the lysate was mixed 1:1 with DNase-free and RNase-free water and the mixture was incubated at 95 °C for 5 min. This boiled culture (0.25 μ l) was used as a template in a 25- μ l PCR reaction with primers flanking the edit

site on the phage genome. These amplicons were indexed and sequenced on an Illumina MiSeq or NextSeq2000 instrument. Sequencing primers are listed in **Supplementary Table 3-3**.

For amplification-free sequencing, extracellular DNA was removed through DNase I treatment, with 20 U of DNase I (NEB, M0303S) for 1 ml of phage lysate, incubated at room temperature for 15 min and then inactivated at 75 °C for 5 min. Phages were then denatured and DNA was extracted using the Norgen Phage DNA isolation kit (Norgen, 46800). The samples were prepared for sequencing in a standard Nanopore workflow. DNA ends were repaired using the NEBNext Ultra II End repair/dA-tailing module (NEB, E7526S). End-repaired DNA was then cleaned up using Ampure XP beads. Barcodes were ligated using the standard protocol for Nanopore barcode expansion kit (Oxford Nanopore Technologies, EXP-NBD196). After barcoding, the standard Oxford Nanopore adaptor ligation, clean-up and loading protocols were followed for ligation sequencing kit 109 and flow cell 106 for the MinION instrument (Oxford Nanopore Technologies, SQK-LSK109 and FLO-MIN106D). Base calling was performed using Guppy Basecaller with high accuracy and barcode trimming settings.

Sanger sequencing of phage plaques was accomplished by picking plaques produced from the full-plate assay described above. Plates were sent to Azenta (Genewiz) for sequencing with one of the MiSeq-compatible primers used to assess the same site. Sequences were analyzed using Geneious through alignment to the region surrounding the edit site on the phage genome.

Enrichment of phage in nonediting hosts

A total of 1–10 million multiplex-edited phages were added to exponentially growing cultures of bacteria at an OD of 0.25 or 0.4 in a volume of 50 ml for $0.0000625 \leq \text{MOI} \leq 0.001$. Phages were propagated through their enrichment host for 16 h shaking at 37 °C. When performing PCR amplification for sequencing, at least double the amount of enriched phage as the number of reads was used as a template (for example, for 1 million reads, ≥ 2 million phages were used as template for the Illumina sequencing PCR).

Editing rate quantification

A custom Python workflow was used to quantify edits from amplicon sequencing data. Reads were required to contain outside flanking nucleotide sequences that occur on the phage genome but beyond the RT-Donor region to avoid quantifying RT-DNA or plasmid. Reads were then trimmed by left and right sequences immediately flanking the edit site. Reads containing these inside flanking sequences in the correct order with an appropriate distance between them (depending on edit type) were assigned to wild type, edit or other. The edit percentage is the number of edited reads over the sum of all reads containing flanking sequences.

A distinct custom Python program was used for quantifying amplification-free Nanopore sequencing data because of the higher error rates in Nanopore sequencing and the lack of a defined region of the genome contained in each read. Reads were aligned using BLAST+ to three reference genomes: wild-type lambda, edited lambda containing the matching edit to the read's experiment and BL21-AI *E. coli*. If reads aligned

to either lambda genomes, the read's alignment coordinates had to be at least 50 nt past the insertion or deletion coordinates, as well as be >500 nt and have >50% of the read mapped to the reference genome as quality scores. If a read aligned to the insertion or deletion point and passed all quality scores, the percentage identity and alignment length over read length were compared to assign the read as either wild type or edited. Coverage of the edit region was 50–1,000× per experimental condition.

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 from the NCBI SRA (PRJNA933262).

Code availability

Custom code to process or analyze data from this study is available from GitHub (https://github.com/Shipman-Lab/Multiplexed_Phage_Recombitrons).

Acknowledgements

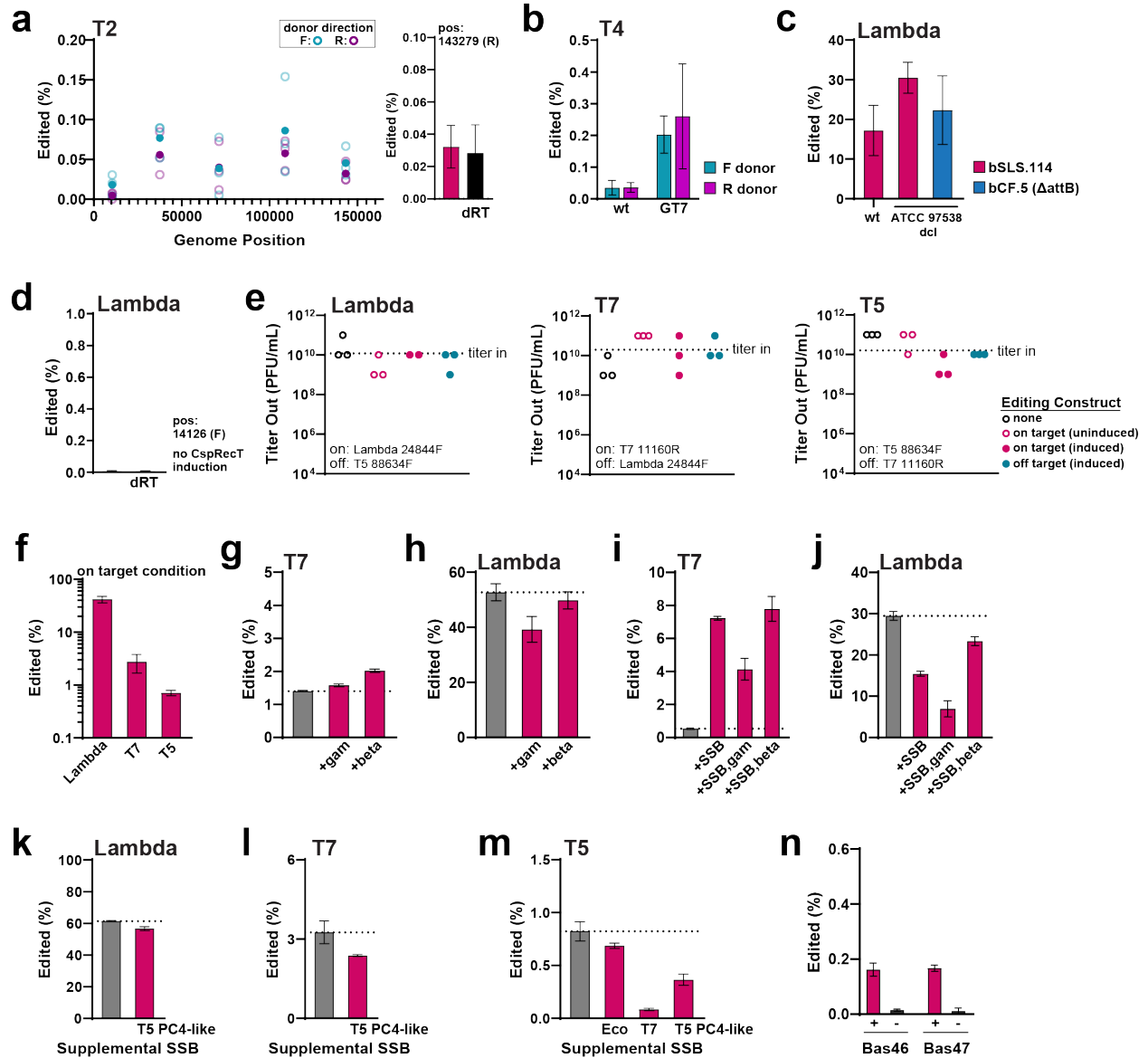
This work was supported by funding from the National Science Foundation (MCB 2137692), the National Institute of Biomedical Imaging and Bioengineering (R21EB031393), the Gary and Eileen Morgenthaler Fund and the National Institute of General Medical Sciences (1DP2GM140917). S.L.S. is a Chan Zuckerberg Biohub, San Francisco investigator and acknowledges additional funding support from the L.K. Whittier

Foundation and the Pew Biomedical Scholars Program. K.D.C. and K.Z. were supported by National Science Foundation Graduate Research Fellowships and University of California, San Francisco Discovery Fellowships. A.G.-D. was supported by the California Institute of Regenerative Medicine scholar program.

Contributions

These authors contributed equally: Chloe B. Fishman, Kate D. Crawford, Santi Bhattarai-Kline, Darshini Poola. C.B.F., S.B.-K. and S.L.S. conceptualized the study and, with K.D.C., K.A.Z. and A.G.-D., outlined the scope of the project and designed experiments. C.B.F. developed the phage handling and editing protocols. Experiments were performed and analyzed by C.B.F. (**Figs 3-1e,f,i, 3-2a-k, 3-4, and Extended Data Figs 3-1b and 3-3**), K.D.C. (**Figs 3-2l,m, 3-3, 3-5, and Extended Data Figs 3-1b and 3-3**), S.B.-K. (**Fig 3-1b-d and Extended Data Fig 3-1a**), D.P. (**Fig 3-5**), K.A.Z. (**Figs 3-1g and 3-3 and Extended Data Figs 3-1n and 3-3**) and A.G.-D. (**Fig 3-2n and Extended Data Fig 3-2b-d**). C.B.F. and S.L.S. wrote the manuscript with input from all authors.

3.6 Supplementary Information

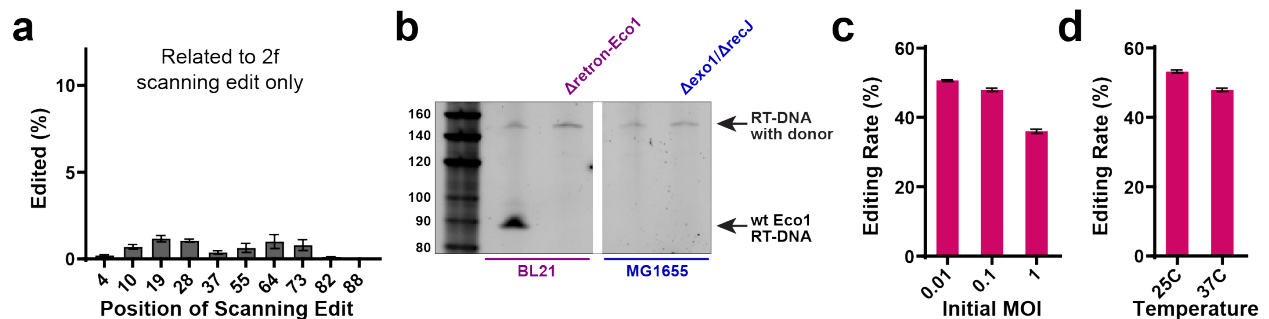


Extended Data Fig 3-1: **Accompaniment to Fig 3-1**

a. Left: Edited phage T2 genomes (%). With forward (blue) or reverse (purple) RT-DNA. Open circles are three biological replicates, closed circles are means. Right: Recombitron editing at site 143,279 (R) (\pm SD) versus a dRT control (unpaired, two-sided *t*-test, $P = 0.77$). **b.** Editing of wild-type and mutant T4 without modified cytosines, shown as in **a** ($N = 3$) (two-way ANOVA, effect of modified bases, $P = 0.0061$). **c.** Editing (%) of lambda and lambda Δ cl in different host strains at site 14,070 (R). Open circles show 3 (wt in bSLS.114), 5 (Δ cl in bSLS.114), and 2 (Δ cl in bCF.5) biological replicates, closed circles show the mean (one-way ANOVA, effect of strain/phage, $P = 0.2421$). **d.** Editing (%) of lambda without induction of CspRecT at site 14,126, shown as in **a** ($N = 3$). **e.** Titer (PFU/mL) of phage lambda, T7, and T5 after propagation through host cells of different conditions, compared to amount of phage added to the culture, without recombitrons (open black circles), with uninduced recombitrons (open pink circles), with induced recombitrons (closed pink circles), and with induced recombitrons targeting a different phage (closed blue) (Figure caption continued on the next page)

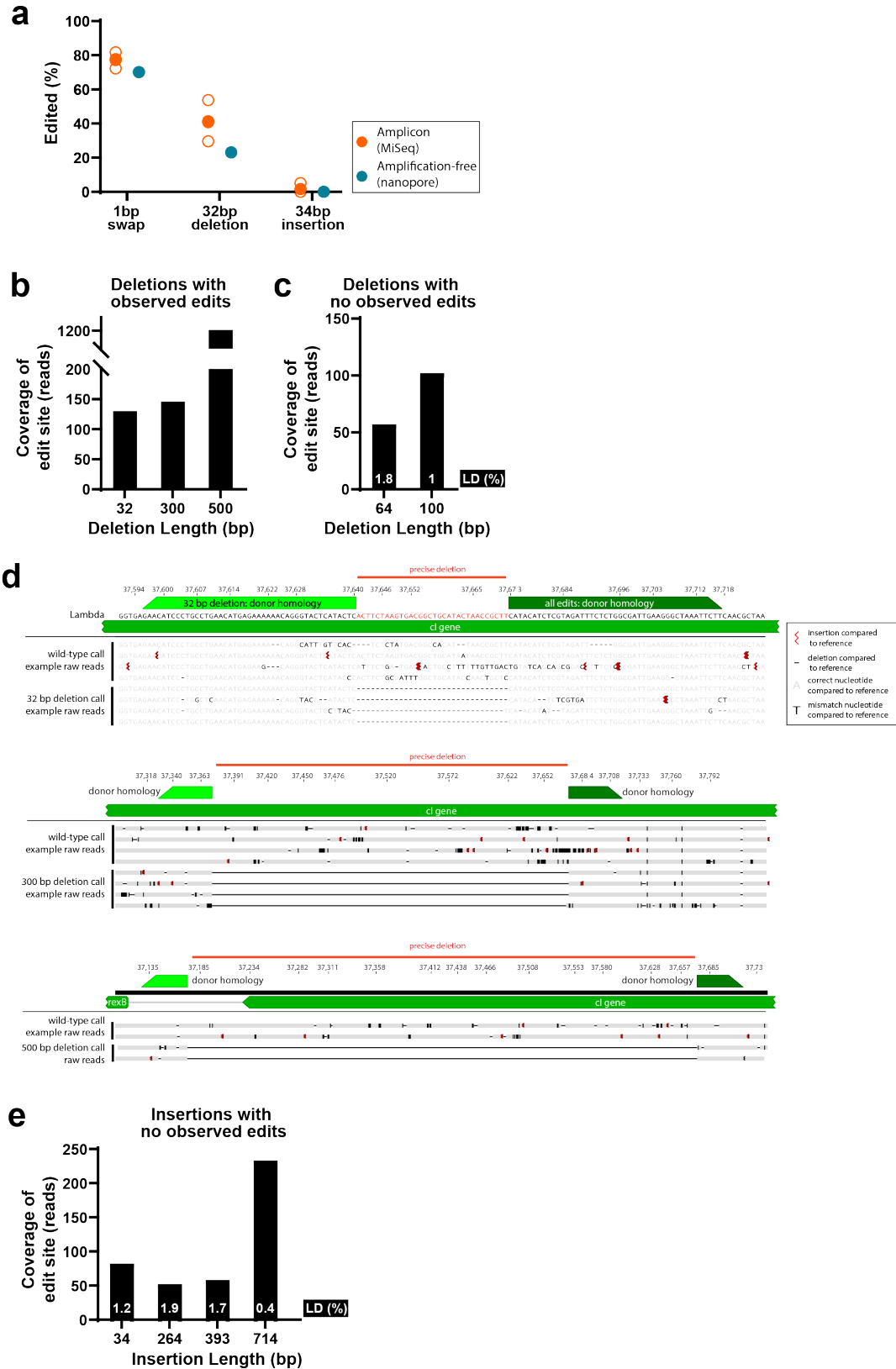
(Figure caption continued from the previous page)

circles). Individual biological replicates are shown. **f.** Editing (%) of lambda, T7, and T5 from the induced, on-target recombitron condition in panel **d**, shown as in **a** ($N = 3$). **g.** Editing (%) of T7 with supplemental expression of lambda genes *gam* or *beta*, shown as in **a** ($N = 3$, one-way ANOVA $P < 0.0001$). **h.** Editing (%) of lambda with supplemental expression of lambda genes *gam* or *beta*, shown as in **a** ($N = 3$, one-way ANOVA $P = 0.0904$). **i.** Editing (%) of T7 with supplemental expression of *E. coli* SSB and lambda genes *gam* or *beta*, shown as in **a** ($N = 3$, one-way ANOVA $P < 0.0001$). **j.** Editing (%) of lambda with supplemental expression of *E. coli* SSB and lambda genes *gam* or *beta*, as in **a** ($N = 3$, one-way ANOVA $P < 0.0001$). **k.** Editing (%) of lambda at site 14126 (F) compared to editing with supplemental expression of T5 SSB, shown as in **a** ($N = 3$). **l.** Editing (%) of T7 at site 22872 (R) compared to editing with supplemental expression of T5 SSB, shown as in **a** ($N = 3$). **m.** Editing (%) of T5 at site 88634 (F) with supplemental expression of *E. coli* SSB, T7 SSB, or T5 SSB, shown as in **a** ($N = 3$). **n.** Editing (%) of phages from the basal collection, Bas46 (A19798T) and Bas47 (A6332G), that contain modified bases with the RT induced (+) or uninduced (-), as in **f** ($N = 3$).



Extended Data Fig 3-2: Accompaniment to Fig 3-2

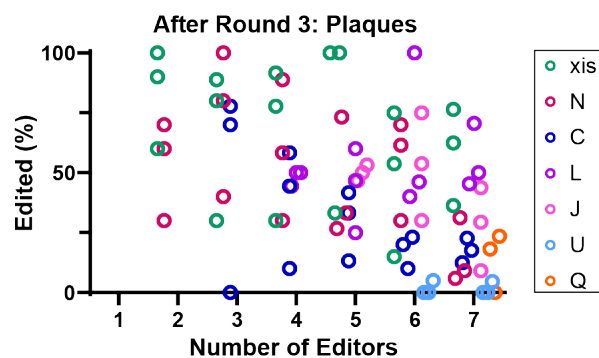
a. Rate of acquiring only the scanning edit in lambda when donors contain both scanning and central edits. (open circles are biological replicates, closed circles are the mean). **b.** PAGE analysis of retron RT-DNA in different *E. coli* strains. **c.** Editing (%) of lambda with editing cultures started at different initial multiplicities of infection (MOI), using MG1655 ($\Delta\text{exo1}, \Delta\text{recJ}$) as the editing host (one-way ANOVA $P < 0.0001$) (open circles are 3 biological replicates, closed circles are the mean). **d.** Editing (%) of lambda with editing cultures incubated at different temperatures, using MG1655 ($\Delta\text{exo1}, \Delta\text{recJ}$) as the editing host (unpaired, two-sided t -test $P = 0.0013$) (open circles are 3 biological replicates, closed circles are the mean).



Extended Data Fig 3-3: Accompaniment to Fig 3-3
(Figure caption continued on the next page)

(Figure caption continued from the previous page)

a. Comparison of edited phages measure by amplicon (Illumina) or amplification-free (Oxford Nanopore) sequencing. Open orange circles represent biological replicates of amplicon data and filled orange circle represents the mean. Filled blue circle represents the aggregate nanopore data from three replicates. **b.** Coverage of the editing site in long-read nanopore sequencing for deletions in which we observe editing. **c.** Coverage of the editing site in long-read nanopore sequencing for deletions in which we do not observe any edits. Estimated limit of detection for these samples is calculated by dividing 100 by the coverage of the site. **d.** Examples of nanopore reads for different deletion conditions. **e.** Coverage of the editing site in long-read nanopore sequencing for large insertions, for which we do not observe any edits. Estimated limit of detection for these samples is calculated by dividing 100 by the coverage of the site.



Extended Data Fig 3-4: **Accompaniment to Fig 3-4**

Editing (%) from Sanger sequencing of plaques at each site from mixed recombitron cultures after 3 rounds of editing. Three biological replicates are shown in open circles for each site, clustered over the number of recombitrons used.

3.7 Supplemental Files

Supplementary_Information_Chapter3.pdf

This PDF file contains:

- Supplementary Fig 3-1: Uncropped gel from **Extended Data Fig 3-2b**
- Supplementary Table 3-1: Statistical details of this study
- Supplementary Table 3-3: Primers used in this study

Supplementary_Table2_Chapter3.xlsx

This Excel file contains recombitron plasmid details.

Chapter 4 Chapter 4 High throughput variant libraries and machine learning yield design rules for retron gene editors

4.1 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 *S. 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.

4.2 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 (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^{1,2} and results in filamentous sequestration of the toxic effector protein¹. A phage-encoded DNA cytosine methyltransferase triggers abortive infection by methylating the Retron-Eco1 reverse-transcribed DNA and results in nucleoside derivative depletion¹. The editron system uses only the reverse transcriptase 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 RT-DNA, or multi-copy single-stranded DNA (msDNA) that has been used as: donor DNA for precise editing in bacteria¹⁻⁶, bacteriophage¹⁻³, plants^{1,2}, and eukaryotes¹⁻⁶; DNA barcodes to record molecular events^{1,2}; DNA containing transcription factor motifs for transcription factor activity attenuation¹; DNA aptamers¹; and DNAzymes for mRNA cleavage¹.

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 ledger¹⁻³. 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 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.

4.3 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 multi-copy single-stranded DNA (msDNA) in cells (**Fig 4-1a**). 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 recording¹⁻³, we set out to systematically understand how variations in ncRNA sequence and structure impact msDNA production in *E. coli*. We constructed a 3,443 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 complementarity 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 Δ Retron-Eco1), and expressed along with the Retron-Eco1 RT for 5 hours, after which we collected msDNA for quantification. All variant sequences are included in **Supplementary Table 4-9-18**.

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 sequence¹⁻³. Briefly, we (1) purified short single-stranded DNA (ssDNA) using a QIAGEN Midiprep Plasmid Plus Kit followed by a Zymo ssDNA Clean & Concentrator Kit, (2) treated the resulting ssDNA with Dbr1 to remove the 2'-5' linkage between the msDNA and ncRNA, (3) extended the debranched ssDNA with a single

polynucleotide using template-independent polymerase (TdT), (4) generated a complementary strand using a primer consisting of the complementary single polynucleotide and an Illumina adaptor, (5) ligated an adaptor to the other end of the now double-stranded msDNA, and lastly (6) 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 (**Fig 4-1b**). 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 4-1c 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-msDNA¹; and around the previously-known UUU putative recognition loop for the Retron-Eco1 reverse transcriptase¹ (**Fig 4-1c, Extended Data Figs 4-1b-e**).

We also analyzed deletions scanning across the Retron-Eco1 ncRNA that varied in length from 1 to 5 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 tolerated than smaller

deletions in the critical region of the P3 stem-loop (**Fig 4-1d, Extended Data Figs 4-1f-j**).

We also assessed 1-, 3-, and 5-nucleotide insertions scanning across the Retron-Eco1 ncRNA (**Fig 4-1e**). While small insertions were slightly more tolerated than small deletions (**Extended Data Fig 4-1g, Extended Data Figs 4-1k-l**), larger insertions in the *msr* region resulted in undetectable levels of msDNA (**Extended Data Fig 4-1m**). 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 **Fig 4-1f**. 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 (**Fig 4-1g**). Breaking

stem-loop P2 complementarity closer to the base reduced msDNA production and restoring complementarity restored msDNA production (**Fig 4-1h**). Breaking stem-loop P3 complementarity reduces msDNA production, but restoring complementarity with an alternate sequence does not restore msDNA production (**Fig 4-1i**). 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 reverse transcriptase recognition (**Fig 4-1j**). 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 (**Fig 4-1k, Extended Data Fig 4-2**).

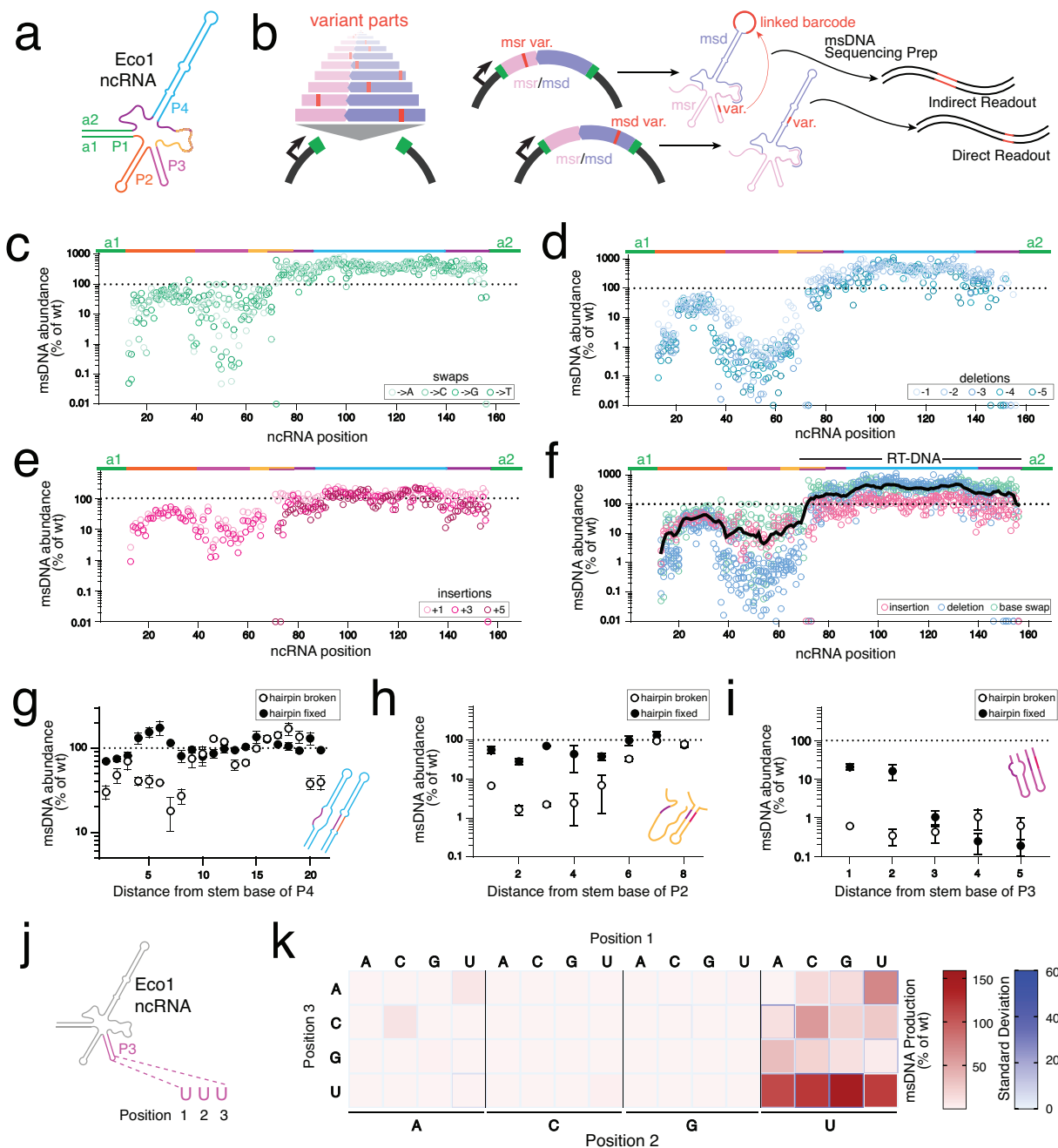


Figure 4-1: msDNA production of Retron-Eco1 variant libraries in *E. coli*.

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*, single-stranded DNA 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 (Figure caption continued on the next page)

(Figure caption continued from the previous page)

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 to 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 to the wild-type stem except position 7 compared to 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 to 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 reverse transcriptase recognition motif UUU in the terminal loop of stem P3. **k.** msDNA production of every permutation of Retron-Eco1 reverse transcriptase 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 with on the red-white colorbar, while the standard deviation is represented by the blue around the squares of the heatmap. 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).

Machine Learning on Libraries Reveals Novel Variables to Increase msDNA Production

Though we tested ~3,400 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) (**Fig 4-2a-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 (**Fig 4-2c**).

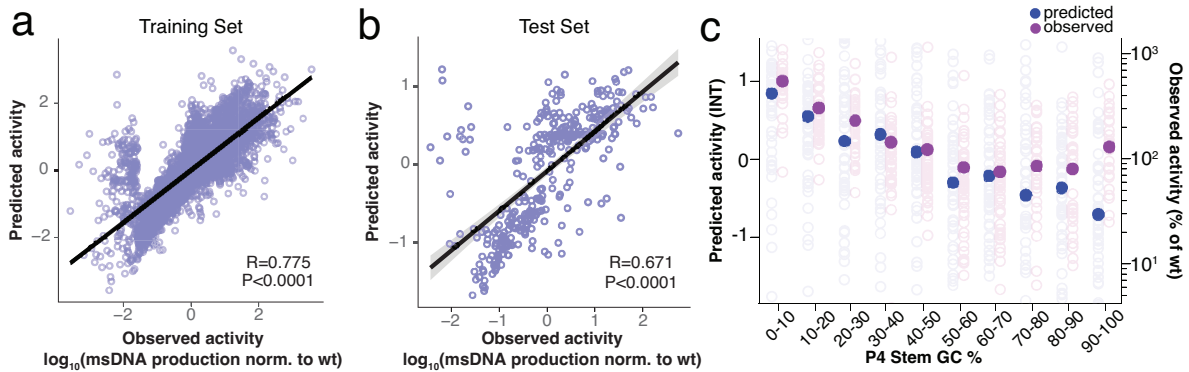


Figure 4-2: Machine learning on variant libraries guides novel predictors of msDNA production.

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 vs. observed activity annotated on the plot. **b.** Machine learning algorithm performance on held-out test data. Each open circle represents an individual ncRNA sequence. Linear regression R and P -values of ML predicted activity vs. observed activity annotated on the plot. **c.** Predicted (blue) and experimentally-determined (purple) 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 $=0.0069$.

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 guide RNA 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 yeast¹ or as an editron¹ 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, guide RNA, 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: 5 donor lengths (54, 64, 78, 94, and 112 nucleotides), 5 homology arm symmetries about the edit site per donor length, msDNA donors that are complementary to the target or non-target strand, and 5 different cut sites (-16, -8, 0, +8, +16 relative to barcode insertion point), leading to 175 donor/gRNA combinations per site (**Fig 4-3a**). We then combinatorially combined these donor/gRNA variants with 25 different ncRNA chassis: wild-type -Eco1 ncRNA, CRISPEY ncRNA⁵⁵, the 13 best-performing structural variants from the *E. coli* variant libraries, and 10 *de novo* predicted ncRNAs from the machine learning algorithm. In all, we tested 4,275 variants per site. All variant sequences are included in **Supplementary Table 4-6-8**.

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 (**Fig 4-3b**). 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 (**Fig 4-3c**). After transformation of the editing libraries into yeast, editing was performed for 48 hours 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 hours 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 hours of editing; for barcodes in the plasmid pool: the reads as summed over samples taken at 0, 24, and 48 hours of editing). This is to integrate the plasmid barcode pool over the entire editing period. Plasmid barcode read count was stable over the 48 hours of editing (**Extended Data Fig 4-3**). 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 **Extended Data Figure 4-4**). 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, 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 **Fig 4-3d**. We then took the median of each biological replicate of each site, based on the distribution on the right of **Fig 4-3d**, 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 to the matched non-target strand donors (**Fig 4-3e**). Both target strand and non-target strand donors have about 50% functional editor variants, as other parameters also influence if an editor is functional (**Extended Data Fig 4-4a**). When examining if 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 if the cut is positioned on the 5' or 3' side of the edit (**Extended Data Fig 4-5**).

We analyzed the effect of cut site positioning relative to insertion point by using Cas9 spacer sequences 8 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 (**Extended Data Fig 4-4b**). 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 to 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 from the insertion point resulted in lower frequency of precise editing (~40-45% efficiency) (**Fig 4-3f**). 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 to the 94 nucleotide donor (**Fig 4-3g**). The percentage of working donors per donor length also increased with donor length (**Extended Data Fig 4-4c**).

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 **Extended Data Fig 4-6**. 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 to centered (**Fig 4-3h**). We observed similar but not identical results at other donor lengths (**Extended Data Figure 4-6**), 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 **Extended Data Fig 4-7**.

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 ML chassis 8 and 9) (**Fig 4-3i**). 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 **Extended Data Figure 4-9**). Specific machine learning chassis structures and sequences can be found in Supplementary Figure 8. We found no evidence of a difference in the percentage of working donors across ncRNA chassis (**Extended Data Figure 4-10**).

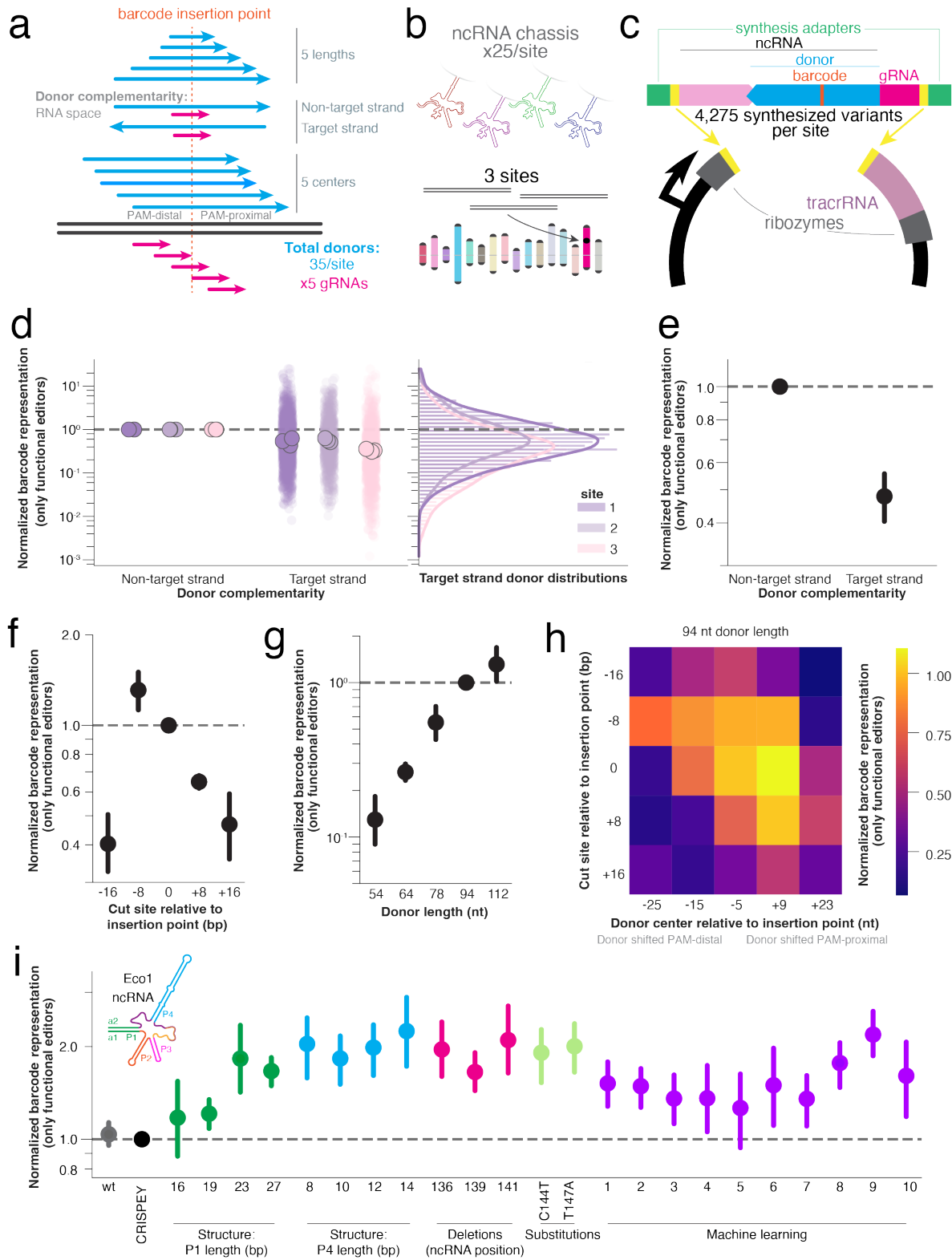


Figure 4-3: **Precise editing of retron *Eco1* editing variant libraries in *S. cerevisiae*.**
(Figure caption continues on the following page)

(Figure caption continued from the previous page)

a. HDR donor variant schematics and gRNA variants, with 5 donor lengths, 2 donor directions relative to the gRNA, and 5 donor centers relative to edit and cut position for a total of 50 donors per editing site. There are 5 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 4,275 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 plotted in different colors, and each biological replicate of a site is summarized by the median (left) of the distribution of variants (right). **e.** Data in Fig 4-3d summarized as the mean of all sites and all biological replicates (closed circle) (\pm standard deviation), with target-strand-homologous donors editing at significantly lower frequencies (1-sample t-test; $P < 0.0001$). **f.** Barcode representation of cut sites normalized to the cut site at the barcode insertion site (\pm standard deviation), with cut sites at -16, +8, and +16 editing at significantly lower frequencies (1-sample t-test, Bonferroni correction for multiple comparisons; $P < 0.0001$, $P < 0.0001$, $P < 0.0001$ respectively, all other comparisons non-significant). **g.** Barcode representation of donor lengths normalized to 94 nucleotide donor length (\pm standard deviation), with donor lengths < 94 nucleotides editing at significantly lower frequencies (1-sample t-test, Bonferroni correction for multiple comparisons; $P < 0.0001$, $P < 0.0001$, $P < 0.01$ respectively, all other comparisons non-significant). **h.** Heat map of normalized barcode representation of cut site vs. 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 (\pm 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 (1-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).

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 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 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 4-5**.

The editron plasmids were transfected into HEK293T cells containing an integrated doxycycline-inducible Cas9, whose expression was induced 24 hours before transfection. Cells were collected three days after transfection and sorted via FACS to only include live and transfected cells, eliminating any variability due to transfection efficiency (**Fig 4-4a**). 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 3 tested gRNAs, indicating highest cutting efficiency (**Fig 4-4b**). 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 (**Fig 4-4c-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 $\alpha 1/\alpha 2$ length increased editing compared to wild-type $\alpha 1/\alpha 2$ length. Excitingly, ML modifications enabled successful editing despite only 30% sequence similarity to wild-type, demonstrating the flexibility of the region (**Fig 4-4e**). Next, we sought to determine the ideal positioning of both the edit and the donor relative to a set cut site. We tested 3 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 (**Fig 4-4f**). 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 (**Fig 4-4g**).

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 towards 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 (**Fig 4-4h**).

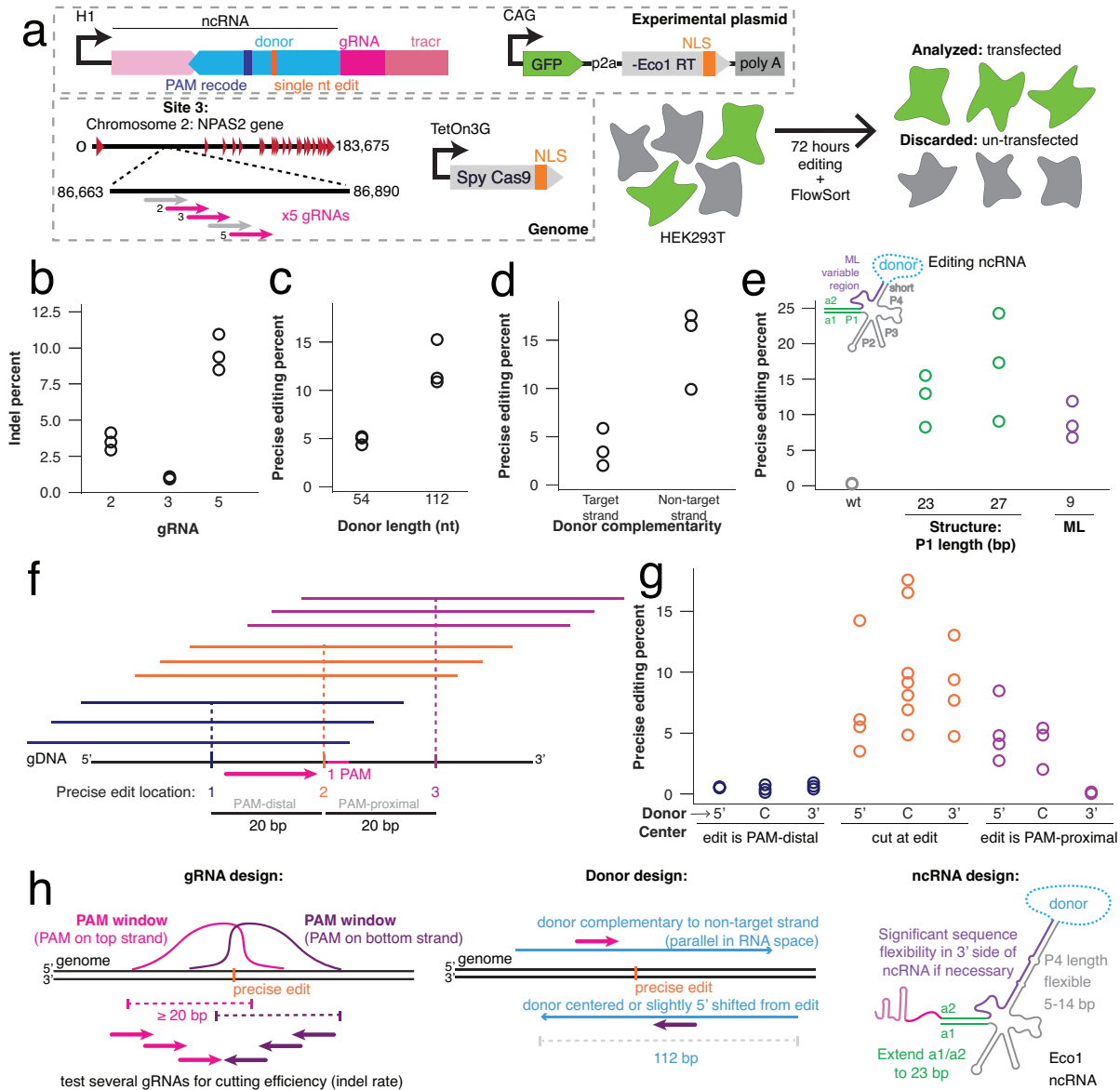


Figure 4-4: Validating yeast editing libraries with individual human variants.

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 4-3**). The HEK293T line also had semi-randomly-integrated *S. pyogenes* Cas9 by PiggyBac transposase under a dox-inducible promoter and a C-terminal NLS. 72 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. (Figure caption continues on the following page)

(Figure caption continued from the previous page)

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 9 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, $P=0.0016$ 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.

4.4 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 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 transcription⁸⁹ and the UAGC sequence which includes the priming guanosine¹. 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 reverse transcriptase 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 resolution¹. 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 insertions¹ than are possible without adding additional proteins, such as Bxb1 recombinase and a recombination donor¹ 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^{1,2}. 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 oligonucleotides¹. 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 results¹⁻³. After cleavage, Cas9 releases the non-target strand, after which a 3'-

to-5' exonuclease, like Klenow, degrades the 3' flap¹. 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 studies¹. This is contrary to other ssODN studies, which find that strand polarity preference depends on cut position relative to the edit¹. 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 claim¹.

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.

4.5 Methods

Biological replicates were taken from distinct samples, not the same sample measured repeatedly. For *E. coli* variant libraries, each biological replicate is an independent electroporation and expression of the libraries into the strain bSLS.114. For *S. 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 4-1**.

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 Wanner¹. This knock-out cassette was amplified from pKD3, adding homology arms to the Retron-Eco1 locus with PCR primers, and electroporated into BL21-AI cells with the Lambda Red recombination machinery (pKD46). After selecting clones on 10 µg ml⁻¹ 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 PqCI sites and a PCR-amplified linear pSCL039⁴². Yeast plasmids containing the three editing sites in the HIS3 site were based off pZS.157¹. These three variants (pSCL194: site 1; pSCL195: site 2; pSCL368: site 3) all contain galactose-inducible Retron-Eco1 RT and *S. 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 transformation¹ of BY4742-. The respective plasmids were linearized using KpnI and transformed into BY4742 for homologous recombination into the HIS3 locus. Clones with selected on SD-HIS media.

All human vectors are derivatives of pSCL.273⁴⁵, 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 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 PqCI. Plasmids were subsequently midi-prepped 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 described⁴².

All strains/lines are listed in **Supplementary Table 4-3**, and all plasmids in **Supplementary Table 4-2**.

Variant library cloning

E. coli variant cloning was done as previously described⁴² using BsaI Type IIS restriction sites and Golden Gate cloning. After high-efficiency cloning and electroporation, variant libraries were miniprepmed 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 midiprepmed 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

E. coli variant libraries were grown overnight and diluted 1:500 into expression media (arabinose and 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 hours 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 minutes, 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 described⁴². Briefly, DNA was purified using a modified miniprep protocol, treated with RNase A/T1 (New England Biolabs), and purified with 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_msdlloop_for_Sequencing_F`, `Eco1_msdlloop_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 4-4**.

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

Yeast gDNA was extracted as previously described⁴². Briefly, cells were lysed in 120 µL lysis buffer (100 mM EDTA pH 8, 50 mM Tris-HCl pH 8, 2% SDS) and boiled for 15 min at 100 C. After cooling the lysate on ice, proteins were precipitated by adding 60 uL of ice-cold 7.5 M ammonium acetate and incubating at -20 C for 10 min. The samples were centrifuged at 17,000g 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 minutes to evaporate all ethanol, and

resuspended in 40 μ L water. Genomic DNA samples for deep-sequencing were then 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 4-4**.

Yeast plasmid DNA was extracted as previously described¹³⁶. The Zymo Yeast Miniprep Kit was scaled up to 500m cells. Briefly, we resuspended yeast in 1 mL digestion buffer and 30 μ L zymolyase, and digested the cell wall for 3 hrs shaking at 900 rpm 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 μ L wash buffer and eluted in 20 μ L pre-warmed ultra-pure nuclease-free H₂O 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 8 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 μ L of the cleaned-up plasmid DNA amplicons for indexing and sequencing on the Illumina MiSeq. Sequencing primers are listed in **Supplementary Table 4-4**.

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 DMEM + GlutaMax supplement (ThermoFisher 10566016) + 10% HI-FBS. 6-well cultures were transiently transfected with 7.32 ug of plasmid per well using Lipofectamine 3000 (ThermoFisher). 24 hrs after transfection, doxycycline was refreshed and cultures were passaged into T-25 flasks to be grown for an additional 48 hrs. Three days after transfection, cells were collected for FACS sorting. DAPI dye was added to stain for live/dead and cells were gated 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.

2 μ l of the gDNA was used as template in 25- μ l 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 described⁴². 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-hr 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-hr genome (editing site) samples, and the 0-, 24-, and 48-hr 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 software used in the analysis of this paper is available on GitHub.

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 will be made available on GitHub prior to peer-reviewed publication: https://github.com/Shipman-Lab/retron_ncRNA_ML_libraries/tree/master and at DOI: 10.5281/zenodo.14058431

Acknowledgements

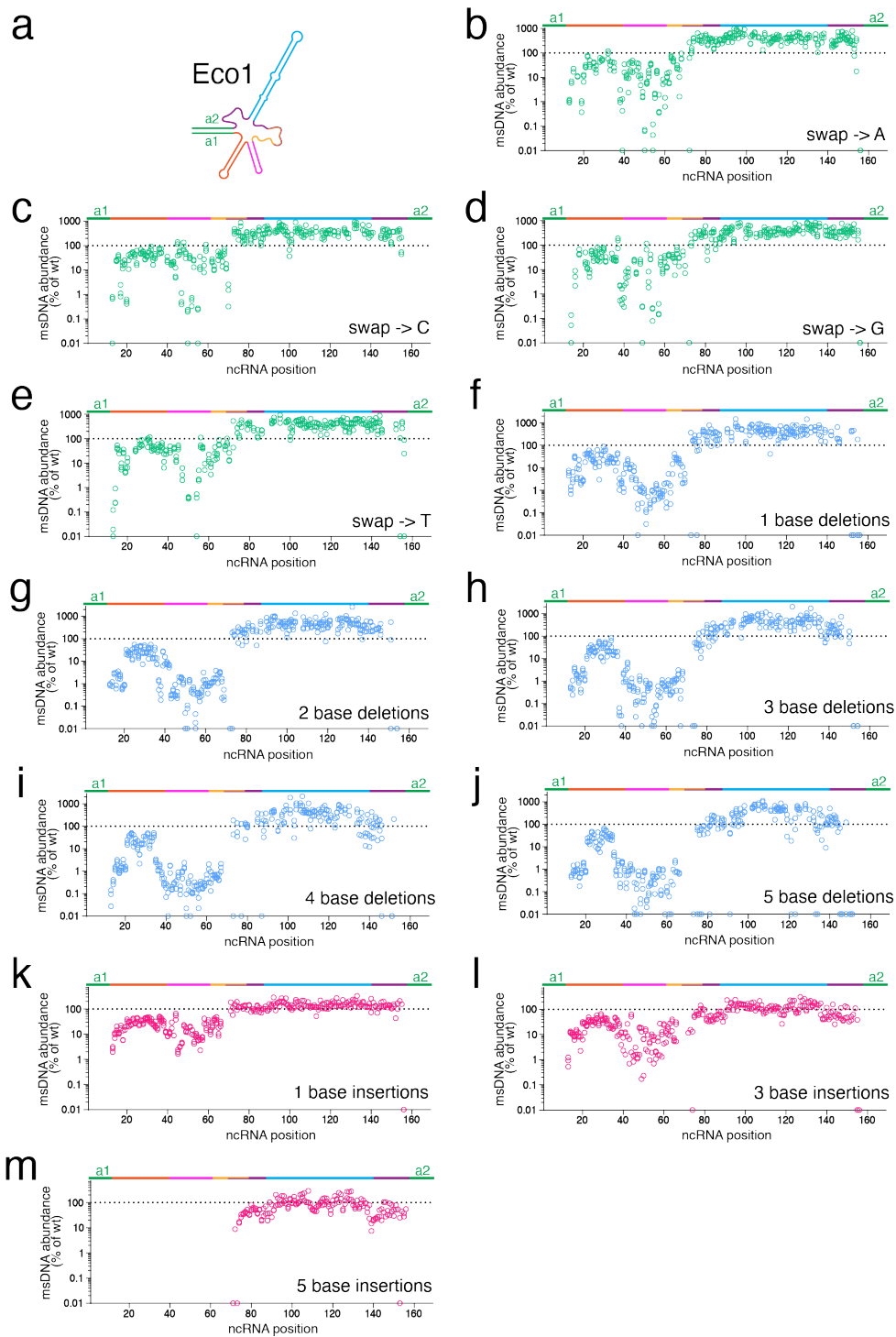
Work was supported by funding from the National Science Foundation (MCB 2137692), the National Institute of Biomedical Imaging and Bioengineering (R21EB031393), and the W.M. Keck Foundation. S.L.S. is a Chan Zuckerberg Biohub - San Francisco Investigator and acknowledges additional funding support from the Pew

Biomedical Scholars Program. K.D.C. is supported by a National Science Foundation Graduate Research Fellowship and a UCSF Discovery Fellowship. We would also like to acknowledge the Gladstone Flow Cytometry Core, which performed the fluorescence-activated cell sorting of the human cells in Figure 4.

Contributions

Work was supported by funding from the National Science Foundation (MCB 2137692), the National Institute of Biomedical Imaging and Bioengineering (R21EB031393), and the W.M. Keck Foundation. S.L.S. is a Chan Zuckerberg Biohub - San Francisco Investigator and acknowledges additional funding support from the Pew Biomedical Scholars Program. K.D.C. is supported by a National Science Foundation Graduate Research Fellowship and a UCSF Discovery Fellowship. We would also like to acknowledge the Gladstone Flow Cytometry Core, which performed the fluorescence-activated cell sorting of the human cells in Figure 4.

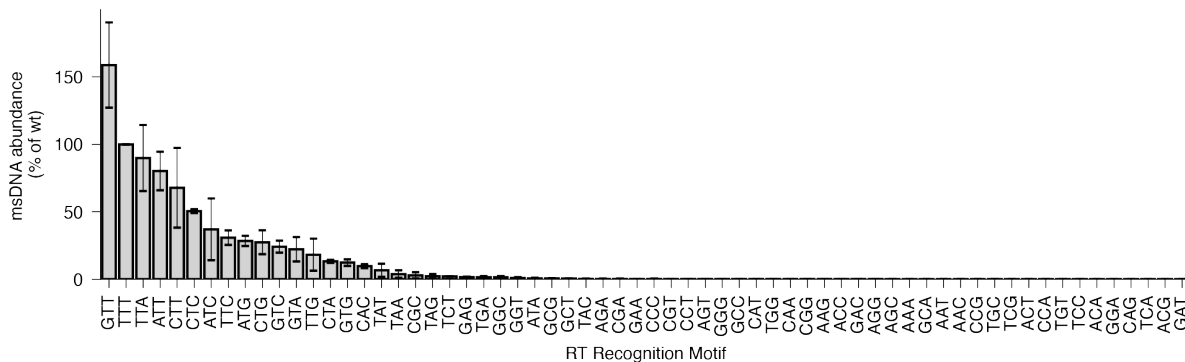
4.6 Supplementary Information



Extended Data Fig 4-1: **Substitution, deletion, and insertion sub-library msDNA production in *E. coli*.** (Figure caption continues on the following page)

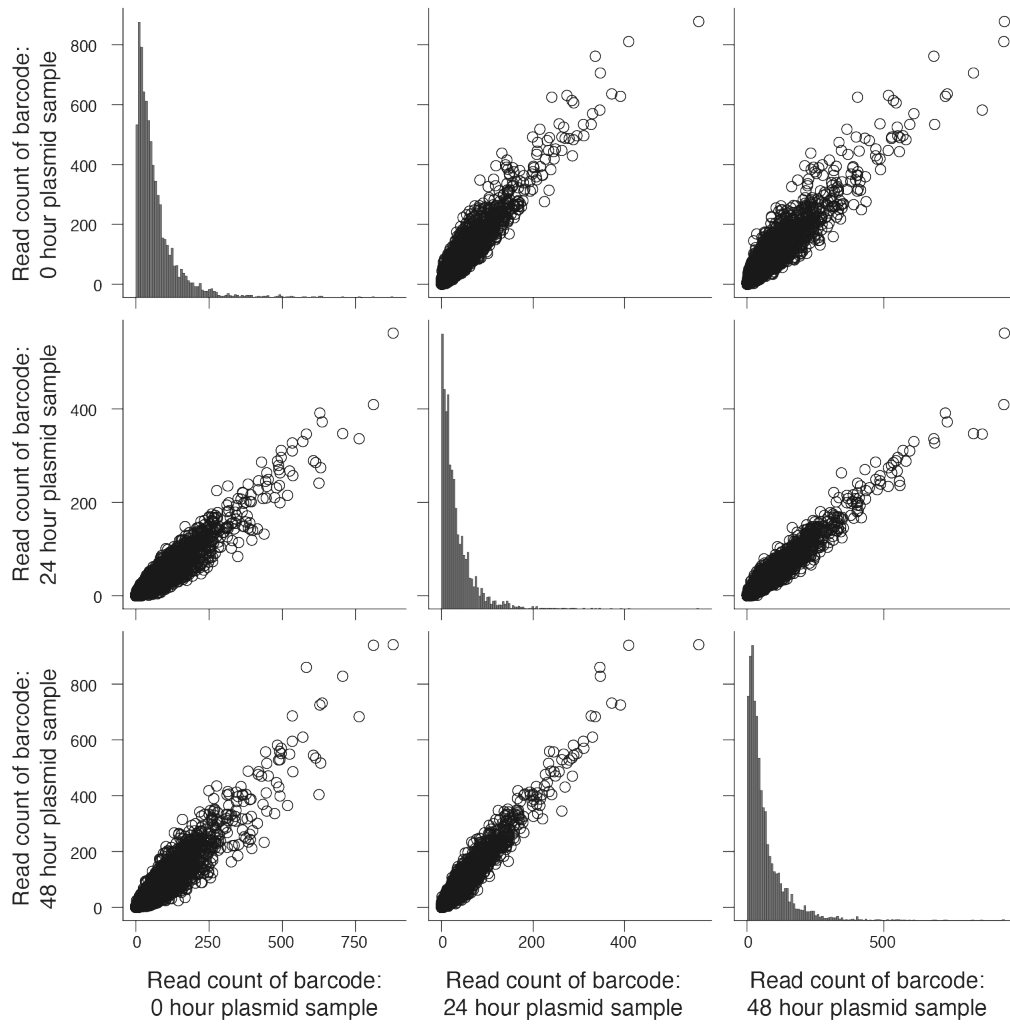
(Figure caption continued from the previous page)

a. Retron-Eco1 ncRNA structure. **b.** msDNA production of N→A nucleotide swap, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **c.** msDNA production of N→C nucleotide swap, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **d.** msDNA production of N→G nucleotide swap, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **e.** msDNA production of N→T nucleotide swap, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **f.** msDNA production of single-base deletions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **g.** msDNA production of two-base deletions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **h.** msDNA production of 3-base deletions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **i.** msDNA production of 4-base deletions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **j.** msDNA production of 5-base deletions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **k.** msDNA production of single-base insertions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **l.** msDNA production of 3-base insertions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate. **m.** msDNA production of 5-base insertions, starting at a specified ncRNA position relative to wild-type msDNA. Each open circle represents an individual biological replicate.



Extended Data Fig 4-2: msDNA production of every permutation of Retron-Eco1 reverse transcriptase recognition motif as a barplot.

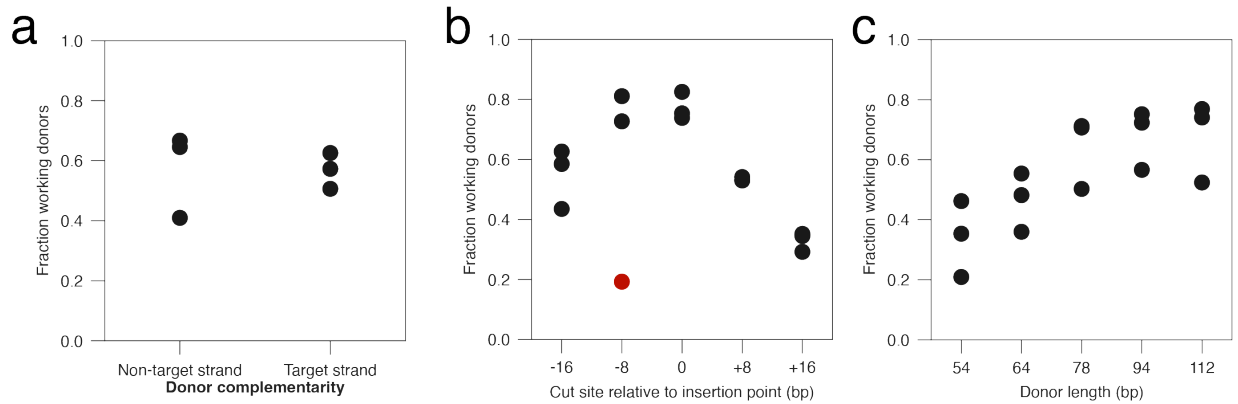
msDNA production is relative to wild-type msDNA abundance. Each bar represents the mean of three biological replicates. There is a significant effect of the RT recognition motif (one-way ANOVA, $P < 0.0001$), with every permutation significantly different than the wild-type UUU ($P < 0.0001$) except UUA and AUU ($P = 0.8991$ and $P = 0.0551$, respectively) (Dunnett's, corrected). Data is the same as is presented in Fig 4-1k.



Site 1: replicate 1 correlation of plasmid read counts over the course of the experiment

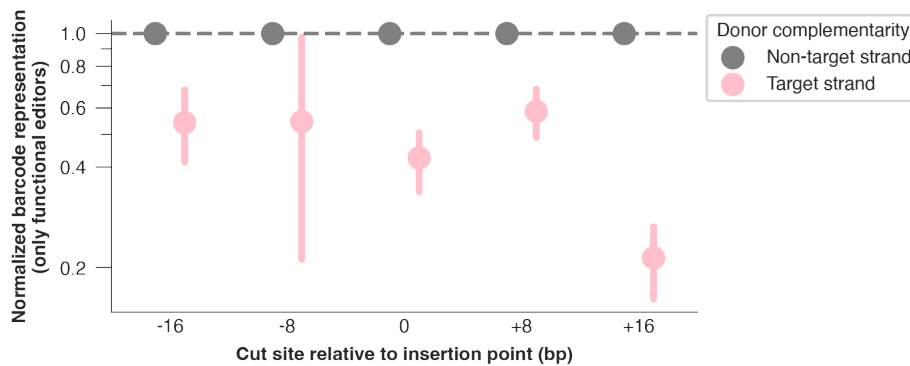
Extended Data Fig 4-3: Correlation in plasmid read counts over an example 48-hr editing window in *S. cerevisiae*.

Correlation between individual plasmid barcode read counts at 0 hr, 24 hr, and 48 hr of editing for the first biological replicate of the site 1 library. Each open circle represents an individual barcode read count.



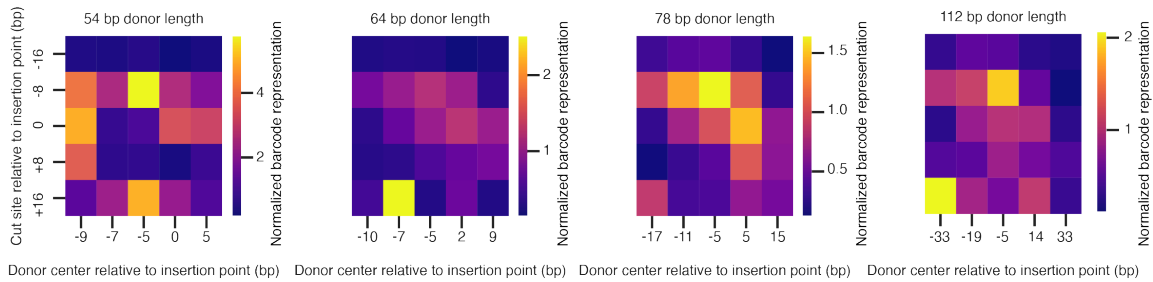
Extended Data Fig 4-4: Fraction working editors for different editing variables.

a. Fraction working donors that are complementary to the non-target strand vs. target strand when reverse-transcribed into donor DNA. Each closed circle represents the mean of the three biological replicates for that site. **b.** Fraction working donors across all tested cut sites. Each closed circle represents the mean of the three biological replicates for that site. The red closed circle represents the cut site excluded from the analysis, as the fraction working donors is below 20%. **c.** Fraction working donors across all tested donor lengths. Each closed circle represents the mean of the three biological replicates for that site.



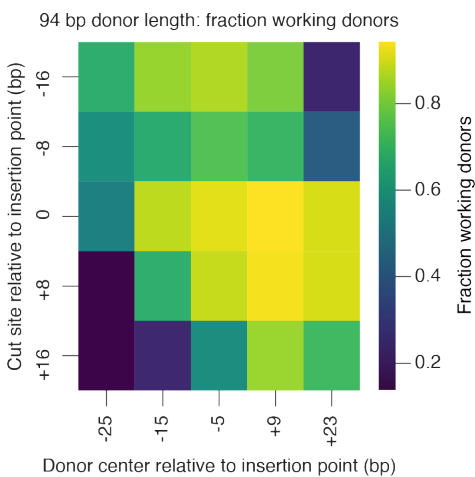
Extended Data Fig 4-5: Normalized barcode representation of target and non-target strand donors broken out by cut site.

All target-strand-homologous gRNA/donor variants' barcode representation normalized against its non-target strand homologous gRNA/donor variant for every cut site (\pm standard deviation), with all other variables held constant (chassis, donor length, and center). The variants for each site are plotted in different colors, and each biological replicate of a site is summarized by the median (left) of the distribution of variants (right). Data is the same as presented in Fig 4-3e.



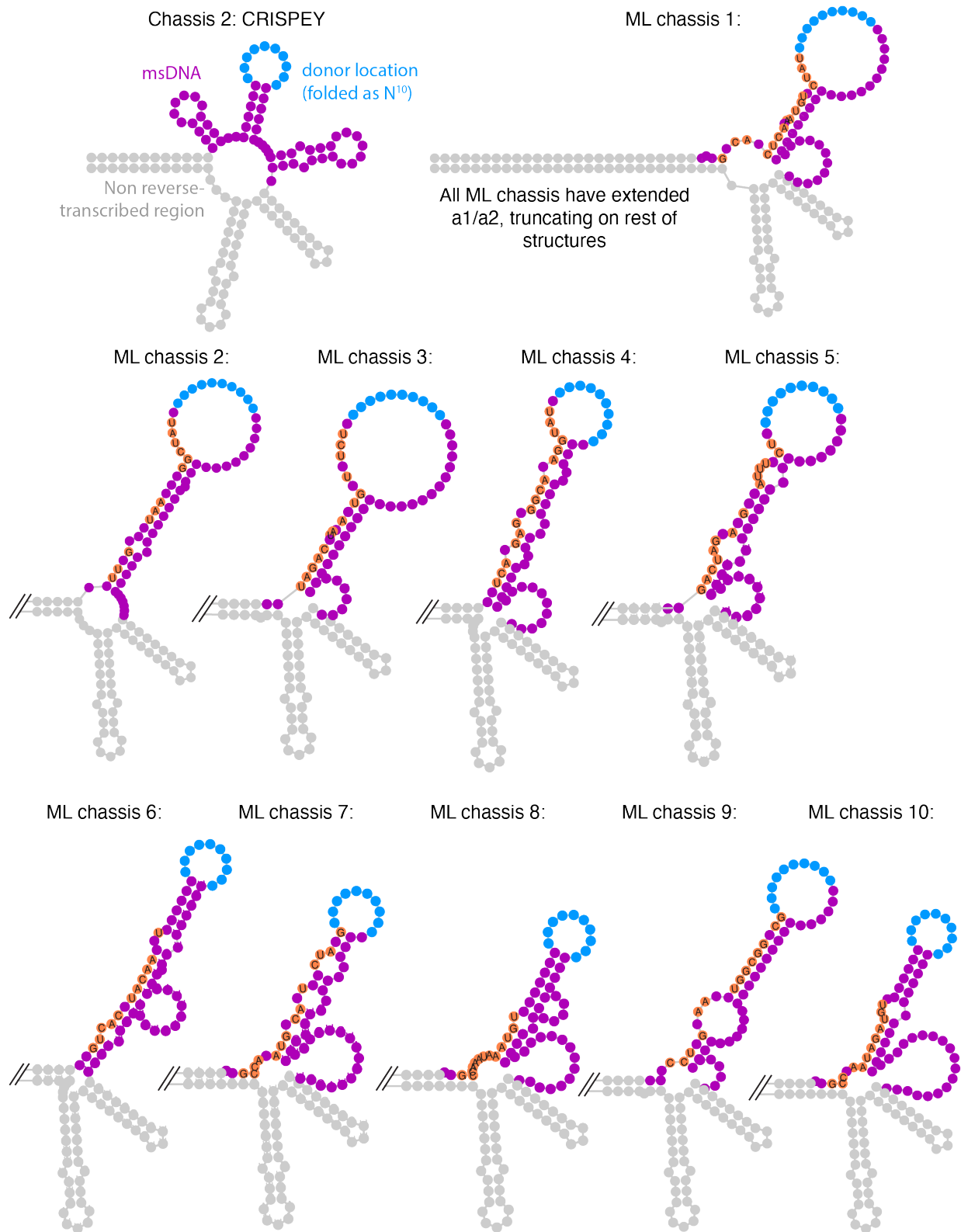
Extended Data Fig 4-6: Normalized barcode representation of donors of varying cut sites vs. donor centers in *S. cerevisiae*.

Heat map of normalized barcode representation of cut site vs. donor center (54, 64, 78, and 112 nucleotide donor length), normalized to the cut site at the barcode insertion site, and donor center of -5 bp from the barcode insertion site. Each square represents the mean of all biological replicates across all sites.



Extended Data Fig 4-7: Standard deviation of normalized barcode representation of donors of varying cut sites vs. donor centers in *S. cerevisiae*.

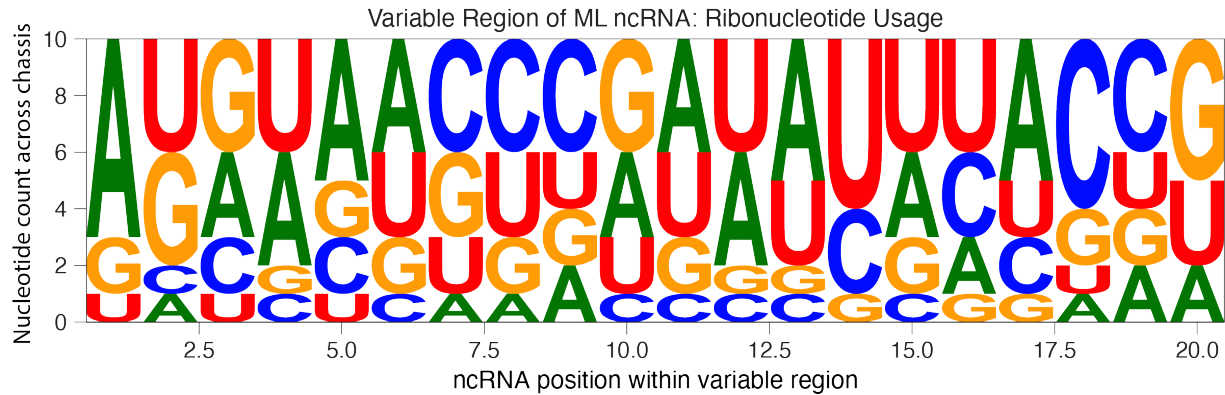
Heat map of the standard deviation of normalized barcode representation of cut site vs. donor center (94 nucleotide donor length), normalized to the cut site at the barcode insertion site, and donor center of -5 bp from the barcode insertion site. Each square represents the standard deviation of all biological replicates across all sites.



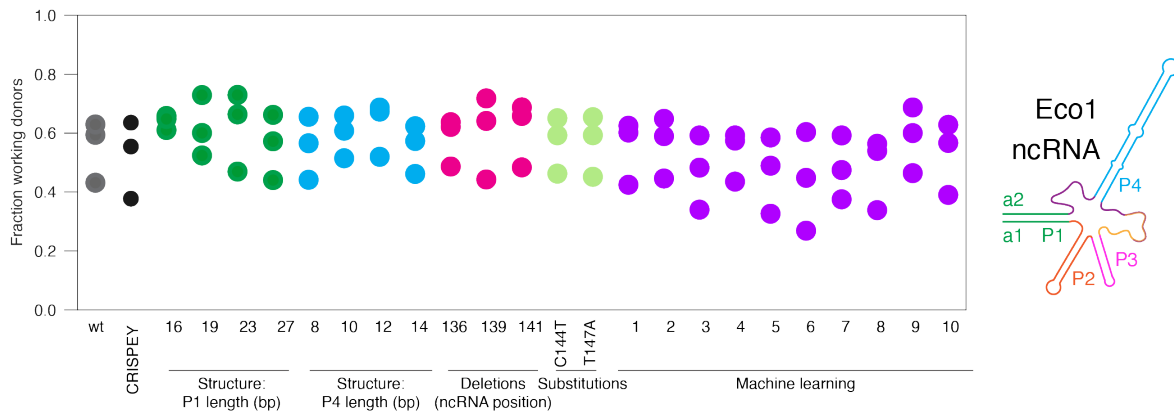
Extended Data Fig 4-8: ncRNA structure and sequence of top machine learning ncRNA chassis.
 CRISPEY and ML chassis were folded using RNAfold (Institute for Theoretical Chemistry, University of
 (Figure caption continues on the following page)

(Figure caption continued from the previous page)

Vienna webtool) using N_{10} to stand in for the variable donor region (light blue). *msr* annotated in grey and *msDNA* annotated in purple. Nucleotides with changes from the CRISPEY reference are highlight in red with the nucleotide identity annotated in black.



Extended Data Fig 4-9: Usage of ribonucleotides in ML ncRNA chassis across variable region. Ribonucleotide height scaled with usage, created by the Python logomaker package.



Extended Data Fig 4-10: Fraction working editors across ncRNA chassis.

Fraction of working donors across all ncRNA chassis. Each closed circle represents the mean of the three biological replicates for that site.

4.7 Supplemental Files

Supplementary_Tables_Chapter4_1-5.xlsx

This PDF file contains:

- Supplementary Table 4-1: Statistical analysis
- Supplementary Table 4-2: Plasmids used in this study
- Supplementary Table 4-3: Strains used in this study
- Supplementary Table 4-4: Primers used in this study
- Supplementary Table 4-5: Donors used in this study

Supplementary_Tables_Chapter4_6-8.xlsx

- Supplementary Table 4-6: Donors used for Site 1 *S. cerevisiae* library
- Supplementary Table 4-7: Donors used for Site 2 *S. cerevisiae* library
- Supplementary Table 4-8: Donors used for Site 3 *S. cerevisiae* library

Supplementary_Tables_Chapter4_9-18.xlsx

- Supplementary Table 9: ncRNAs used for **Fig 4-1c,f** and **Extended Data Fig 4-1b-e** (*msd*)
- Supplementary Table 10: ncRNAs used for **Fig 4-1c,f** and **Extended Data Fig 4-1b-e** (*msr*)
- Supplementary Table 11: ncRNAs used for **Fig 4-1d,f** and **Extended Data Fig 4-1f-j** (*msd*)
- Supplementary Table 12: ncRNAs used for **Fig 4-1d,f** and **Extended Data Fig 4-1f-j** (*msr*)

- Supplementary Table 13: ncRNAs used for **Fig 4-1e,f** and **Extended Data Fig 4-1k-m** (*msd*)
- Supplementary Table 14: ncRNAs used for **Fig 4-1e,f** and **Extended Data Fig 4-1k-m** (*msr*)
- Supplementary Table 15: ncRNAs used for **Fig 4-1g**
- Supplementary Table 16: ncRNAs used for **Fig 4-1h**
- Supplementary Table 17: ncRNAs used for **Fig 4-1i**
- Supplementary Table 18: ncRNAs used for **Fig 4-1k**

Chapter 5 References

1. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
2. Eden, E., Geva-Zatorsky, N., Issaeva, I., Cohen, A., Dekel, E., Danon, T., Cohen, L., Mayo, A. & Alon, U. Proteome Half-Life Dynamics in Living Human Cells. *Science* **331**, 764–768 (2011).
3. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).
4. Nakata, A., Amemura, M. & Makino, K. Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J. Bacteriol.* **171**, 3553–3556 (1989).
5. Mojica, F. J. M., Juez, G. & Rodríguez-Valera, F. Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified *Pst* I sites. *Mol. Microbiol.* **9**, 613–621 (1993).
6. Mojica, F. J. M., Ferrer, C., Juez, G. & Rodríguez-Valera, F. Long stretches of short tandem repeats are present in the largest replicons of the *Archaea Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol. Microbiol.* **17**, 85–93 (1995).
7. Jansen, Ruud., Embden, Jan. D. A. V., Gaastra, Wim. & Schouls, Leo. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).

8. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).
9. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
10. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J. Mol. Evol.* **60**, 174–182 (2005).
11. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
12. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).
13. Marraffini, L. A. & Sontheimer, E. J. Invasive DNA, Chopped and in the CRISPR. *Structure* **17**, 786–788 (2009).
14. Mali, P. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
15. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. & Doudna, J. RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).

16. Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).
17. Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
18. Gelman, D., Yerushalmy, O., Alkalay-Oren, S., Rakov, C., Ben-Porat, S., Khalifa, L., Adler, K., Abdalrhman, M., Copenhagen-Glazer, S., Aslam, S., *et al.* Clinical Phage Microbiology: a suggested framework and recommendations for the in-vitro matching steps of phage therapy. *Lancet Microbe* **2**, e555–e563 (2021).
19. Würstle, S., Lee, A., Kortright, K. E., Winzig, F., An, W., Stanley, G. L., Rajagopalan, G., Harris, Z., Sun, Y., Hu, B., *et al.* Optimized preparation pipeline for emergency phage therapy against *Pseudomonas aeruginosa* at Yale University. *Sci. Rep.* **14**, 2657 (2024).
20. Strathdee, S. A., Hatfull, G. F., Mutalik, V. K. & Schooley, R. T. Phage therapy: from biological mechanisms to future directions. *Cell* **186**, 17–31 (2023).
21. Pires, D. P., Cleto, S., Sillankorva, S., Azeredo, J. & Lu, T. K. Genetically Engineered Phages: a Review of Advances over the Last Decade. *Microbiol. Mol. Biol. Rev.* **80**, 523–543 (2016).
22. Kiro, R., Shitrit, D. & Qimron, U. Efficient engineering of a bacteriophage genome using the type I-E CRISPR–Cas system. *RNA Biol* **11**, 42–44 (2014).

23. Box, A. M., McGuffie, M. J., O'Hara, B. J. & Seed, K. D. Functional Analysis of Bacteriophage Immunity through a Type I-E CRISPR-Cas System in *Vibrio cholerae* and Its Application in Bacteriophage Genome Engineering. *J. Bacteriol.* **198**, 578–590 (2016).
24. Bari, S. M. N., Walker, F. C., Cater, K., Aslan, B. & Hatoum-Aslan, A. Strategies for editing virulent staphylococcal phages using CRISPR–Cas10. *ACS Synth Biol* **6**, 2316–2325 (2017).
25. Huss, P., Meger, A., Leander, M., Nishikawa, K. & Raman, S. Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. *eLife* **10**, e63775 (2021).
26. Adler, B. A. Broad-spectrum CRISPR–Cas13a enables efficient phage genome editing. *Nat Microbiol* **7**, 1967–1979 (2022).
27. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429–432 (2013).
28. Strotskaya, A. The action of *Escherichia coli* CRISPR–Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic Acids Res* **45**, 1946–1957 (2017).
29. Rong, Z., Zhu, S., Xu, Y. & Fu, X. Homologous recombination in human embryonic stem cells using CRISPR/Cas9 nickase and a long DNA donor template. *Protein Cell* **5**, 258–260 (2014).

30. Roth, T. L., Puig-Saus, C., Yu, R., Shifrut, E., Carnevale, J., Li, P. J., Hiatt, J., Saco, J., Krystofinski, P., Li, H., *et al.* Reprogramming human T cell function and specificity with non-viral genome targeting. *Nature* **559**, 405–409 (2018).
31. Chen, P. J. & Liu, D. R. Prime editing for precise and highly versatile genome manipulation. *Nat. Rev. Genet.* **24**, 161–177 (2023).
32. 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.* Phage-assisted evolution and protein engineering yield compact, efficient prime editors. *Cell* **186**, 3983-4002.e26 (2023).
33. Anzalone, A. V., Gao, X. D., Podracky, C. J., Nelson, A. T., Koblan, L. W., Raguram, A., Levy, J. M., Mercer, J. A. M. & Liu, D. R. Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat. Biotechnol.* **40**, 731–740 (2022).
34. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
35. Kim, Y., Oh, H.-C., Lee, S. & Kim, H. H. Saturation profiling of drug-resistant genetic variants using prime editing. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02465-z.
36. Belli, O., Karava, K., Farouni, R. & Platt, R. J. Multimodal scanning of genetic variants with base and prime editing. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02439-1.
37. Yee, T. & Inouye, M. DNA Isolated from a *Myxococcus xanthus*.

38. Inouye, M. The first demonstration of the existence of reverse transcriptases in bacteria. *Gene* **597**, 76–77 (2017).
39. Mestre, M. R., González-Delgado, A., Gutiérrez-Rus, L. I., Martínez-Abarca, F. & Toro, N. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res.* **48**, 12632–12647 (2020).
40. Bobonis, J. Bacterial retrons encode phage-defending tripartite toxin–antitoxin systems. *Nature* **609**, 144–150 (2022).
41. Millman, A. Bacterial retrons function in anti-phage defense. *Cell* **183**, 1551–1561 (2020).
42. Lopez, S. C., Crawford, K. D., Lear, S. K., Bhattarai-Kline, S. & Shipman, S. L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022).
43. González-Delgado, A., Lopez, S. C., Rojas-Montero, M., Fishman, C. B. & Shipman, S. L. Simultaneous multi-site editing of individual genomes using retron arrays. Preprint at <https://doi.org/10.1101/2023.07.17.549397> (2023).
44. Liu, W., Zuo, S., Shao, Y., Bi, K., Zhao, J., Huang, L., Xu, Z. & Lian, J. Retron-mediated multiplex genome editing and continuous evolution in *Escherichia coli*. *Nucleic Acids Res.* **51**, 8293–8307 (2023).
45. Khan, A. G., Rojas-Montero, M., González-Delgado, A., Lopez, S. C., Fang, R. F., Crawford, K. D. & Shipman, S. L. An experimental census of retrons for DNA production and genome editing. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02384-z.

46. Lim, H., Jun, S., Park, M., Lim, J., Jeong, J., Lee, J. H. & Bang, D. Multiplex Generation, Tracking, and Functional Screening of Substitution Mutants Using a CRISPR/Retron System. *ACS Synth. Biol.* **9**, 1003–1009 (2020).
47. Ellington, A. J. & Reisch, C. R. Efficient and iterative retron-mediated in vivo recombineering in *Escherichia coli*. *Synth. Biol.* **7**, ysac007 (2022).
48. Fishman, C. B., Crawford, K. D., Bhattarai-Kline, S., Poola, D., Zhang, K., González-Delgado, A., Rojas-Montero, M. & Shipman, S. L. Continuous multiplexed phage genome editing using recombitrons. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02370-5.
49. Goren, M. G., Mahata, T. & Qimron, U. An efficient, scarless, selection-free technology for phage engineering. *RNA Biol.* **20**, 830–835 (2023).
50. Molla, K. A., Shih, J., Wheatley, M. S. & Yang, Y. Predictable NHEJ Insertion and Assessment of HDR Editing Strategies in Plants. *Front. Genome Ed.* **4**, 825236 (2022).
51. Jiang, W., Sivakrishna Rao, G., Aman, R., Butt, H., Kamel, R., Sedeek, K. & Mahfouz, M. M. High-efficiency retron-mediated single-stranded DNA production in plants. *Synth. Biol.* **7**, ysac025 (2022).
52. Zhao, B., Chen, S.-A. A., Lee, J. & Fraser, H. B. Bacterial Retrons Enable Precise Gene Editing in Human Cells. *CRISPR J.* **5**, 31–39 (2022).
53. Kong, X. Precise genome editing without exogenous donor DNA via retron editing system in human cells. *Protein Cell* 13238-021-00862–7 (2021).
54. Schubert, M. G. High-throughput functional variant screens via in vivo production of single-stranded DNA. *Proc Natl Acad Sci USA* **118**, e2018181118, (2021).

55. Sharon, E., Chen, S.-A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K. & Fraser, H. B. Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).
56. Ramirez-Chamorro, L., Boulanger, P. & Rossier, O. Strategies for Bacteriophage T5 Mutagenesis: Expanding the Toolbox for Phage Genome Engineering. *Front. Microbiol.* **12**, 667332 (2021).
57. Bhattarai-Kline, S. Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* **608**, 217–225 (2022).
58. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, (2014).
59. Lee, G. & Kim, J. Engineered retrons generate genome-independent protein-binding DNA for cellular control. Preprint at <https://doi.org/10.1101/2023.09.27.556556> (2023).
60. Vibhute, M. A., Machatzke, C., Bigler, K., Krümpel, S., Summerer, D. & Mutschler, H. Intracellular Expression of a Fluorogenic DNA Aptamer Using Retron Eco2. Preprint at <https://doi.org/10.1101/2024.05.21.595248> (2024).
61. Liu, J., Cui, L., Shi, X., Yan, J., Wang, Y., Ni, Y., He, J. & Wang, X. Generation of DNAzyme in Bacterial Cells by a Bacterial Retron System. *ACS Synth. Biol.* **13**, 300–309 (2024).
62. Luo, D. & Saltzman, W. M. Synthetic DNA delivery systems. *Nat Biotechnol* **18**, 33–37 (2000).

63. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, 04766 (2014).
64. Paquet, D. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).
65. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol* **36**, 765–771 (2018).
66. Devkota, S. The road less traveled: strategies to enhance the frequency of homology-directed repair (HDR) for increased efficiency of CRISPR/ Cas-mediated transgenesis. *BMB Rep* **51**, 437–443 (2018).
67. Anzalone, A. V. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
68. Bobonis, J. Phage proteins block and trigger retron toxin/antitoxin systems. (2020) doi:10.1101/2020.06.22.160242.
69. Gao, L. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077–1084 (2020).
70. Mirochnitchenko, O., Inouye, S. & Inouye, M. Production of single-stranded DNA in mammalian cells by means of a bacterial retron. *J. Biol. Chem.* **269**, 2380–2383 (1994).
71. Lampson, B. C., Inouye, M. & Inouye, S. Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res* **110**, 491–499 (2005).

72. Simon, A. J., Ellington, A. D. & Finkelstein, I. J. Retrons and their applications in genome engineering. *Nucleic Acids Res* **47**, 11007–11019 (2019).
73. Lampson, B. C. Reverse transcriptase in a clinical strain of *Escherichia coli*: production of branched RNA-linked msDNA. *Science* **243**, 1033–1038 (1989).
74. Lampson, B. C., Inouye, M. & Inouye, S. Reverse transcriptase with concomitant ribonuclease H activity in the cell-free synthesis of branched RNA-linked msDNA of *Myxococcus xanthus*. *Cell* **56**, 701–707 (1989).
75. Lim, D. & Maas, W. K. Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA–RNA compound in *E. coli* B. *Cell* **56**, 891–904 (1989).
76. Miyata, S., Ohshima, A., Inouye, S. & Inouye, M. In vivo production of a stable single-stranded cDNA in *Saccharomyces cerevisiae* by means of a bacterial retron. *Proc. Natl. Acad. Sci.* **89**, 5735–5739 (1992).
77. Chappell, S. A., Edelman, G. M. & Mauro, V. P. Ribosomal tethering and clustering as mechanisms for translation initiation. *Proc Natl Acad Sci USA* **103**, 18077–18082 (2006).
78. Wannier, T. M. Improved bacterial recombineering by parallelized protein discovery. *Proc Natl Acad Sci USA* **117**, 13689–13698 (2020).
79. Aronshtam, A. & Marinus, M. G. Dominant negative mutator mutations in the *mutL* gene of *Escherichia coli*. *Nucleic Acids Res* **24**, 2498–2504 (1996).
80. Nyerges, Á. A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species. *Proc Natl Acad Sci USA* **113**, 2502–2507 (2016).

81. Wang, H. H. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
82. Zhang, Y. A gRNA–tRNA array for CRISPR–Cas9 based rapid multiplexed genome editing in *Saccharomyces cerevisiae*. *Nat Commun* **10**, 1053 (2019).
83. Liu, J.-J., Orlova, N., Oakes, B. L., Ma, E., Spinner, H. B., Baney, K. L. M., Chuck, J., Tan, D., Knott, G. J., Harrington, L. B., *et al.* CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218–223 (2019).
84. Knapp, D. Decoupling tRNA promoter and processing activities enables specific Pol-II Cas9 guide RNA expression. *Nat Commun* **10**, 1490 (2019).
85. Rogers, J. K., Guzman, C. D., Taylor, N. D., Raman, S., Anderson, K. & Church, G. M. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Res.* **43**, 7648–7660 (2015).
86. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci.* **97**, 6640–6645 (2000).
87. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
88. Baker Brachmann, C., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P. & Boeke, J. D. 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 (1998).
89. Tian, S. & Das, R. Primerize-2D: automated primer design for RNA multidimensional chemical mapping. *Bioinformatics* **33**, 1405–1406 (2017).

90. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).
91. Łusiak-Szelachowska, M. & Górski, A. Phage therapy in Poland— a centennial journey to the first ethically approved treatment facility in Europe. *Front Microbiol* **11**, 1056 (2020).
92. O'Neill, J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of the nations. (2014).
93. Chan, B. K., Stanley, G., Modak, M., Koff, J. L. & Turner, P. E. Bacteriophage therapy for infections in CF. *Pediatr. Pulmonol.* **56**, (2021).
94. Schooley, R. T. Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant *Acinetobacter baumannii* infection. *Antimicrob Agents Chemother* **61**, 00954–17 (2017).
95. Gencay, Y. E. Engineered phage with antibacterial CRISPR–Cas selectively reduce *E. coli* burden in mice. *Nat. Biotechnol.*
96. Dedrick, R. M. Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant *Mycobacterium abscessus*. *Nat Med* **25**, 730–733 (2019).
97. Mahler, M., Costa, A. R., Beljouw, S. P. B., Fineran, P. C. & Brouns, S. J. Approaches for bacteriophage genome engineering. *Trends Biotechnol* **41**, 669–685 (2023).
98. Ando, H., Lemire, S., Pires, D. P. & Lu, T. K. Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. *Cell Syst.* **1**, 187–196 (2015).

99. Nozaki, S. Rapid and accurate assembly of large DNA assisted by in vitro packaging of bacteriophage. *ACS Synth Biol* **11**, 4113–4122 (2022).
100. Emslander, Q. Cell-free production of personalized therapeutic phages targeting multidrug-resistant bacteria. *Cell Chem Biol* **29**, 1434–1445 (2022).
101. Palka, C., Fishman, C. B., Bhattarai-Kline, S., Myers, S. A. & Shipman, S. L. Retron reverse transcriptase termination and phage defense are dependent on host RNase H1. *Nucleic Acids Res.* **50**, 3490–3504 (2022).
102. Mosberg, J. A., Lajoie, M. J. & Church, G. M. Lambda red recombineering in *Escherichia coli* occurs through a fully single-stranded intermediate. *Genetics* **186**, 791–799 (2010).
103. Nyerges, Á. Conditional DNA repair mutants enable highly precise genome engineering. *Nucleic Acids Res* **42**, e62, (2014).
104. Ellis, H. M., Yu, D., DiTizio, T. & Court, D. L. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc Natl Acad Sci USA* **98**, 6742–6746 (2001).
105. Weigele, P. & Raleigh, E. A. Biosynthesis and function of modified bases in bacteria and their viruses. *Chem Rev* **116**, 12655–12687 (2016).
106. Bryson, A. L. Covalent modification of bacteriophage T4 DNA inhibits CRISPR–Cas9. *mBio* **6**, e00648, (2015).
107. Fleischman, R. A., Cambell, J. L. & Richardson, C. C. Modification and restriction of T-even bacteriophages. In vitro degradation of deoxyribonucleic acid containing 5-hydroxymethylctosine. *J Biol Chem* **251**, 1561–1570 (1976).

108. Weigel, C. & Seitz, H. Bacteriophage replication modules. *FEMS Microbiol Rev* **30**, 321–381 (2006).
109. Wolfson, J., Dressler, D. & Magazin, M. Bacteriophage T7 DNA replication: a linear replicating intermediate (gradient centrifugation–electron microscopy–E. coli–DNA partial denaturation). *Proc Natl Acad Sci USA* **69**, 499–504 (1972).
110. Bourguignon, G. J., Sweeney, T. K. & Delius, H. Multiple origins and circular structures in replicating T5 bacteriophage DNA. *J Virol* **18**, 245–259 (1976).
111. Hochschild, A. & Lewis, M. The bacteriophage lambda CI protein finds an asymmetric solution. *Curr Opin Struct Biol* **19**, 79–86 (2009).
112. Tal, A., Arbel-Goren, R., Costantino, N., Court, D. L. & Stavans, J. Location of the unique integration site on an Escherichia coli chromosome by bacteriophage lambda DNA in vivo. *Proc Natl Acad Sci USA* **111**, 7308–7312 (2014).
113. Filsinger, G. T. Characterizing the portability of phage-encoded homologous recombination proteins. *Nat Chem Biol* **17**, 394–402 (2021).
114. Hernandez, A. J. & Richardson, C. C. Gp2.5, the multifunctional bacteriophage T7 single-stranded DNA binding protein. *Semin Cell Dev Biol* **86**, 92–101 (2019).
115. Werten, S. Identification of the ssDNA-binding protein of bacteriophage T5: implications for T5 replication. *Bacteriophage* **3**, e27304, (2013).
116. Maffei, E., Shaidullina, A., Burkolter, M., Heyer, Y., Estermann, F., Druelle, V., Sauer, P., Willi, L., Michaelis, S., Hilbi, H., *et al.* Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection. *PLOS Biol.* **19**, e3001424 (2021).

117. Marinelli, L. J. BRED: a simple and powerful tool for constructing mutant and recombinant bacteriophage genomes. *PLoS ONE* **3**, 3957 (2008).
118. Parson, K. A. & Snustad, D. P. Host DNA degradation after infection of *Escherichia coli* with bacteriophage T4: dependence of the alternate pathway of degradation which occurs in the absence of both T4 endonuclease II and nuclear disruption on T4 endonuclease IV. *J Virol* **15**, 221–224 (1975).
119. Warner, H. R., Drong, R. F. & Berget, S. M. Early events after infection of *Escherichia coli* by bacteriophage T5. Induction of a 5'-nucleotidase activity and excretion of free bases. *J Virol* **15**, 273–280 (1975).
120. Dunne, M. Reprogramming bacteriophage host range through structure-guided design of chimeric receptor binding proteins. *Cell Rep* **29**, 1336–1350 (2019).
121. Yehl, K. Engineering phage host-range and suppressing bacterial resistance through phage tail fiber mutagenesis. *Cell* **179**, 459–469 (2019).
122. Jeong, H., Kim, H. J. & Lee, S. J. Complete Genome Sequence of *Escherichia coli* Strain BL21. *Genome Announc.* **3**, e00134-15 (2015).
123. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
124. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G. & Sorek, R. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
125. *Bacteriophages*. vol. 501 (Humana Press, Totowa, NJ, 2009).

126. Epp, C., Pearson, M. L. & Enquist, L. Downstream regulation of int gene expression by the b2 region in phage lambda. *Gene* **13**, 327–337 (1981).
127. Rajagopala, S. V., Casjens, S. & Uetz, P. The protein interaction map of bacteriophage lambda. *BMC Microbiol.* **11**, 213 (2011).
128. Carabias, A., Camara-Wilpert, S., Mestre, M. R., López-Méndez, B., Hendriks, I. A., Zhao, R., Pape, T., Fuglsang, A., Luk, S. H.-C., Nielsen, M. L., *et al.* Retron-Eco1 assembles NAD⁺-hydrolyzing filaments that provide immunity against bacteriophages. *Mol. Cell* **84**, 2185-2202.e12 (2024).
129. Wang, Y., Wang, C., Guan, Z., Cao, J., Xu, J., Wang, S., Cui, Y., Wang, Q., Chen, Y., Zhang, D., *et al.* Defense mechanism of a bacterial retron supramolecular assembly. Preprint at <https://doi.org/10.1101/2023.08.16.553469> (2023).
130. Inouye, S., Hsu, M.-Y., Xu, A. & Inouye, M. Highly Specific Recognition of Primer RNA Structures for 2'-OH Priming Reaction by Bacterial Reverse Transcriptases. *J. Biol. Chem.* **274**, 31236–31244 (1999).
131. Inouye, M., Ke, H., Yashio, A., Yamanaka, K., Nariya, H., Shimamoto, T. & Inouye, S. 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 (2004).
132. 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.* Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* **41**, 9049–9061 (2013).

133. Liang, X., Potter, J., Kumar, S., Ravinder, N. & Chesnut, J. D. Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. *J. Biotechnol.* **241**, 136–146 (2017).
134. Wang, Y., Mallon, J., Wang, H., Singh, D., Hyun Jo, M., Hua, B., Bailey, S. & Ha, T. Real-time observation of Cas9 postcatalytic domain motions. *Proc. Natl. Acad. Sci.* **118**, e2010650118 (2021).
135. Paix, A., Folkmann, A., Goldman, D. H., Kulaga, H., Grzelak, M. J., Rasoloson, D., Paidemarry, S., Green, R., Reed, R. R. & Seydoux, G. Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proc. Natl. Acad. Sci.* **114**, (2017).
136. Muller, R., Meacham, Z. A., Ferguson, L. & Ingolia, N. T. CiBER-seq dissects genetic networks by quantitative CRISPRi profiling of expression phenotypes. *Science* **370**, eabb9662 (2020).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

1E21EBFB4EC14C0... Author Signature

12/10/2024
Date