Title

Characterisation and engineering of prokaryotic innate and adaptive immune systems

Author

Lopez, Santiago Caetano

Publication Date

2024

Characterisation and engineering of prokaryotic innate and adaptive immune systems

by

Santiago Caetano Lopez

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:
Hani Goodarzi
Hani Goodarzi

Signed by:
Jennifer Doudna
6AA02EBB9EAD40D...
Jennifer Doudna

Committee Members

A Jorge Arévalo;

To Christina, Rafa, Cy, Za and Ri;

A mis padres, Marielle y Jaime, y a mi hermano, Lorenzo.

ACKNOWLEDGEMENTS

A Ph.D. is a journey through and exercise in many matters, the extent of which I could not possibly have been prepared for prior to embarking on it. I would like to thank my advisor, Seth Shipman, for taking a gamble on me as he started his own independent research endeavour, infusing me with confidence, and later with freedom and support to set me on my path to independence. Being privy to witness first-hand the lab's inception, challenges, first successes, and highs & lows along the way has been eye-opening. I feel fortunate to have had this exposure in such an intellectually freeing and supportive environment.

Similarly, I am beyond grateful to the UC Berkeley Graduate Division for naming me a Graduate Fellow prior to my official acceptance: while certainly an undeserved accolade, it likely tipped the scales in my favour and led the UC Berkeley – UC San Francisco Bioengineering Graduate Program to grant me the opportunity of a lifetime.

I am deeply indebted to Aaron Streets for his role as unofficial advisor and mentor in ways that vastly exceed the scientific realm. His support and ability to conjure the right words during a particularly difficult moment in my Ph.D. rank highly among the reasons I am writing these words, and did not give up then and there. I wish everyone the fortune of having an Aaron in their corner. I hope to be able to pay it forward and be someone else's Aaron down the line.

To current and past committee members and unofficial mentors, Liana Lareau, Hani Goodarzi, Martin Kampmann and Jennifer Doudna; to Nick Altemose, Ben Adler, Sukrit Silas, Alex Chavez, Xiao Guo, Surge Biswas; Andrés Guillén and Josué Ortega: I

am truly grateful for the scientific, career and emotional support. I can only hope to someday be a fraction of the mentor you have shown yourselves to be.

To the incredible group of people in the Shipman Lab, past and present, thank you – for the support, companionship, and for all of the good times. To Rebecca and Sierra.

To the utterly indispensable and unbelievably hardworking Gladstone community of aides, assistants and others working behind the scenes, nearly always multiple shifts across different jobs to ensure that we can come to work every day and focus solely on our research: Tony, Vivian, Lyle, Arturo, Joshua, Faith, José, Jorge and countless others whom I've had the privilege to meet during my time here. There is simply no way we could undertake any of our daily research without each and every one of you.

A most heartfelt thank you to my friends here, across the country and around the world for their love, support, patience and understanding throughout this process: to Alison and Patty, Amy, Owen and Brittney, Kwasi and María, Moz, Jacob, Vanessa, Nadia, Reuben, Rodrigo and Sath; Adriana, Jovanka, Fabiola, Eduardo, Jara, Marcelo, Matías, Mario, Pablo, Cami, Joaquín, Paloma; à Marion; Alonso, Sebas, Maëva, Barbara, Julian, and countless more.

To Christina and Rafa, thank you for your unconditional love and validation, your scientific rigour and immense heart; thank you for helping me become a better person.

Thank you, Hon, amor, for making every day an adventure. And, finally, a most special thanks to Marielle and Jaime, my parents, and Lorenzo, my brother, for being my biggest supporters and advocates, my moral compass; for keeping me grounded; for instilling in me the inherent value of dedicating one's life to making the world a better place; and for leading by example. For the unwavering support, love and encouragement.

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. S.L.S., S.C.L. and K.D.C. conceived the study. S.C.L., K.D.C., S.K.L., S.B.-K. and S.L.S. designed the work, performed experiments, analysed the data and wrote the manuscript.

Chapter 3:

González-Delgado A*, **Lopez SC***, et al. (2024), Simultaneous multi-site editing of individual genomes using retron arrays. *Nat Chem Biol*, https://doi.org/10.1038/s41589-024-01665-7. S.L.S., A.G.-D. and S.C.L. conceived the study. A.G.-D., S.C.L. and M.R.M. cloned plasmids used in this study; A.G.-D. performed all experiments in *E. coli*; S.C.L. performed all experiments in *S. cerevisiae* and human cultured cells. M.R.M. and C.B.F. performed NGS prepped and ran the sequencing libraries; A.G.-D., S.C.L. and S.L.S. designed the work, analysed the data and wrote the manuscript.

Chapter 4:

Lopez SC, Lee Y, Zhang K, Shipman SL (2024), SspA is a transcriptional regulator of CRISPR adaptation in E. coli. In review, bioRxiv https://doi.org/10.1101/2024.05.24.595836. S.C.L. conceived the study, and with S.L.S., outlined the scope of the project and designed experiments. All experiments were performed and analysed by S.C.L., with technical assistance from Y.L. and K.A.Z for **Figure 4-1e** (Y.L.) and **Figure 4-4f** (K.A.Z.). S.C.L. and S.L.S. wrote the manuscript, with input from all authors.

"Mais à quelle fin ce monde a-t-il donc été formé ?" dit Candide.

"Pour nous faire enrager", répondit Martin.

— Voltaire, *Candide, ou l'Optimisme*

Characterisation and engineering of prokaryotic innate and adaptive immune systems

Santiago Caetano López

ABSTRACT

For the past 3.7 billion years, Earth has been the setting of perhaps the grandest and still ongoing genomic, evolutionary and ecological experiment: the war between bacteria and their viral parasites. Given the unfathomably large global viral and bacterial reservoirs, the former outnumbering the latter at a ratio of 10:1, it is estimated that viral transductions happen in the order of 2×10^{16} times per second. A sizeable fraction of these infections leads to prokaryotic death, and the subsequent estimated daily turnover of 15% of Earth's biomass. However, bacteria have not stood around idly: they have developed weapons of their own to fend against their viral invaders, in the form of immune systems.

Over the course of evolutionary history, bacteria have developed multiple lines of defence to fend off infections, in the form of innate and adaptive immune systems. These immune systems, in turn, have been domesticated by researchers to develop novel biotechnological tools. Chapters 2 and 3 of this dissertation detail the repurposing of one of these innate immune systems, the bacterial retron, for precise genome editing in human cells, and its further engineering to enable multiple precise edits on individual genomes across the tree of life.

Chapter 4 presents a study of the only known bacterial adaptive immune system, the CRISPR-Cas defence system. There, I attempt to discover novel host factors required for CRISPR adaptation, the process by which bacteria create immune memories of infection, and characterise SspA as a novel transcriptional regulator of the process.

Taken together, this dissertation contends that bacterial immune systems are inseparable and cannot be properly understood in isolation of their cellular contexts, and argues for a more systems-biological understanding of their regulation and embeddedness within broader cell metabolism.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abi	Abortive infection
aTc	Anhydrotetracycline
Carb	Carbenicillin
Cas	CRISPR-associated
Cascade	CRISPR-associated complex for antiviral defence
CAST	CRISPR-associate Transposase
Cm	Chloramphenicol
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRi/a	CRISPR interference / activation
crRNA	CRISPR RNA
dCas9	dead Cas9
(ss/ds) DNA	(single-stranded / double stranded) Desoxyribonucleic acid
DSB	Double-stranded break
DUF	Domain of unknown function
HR / HDR	Homologous recombination / Homology Directed Repair
IHF	Integration Host Factor
Indel	insertion / deletion
IPTG	Isopropyl β- d-1-thiogalactopyranoside
Kan	Kanamycin
KO	Knock-out
L-Ara	L-Arabinose
MGE	mobile genetic element
mRNA	messenger RNA
msDNA	multicopy single-stranded DNA
mTol	m-Toluic acid
ncRNA	non-coding RNA
NHEJ	Non-Homologous End Joining
ORF	Open reading frame
Ori	origin of replication
PAGE	Polyacrylamide gel electrophoresis
PAM	Protospacer Adjacent Motif

(RT-)(q)PCR	(Reverse-Transcription) (quantitative) Polymerase Chain Reaction
pegRNA	prime editing guide RNA
Phage	Bacteriophage
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
PolA	DNA Polymerase I
R-M	Restriction-Modification
RBS	Ribosome binding site
RNA	Ribonucleic acid
RNAP	RNA polymerase
RT	Reverse transcriptase
RT-DNA	reverse-transcribed DNA
sgRNA	single guide RNA
Spec	Spectinomycin
SSAP	Single-stranded Annealing Protein
SSB	Single-stranded Binding Protein
SspA/B	Stringent Starvation Protein A / B
tracrRNA	Trans-activating CRISPR RNA
tRNA	transfer RNA
UTR	Untranslated region
WT	Wild-type

Chapter 1 Introduction

As Engels presciently proposed in his unfinished *Dialectics of Nature*[1], much of the natural world can be understood through the lens of revised Hegelian dialectics, which are concerned with "the most general laws of change and development in nature, society and thought"[2]. Dialectics brings forward, to quote Stephen Jay Gould, “a holistic vision that views change as interaction among components of complete systems, and sees the components themselves ‘…’ as both products and inputs to the system”[3]. One such natural system, and a major force shaping global ecosystems[4], is the ongoing conflict between prokaryotes and their viral predators (bacteriophages, or phages). These two parties exist in an intimate state of interdependence. Their warfare can be seen as a dialectical process. First, it encompasses the deep interdependence of both actors as “interpenetrating opposites”. Second, as Engels noted when referring to the evolved quasi-symbiotic relationship between plants and animals, “the negation of the negation *really does take place* in [both kingdoms of] the organic world”[5]; in our case, prokaryotes and viruses alternate the upper hand at different points in history, without reversal to prior geno- or phenotypic states. Lastly, the accumulation of mutations and other minute modifications in either prey or predator crystallises into a qualitative innovation, to temporarily best their adversary: in turn novel defence systems created the prokaryotes, and offence or anti-defence systems invented by the viruses.

It is this latter case of alternating prokaryotic and viral innovations to claim the upper hand that this dissertation is concerned with. It has been estimated that, given the unfathomably large global bacterial and viral reservoirs, viral transductions happen in the order of 2×10^{16} times per second[6]. A sizeable fraction of these infections lead to

prokaryotic death, and the subsequent estimated daily turnover of 15% of Earth's biomass[4,7]. However, prokaryotes have developed weapons of their own to fend against their viral invaders, in the form of immune systems.

1.1 Prokaryotic immune systems

Studies of marine viral particles have estimated that these outnumber their bacterial counterparts at a ratio approaching 10:1[8]. Thus, it is unsurprising that bacteria have developed multiple lines of defence to fend off infections[7,9]. In many ways, these defence lines mirror the steps that viruses perform to infect and take over their hosts, ranging from attachment to the bacterial cell wall[10] to appropriation of the bacterial nucleic acid pool for their own metabolism and replication. Correspondingly, bacteria have evolved a range of innate defences, ranging from modifications to their cell wall[11,12] to depletion of intracellular nucleotide levels[13,14] and degradation of phage nucleic acids[15]. If these are insufficient to prevent infection, altruistic Abi systems induce cell suicide, preventing phage multiplication and escape[16]. These systems, along with a wealth of recently discovered new systems, form the core of prokaryotic innate immunity.

1.2 Prokaryotic innate immunity

Avoiding phage entry

Bacterial escape from viral destruction was described a century ago in experiments that, mirroring those still performed today in modern laboratories, consisted simply of mixing different ratios of bacteria and viruses, and counting the number of bacteria that survived the encounter, isolating these survivors, and attempting to isolate the resistance factor[17,18]. Researchers observed a population of resistant bacteria that arose consistently after phage challenge and hypothesised that resistant bacteria were a result of mutations in sensitive cells, independent of the viral challenge; and further, observing that "physiological changes" (or, phenotypic changes, such as colony morphology) sometimes correlated to resistance, suggested that changes in bacterial surface structures could be a mode of viral resistance[17,18]. Since then, studies have confirmed these initial phenotypically-driven observations, uncovering the mechanisms of phage adsorption on cell surface receptors and bacterial resistance mechanisms to prevent this, such as missense mutations to the phage receptor[11,12,19–23], flagellum[24], pili[25] and fimbriae[24,26–30] genes. The molecular details underpinning these phenotypic changes are also diverse, and include genomic rearrangements mediated by site-specific recombination[28–30], slipped-strand mispairing of repetitive sequences during DNA replication[31,32] and epigenetic modifications through DNA methylation[33].

Restriction-Modification (R-M) systems

A few years after the discovery of phage restricting bacteria, it was discovered that restriction endonucleases, present in the bacteria, provided their host with immunity to

phage infection[34,35]. Interestingly, the term “restriction” arises from the observation that there are bacterial factors that restrict the growth of phages in these cells[34]; it was only a few years later that enzymes, in the form of “highly specific ‘restriction enzyme’”[35] were proposed as DNA targeting and modifying agents underpinning the restricting process[15,35–38]. Soon after the identification of restriction enzymes, it was suggested and confirmed that host modification (and thus immunity to self-restriction) was a result of host DNA modification in the form of nucleotide methylation[15,39–41]. Those systems were thus baptised Restriction-Modification (R-M) systems, and they are an essential line of defence against phages[19,20]. They are now canonically known to be composed at minimum of two genes, one encoding a sequence-specific restriction endonuclease (R) and a DNA methylase (M) responsible to methylating the endogenous genome to protect if from self-cutting. Incidentally, it did not escape some researchers’ attentions that phages could also co-opt these DNA modification strategies to escape restriction[42], foreshadowing the discovery of phage-encoded host defence evasion mechanisms a few decades later.

At this point in the mid 1970s, it was beginning to dawn on biologists that the bacterial-phage arms race was not only a realm for the fundamental microbiologists (and often biophysicists or reformed physicists, such as Max Delbrück and Seymour Benzer), but also a treasure trove of novel tools to advance research in molecular biology[15]. The 1978 Nobel Prize in Medicine was thus awarded to Werner Arber, Dan Nathans and Hamilton Smith, for the discovery of “restriction enzymes and their application to problems of molecular genetics”.

Retrons

About a decade later, Masayori Inouye's at SUNY Stony Brook group made a surprising discovery. While mapping polymorphisms between isogenic strains of *Myxococcus xanthus* by digesting their chromosomal DNA with restriction enzymes followed by gel electrophoresis[43], they observed very stable single-stranded DNA of 162 bases long, existing in high copy number, 500-700 copies per cell: they called it msDNA for multicopy single-stranded DNA[44]. Further study of this short satellite DNA revealed that it was branched to an RNA molecule through a 2-5' phosphodiester linkage, resulting in a covalently bound msDNA-RNA duplex, with double-stranded structures of DNA and RNA at their respective 3' ends[45,46]. Incredibly, they also found that an operonically-encoded reverse transcriptase (RT) was required for the synthesis of msDNA, showing for the first time the existence of RTs in bacteria[47–49]. These genetic elements were called retrons[43], and subsequently discovered and characterised in *E. coli*[50] and several other bacteria. The mechanism for msDNA production is as follows: the retron operon encodes an RNA that encodes for both the RT and a structured non-coding (untranslated) RNA. The RT recognises conserved regions of this structured RNA and partially reverse transcribes it into DNA, forming the aforementioned 2-5' linkage between the msDNA and RNA template.

The cellular function of the retron remained elusive for the next three decades, until in 2020, studies from three independent groups reported that retrons are toxin-antitoxin systems that confer phage resistance to bacteria by mediating an Abi response[51–53]. More recently, the structures of some retron RTs in complex with their ncRNAs and accessory effector proteins were solved[54,55], providing further support for

their proposed physiological role. The current model, at least for Retron-Eco1, is that in the absence of phage, the msDNA acts as an antitoxin by stabilising filaments of Retron-Eco1 to cage and neutralise the N-glycosidase effector. Though the exact events that trigger Retron-Eco1 activation as a result of phage infection remain unclear, it appears that activation of the retron leads to NAD+ depletion through phage sequestration or otherwise modification of the msDNA: this would likely impact host and phage metabolism, resulting in cell arrest and Abi[55].

Abortive Infection (Abi) systems

Thus, as of 2020, retrons became part of the growing set of systems enabling Abi mechanisms. As illustrated by Retron-Eco1, Abi systems are composed of at least two modules: a sensor of phage infection, and effector that causes to cell toxicity through multiple pathways[16,55,56], and either cell metabolic arrest or death. In terms of infection sensing, Abi systems have been reported to sense intermediates of phage genome replication[57] and transcription[58,59], as well as phage protein expression[60–62].

Some effectors, like the retron N-glycosidase effector, deplete the intracellular pool of NAD+; this is also the case in the Thoeris system[63,64] and in some toxin-antitoxin[65]. Other effectors cause cell death by cause membrane degradation (e.g. CBASS[66,67]) or depolarisation (e.g. Rex[68]) or degradation of host DNA[69], tRNAs[70], or mRNAs[59].

Studies of bacterial immunity in the modern age

Arguably the most important breakthrough in recent years has been the adoption of high-throughput computational and experimental approaches to mine metagenomic

and metatranscriptomic databases, which has jump-started the current wave of discoveries of novel anti-phage defence systems – a wave which proceeds at a frantic pace. This began with an observation by Kira Makarova in Eugene Koonin's group, who noted that defence-associated genes tend to cluster together, in what they called "defence islands"[71]. Indeed, one solution to the fitness trade-off between the benefits of encoding multiple lines of anti-phage defence vs. the burden (e.g., autoimmunity) of encoding these defence systems is to cluster them together and mobilise them on short evolutionary time-scales. It should be noted that a different yet complementary solution to this trade-off has been proposed in the form of a bacterial pan-immune system, where the available defence arsenal exists as a shared resource for a population of bacteria, rather than by individual cells – no strain encodes all systems, but as they exist as a part of a population, the "pan-genome" contains all of the immune reservoir needed for survival of the population, thus offsetting the individual cellular burden of carrying multiple systems while keeping them readily available by horizontal gene transfer[72].

In any case, the concept of defence islands led to the realisation that these genetic regions also happened to be enriched for uncharacterised but highly conserved gene clusters: their existence within defence islands, in proximity to known defence systems, were enough evidence to propose these to be novel bacterial immune systems.

With these conceptual tools, Rotem Sorek's group pioneered computational mining of genomic databases using defence island "guilt-by-association", paired with protein domain predictions and gene cluster conservation analyses, to predict the existence of new defence systems within defence islands. This approach and variations on this initial theme, paired with deployment on larger metagenomic databases, has proven fruitful,

enabling the discovery of a wealth of previously unsuspected anti-phage defence systems[51,73–79].

While the particulars of these newly discovered systems have been reviewed elsewhere[20,72,76,80–82], it is perhaps worth nothing four lessons from these discoveries. First, the increasing availability of curated, large datasets of metagenomes and metatranscriptomes, coupled with developments in computational packages to query these databases[83] has uncovered structural and functional conservation between these "ancestral" immune components and those found in eukaryotes (including, but not limited to Argonautes, Viperins[84], and TIR domains). Consequently, the concept of immune modules was coined to describe the conserved immune building blocks that recur across the tree of life[85].

Second, the aforementioned analysis of vast amounts of meta- & pan-genomic data has suggested that there exist some broad rules for the genomic organisation, distribution and functional hierarchisation of defence systems[86]. Defence systems do not distribute randomly across genomes: rather, they tend to co-occur with mobile genetic elements[74], occurring in modular and extensive modules that allow exchange of immune modules[85,87]. In terms of hierarchisation, bacteria have a tendency to encode first and second lines of defence: the first line is typically R-M and CRISPR-Cas, while the second line is often of the Abi type[16].

Third, while these immune modules appear to be modular and independent of other defence systems encoded in the hosts' main or accessory genomes (current estimates are that prokaryotes encode on average 6 anti-phage systems per genome[86]), there is mounting evidence for cooperation and synergy between anti-phage defence

systems encoded in single bacteria[88,89]. This, in turn, raises the issue of regulation and coordination of these defence systems, which has long been recognised to be an open issue in the field[88,90].

Lastly, while novel anti-phage defence systems are being reported and characterised at a dizzying pace (more than a hundred and counting[76]), we are more than certainly under-sampling the true depth of what Nature has created[91]. Deeper sampling & sequencing efforts, paired with re-thought computational algorithms, are sure to uncover biology of beauty and beauty such that we could not begin to dream of designing ourselves[92].

* * *

In 1987, while Masayori Inouye's group was hard at work on characterising the *M. xanthus* retron, Atsuo Nakata's group reported another strange observation. They found "an unusual structure [found] in the 3' end flanking region of *iap*"[93], where they noticed that "[F]ive highly homologous sequences of 29 nucleotides were arranged as direct repeats with 32 nucleotides as spacing."[93] They noted finally that "no sequence homologous to these has been found elsewhere in procaryotes, and the biological significance of these sequences is not known."[93] It would take another twenty years for the function of these mysterious sequences to be elucidated, and for researchers to reveal that are an essential component of the second arm of prokaryotic adaptive immunity, CRISPR-Cas.

1.3 Prokaryotic adaptive immunity

CRISPR-Cas is the sole member of the bacterial adaptive immune system family. As the history of the discovery and characterisation of CRISPR-Cas is now a matter for the history books and has been extensively reviewed elsewhere[94–96], this section will describe the phylogeny and classification of CRISPR-Cas systems, followed by an outline of the core principles underlying CRISPR-Cas adaptation. For further phylogenetic and evolutionary perspectives, readers are directed to the excellent work of Kira Makarova and Eugene Koonin[97].

A primer on CRISPR

CRISPR-Cas functions as an adaptive immune system in bacteria and archaea, mediating sequence-specific immunity against viruses and mobile genetic elements. It does so by using captured short nucleic acid sequences, derived from these invaders, stored as immune memories into clustered, regularly interspaced, short palindromic repeats (CRISPR) loci. Immune memories function as complex with Cas proteins in ribonucleoprotein-mediated interference with foreign nucleic acids[98–125].

At the core of this defence system lies the CRISPR array, composed of alternating endogenous short Repeats and spacer sequences acquired from mobile genetic elements, typically phages or plasmids, by Cas proteins during a process known as CRISPR adaptation[102]. The array is then transcribed as a long RNA precursor, and further processed into small CRISPR RNAs (crRNAs) by Cas proteins, with each crRNA containing a spacer and part of the Repeat sequence[101,123]. Thus, crRNAs contain both a structural element for complex formation with Cas effector proteins and the spacer

sequence which directs the Cas-crRNA ribonucleoprotein complex to the nucleic acids of the invader carrying the complementary sequence. crRNA binding induces the formation of an active CRISPR interference complex, which then scans the genomes for the complementary spacer sequence. If matching occurs and the sequence of the match is flanked by a protospacer adjacent motif (PAM, see below), the complex can either cleave the target sequence or recruit additional effectors to carry out more generalised nucleic acid destruction of host and invader[114]. This process is known as CRISPR interference[124,126].

An impossibly terse phylogeny of CRISPR-Cas systems

CRISPR-Cas systems are thus characteristically composed of a CRISPR array and Cas proteins, organised in loosely-defined modules, which in turn participate in CRISPR adaptation and interference. The broadest categorisation of CRISPR system relies on the type of Cas effector complex utilised in CRISPR-Cas defence: Class 1 systems harness a multi-subunit effector complex, while Class 2 systems are known for their single-subunit effector protein[97,127]. In turn, within each class, a myriad of CRISPR-Cas system types and subtypes exists, and this categorisation relies on the presence of a unique signature protein: as of this writing, CRISPR-Cas Class 1 systems are recognised to contain Types I, III and IV, totalling 16 subtypes[127]. Type I systems use the multi-subunit effector complex Cascade (CRISPR-associated complex for antiviral defence), in association with the nuclease Cas3, to identify, target and destroy the invading DNA. Subtypes within Type I systems vary in their *cas* gene order and composition (e.g., absence/presence/fusions of *cas4*). Type III systems also perform

CRISPR interference using a multi-subunit effector complex, though in contrast with Type I systems, Cas10 proteins are harnessed to form Csm (Type III-A) and Cmr (Type III-B) effector complexes that recognise nascent RNA transcripts, leading to nicking of the transcribed DNA, induction of cyclic oligoadelynate by Cas10, and ultimately activation of unspecific RNAse activity of Csm6[59,128–130]. This RNAse activity degrades both host and invader RNA, inducing host Abi response[59,128–130]. CRISPR-Cas Class 2 systems include Types II, V and VI, which encompass 17 subtypes[127]. Class 2 system effectors are Cas9 in Type II, Cas12 in Type V and Cas13 in Type VI: these effectors differ in their nuclease domain composition, nucleic acid target preference, and strand targeting mechanisms[114,131–133].

Setting aside the field-specific jargon, two interesting observations emerge from the phylogeny of CRISPR systems: first, degree to which these CRISPR adaptation and interference modules overlap and subsequently functionally interact remains unclear, as it has recently been reported that components of the CRISPR adaptation module may play important roles in the interference process[134,135], and conversely, reports have shown that effector Cas proteins aid the adaptation Cas proteins in spacer selection in some CRISPR systems[136]. Second, the identification of *cas* gene-lacking CRISPR systems has been reported. Although initially puzzling, subsequent experimental characterisation of these systems has uncovered cellular functions for CRISPR systems beyond adaptive immunity, including gene regulation, virulence, signalling, and mobilisation of other genetic elements[137,138]. Examples include the co-optation of nuclease-deficient Type I-F systems by some transposons for RNA-guided integration[139–141], and the loss of virulence upon loss of Type II effectors in diverse bacterial species[142–144].

CRISPR adaptation

Despite the high diversity of CRISPR-Cas effector Cas proteins, accessory factors and, by extension, mechanisms mediated by these complexes, the CRISPR adaptation module responsible for the acquisition of new spacers is conserved across nearly all CRISPR systems[97,127]. At the core of CRISPR adaptation is a protein complex formed by Cas proteins Cas1 and Cas2, called the Cas1-Cas2 integrase complex. Some CRISPR systems contain additional accessory proteins or Cas-fused domains that enhance or extend the functionality of the integrase complex: two notable examples are RT-Cas1 fusions that enable integration of RNA spacers[145–148], and DnaQ-like domains that assist in spacer trimming and directional spacer integration[149,150], though it is likely that many more remain to be characterised[151]. Studies spanning the past decade and a half have shed light on the mechanisms underpinning spacer acquisition, as well as some of the key host factors that enable the Cas1-Cas2 integrase complex to update the immune memory bank[108,149,152–160].

The molecular mechanisms of CRISPR adaptation have been worked out most extensively for the *E. coli* Type I-E system, and these will be outlined below; differences with other CRISPR systems will be noted. The preferred substrate of the CRISPR Cas1-Cas2 integrases appear to be double-stranded DNA fragments[153]. These are produced by foreign DNA degradation, helicase-nuclease enzymatic complexes of the RecBCD complex[161], AddAB[162], or the CRISPR-Cas effector complex in a process called primed adaptation[98,163,164], as well as from the replicating bacterial and phage genomes[165]. Before a spacer can be integrated into the CRISPR array, it must undergo trimming by Cas4[166], DnaQ[167] or other host exonucleases[168], generating free 3’ OH groups required

as substrates for spacer integration[149,167]. In Type I systems, though the integrase complex shows intrinsic affinity for the Leader-Repeat junction of the CRISPR array *in vitro*[153,169], *in vivo* activity necessitates the Integration Host Factor (IHF)[170–173], which accommodates the integrase complex by generating a bend in the Leader sequence that allows it to form stabilising contacts with the DNA[170,172]. In contrast, Type II systems generally lack IHF, and are thought to rely on intrinsic sequence specificity for the Leader-Repeat junction[98,169,174]. After the CRISPR integrase complex docks at the CRISPR array, it catalyses two consecutive nucleophilic attacks, adding a new spacer at the Leader-Repeat junction[152,158,172,174–177]. The two nucleophilic are catalysed by the catalytic Cas1 subunits of the Cas1-Cas2 integrase complex between the free 3' OH residues of the captured prespacer and the Leader-adjacent and Leader-distal end of the first repeat (first and second nucleophilic attacks, respectively, in Type I systems; the order of the nucleophilic attacks is reversed in Type II systems)[98,178]. This process generates staggered double-strand breaks at either end and on opposite strands of the Repeat, which are presumably repaired by host polymerases, either in coordination with genome replication or transcription[158]. The outcome is a duplication of the Repeat and an expanded CRISPR array.

PAMs, or one way to avoid self-immunity

Given that the crRNA generated by transcription of the CRISPR array matches the DNA sequence of the CRISPR array and would lead to self-targeting, prokaryotes have added an additional sequence requirement to avoid self-immunity. Indeed, genome cleavage by effector complex from CRISPR-Cas systems that target DNA require that the

target DNA contain a protospacer adjacent motif (PAM) flanking the spacer[106,109,111,179]. Thus, PAM-based target discrimination helps cells avoid accidental recognition and self-targeting of their CRISPR array by the CRISPR effector complex. In Type I-E systems, the PAM also provides directionality to spacer integration by biasing the first nucleophilic attack to occur Leader-adjacent[98,178,180]: PAM trimming takes place after the first nucleophilic attack, and is mediated by either Cas1 or an accessory adaptation protein[98,178,180]. Finally, some CRISPR adaptation modules have evolved PAM selection via their effector complexes (e.g. Cas9) or accessory Cas proteins such as Cas4[98,178,180].

1.4 Engineering prokaryotic immune systems

Although the first prokaryotic immune to be repurposed for biotechnological uses is the R-M system, with restriction enzymes still playing an immense role in run-of-the-mill laboratory settings, the current paradigm permeating biological research is that of harnessing CRISPR-based tools for genetic manipulations and perturbations. This has blazed the trail for fundamental discovery and holds considerable therapeutic promise. Arguably the most utilised of these tools is the *S. pyogenes* interference effector protein Cas9 (SpCas9), which when paired with a single guide RNA (sgRNA)[114], can induce RNA-guided, targeted double-stranded DNA breaks[181] – perhaps the most well-known early breakthrough application of this technology was human cell line editing[182–184]. Since then, a number of technologies have emerged as additional entries in the growing biotechnological toolset, spanning from potent, genome-wide transcriptional activators and repressors to CRISPR-associated transposon-mediated large cargo insertion in human cells. This section will briefly outline some of these tools and offer a comment on what lies ahead.

CRISPR-Cas9 genome editing

Genome editing, which encompasses precise elimination, replacement, or modification of sequences within a genome, holds significant potential for diverse fundamental and practical uses. SpCas9-induced double-strand breaks (DSBs) in the genome can lead to a variety of outcomes, depending on the DNA repair pathway selected. Initial reports of SpCas9-mediated editing harnessed its precise targeting ability to perform both non-homologous end-joining (NHEJ)-caused knock-outs of genes and,

by co-delivering an editing template containing a donor DNA fragment sharing homology to the target loci, precise edits by homologous recombination (HR) in human cells[182–190], with rapid scaling from single loci to genome-wide knockout screens[191,192]. Beyond human cells, SpCas9 was quickly adopted as a workhorse editor as well, with proof-of-concept studies in other model organisms such as *C. elegans*[193–195], *Drosophila*[196,197], zebrafish[198–200], *Xenopus*[201,202], mice[203–205], a variety of plants[206–209], as well as yeast[210] and bacteria[211].

Even at this early stage it began clear that SpCas9's targeting capabilities were a double-edged sword: on one hand, it was a wonderful tool to cleave DNA, but without the appropriate DNA repair pathway harnessed, precise edits were plagued and far outnumbered by insertions or deletions (indels)[188,189]. One solution that researchers found was to deactivate one of the cutting domains of SpCas9, turning it into a nickase (nCas9)[114,182] – the enzymatic activity of nCas9 avoids the formation of DSBs and appeared to stimulate HR DNA repair in human cells[183,184,189]. This nCas9 philosophy was later revisited in further iterations of Cas9-based technologies, such as prime editing.

While SpCas9 remains by far the most widely used CRISPR effector for genome editing, its large size (4107 nucleotides) poses formidable challenges for delivery. Thus, in recent years, the massive increase in publicly available genomic sequences has fuelled the discovery of countless SpCas9 homologues: the criteria was to search for and characterize Cas9 enzymes that are highly active, have simple PAM requirements, low off-target cleavage rates, and are compact enough to be packaged into viral capsids[212]. Several novel variants have been found and some characterised[151,213–216], though most have yet to be engineered for mammalian cell applications. Other attempts to minimise

SpCas9 and increase its PAM range have involved high-throughput screens[217–219], structural-guided mutagenesis[220,221], directed evolution[222,223] and protein engineering[224,225].

Cas nucleases can initiate HDR using double-stranded DNA (dsDNA) or single-stranded DNA (ssDNA) donor templates, but especially in higher eukaryotes, have a tendency to cause a large fraction of undesired edits at the target site through NHEJ-mediated DNA repair rather than HDR[226]. Strategies to enhance HDR efficiency include inhibiting proteins involved in nonhomologous end-joining, utilizing small molecules, gene silencing, or overexpressing proteins known to promote HDR pathways[227–236]. However, implementing these strategies in vivo remains challenging. Besides the described DNA repair pathways, DSB formation can also result in unintended genomic alterations like translocations, large deletions, and activation of cellular responses such as p53 signalling[237–240]. Various enhancements to HDR techniques include optimizing DNA donor template designs, ensuring colocalization of donor DNA with nuclease-induced DSBs, synchronizing cell cycles, or employing adeno-associated virus (AAV) genomes as donor templates, all of which can improve editing efficiencies[234,241–252]. Despite these advancements, indels continue to predominate in edited products initiated by Cas nucleases, particularly in non-dividing cells[226].

The paradigm of fusing domains or entire proteins to Cas9 for arbitrary site-specific activity conferred by the fused domain was widely recognised as a potentially significant tool in itself. Indeed, despite lack of hindsight at the time, it was already hypothesised that these could generate powerful epigenetic modifiers, which became a reality only a couple of years later[253–257]. The idea of Cas9 fusion proteins also contributed to the genesis of

the most recent generation of genome modifying tools, notably base and prime editors, which are discussed below.

Cas9-mediated sequence-specific control of gene expression

In addition to Cas9's ability to cleave DNA, it was discovered that targeting a nuclease-inactive Cas9 (i.e., having both of its catalytic domains inactivated) to a genomic location could influence local transcriptional activity at that locus[258,259]. Cas9's influence would generally lead to transcriptional repression, due to Cas9 sterically interfering with RNA polymerase access to promoter regions (a tool creatively named CRISPR interference, or CRISPRi). However, it was discovered that by fusion or otherwise recruitment of transcriptional activators to the Cas9 protein, activation of gene expression was possible (CRISPRa)[260–262] – crucially, as with the nuclease-proficient Cas9, both CRISPRi and CRISPRa worked across model organisms[263–266].

It wasn't long before both CRISPRi and CRISPRa were scaled to querying the whole genome, enabling the study of the gene ←→ phenotype link[267]. The rationale for performing and interpreting genome-wide screens is the following: a library of sgRNAs targeting loci across the genome is constructed, and delivered to cells. These cells are then subjected to a phenotypic challenge, such as cell viability or sensitivity to a selective pressure (e.g., drug), and the impact of sgRNAs (and, by proxy, of the genes they target) can be assessed by quantifying the relative frequencies of cells carrying each sgRNA at the initial time point and after exposure of cells to the selective pressure. sgRNAs are thus considered "enriched" or "depleted" under certain conditions, which allows researchers to infer their importance in cell fitness under the chosen selective

pressure[268,269]. Genome-scale CRISPRi/a screens have been applied to studying cell and diseases states[270–273], non-coding RNAs[274], and for chemical compound characterisation[275–279]. Noteworthy technical improvements to genome-wide screen have been the development of paired CRISPRi/a[275,280,281] and combinatorial screens for the construction of genetic interaction maps[282–287], pathways and inference of protein complexes[288]; the use of these screens to improve existing genome editing tools[289,290]; and the combination of CRISPRi/a (or live Cas9, in the case of Perturb-Seq[291–296]) screens with single-cell readouts (e.g., single-cell RNA sequencing) to obtain richer and more granular datasets to interpret each genetic perturbation[297–301]. Importantly, genome-scale CRISPRi/a screens also across model organisms[302–308] – Chapter 4 demonstrates an application of a genome-wide CRISPRi screen in *E. coli* to identify host factors involved in CRISPR adaptation.

Genome editing and transcriptional control beyond Cas9

Beyond Cas9, other CRISPR interference systems have been repurposed as potent gene editors and transcriptional modifiers, utilising their intrinsic immune-derived properties to address some of the shortcomings of using Cas9.

For instance, using Cas9 to perform multiple genetic perturbations in a single cell requires laborious cloning for individual expression of the sgRNAs from their own promoters, as in its native context, Cas9 relies on other host enzymes to process the CRISPR array transcript encoding the crRNAs. In contrast, Cas12, a member of the Type V CRISPR-Cas family, offers inherent multiplexing capabilities as it possesses RNase activity that allows it to process the CRISPR array transcript into multiple crRNAs, as well

as performing the function of RNA-targeted DNA interrogation. Studies using Cas12 homologues have shown that Cas12 enzymes are potent tools for multiplex genome editing[309,310]; further, Cas12 enables CRISPRi libraries to be run with dozens of crRNAs, allowing the exploration of higher-order combinations of regulatory elements[311–317]. This concept of multiplexed genome targeting via arrayed modules was the inspiration for the work described in Chapter 3.

Akin to the search for smaller and more broad-range Cas9 homologues, efforts have been made to utilise other Cas12 variants with distinct properties as molecular tools. Some of these variants, such as the members of the Cas12f-j families, are notably smaller than typical CRISPR effectors: these range from 400 to 700 amino acids, and have been shown to be useful for DNA targeting and transcriptional regulation[318–326].

Type VI CRISPR systems are distinct in that they exclusively target RNA, in contrast to the DNA-centric endonucleases Cas12 and Cas9. Cas13, a highly-specific RNA-guided RNase, causes collateral, non-specific RNase activity upon target RNA recognition[132,327], which can induce an Abi response in bacteria[131,132]. Cas13 homologues have been applied to RNA knockdown[328–333] and editing[334], as well as bacteriophage editing[335]. Cas13d has also been repurposed for both pre-crRNA processing and nucleic acid interference in pooled screens, where Cas13d's native RNA-targeting role has been used to run combinatorial genetic screens by targeting transcripts directly, without having to use a fusion protein (c.f. dCas9 approaches)[336–339]. Beyond their applications to RNA targeting through their nuclease activity, catalytically-inactivated forms of Cas13 (dCas13) have been employed at the transcriptome level: their applications span live cell imaging, RNA transcript splicing regulation, translational regulation and RNA editing[328,340–345].

Finally, Type III systems also target RNA. Though they are larger, multi-subunit complexes, some effector complexes lack the collateral RNA targeting activity, thereby enabling RNA targeting without the potential toxicity in some cell types[340,346,347].

CRISPR-Cas for phage engineering

CRISPR-Cas systems, having evolved to target and destroy phages, are well-suited to the task of counter-selection, or depleting for wild-type phages in a mixed population of edited and wild-type phages. After an initial round of homologous recombination generates a mixture of wild-type and recombinant phages, counter-selection with CRISPR-Cas is employed: this involves propagating the phage mixture on a host strain equipped with a CRISPR-Cas system that targets and eliminates wild-type phage genomes, thereby enriching the population with recombinant phages. Various CRISPR-Cas systems have demonstrated effective discrimination against wild-type phages in these approaches[335,348–360], though phages are infamously known for evolving strategies to escape counter-selection.

CRISPR-Cas for proximity labelling

CRISPR-Cas effectors have been engineered for *in vivo* proximity labelling of proteins that interact with DNA or RNA, by fusing nuclease-deficient variants of DNA or RNA-targeting effectors to labelling domains and guiding them to the locus of interest[361–363]. Labelling domains, usually either biotin ligases, horseradish peroxidases or engineered ascorbate peroxidases, attach a covalent biotin tag to proteins located near the locus targeted by the Cas effector, within live cells – the enzymes then convert a

supplemented substrate into a highly reactive biotinylated intermediate that subsequently transfers biotin into amino acid side chains that are in close proximity to the labelling complex. Spatial control is challenging and is chiefly achieved by physically linking the labelling enzyme to the Cas effector and careful control of the duration of labelling. Following the labelling reaction, cells are lysed, biotinylated proteins isolated using streptavidin beads and analysed by mass spectrometry. Using a dCas9-driven complex[364], this approach has been applied to study the proximal proteomes of chromatin complexes[365–369], telomeres[370,371], centromeres[372], cis-regulatory elements[170,367,373,374]. RNA-protein interactomes have also been studied using various dCas13-driven labelling complexes[375–382].

CRISPR-Cas for lineage tracing and recording molecular events

CRISPR-Cas systems have been used to develop novel approaches to molecular recording and cell lineage tracing. Instead of reading out the effects of perturbing the cells' RNA or DNA, these methods focus on deciphering the history and decision-making of these cells through in-genome information storage. Subsequent sequencing can reveal the lineage relationships between cells within a population and stimuli sensed / processes undergone by the cells and their predecessors.

Cas9 has been used for lineage tracing by harnessing its inherent tendency to cause random indels, which generated and stored in arrays of tandem target sequences. These indels serve as barcodes that are inherited through cell divisions, enabling reconstruction of cell lineages along developmental pathways through sequencing of the arrays. A similar approach uses self-targeting (or homing) sgRNAs that cause re-targeting

of the sgRNA locus itself, thus consolidating the system by merging the sgRNA locus with the target site and enabling retargeting and evolvability of barcodes[383–402].

As an endogenous molecular recorder of phage infections, the CRISPR Cas1-Cas2 integrase complex sequentially accumulates spacers as memory of past exposures to MGEs. While the complex was initially used to store arbitrary exogenously-delivered DNA fragments into CRISPR arrays[403,404], subsequent studies showed that intracellular DNA could be used to record the exposure of cells to biological stimuli[405–409]. Further, we have shown that, by using the retron to produce transcription-derived barcodes, we could use the *E. coli* Type I-E Cas1-Cas2 integrase complex to capture the retron-derived DNA barcodes, log them into CRISPR arrays, and thus reconstruct the order of two transcriptional events[410,411]. Similar approaches, but using fused RT-Cas1 CRISPR adaptation systems[147], have been used to capture spacers directly from the RNA pool, thereby recording transcriptional changes in the CRISPR arrays of populations of bacteria[412–414]. Despite its promises as a powerful tool for both lineage tracing and *in vivo*, non-invasive recording of transcriptional events in cells, CRISPR adaptation-dependent approaches have thus far been limited to use in bacteria. Though it remains unclear why attempts at heterologous expression and use of CRISPR adaptation systems outside of prokaryotes (and particularly, in eukaryotes) has failed, I hypothesised that we lack the full set of host factors required to reconstitute functional CRISPR adaptation outside of its native host. This is part of the motivation behind the work presented in Chapter 4.

The next frontier, part 1: solving the DNA repair template delivery issue

Conventional RNA-guided nucleases are employed across a wide spectrum of editing strategies involving double-strand break (DSB)-induced homology-directed repair (HDR). However, their in vivo use has encountered significant limitations since the HDR pathway is restricted to dividing cells and necessitates the use of an exogenous donor template[183,184]. To address the issue of exogenous DNA delivery, David Liu's group developed two technologies: base editors and prime editors.

Base editors introduce precise point mutations without relying on DSBs or exogenous donor templates. This is achieved through the deployment of single-stranded DNA deaminase enzymes fused to either catalytically inactive or nickase Cas effectors[415–420]. Base editors are broadly categorised into two main classes: cytosine base editors (CBEs), which convert C·G pairs to T·A pairs, and adenine base editors (ABEs), which convert A·T pairs to G·C pairs[415,416]. ABEs and CBEs have been extensively applied across diverse cell models and organisms for the purpose of introducing and correcting various transition point mutations[415–420]. Since their initial report, considerable research efforts have been devoted to refining base editors through multiple strategies, including the evolution of deaminase domains, swapping of deaminase domains, use of repair mechanism inhibitors to inhibit competing base excision repair, optimization of linkers and nuclear localization signals (NLS), and expanding PAM compatibility by swapping Cas effector homologues[421] – for instance, Cas12a proteins, with their distinct T-rich PAMs compared to the purine-rich PAMs of Cas9[422], significantly expand the range of genome targeting when paired with base editors; and the discovery of smaller Cas effector variants[318,319,321,322,423] facilitate the development of more compact and *in vivo* deliverable

base editors[424]. Despite the promises of base editing, it is limited to installing point mutations and afflicted by off-target edits and “bystander editing”, which refers to additional, undesired edits in the base editor target window[417,425,426].

To address the bystander editing limitation and extend the editing capabilities of base editors, prime editors have recently emerged as powerful tools to install all single-base edits, as well as small insertions small deletions while minimizing indel formation. Prime editors are comprised of a reverse transcriptase (RT) fused with Cas9 nickase (H840A mutant), which nicks the non-target strand. They utilise an sgRNA containing an extension that encodes the desired edit, along with a primer binding site complementary to the nicked target strand (name pegRNA). This setup allows for priming and extension by the RT, with the extension also encompassing a sequence complementary to the genome target for directing RT-mediated editing. Following RT extension, a redundant 3′ DNA flap forms alongside the 5′ unedited genome flap, necessitating cellular DNA repair pathways to remove the 5′ flap and incorporate the edited 3′ flap. These reactions yield heteroduplex DNA where one strand carries the desired edit while the other retains the wild-type sequence; successful resolution of the non-edited strand by DNA repair pathways results in fixation of the prime edit in the genome[289,421,427–430]. As with base editing, prime editing has the added benefit of functioning in non-dividing cells. The initial nick in the target strand augments adoption of the desired edit via genome repair, though it has also been associated with large genomic inversions and deletions[431]; incidentally, this is an issue that has plagued base editors, which also rely on a nicking Cas9 variant[432].

In recent years, retrons have emerged as promising tools for genome editing due to their ability to generate high-copies of bespoke intracellular DNAs in diverse host

organisms. Indeed, shortly after their discovery and characterisation in *M. xanthus*, Masayori Inouye's group discovered that a retron's reverse-transcribed region (msd) is almost entirely sequence-flexible[433], and that they were capable of producing RT-DNA in eukaryotic cells, first in *S. cerevisiae*[434] and later in cultured mouse cells[435].

Prior to this dissertation, retrons had been used prokaryotic cells for gene silencing and genome editing[436,437], and in *S. cerevisiae* for gene editing[438]. However, at the time, it was unclear what their true biotechnological potential was, as retron-mediated precise genome editing in human cells had yet to be demonstrated; and, although thousands of retrons had been phylogenetically predicted[439], only 16 had been fully characterised and validated experimentally. Since, then we and others have shown that retrons are powerful gene editors across the tree of life, with recent reports of precise editing rates comparable to the latest generation of prime editors[440–443]; we and others have also experimentally characterised hundreds of new, bioinformatically-predicted retrons[441,442], and vastly expanded the biotechnological toolset with novel uses for retrons[410,411,444–453]. Some of this work constitutes the core of Chapters 2 and 3 of this dissertation.

The next frontier, part 2: targeted, large cargo insertions

Despite the benefits of the Cas nuclease-mediated genome editing technologies described above, one major limitation to nearly all editing platforms is that they are constrained to relatively small insertion sizes (e.g., ~50-100bp for prime editing). This represents a major hurdle for biomedical research, limiting fields from fundamental, research-oriented cell engineering to the development of novel gene therapies.

Workarounds have been developed for targeted gene insertion, combining prime editing and serine integrases: prime editing is used to insert a recombination site for serine integrases, which in turn are capable of inserting large cargo, with activity in non-dividing cells and low indel rates[454,455]. However, these systems massively exceed the size limits for conventional viral delivery and even mRNA synthesis.

One exciting development has been the *in vitro*, and most recently *in vivo* characterisation of CRISPR-associated transposases (CASTs) as precise large (multi-kilobase) cargo delivery tools. These were initially identified by computational analyses of phylogenetically diverse, complete but nuclease-deficient CRISPR-Cas systems from Type I-B, Type I-D, Type I-F, or Type V-K that also encode transposase-specific genes: this led to the prediction that Cas domains function as modules for targeting nucleic acids, possibly in collaboration with transposases associated with these loci[97,141,456–459]. Recently, several Tn7-like transposon variants have been engineered to enable CRISPR-associated transposon-mediated genomic integrations in bacteria and human cells[139,460–472].

The mechanism of these CRISPR-associated transposases has been determined and sheds light as much on the fascinating biology at play as on the inherent limitations to these systems for targeted DNA delivery. All CAST systems share the conserved TnsB transposase, essential for coordinated strand transfer reactions during transposition, alongside accessory factors like TnsC and TniQ[97,141,456–459]. However, the DNA targeting mechanism varies, and is akin to that of that of conventional RNA-guided Cas complexes: type I CASTs use the RNA-guided Cascade complex for target recognition, while type V-K CASTs use the Cas12k effector[97,141,456–459]. CAST systems vary significantly in their

number of accessory components, as well as their transposition outcomes, such as genome-wide fidelity and targeting efficiency[97,141,456–459].

The RNA-guided DNA transposition requires the coordination of two molecular machines: in Type I-F CAST systems, the transposase complex TnsABC consists of the TnsA endonuclease, TnsB transposase, and TnsC ATPase; and the TniQ-Cascade complex, is composed of a crRNA guide itself in complex with proteins TniQ and Cas6-8 (Cascade), responsible for RNA-guided DNA targeting[460,464,467]. The transposase complex specifically recognises the left end (LE) and right end (RE) motifs flanking the transposon, catalysing its excision from the donor locus by cleaving each end of the transposon DNA[460,464,467]. The RNA-guided DNA targeting complex uses the crRNA to locate and bind its DNA target in a PAM-dependent fashion, thereby mobilising the transposase complex to the target locus, where accessory transposase proteins define the target ends for cargo insertion[460,464,467]. The transposase complex subsequently ligates the free DNA ends to the target locus, generating gaps at the junctions[460,464,467]. It is the repair of these gaps that causes the characteristic 5bp target site duplications (TSDs) flanking the inserted payload[460,464,467] – this by definition makes the editing outcome of these technologies non-scarless, which can be a drawback for certain applications that require it. Interestingly, functional reconstitution of Cas RNA-guided transposition in human cells required the expression of an additional host factor, one not initially predicted to play a part in the process – as the CRISPR Cas1-Cas2 adaptation complex is thought to have emerged partly through domestication of recombinase/transposase proteins, this gave more support to my hypothesis that

additional, non-CRISPR-Cas related host factors are required to heterologous reconstitution of the CRISPR adaptation process outside of the natural host.

1.5 Context and scope of this dissertation

I began my Ph.D. at a time where the consensus in the CRISPR-Cas field seemed to be that new, fundamental discoveries in the field would now be few and far between, and that our collective efforts should go towards biotechnological applications and CRISPR-based tool development (Time has since proven both postulates wrong); and coincidentally, the field's attention was starting to turn towards novel, recently-(re)discovered bacterial defence systems[73].

One of the rediscovered defence systems was the bacterial retron. I started my Ph.D. working to expand retron-based tools for precise genome editing, focusing initially on developing retron-based gene editing to human cells[440] and further extending the technology to perform multiplexed retron-based genetic modifications on individual genomes[444], detailed in Chapters 2 and 3, respectively. I also assisted efforts to investigate the design rules for retron-based editors through high-throughput library screens in yeast[443], and helped establish a platform for pooled screening of novel, bioinformatically-predicted retrons to identify high-performing precise retron-based human editors[442].

In the spirit of expanding CRISPR-based tools, I assisted in a project that developed temporally-ordered transcriptional recorders, using retrons to generate barcodes of transcriptional events, and harnessing the CRISPR Cas1-Cas2 integrase complex to acquire these retron-derived barcodes into CRISPR arrays[410]; and later developing a computational framework to reconstruct the order of barcoded transcriptional events from high-throughput sequencing data[411].

This time was edifying in many ways, personally and scientifically – during our efforts at heterologous reconstitution of CRISPR adaptation in eukaryotic cells, I came to three realisations.

First, that we were far from a thorough, cellular and context-aware understanding of CRISPR-Cas systems.

Second, that our biotechnological tools would only ever be as good as our fundamental understanding of the underlying biology of the parts used is.

Lastly, living through a few "gold rushes" in my short scientific career ("tail end" of CRISPR, anti-phage system discovery, next-generation precise genome editors, etc…), has made me develop a love for identifying recurrent and long-standing conceptual blind spots that plague "modern" research, and in particular, those oftentimes cementing the base of the pillars upon which these proverbial scientific hype trains steam ahead. Disembarking the hype trains for a more leisurely, sinuous but ultimately rewarding path to discoveries is infinitely more gratifying and likely to bring about serendipity. As Candide notes in Voltaire's *Candide*, one must cultivate one's own garden[473], and this is my way of doing so.

The first two realisations led me to embark on a multi-pronged project to cast a wide net for novel host factors involved in the *E. coli* Type I-E CRISPR adaptation process, which has helped uncover SspA as a transcriptional regulator of CRISPR adaptation, and help lift the veil on the exquisitely complex regulation underlying CRISPR-Cas defence, and probably most other bacterial immune systems as well[474].

The last realisation is, in a sense, the main personal scientific takeaway from my time in graduate school, and one that I will take into the next steps of my scientific journey.

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 defence. However, little effort has been directed at optimizing retrons to produce designed sequences. Here, we identify modifications to the retron non-coding RNA that result in more abundant reverse transcribed 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 harboured, 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 synthesised or assembled in a tube, 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 recipients[475]. These technical challenges likely contribute to low rates of precise editing as well as unintended editing that occurs in the absence of template DNA[239,242,247]. 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 concentration[476]. 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[430,437,438,447]. This reverse transcribed DNA (RT-DNA) is produced in cells from plasmids, transgenes, or viruses, benefiting from transcriptional amplification to create high cellular concentrations that overcome inefficiencies in genome editing. One retroelement class that has been useful in this regard are bacterial retrons[437,438,447], which are elements involved in phage defence[51–53]. Retrons are attractive as tools for biotechnology due to their compact size, tightly defined sites of RT initiation and termination, lack of known host factor

requirements, and lack of transposable elements. Indeed, retron-generated RT-DNA has demonstrated utility in bacterial[437,447] and eukaryotic[438] genome editing.

Despite the potential of the retron as component of molecular biotechnology, it has so far been modified only 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 in editing efficiency. Further, most work with retron has been carried out in bacteria, with only one functional demonstration of RT-DNA production in yeast[438], and only a brief description of reverse transcription in mammalian cells (NIH3T3 mouse cells)[435]. 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 *E. coli* to show that extension of complementarity in the a1/a2 region of the retron non-coding RNA (ncRNA) increases production of RT-DNA. This effect generalised 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), a non-coding RNA (ncRNA) that is both the primer and template for the reverse transcriptase, and one or more accessory proteins[477] (**Figure 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 48-163 bases[1478]. The ncRNA can be sub-divided into a region that is reverse transcribed (msd) and a region that remains RNA in the final molecule (msr), which are partially overlapping[47–50].

One of the first described retrons was found in *E. coli*, Eco1 (previously ec86)[50]. In BL21 cells, this retron is both present and active, producing RT-DNA that can be detected at the population level, which is eliminated by removing the retron operon from the genome (**Figure 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 (**Figures 2-1c, d**). In *E. coli* lacking an endogenous retron, overexpression of the ncRNA and RT from a plasmid yielded an ~8-10 fold enrichment of the RT-DNA/plasmid region over the plasmid alone, which is evidence of robust reverse transcription (**Figure 2-1d**).

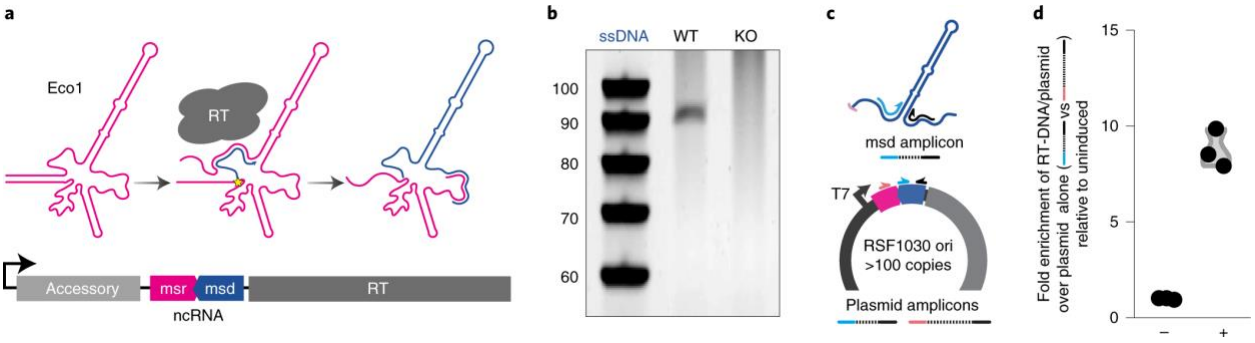


Figure 2-1: Bacterial retrons enable RT-DNA production.

***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.*

Given that the 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 set out to identify aspects of the ncRNA that could be modified to produce more abundant RT-DNA. To do this, we synthesised variants of the Eco1 ncRNA and cloned them into 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 (**Figure 2-2a**). Cells harbouring these library vector sets were grown overnight and then diluted and ncRNA expression was induced during growth for 5 hours.

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 harbouring different retron variants by

isolating cellular nucleic acids, treating that population with an RNase mixture (A/T1), and then isolating single-stranded DNA from double-stranded DNA using a commercial column-based kit. We then sequenced the RT-DNAs, comparing 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 prep each RT-DNA without biasing toward any variant. This involved tailing purified RT-DNA with a string of polynucleotides using a template-independent polymerase (TdT), and then generating a complementary strand via an adapter-containing, inverse anchored primer. Finally, we ligated a second adapter to this double-stranded DNA and proceed to indexing and multiplexed sequencing (**Extended Data Figures 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 (**Figure 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.

In a second library, we investigated the effect of increasing the loop length at the top of what becomes the RT-DNA stem (**Figure 2-2c**). To do this, we created five random sequences of 70 bp each. We then synthesised 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 normalised 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 neighbouring 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 hypothesised plays a role in initiating reverse transcription (**Figure 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 (**Figure 2-2e**). We amplified these barcodes directly from the purified RT-DNA for sequencing (**Figure 2-2e**) or prepared the RT-DNA using the TdT extension method described above (**Extended Data Figure 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 (**Figure 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.

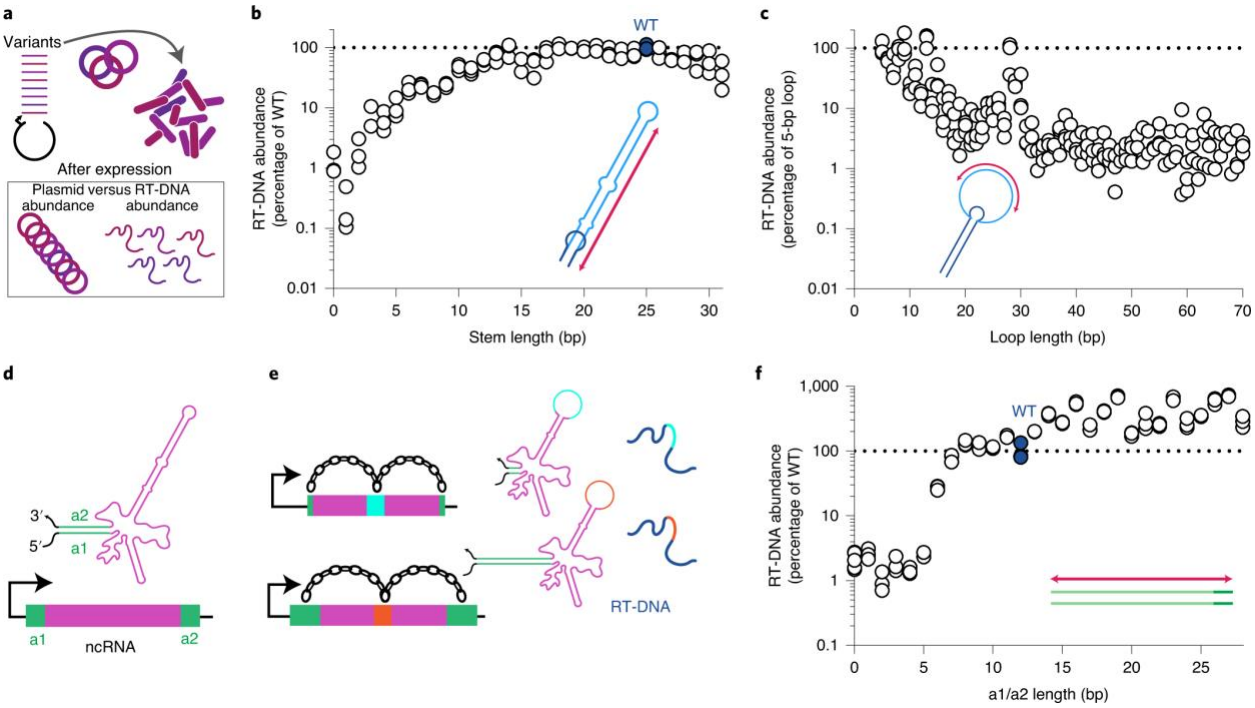


Figure 2-2: Modifications to retron ncRNA affect RT-DNA production.

***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.*

RT-DNA production in eukaryotic cells

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 arrangement[434]. 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[434,479]. 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 (**Figure 2-3a**).

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 (**Figure 2-3b** and **Extended Data Figure 2-2a**). We then extended this analysis to another retron, Eco2[48]. 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 (**Figure 2-3c** and **Extended Data Figure 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 Figure 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 (**Figure 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 (**Figure 2-3e**), from a tightly regulated promoter (**Extended Data Figure 2-2c**). By contrast, Eco2 produces detectable RT-DNA, with both a WT and extended a1/a2 region (**Figure 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.

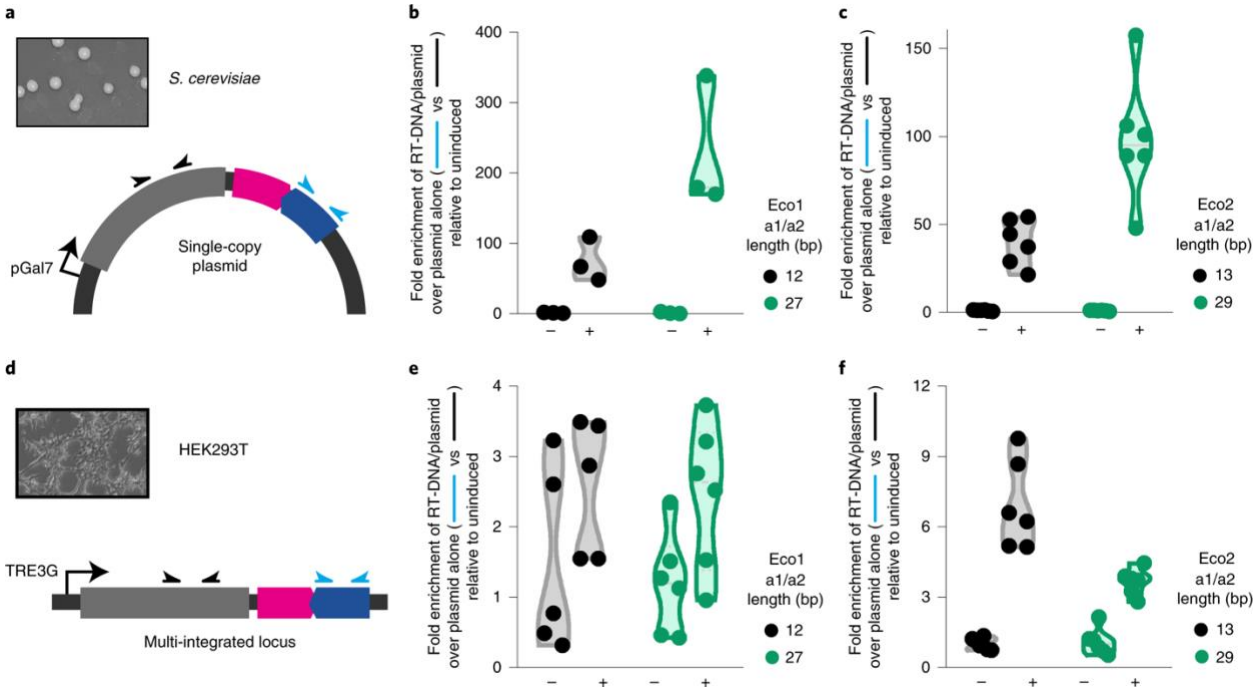


Figure 2-3: RT-DNA production in eukaryotic cells.

***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):*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

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.*

Improvements extend to applications in genome editing

In prokaryotes, retron-derived RT-DNA can be used as a template for recombineering[437,447]. 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 (**Figure 2-4a**). When RT-DNA from this modified ncRNA is produced along with a single-stranded annealing protein (for example, λ Redβ), 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 cells[447]. Therefore, we asked whether increasing RT-DNA abundance using retrons with extended a1/a2 regions could increase the rate of editing in relatively unmodified strains.

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 (**Figures 2-4b,c**). Finally, we expressed each version

of the ncRNA along with CspRecT, a high-efficiency single-stranded annealing protein[480], and mutL E32K, a dominant-negative mutL that eliminates mismatch repair at sites of single-base mismatch[481,482], 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 (**Figure 2-4d**). To test whether the effect of the a1/a2 extension was locus-specific or generalised across genomic sites, we tested an additional three loci[483] 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 Figure 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 cells[438]. 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 (**Figure 2-4e**). We tested our modified ncRNAs using an architecture that was otherwise unchanged from a previously described version[438]. 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 (**Figures 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 (**Figures 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 (**Figure 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 Figure 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 Figures 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 retrons to edit four additional loci[484] in yeast (*TRP2*, *FAA1*, *CAN1* and *LYP1*). We found that for three of the four additional loci, the extended a1/a2 retrons yielded higher rates of precise editing, whereas one site showed lower, but still substantial, rates of editing with the extended version (**Extended Data Figure 2-5**). Overall, across the nine sites tested in bacteria and yeast, the a1/a2 extension improved editing rates at eight sites.

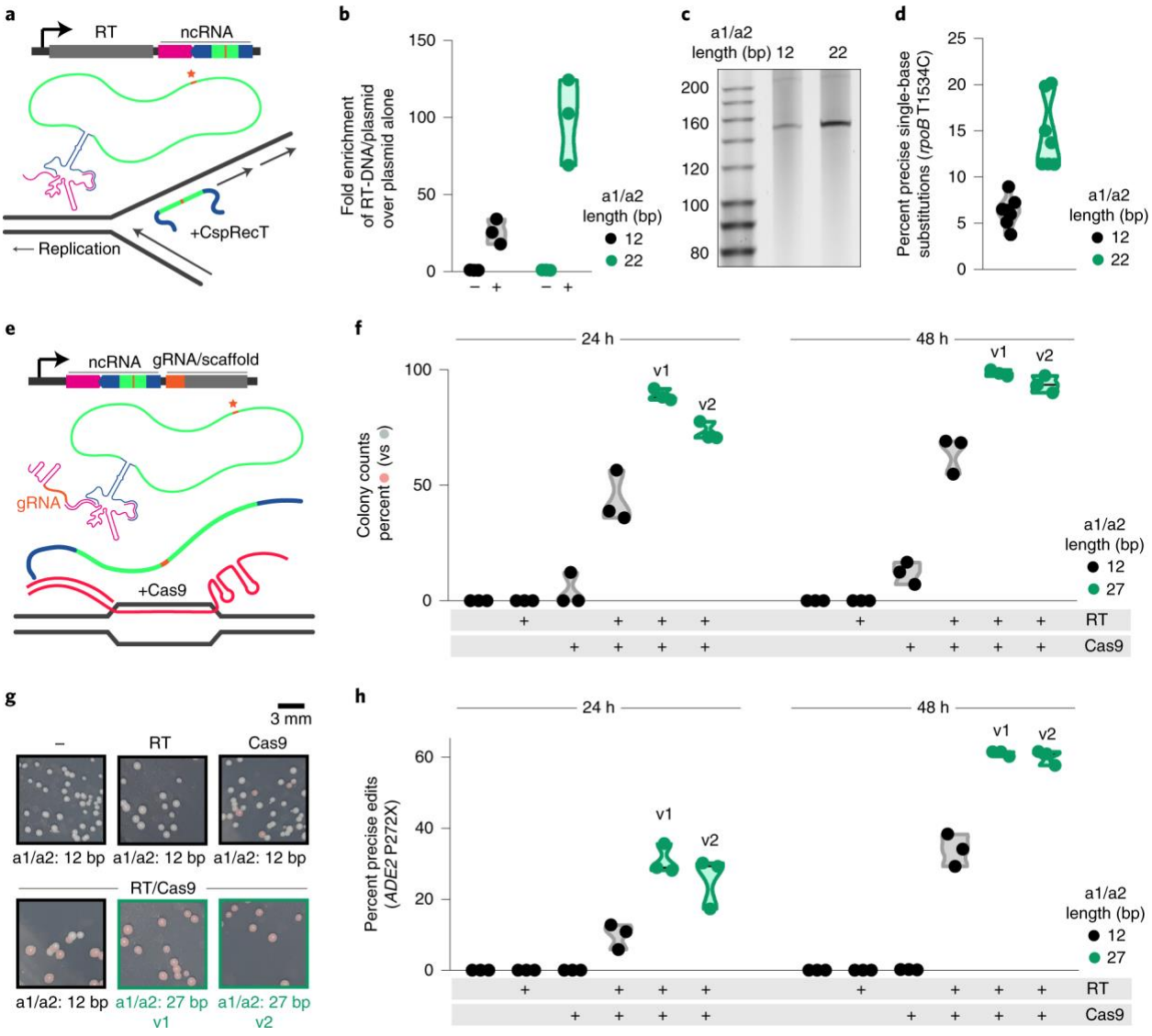


Figure 2-4: Improvements extend to applications in genome editing.

***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 next 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.*

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 P2A[485] 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 (**Figure 2-5a**).

We then created two HEK293T cell lines that each harboured one of two integrating cassettes: Cas9 alone or Cas9-P2A-Eco1RT (**Figure 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 cells[486]. 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 (**Figure 2-5b**). Six genomic loci (*HEK3*, *RNF2*, *EMX1*, *FANCF*, *HEK4*[430] and *AAVS1*[184] 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 (**Figure 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 Figure 2-6**).

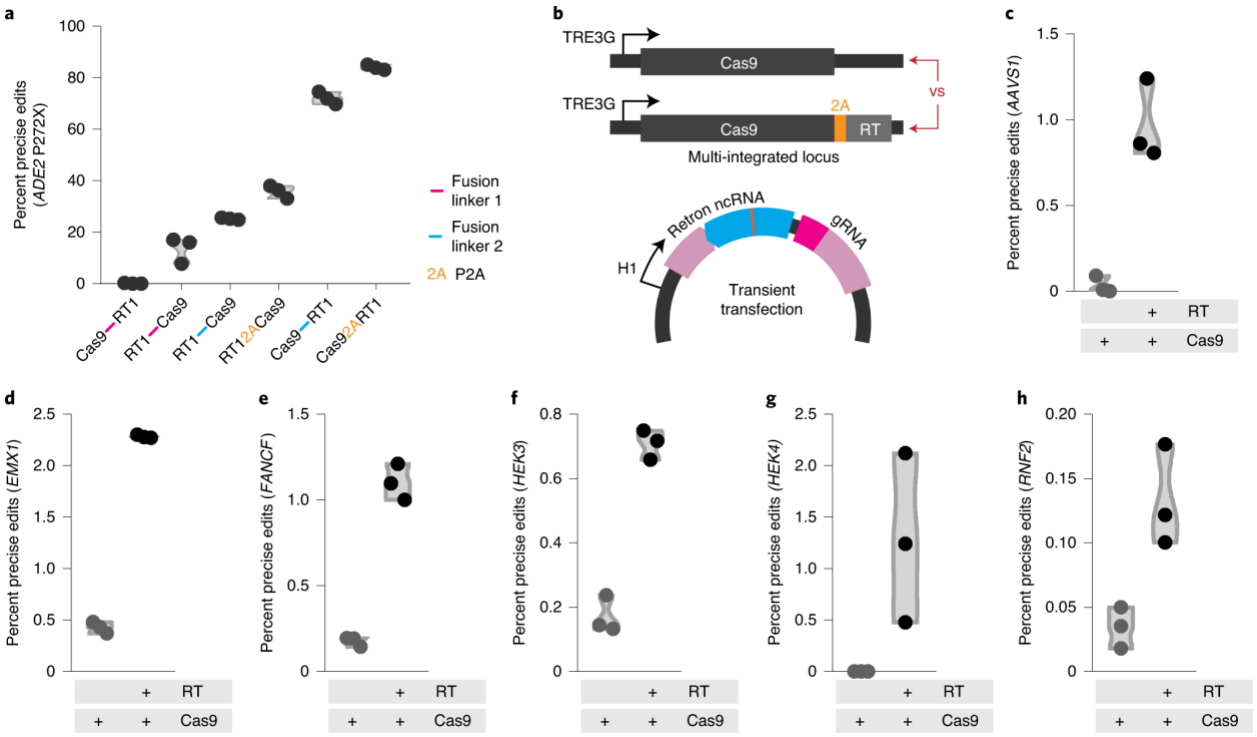


Figure 2-5: Precise editing by retrons extends to human cells.

***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 section. 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.*

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 optimise 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[437,438,447], 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 studies[487], 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-mphR[488] 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 synthesised by Twist in pET-21(+).

Bacterial experiments were performed in BL21-AI cells or a derivative of BL21-AI cells. These cells harbour 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 Wanner[489] 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 μg 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[438], 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)[430] and P2A (ATNFSLLKQAGDVEENPGP)[485].

The second set of plasmids built for the genome-editing experiments were based off of pZS.165[438], 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-synthesised 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 retrons 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 retrons are based off of pSCL.039 and were generated similar to the WT-length a1/a2 retron-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 retron systems, were derived from pSCL.002. IDT-synthesised gBlocks encoding a mammalian codon-optimised Eco1RT and ncRNA (WT), a dead Eco1RT and ncRNA (WT) and a human codon-optimised 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-optimised Eco1RT and ncRNA (extended a1/a2: 27 bp). Similarly, pSC.0L17 was used to generate pSCL.034 by PCR, which carries a mammalian codon-optimised Eco2RT and ncRNA (extended a1/a2: 29 bp).

All yeast strains were created by LiAc/SS carrier DNA/PEG transformation[490] of BY4742[491]. Strains for evaluating the genome-editing efficiency of various retron ncRNAs were created by BY4742 integration of plasmids pZS.157, pSCL.004, pSCL.005 or

pSCL.006 using KpnI-linearised 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-linearised 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, synthesised gBlocks encoding human codon-optimised 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 µg of retron expression plasmids (pKDC.015, pKDC.018, pKDC.019, pKDC.020 or pKDC.031) and 4.2 µg 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 assembly[492]) 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 studies[241]. 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 µg of protein expression plasmids (pSCL.139 and pSCL.140) and 4.2 µg of PiggyBac transposase plasmid (pCMV-hyPBase). 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 are available on Addgene: https://www.addgene.org/Seth_Shipman/

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 analysed 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 $2^{-\Delta\Delta Ct}$ 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 µl of culture was collected,

mixed with 25 µl of water and incubated at 95 °C for 5 min. A volume of 0.3 µl of this boiled culture was used as a template in 30-µl 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 µl) of the uninduced and induced cultures were collected for qPCR analysis. For qPCR sample preparation, the aliquots were centrifuged, resuspended in 50 µl of water and incubated at 100 °C for 15 min. The samples were then briefly centrifuged and placed on ice to cool, and 50 µl of the supernatant was treated with Proteinase K by combining with 29 µl of water, 9 µl of CutSmart buffer and 2 µl 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 (~21,000*g*). The supernatant was collected and used as a template for qPCR reactions consisting of 2.5 µl of template in 10-µl KAPA SYBR FAST qPCR reactions.

For mammalian experiments, retron expression in stable HEK293T cell lines was induced using 1 µg 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 analyse 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 analysed 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 analyse 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 µl 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 µl of TE and treated with 5 µl of RNAse A/T1 + 30 µl 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 analysed on Novex 10% TBE-Urea gels as described above.

Variant library cloning

Eco1 ncRNA variant parts were synthesised by Agilent. Variant parts were flanked by BsaI 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 synthesised 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 BsaI 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 miniprepped 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_Dataset_Chapter2.xlsx.**

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 normalised 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[480] for 16 h with shaking at 37 °C. After expression, a volume of 25 µl of culture was collected, mixed with 25 µl of water and incubated at 95 °C for 5 min. A volume of 0.3 µl of this boiled culture was used as a template in 30-µl 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-µl aliquots of the culture for each timepoint in PCR strips and performed a genomic preparation as follows. The pellets were resuspended in 120 µl of lysis buffer (see above), heated at 100 °C for 15 min and cooled on ice. Protein precipitation buffer (60 µl; 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 µl of ice-cold 70% ethanol and dissolved in 40 µl of water. gDNA (0.5 µl) was used as template in 10-µl reactions with primers flanking the edit site in *ADE2*, which additionally contained adapters for Illumina sequencing preparation (see **Supplementary Table 2-3** 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 µl 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 µg 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 µg 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 µl of ultra-pure, nuclease-free water. Then, 0.5 µl of the gDNA was used as template in 12.5-µl PCR reactions with primer pairs to amplify the locus of interest, which also contained adapters for Illumina sequencing preparation (see **Supplementary**

Table 2-3 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 µl 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.

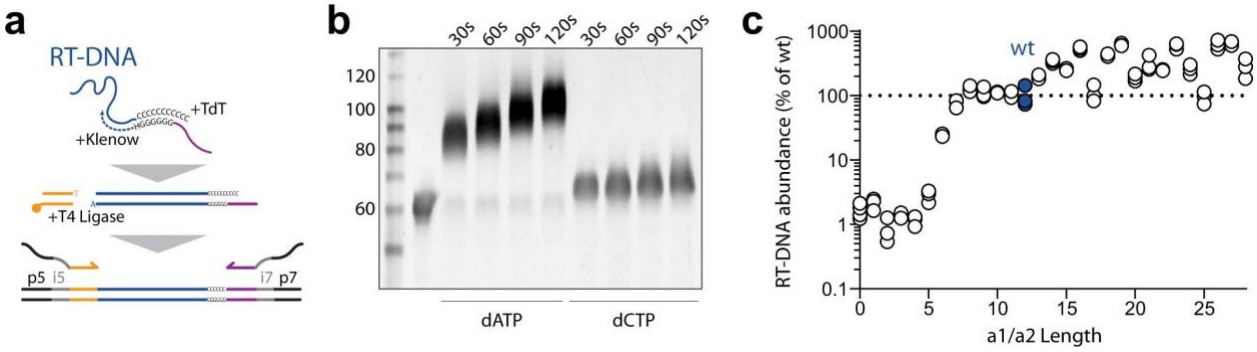
Data availability

All data supporting the findings of this chapter are available within the chapter and accompanying **Supplementary Information** and **Supplemental Files**. Sequencing data associated with this study are available through the NCBI BioProject database under accession number PRJNA770365.

Code availability

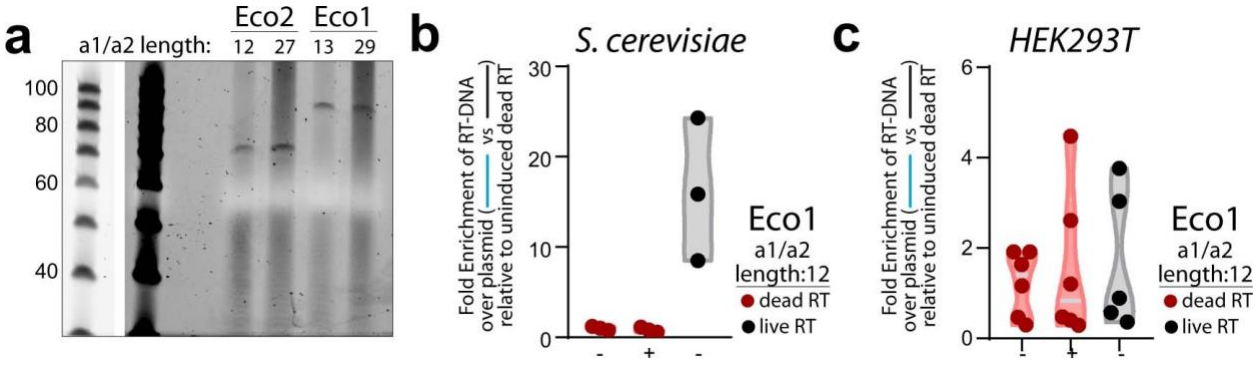
Custom code to process or analyse data from this study is available on GitHub at https://github.com/Shipman-Lab/retron_architectures.

2.6 Supplementary Information



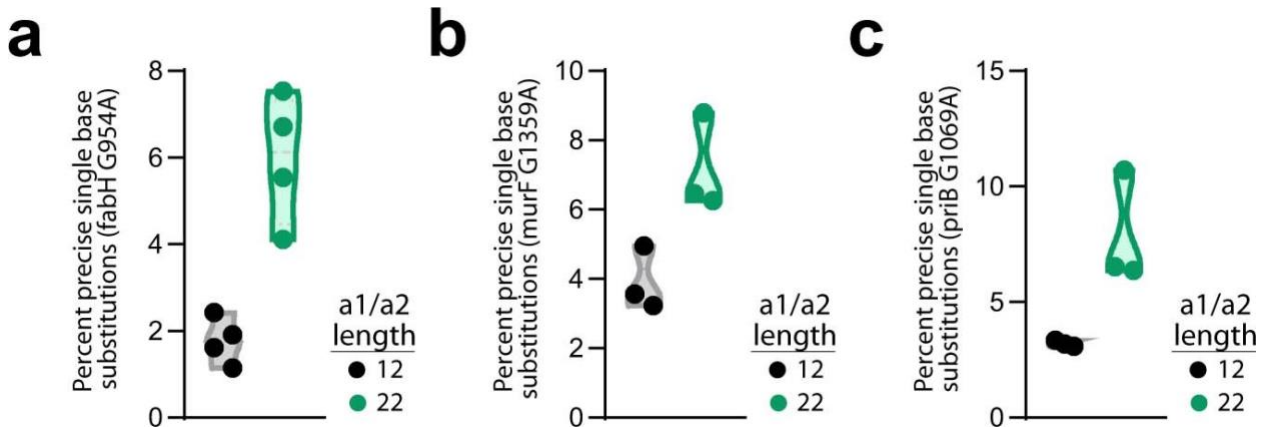
Extended Data Figure 2-1: RT-DNA sequencing prep

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 ***Figure 2-2.***

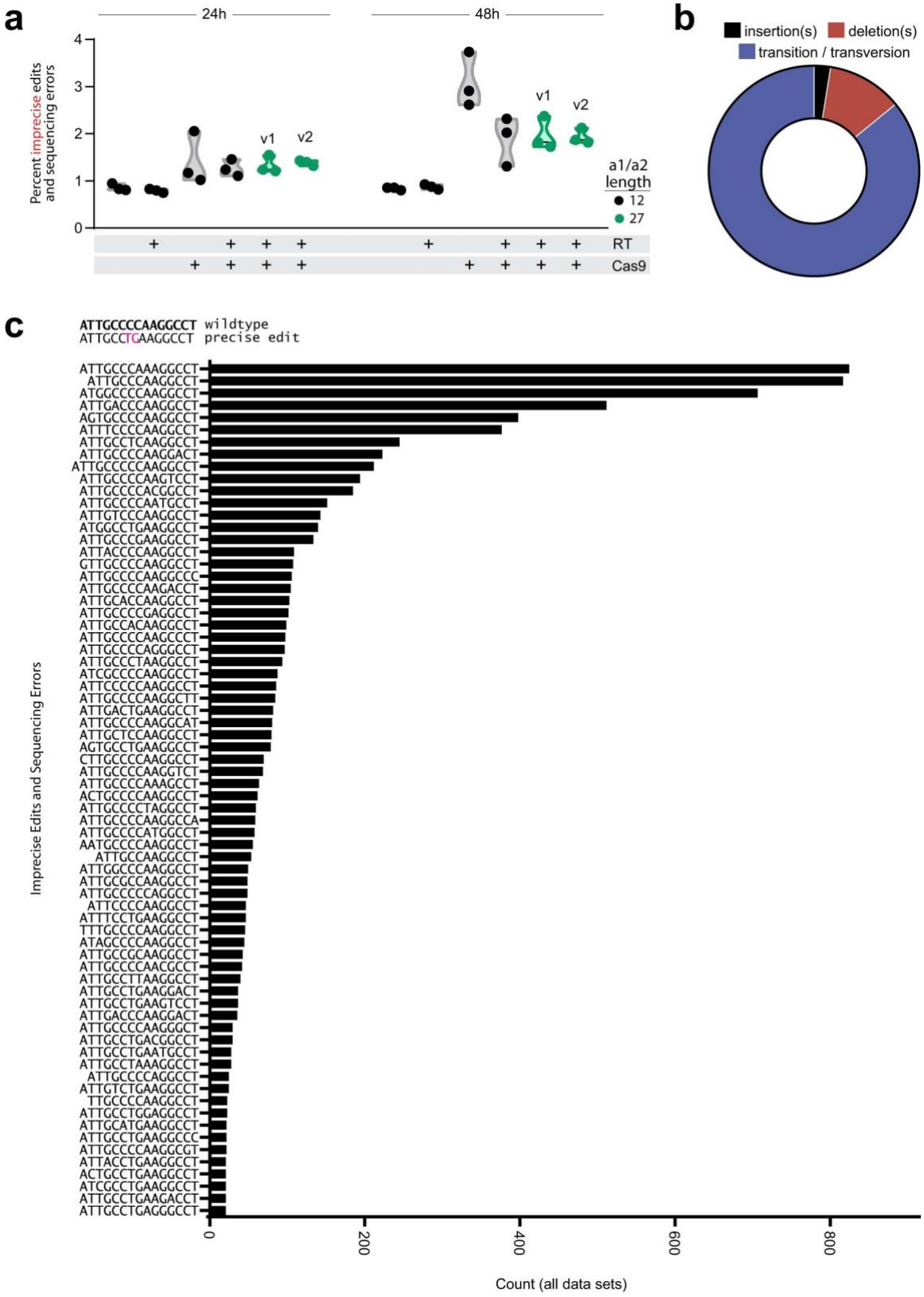


Extended Data Figure 2-2: RT-DNA production in eukaryotic cells.

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. I****dentical analysis as in b, but for Eco1 in HEK293T cells. Related to* ***Figure 2-3.***



Extended Data Figure 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.* ***Figure 2-4a-d.***

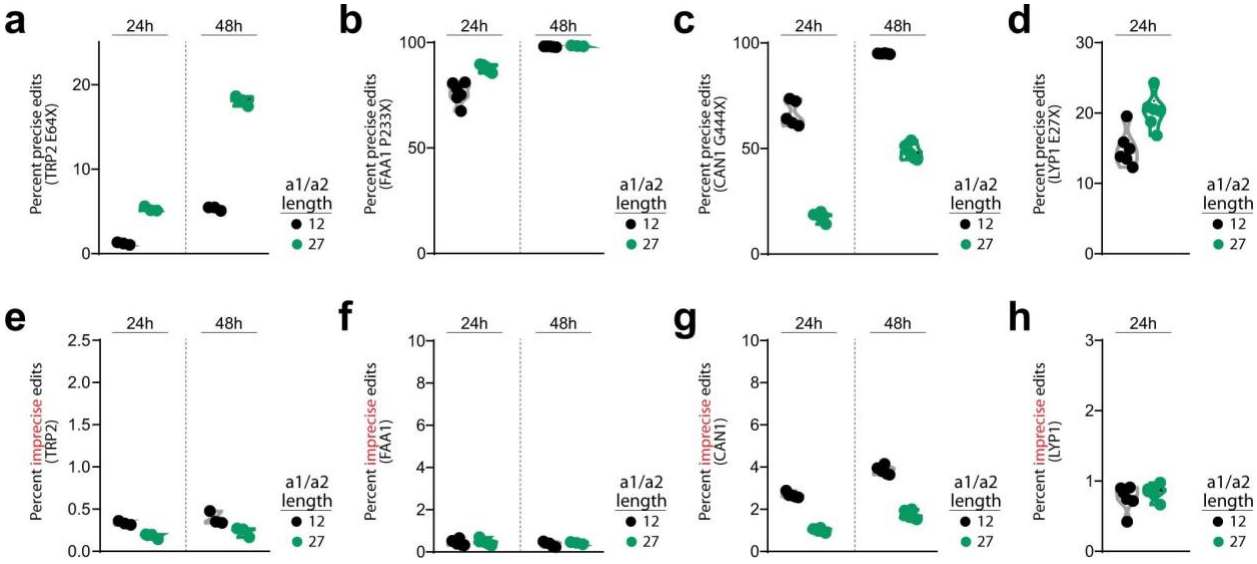


Extended Data Figure 2-4: Imprecise editing profile of the yeast ADE2 locus.

(Figure caption continued on the next page)

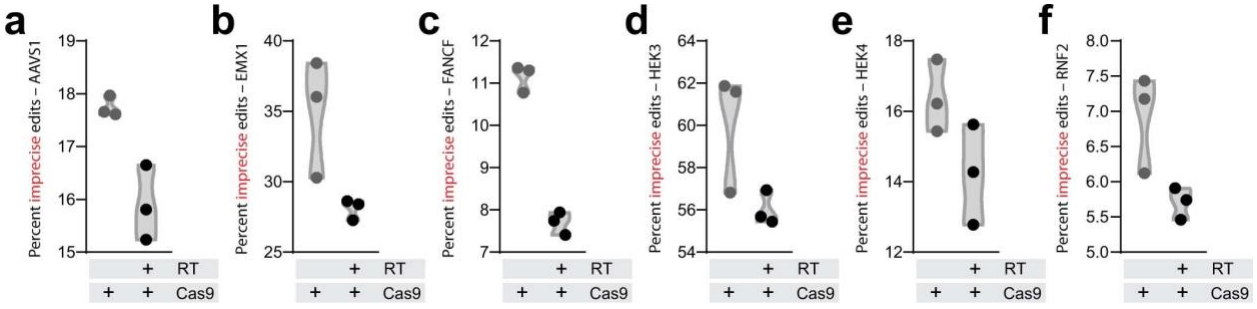
(Figure caption continued from the next page)

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 **Figure 2-4h.



Extended Data Figure 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 **Figure 2-4e-h.



Extended Data Figure 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 **Figure 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 Simultaneous multi-site editing of individual genomes using retron arrays

3.1 Abstract

During recent years the use of libraries-scale genomic manipulations scaffolded on CRISPR gRNAs have been transformative. However, these existing approaches are typically multiplexed across genomes. Unfortunately, building cells with multiple, non-adjacent precise mutations remains a laborious cycle of editing, isolating an edited cell, and editing again. The use of bacterial retrons can overcome this limitation. Retrons are genetic systems composed of a reverse transcriptase and a non-coding RNA (ncRNA) that contains an msd, which is reverse transcribed to produce multiple copies of single-stranded DNA. Here, we describe a technology – termed a multitron – for precisely modifying multiple sites on a single genome simultaneously using retron arrays, in which multiple donor-encoding DNAs are produced from a single transcript. The multitron architecture is compatible with both recombineering in prokaryotic cells and CRISPR editing in eukaryotic cells. We demonstrate applications for this approach in molecular recording, genetic element minimization, and metabolic engineering.

3.2 Introduction

Multiplexing – the act of consolidating multiple discrete elements into a single composite channel – has enabled genomic technologies to scale toward the complexity of the biology we hope to understand. Today, one might use multiplexed DNA synthesis to make a library of distinct CRISPR gRNAs on a single synthesis chip, then use multiplexed experimental design to clone and transfect that library of gRNAs across cells in a single culture, and finally use multiplexed sequencing to analyse the effect of the perturbation on a single sequencing flow-cell[493,494]. This now-standard multiplexed gRNA workflow has allowed scientists run experiments across every gene in parallel with barely more effort than they might have previously put into determining the effect of single gene. However, the typical multiplexing of a gRNA library precludes an important level of analysis: it is implemented across cells, where a single edit is made per genome, and thus cannot be used to study the interaction of mutations within a genome.

Technologies for multiplexing *within* genomes – where multiple distinct, non-adjacent edits are made using a single, consolidated editor – are much more limited. Yet, applications for multiplexing within genomes abound in both fundamental biology (e.g. studying epistasis, long-range gene regulation, and genome organization) and biotechnology (e.g. metabolic engineering, molecular recording, and genome minimization). These complex applications require precise mutations, not genomic scars or transcriptional perturbations. Precision is essential to understand combinatorial genome complexity, such as probing compensatory mutations across genes in a complex or interrogating enhancer-promoter interactions, and is necessary to build nuanced

technological advances, such as ribosome-dependent tuning of gene expression in a metabolic pathway.

In bacteria, the most commonly used approach to introduce combinatorial, precise mutations is MAGE (multiplexed automated genome engineering), which relies on single stranded DNA (ssDNA) recombineering[483,495,496]. A eukaryotic version of this technology has been developed to extend this approach to yeast[497]. However, MAGE is limited by its requirement for numerous labor-intensive recombineering cycles required to attain efficient combinatorial editing rates, and by its reliance on exogenously-delivered oligonucleotides that leave no trackable plasmid element for phenotyping by proxy[498]. Base-editing (BE) and prime-editing (PE)[415,416,430] are two other precise editing approaches that can be multiplexed[499–504]. Base-editors are the simplest to multiplex using tandem gRNAs, but are limited to single base mutations of a defined type (either A•T-to-G•C or C•G-to-T•A)[499,501,502]. Prime-editors have also been multiplexed, but the complexity of the editing elements grows quickly with additional sites. In bacteria, multiplexed prime editing requires a three plasmid system, and multiple edits occur on the same genome in less than 1% of cells[500], while systems built for human and plant cells require two gRNAs per site in addition to the editing template, which can create issues with the assembly of multiplexed plasmids[502–504].

Another way to introduce precise mutations that is compatible with both prokaryotic and eukaryotic editing is to produce editing donors inside a cell using modified retrons. Retrons are bacterial tripartite systems that have been shown to provide phage defence[51–53,439]. Two of the components of the retron operon are a reverse transcriptase and a small (200-300 base), structured non-coding RNA (ncRNA). The reverse

transcriptase recognises and partially reverse transcribes the ncRNA into a single-stranded DNA fragment that is present at the abundance of a cellular transcript[44,47,49,52,478].

We and others have previously shown that the retron ncRNA can be modified to encode an editing donor to precisely edit the genomes of bacteria, phage, plant, yeast, and even human cells[437,438,440,445,447,505,506]. However, these retron-derived editors have only been used to edit genomic positions one at a time. Here, we describe a substantial modification of the retron ncRNA to produce multiple editing donors simultaneously from a single transcript after reverse transcription. We show that these multiplexed, arrayed retron elements – termed multitrons – can be paired with single-stranded annealing proteins to edit prokaryotic genomes and with CRISPR components to edit eukaryotic genomes[438,440,447,505]. We demonstrate utility with proof-of-concept applications in molecular recording, multiplexed deletions, and metabolic engineering.

3.3 Results

Multiplexed editing from multiple donors in a retron msd

The use of retrons in bacterial recombineering was originally developed for applications in molecular recording[437], and has more recently been optimised to install single targeted edits and interrogate biology[438,440,447,505]. To do so, a retron ncRNA – which can be divided into two regions: an msr (multicopy single-stranded RNA) that is not reverse transcribed and an msd (multicopy single-stranded DNA) that is reverse transcribed – is modified to encode an editing donor within the msd region. This modified ncRNA is expressed in cells along with a retron reverse transcriptase (e.g. retron Eco1-RT) that reverse transcribes the retron msd to produce an editing donor (RT-Donor). An overexpressed single-stranded annealing protein (SSAP, e.g. CspRecT) and the host single-stranded binding protein (SSB) promote annealing of the RT-Donor to the lagging strand of a replicating chromosome to install the edited sequence[480,507].

We aimed to further modify retrons to create multitron editors, capable of multiplexed editing of a single genome from a consolidated retron element generating multiple RT-Donors per transcript. Recombineering via oligonucleotide donors is most efficient with donors between 70 and 90 bases long[483], which is also the ideal range for retron recombineering donors[445,447]. Yet, retron RTs are capable of reverse transcribing much longer RT-Donors, even up to an entire gene length[438]. Thus, we initially tested a multitron architecture that encodes multiple 70 bp donors end-to-end within a single msd loop (**Figure 3-1a**) using the two tandem donors to make point mutations in both the *rpoB* and *gyrA* genes in *E. coli*. We tested two versions of this multitron with the donors in each of the possible orders in the msd as well as a control *rpoB* singleplex editor. Both tandem

multitron variants edited both sites, and editing rates for *rpoB* were comparable in the singleplex versus multitron configurations (**Figure 3-1b**).

When comparing the two multitron versions, we noticed that the site edited by the first donor in the multitron tended to have a higher editing rate than the site edited by the second donor. The donor in position one is reverse transcribed first, so the editing difference could be due to a small effect of RT processivity, or due to a positional effect of the donors after reverse transcription. To distinguish between these possibilities, we compared the relative editing efficiencies at each site using the multitrons versus synthetic oligonucleotides of the same sequence as the tandem RT-Donors. Unlike RT-Donors produced by multitrons, oligonucleotide donors had similar relative editing rates across the sites independent of their donor position (**Figure 3-1c**), consistent with an effect of RT processivity.

We next tested three donor multitrons in the tandem msd architecture, using a third donor targeting *lacZ* on the leading strand (less effective than targeting the lagging strand). All three sites were edited in each of the three permutations of donor order (**Figure 3-1d**), with the same positional bias for higher editing at the 5' end of the RT-Donor (**Figure 3-1e**). Although the positional bias is a bug in our intended design, we wondered whether it could be exploited to create a range of editing efficiencies for analogue molecular recording. Retrons have previously been used as analogue molecular recorders capable of detecting the magnitude and duration of a specific input by accumulating precise mutations in the genome[437]. These analogue molecular recorders are, however, limited to operating in the linear range of the interaction between reporter and editing efficacy. We reasoned that using a tandem multitron could add

robustness by expanding the dynamic range of a recording across multiple sites. We constructed another multitron encoding three lagging donors (*gyrA*, *priB*, *rpoB*) driven by an m-toluic acid (mTol)-inducible promoter. Here too, we found that the editing rates were inversely proportional to the order of donor reverse transcription at maximal induction (**Figure 3-1f**). As a result, the editing rates for each site saturate at different mTol concentrations when used as an analogue recorder of mTol (**Figure 3-1g**), effectively increasing the dynamic range of the recorder.

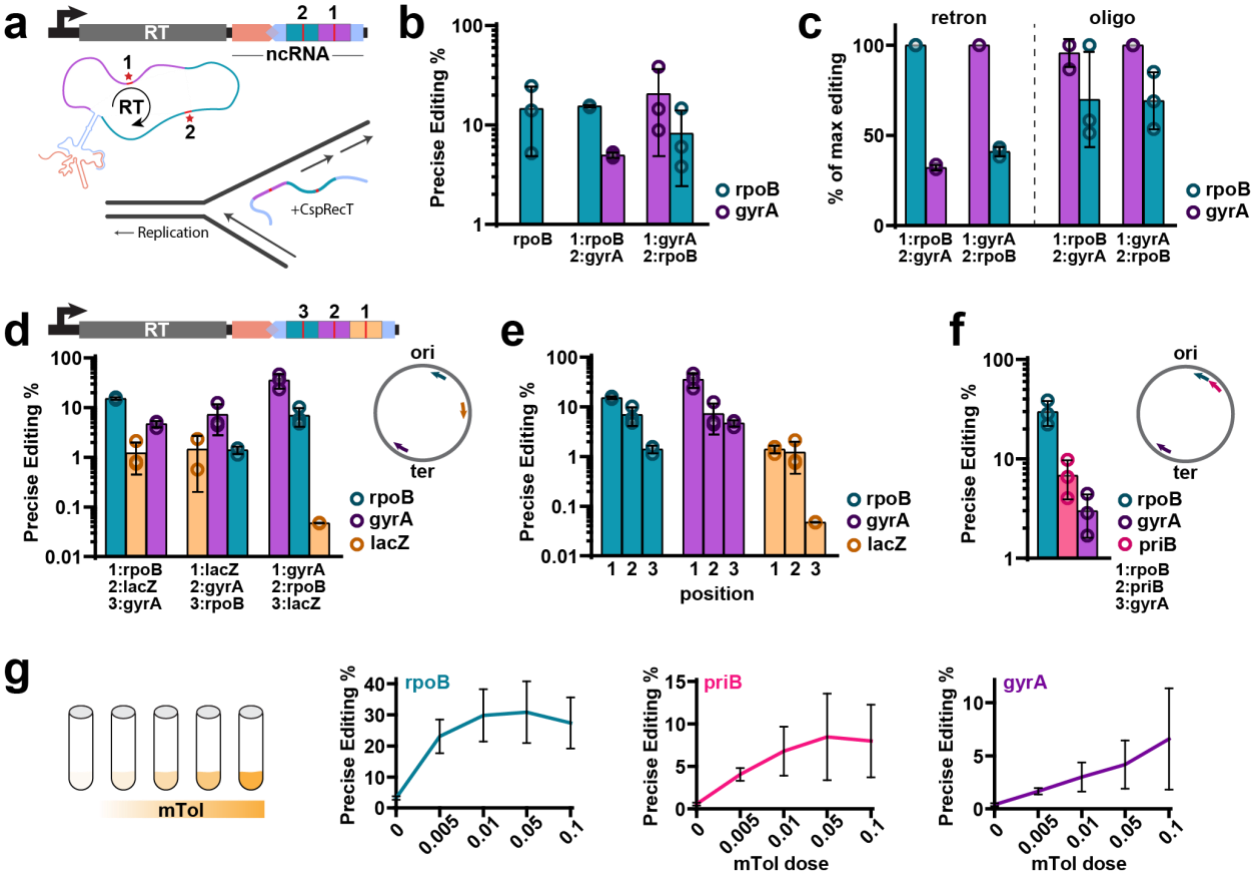


Figure 3-1: Encoding several donors in a retron msd enables multiplexed retron recombineering.

a. *Top: schematic of the retron recombineering operon with two donors encoded within the msd. Donor labels indicate the order in which the donor is reverse transcribed. Bottom: schematic of the retron recombineering process.* ***b.*** *Quantification of precise editing rates of the rpoB locus alone and both rpoB and gyrA loci in bacteria. The order in which the donors are reverse transcribed is indicated. For* ***b, c, d, e,*** *(Figure caption continued on the next page)*

(Figure caption continued from the next page)

*and **f**, data were quantified by sequencing after 24h of editing, circles show each of the three biological replicates, bars are mean ±SD (one-way ANOVA, effect of condition on rpoB editing P=0.2616). **c**. Comparison of donor order for retron-encoded donors versus oligonucleotide donors. Editing is shown as percent of maximum precise editing for each condition. Retron, but not oligonucleotide, is influenced by position effects (one-way ANOVA effect of conditions P<0.0001; Tukey's corrected effect of retron order P<0.0001, oligo order P=0.9842). **d**. Top: schematic of the retron recombineering cassette with 3 donors encoded in the msd. Numbers above indicate order of reverse transcription. Bottom: quantification of precise editing rates of bacterial rpoB, gyrA, and lacZ loci. Right: schematic indicating donor position and strand with respect to the origin of replication (lagging strand for rpoB and gyrA donors and leading strand for the lacZ donor). **e.** Replot of the data in d, illustrating effect of position on editing at each site (two-way ANOVA effect of position P<0.0001). **f.** Quantification of precise editing rates for rpoB, gyrA and priB, in the architecture shown in d. Right: schematic of donor position and strand respect to the origin of replication. All donors are in the lagging strand (one-way ANOVA, effect of editing site P=0.0015). **g**. Use of multiplexed retron recombineering to improve analogue molecular recording technologies. (left) Increasing amounts of m-toluic acid (mTol) are recorded using a retron-derived analogue recorder; (right) quantification of precise editing rates for rpoB, gyrA and priB loci using different amounts of mTol. Error bars are ±SD for three biological replicates. Additional statistical details in **Supplementary Table 3-1**.*

Improved Multiplexed Editing Using Donors in Retron Arrays

To overcome the effect of donor position inside a single msd loop, we engineered a different version of the multitron architecture composed of an ncRNA array with multiple msr-msd regions in tandem, each one containing a distinct donor to edit a unique target site (**Figure 3-2a**). With this arrayed ncRNA multitron, the retron RT has different substrates available within a transcript to generate multiple RT-donors independently, each at the same distance from an internal RT priming site. We tested the ability of this arrayed ncRNA multitron to edit *rpoB* and *gyrA* versus singleplex retron editors, and found that the arrayed ncRNA multitron performed as well or better than the singleplex versions (**Figure 3-2a**). However, this arrayed ncRNA created a new constraint. The length of the ncRNA donor unit is 229 bp and the arrayed design adds 109 bp of direct repeat for each additional editor due to msr duplication, both of which pose challenges for the synthesis and assembly of new multitron plasmids.

Therefore, we engineered a third multitron version composed of an msd array rather than an ncRNA array. In this case, each msd encodes a distinct donor as in the

previous version, but the msr is expressed in trans as a separate transcript (**Figure 3-2b**). This trans msr arrangement was previously shown to be a tolerated modification for reverse transcription of endogenous retron msds[508]. In practice, this reduces the editing unit to 149 bp and reduces the length of the longest direct repeat to 74 bases. The trans msr can interact with any of the arrayed msds, again keeping the donor at a constant distance from the site of RT priming (**Figure 3-2c**).

We tested editing by the arrayed msd multitron versus singleplex editors and found no difference in editing rates at either site (**Figure 3-2b**). The trans msr arrangement in fact yielded consistently higher editing rates than the endogenous retron ncRNA architecture in both singleplex and multiplexed forms throughout this project. Although the msd array and msr/RT transcript contain no terminator between them and could potentially be transcribed as a single unit rather than the intended trans arrangement, we found both sites could be edited at a similar efficiency when using a plasmid containing a terminator between the msd array and the msr (**Extended Data Figure 3-1**).

To test whether donor position inside the msd array multitron affects editing, we constructed three multitron variants with donors to edit *priB*, *rpoB* and *gyrA* genes in each possible order. All three sites were edited by each multitron variant (**Figure 3-2d**), and there was no effect of donor position using arrayed msds (**Figure 3-2e**). Finally, to push the limits of within-genome multiplexing, we constructed an arrayed msd multitron to simultaneously edit 5 target sites (*hda*, *fbaH*, *priB*, *rpoB* and *gyrA*). Editing rates ranged from 5 to 25% for each site, illustrating that arrayed msd multitrons are a potent tool for multiplexed genome editing technologies (**Figure 3-2f**).

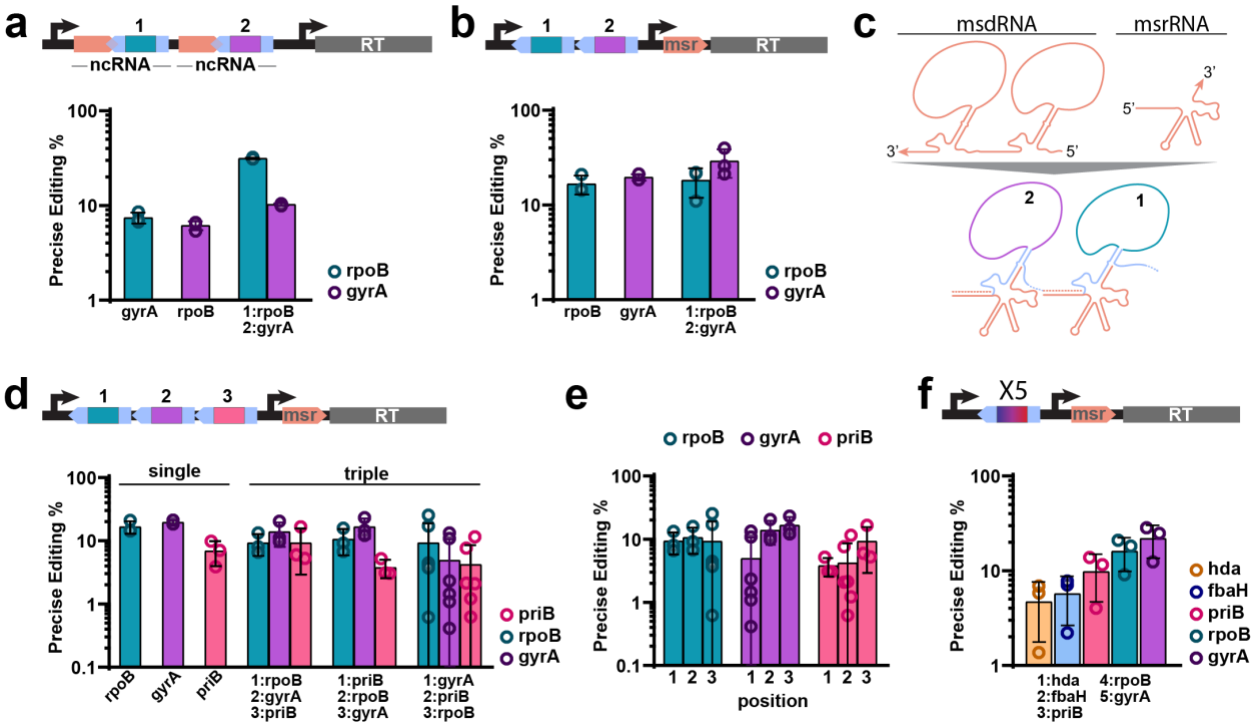


Figure 3-2: Improved multiplexed editing using donors in arrayed retron msds.

a. *Top: schematic of retron recombineering using 2 independent ncRNAs. Each msd region (blue) encodes a different donor (1 and 2). Bottom: quantification of precise editing rates for precise editing of gyrA or rpoB alone or simultaneously (unpaired, two-tailed t-test, singleplex versus multiplex, rpoB P<0.0001, gyrA P=0.0006).* ***b.*** *Top: Schematic of retron recombineering using an msd array with a single msr sequence in trans. Bottom: quantification of precise editing rates for precise editing of rpoB or gyrA alone or simultaneously (unpaired, two-tailed t-test, singleplex versus multiplex, rpoB P=0.7312, gyrA P=0.1702).* ***c.*** *Top: schematic of arrayed msd and msr transcription products. Arrayed msd is transcribed as a single transcript. Bottom: schematic of RT-DNA production using as template an arrayed msd. 1 and 2 indicates the number of the msd in the arrayed msd.* ***d.*** *Top: schematic of 3x arrayed msd. Bottom: quantification of precise editing of rpoB, gyrA or priB edits alone or simultaneously.* ***e.*** *Replot of the data in d, illustrating the effect of position on editing at each site (two-way ANOVA, effect of position P=0.1138).* ***f.*** *Quantification of precise editing using a 5x arrayed msd to edit hda, fbaH, priB, rpoB and gyrA. Data in* ***a, b, d, e*** *and* ***f*** *were quantified by Illumina sequencing after 24h of editing, circles show each of the three biological replicates, bars are mean ±SD (one-way ANOVA effect of editing site P=0.016). Additional statistical details in* ***Supplementary Table 3-1****.*

Increasing Limits of Deletion Size Using Nested Multitrons

One benefit of using retron-derived donors is that they support a broad range of precise mutations, including insertions, deletions and replacements. However, when recombineering with either retron RT-Donor or oligonucleotide donors, the efficiency of inserting and deleting base pairs is inversely related to the size of the edit[445,483]. This is

presumably intrinsic to the mechanism of recombineering, a result that we replicated here using RT-Donors to delete 1 to 100 bp, finding a declining efficiency with deletion size whether using an endogenous ncRNA architecture or the trans msr architecture (**Figure 3-3a**).

We wondered whether we could overcome this limitation on deletion efficiency at larger sizes by using arrayed msd multitrons encoding a series of nested deletion donors. A nested deletion series consists of multiple donors intended to make deletions of increasing size progressively at same locus. If the smallest deletion succeeds, it creates a smaller target size for a previously disfavoured large deletion. We explored nested deletions by first comparing the editing efficiency of single 25 and 50 bp deletions in the *lacZ* gene with simultaneous deletions of overlapping 25 and 50 bp at the same location using a multitron (**Figure 3-3b**). The 50 bp deletion was not significantly less efficient than the 25 bp deletion using singleplex retron donors so, unsurprisingly, the rate of 50 bp deletions by the multitron version was not significantly increased. However, the rate of the 25 bp deletion was decreased by the multitron, suggesting that 25 bp deletions were being converted into 50 bp deletions.

Next, we tested a multitron containing a 25, 50, and 100 bp nested deletion donor series (**Figure 3-3c**). In this case, the previously disfavoured 100 bp deletion was significantly more efficient using the multitron series than using the singleplex deletion donor. In fact, this strategy created a 100 bp deletion in ~42% of genomes, overcoming an intrinsic inefficiency in recombineering deletions. Furthermore, the multitrons generated a heterogeneous population of genetic elements with different deletions sizes

that could be used to probe functional domains of a target gene or miniature versions of a protein of interest.

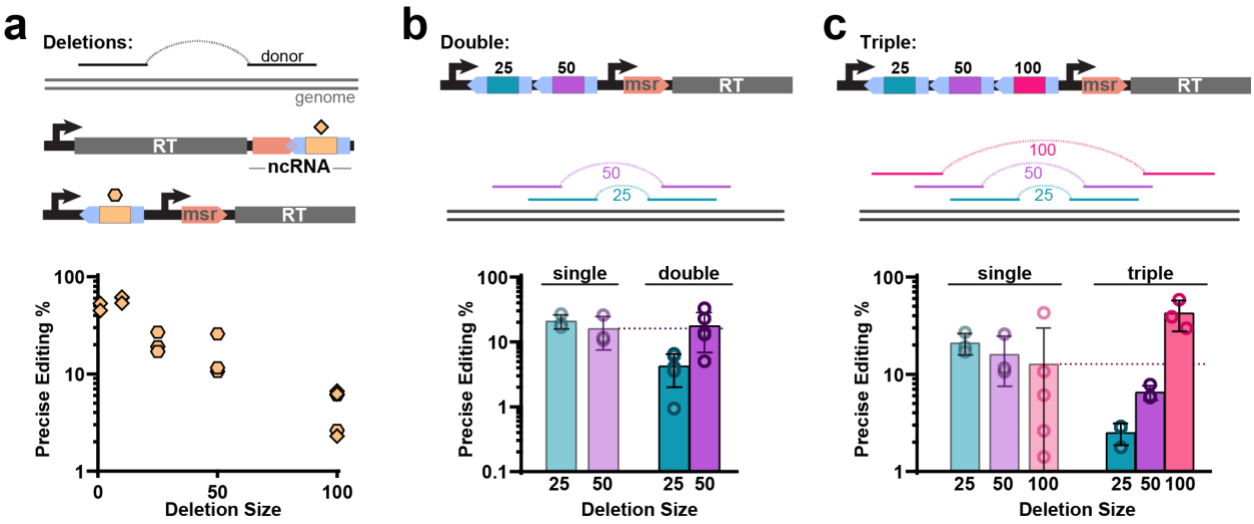


Figure 3-3: Increasing limits of deletion size using nested deletion donor arrays.

a. *Top: Schematic of genome deletions using retron recombineering. Middle: schematic of a standard retron cassette to make deletions (top) with the donor represented by a diamond and an arrayed msd retron cassette with the donor represented by a hexagon. Bottom: quantification of precise editing rates for a single deletion of 1 bp, 10 bp, 25 bp, 50 bp or 100 bp deletions by Illumina sequencing after 24h of editing. Diamonds show deletions from a standard architecture and hexagons show deletions using an arrayed architecture (one-way ANOVA, effect of deletion size P<0.0001).* ***b.*** *Top: Schematic of arrayed msd retron cassette with two donors to make 25 and 50 bp deletions. Middle: Schematic of a nested deletion strategy using two donors to delete 25 bp and 50 bp. If the 25 bp occurs first, the 50 bp deletion becomes a 25 bp deletion. Bottom: Quantification of precise editing rates for single 25 and 50 bp deletions, and for the nested 50 bp deletion (unpaired, two-tailed t-test, 25 base deletion, single vs multi P=0.0006, 50 base deletion, single vs multi P=0.8393).* ***c.*** *Top: Schematic of arrayed msd retron cassette with three donors to make 25, 50 bp and 100 bp deletions. Middle: Schematic of a nested deletion strategy using three donors to delete 25 bp, 50 bp and 100 bp. Bottom: Quantification of precise editing rates for single 25 bp, 50 bp and 100 bp deletions, and for each deletion using the nested strategy (unpaired, two-tailed t-test, singleplex versus multiplex 100bp deletion, P=0.0485). Data in b and c were quantified by Illumina sequencing 24h after of editing, circles show each of the three biological replicates, bars are mean ±SD. Additional statistical details in* ***Supplementary Table 3-1****.*

Multiple Edits in an Individual Genome Using Multitrons

Up to this point, editing has been quantified by bulk sequencing of each individual locus, with the assumption that edits accumulate on genomes according to the product of the rates at each site. We next aimed to explicitly test that assumption. First, we designed a multitron editor producing three, non-overlapping msd donors, each targeting a single

gene, *gyrA*, in a genome window of 300 bp **(Figure 3-4a)**. All 70 bp donors target the lagging strand. With this narrow editing window, we were able to analyse recombineering efficiencies for individual sites as well as combinatorial edits from an amplicon of the locus. Sequencing revealed editing rates of 8 to 25% across the sites, comparable to previous experiments **(Figure 3-4a)**. From this individual site data, we calculated an expected frequency that we should find the various double edits and the triple edits among genomes, based on the product of rates at each site (**Figure 3-4b**). We compared this to the real frequency of each double combination and the triple edit in our sequencing data and found that the expected and real rates were matched (**Figure 3-4b**). Here, the double edits were present in 1.4-7.1% of genomes and the triple edit was present in ~0.77% of genomes.

To test the accumulation of multiple edits on individual genomes in a more practical scenario, we decided to isolate multiply edited clones using a single editing plasmid that can be easily removed after editing. To do this, we combined the five molecular elements required for multitron recombineering – msd array, msr, RT, RecT, and dominant negative mutL (to suppress mismatch repair for single base mutations) – onto a single plasmid with RSF1010 origin of replication (**Extended Data Figure 3-2a**). However, initial testing of this architecture yielded editing rates for the *rpoB* gene were ~5x lower using the single plasmid compared to the previous two plasmid system (~5% and ~25%, respectively). To increase recombineering efficiency, we added an *E. coli* optimised ribosome binding site (RBS) immediately upstream of only the RT gene or both the RT and the CspRecT genes, both of which increased editing rates but still fell short of the level achieved by the two-plasmid system (**Extended Data Figure 3-2a**).

We next changed the origin of replication for the single plasmid system, opting for a temperature-sensitive origin (oriR101) so that the plasmid becomes curable after editing by moving from a permissive temperature (30°C) to a non-permissive temperature (37°C)[453,489]. Interestingly, the editing rates using this single plasmid finally reached comparable levels to those of the previous the two-plasmid system (**Extended Data Figure 3-2a**). This improvement in editing was not due to an effect of temperature, as we found similar editing rates with a temperature-insensitive version at both 30°C and 37°C (**Extended Data Figure 3-2b)**. An alternative possibility that is consistent with the data could be the effect of the different inducers used with the different plasmid backbones: m-toluic acid for RSF1010 derived plasmid and arabinose for the oriR101 derived plasmid. We find that increasing concentrations of m-toluic acid have a negative effect on bacterial growth (**Extended Data Figure 3-2c**). We do not exclude an additional effect of the plasmid copy number. Next, we optimised arabinose concentration (**Extended Data Figure 3-d**). Finally, we also studied the stability of the genetic system with retrons arrays of different length using a 5-day protocol in the presence or absence of the inducer **(Extended Data Figure 3-e,f).** Sequencing of the whole retron array harbouring 2, 3 or 5 msds with different donors revealed that in most cases more than 80% of the colonies preserve an intact retron array after 5 days showing the robustness of the multitron technology.

With curable, single-plasmid parameters optimised, we next attempted to isolate clones that were simultaneously edited at distant regions of an individual genome (*fbaH* and *hda*). We found substantial editing of each target (~20%) and additionally found that the efficiency of editing could be increased to ~45% with an additional day of editing,

demonstrating the continuous nature of this approach **(Figure 3-4c)**. Following editing, we cured the temperature sensitive editing plasmid from 96 individual colonies (48 after 24 hours and 48 after 48 hours) and sequenced the editing loci from each colony. The overall rates of editing at both sites and time points from the individual colonies closely matched the bulk sequencing data (**Figure 3-4c**). We also calculated the expected frequency of finding doubly edited colonies based on the product of the bulk rates at each site and found that the real frequency of doubly edited colonies (~4% after 24 hours and ~22% after 48 hours) was exactly reflected in the real colony sequencing (**Figure 3-4d**).

We also investigated the background mutation rate of multitrons to evaluate the usefulness of the method when fidelity is required. Specifically, we measured the accumulation of local and global off-target mutations in *E. coli* bMS.346 genome in the presence or absence of RT activity. First, we constructed a dead RT version of the multitron targeting *fbaH* and *hda* genes which showed eliminated effective precise editing (**Extended Data Figure 3-3a**). Local off-target mutations were quantified by analyzing the 70 bp homology window of *fbaH* and *hda* donors in the chromosome for unintentional mutations. We found no difference in mutation frequency in the donor window in the live versus dead RT condition (approximately $5x10^{-5}$ errors/base, consistent with Illumina sequencing error; **Extended Data Figure 3-3b**). Global off-target mutations were measured by comparing whole-genome sequencing of colonies after recombineering against with the parental strain. We found three mutations across the colonies in the live RT version (one of which appears to be a longer homologous recombination event between the plasmid *araC* and the genome *araC*) versus two mutations across the colonies in the dead RT version (**Supplementary Table 3-2**). The number of mutations

is below what has been found previously with CspRecT alone (4 off-target mutations per genome)[480], so we conclude that the retron component is not adding substantively to off-target mutations.

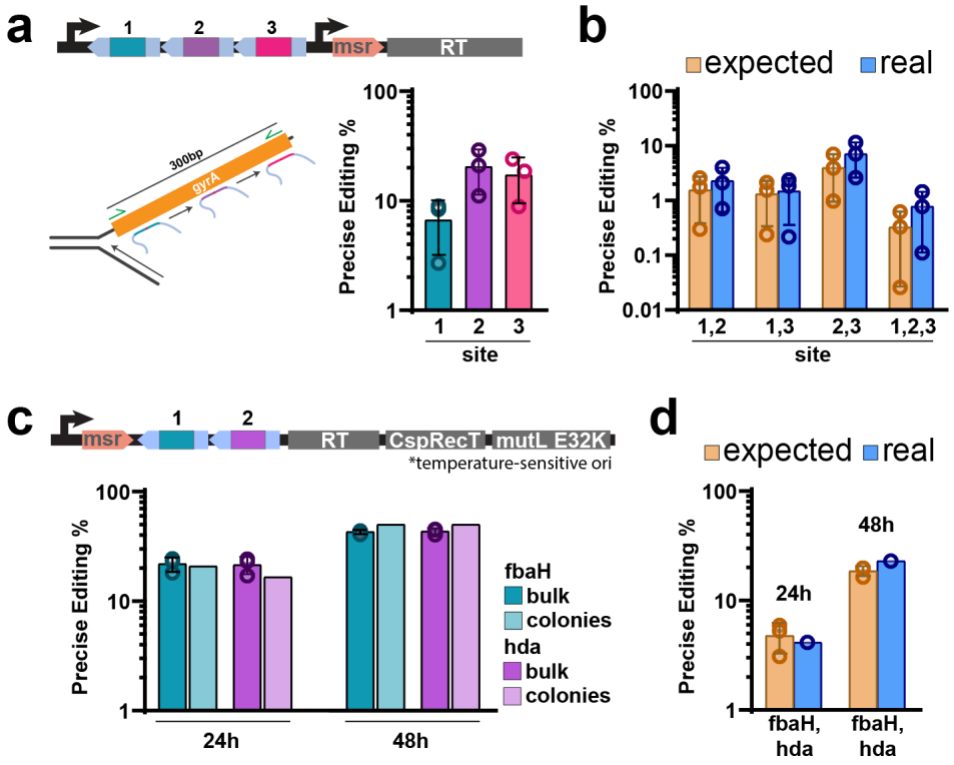


Figure 3-4: Multisite editing of individual bacterial genomes using multitrons

a. *Top: Schematic of retron recombineering using an msd array encoding 3 donors with a single msr sequence in trans. Bottom: (left) schematic of the multitron recombineering process at this locus. All retron donors are able to target the lagging strand of gyrA gene during bacterial replication in a chromosomal window of 300 bp. Green arrows represent the primers used to amplify the target region. (right) quantification of precise editing rates of individual target sites along the gyrA gene, circles show each of the three biological replicates, bars are mean ±SD.* ***b.*** *Quantification of expected (product of bulk rates at each indicated site) and real precise editing rates of double and triple combinatorial edits in the gyrA locus of an individual genome. Circles show each of the three biological replicates, bars are mean ±SD (two-way ANOVA, expected vs real, P=0.0765).* ***c.*** *Top: Schematic of single-plasmid, temperature sensitive multitron architecture. Below: Editing rates for each indicated site at each time point from bulk (Illumina amplicon sequencing) and individual colony sequencing. Circles show each of the three biological replicates, bars are mean ±SD. Mean colony sequencing rates are indicated with a bar.* ***d.*** *Quantification of expected (product of bulk rates at each indicated site) and real precise editing rates of double edits in individual genomes. Circles show each of the three biological replicates, bars are mean ±SD (two-way ANOVA, expected vs real, P=0.2734). Colony sequencing represented by a single point.*

Metabolic Engineering in Bacterial Genomes Using Multitrons

We next pushed toward a proof-of-concept use of multitrons in metabolic engineering by modifying bacterial genomes. First**,** we next assessed the ability of re-optimised, temperature-sensitive arrayed msd multitrons to simultaneously edit five positions (*hda, fbaH, priB, rpoB* and *gyrA*). All sites were precisely edited after 24h, and editing continued to increase over the next 24h following a passage, illustrating the continuous nature of the retron-derived editing (**Figure 3-5a**).

To test multitrons in the context of metabolic engineering, we chose to focus on increasing production of lycopene by modifying genes in its biosynthetic pathway (**Figure 3-5b**). We selected eight bacterial genes which have been shown to affect lycopene yield[483,509–511] (**Figure 3-5b**)**.** Five of them (*dxs, idi, ispA, ispC, rpoS*) were subjected to modification of their RBS regions to enhance their similarity to the canonical Shine-Dalgarno sequence (TAAGGAGGT)[512]. The other three genes (*gmpA, gdhA, fdhF*) were specifically targeted for inactivation by the introduction of premature stop codons within their open reading frames.

We established a general workflow for metabolic engineering using multitrons (**Figure 3-5c**). The multitron plasmid (MP) was generated using a one-pot golden gate approach[513] to clone arrayed msds encoding different donors. The MP was next transformed into the bacterial host harbouring the lycopene plasmid (LP, a plasmid containing three essential genes (*crtE, crtI, crTB*) required for lycopene production[514]. Editing cycles were carried out at the permissive temperature (30°C), with dilutions of the culture after every cycle. Editing targets were sequenced in bulk using Illumina MiSeq to determine overall efficiencies. In parallel, cells were plated at 37°C to cure the MP. Finally,

red colonies (indicative of lycopene) from the plates were selected for further quantification of lycopene production levels (**Figure 3-5c**).

In total, we tested six different arrayed msd multitrons across this workflow, containing target gene donors in combinations that have been have been shown to increase lycopene yield[483]. Editing rates were measured after cycles 1 and 3 of editing (24h and 72h, respectively) showing values that increase with time (**Figure 3-5d**). After 72h of editing, the precise editing rates when making one or two mutations ranged from 10 to 40%. When making three or five mutations, editing rates were lower, which could be due to the known negative fitness effect[483] of these mutations on the bacterial growth (**Figure 3-5d**).

We measured relative lycopene production from 84 isolated red colonies after plating cultures on LB agar plates after editing (**Figure 3-5e**)**.** In each case other than the control, individual colonies produced variable amounts of lycopene, likely resulting from the intended genotypic diversity generated by the editing. As an example, the most productive isolate after RBS optimization of *dxs* and *idi* genes increased lycopene production by more than 400% of control values, there was a second production cluster around 300% of control, and a final cluster around 200% of control (**Figure 3-5e**). We reasoned that these three different clusters may represent a single *dxs* mutation, a single *idi* mutation, and both together. To test that hypothesis, a representative of each cluster was selected and re-streaked for colonies, which were re-measured for lycopene and Sanger sequenced. Indeed, that the best producing isolate carried RBS mutations of both *dxs* and *idi* genes, second-best had only the *dxs* mutation, and the third-best had only the *idi* mutation (**Figure 3-5f**)**.** This proof-of-concept was achieved with a single cloning

reaction (one-pot Golden Gate) to generate a single plasmid and one course of editing, creating both single mutants and the double mutant. To generate this same result without multiplexing would require cloning two distinct editors for each of the sites, running parallel editing, genotyping, and quantification on each single edit. Then, curing the plasmid from an edited clone, adding the opposite plasmid to make the other edit, running another editing course, and finally quantifying the double mutant. Thus, a multiplexed experiment generates a diversity of genotypes and corresponding phenotypes across multiple sites simultaneously.

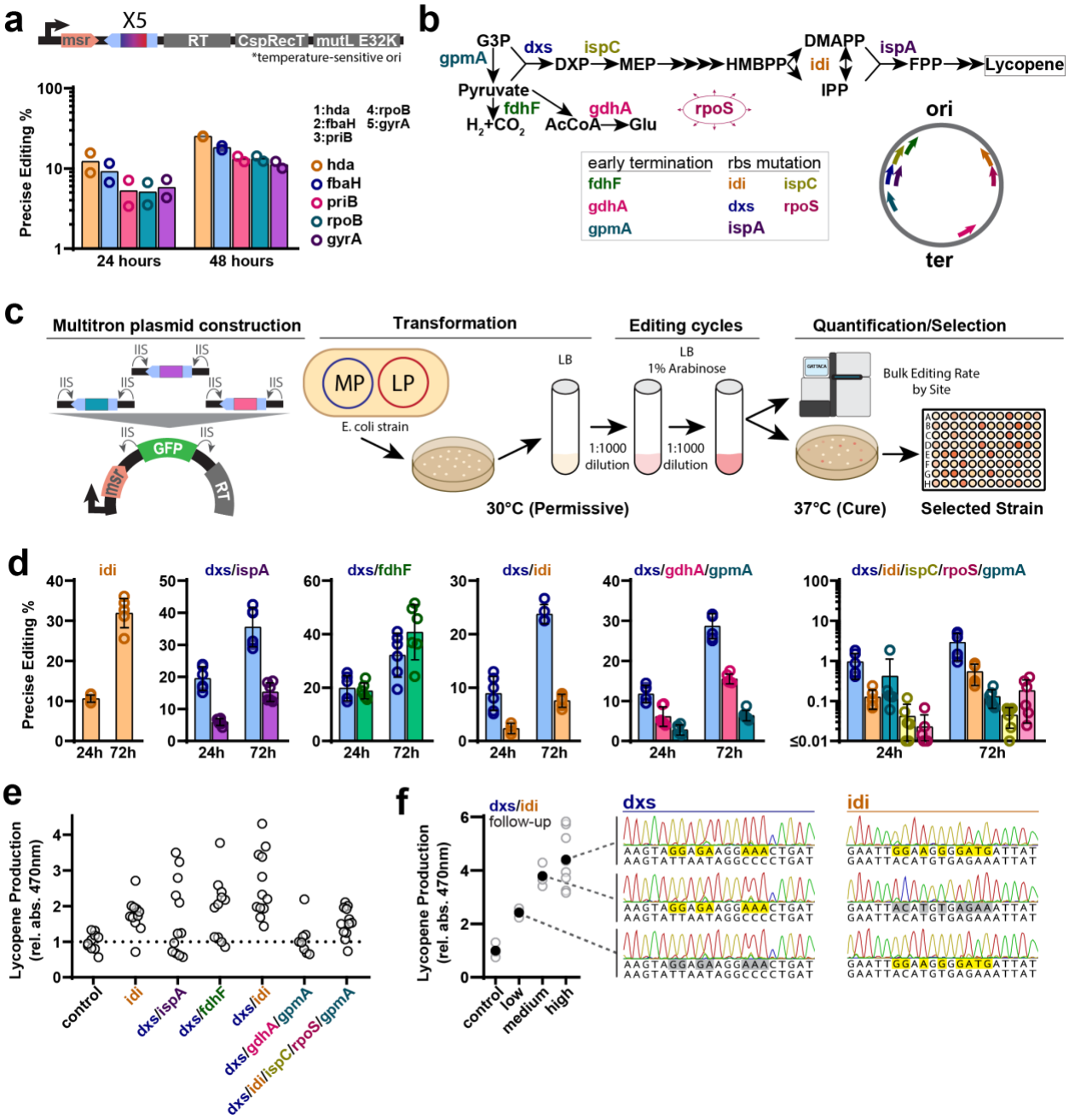


Figure 3-5: Metabolic engineering using multitrons.

a. *Top: architecture of the multiplexed retron recombineering cassette in the temperature sensitive plasmid. The operon is composed of a single msr followed by 5x arrayed msds with donors and the genes encoding the RT, the CspRecT and the dominant negative MutLE32K. Bottom: quantification of precise editing rates using a 5x arrayed msd to edit hda, fbaH, priB, rpoB and gyrA by Illumina sequencing 24h and 48h after of editing (two-way ANOVA, effect of expression time P<0.0001). Circles show each of the three biological replicates, bars are mean ±SD. The order of the donors in the arrayed msd is indicated.* ***b.*** *Top: Schematic of the lycopene biosynthesis pathway, with key genes to increase lycopene production highlighted. Bottom: Schematic of metabolic engineering of lycopene biosynthesis pathway using multiplexed retron recombineering.* ***c.*** *The donors are cloned into a temperature sensitive backbone using a golden gate*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

assembly protocol. Single colonies are grown for 24 h and then induced with arabinose. This cycle is repeated by making 1:1000 dilutions for several days. Editing rates are measured by Illumina sequencing and cultures are plated to select individual colonies based on color for quantification of lycopene production. ***d.*** *Quantification of precise editing rates using different recombitron plasmids containing a variable number of donors to edit genes in the lycopene pathway, quantified by Illumina sequencing after 24h and 72h. Circles show each of the six biological replicates, bars are mean ±SD.* ***e.*** *Quantification of lycopene production in single colonies. Lycopene production was normalised against the average production of the control, which contains the pAC-LYC but was not exposed to the recombineering process. Each point represents a colony (n=12).* ***f.*** *Quantification of lycopene production from colonies re-isolated from samples in the low (~2X control), medium (~3X control), and high (~4x control) production clusters of the dxs/idi condition. Open circles are individual colony values (3 biological replicates or the control, low, and medium groups; eight biological replicates for the high group) and closed circles are the mean. Sanger sequencing examples to the right illustrate the genotype of each subset (all individual colonies within a condition have identical genotypes). Additional statistical details in* ***Supplementary Table 3-1****.*

Multitrons with CRISPR Editing in Eukaryotic Cells

Given the success of the arrayed msd multitron in recombineering, we next sought to expand the utility of this technology to eukaryotic cells. Retron RT-Donors have been used in *S. cerevisiae* in combination with CRISPR Cas9 and gRNAs to install precise mutations via templated repair of a cut site[438,440] (**Figure 3-6a; Extended Data Figure 3-4a**). The architecture of the donor element in yeast is typically a retron ncRNA fused to a CRISPR gRNA and scaffold, all surrounded by ribozymes to excise the editing elements from an mRNA. Given the goal of engineering a eukaryotic msd array, the relatively large, structured ribozymes present a potential engineering hurdle if they need to be multiply duplicated. Therefore, we first tested replacement of the ribozymes with Csy4 recognition sites and Csy4 nuclease expression[515] by comparing a singleplex retron-derived precise editor of the *ADE2* locus in the standard arrangement using ribozymes against an alternate version in which the flanking ribozymes were replaced by Csy4 recognition sites. In both cases, we tested editing with or without the inclusion of a Csy4 gene in an integrated, inducible, genomic cassette that also expresses the retron RT and Cas9. We found, as expected, no effect of Csy4 expression on the ribozyme version of the precise

editor, but a dramatic effect of Csy4 expression on the alternate version with Csy4 sites. Precise editing nearly matched the efficiently of the ribozyme version with Csy4 expression, but was sharply reduced in its absence, indicating that processing of the non-coding elements is required and can be achieved using Csy4 (**Figure 3-6a**).

We next tested a eukaryotic multitron based on an array of ncRNA/gRNAs targeting *ADE2* and *FAA1* for precise mutations of three base pairs each. For each site, the ncRNA encoding the donor for the site was fused to the gRNA for the same site. The two sites were separated by a Csy4 recognition site and the double ncRNA/gRNA array was surrounded by ribozymes (**Figure 3-6b**). Both sites were edited to nearly 100% in the presence of Csy4 expression, and we observed low indel rates that were similar between the ribozyme- and Csy4-processed cassette (**Extended Data Figures 3-4a,b**). In the absence of Csy4, in contrast, the *FAA1* site was edited to nearly 100%, while the *ADE2* editing was sharply reduced. In our multitron, the *ADE2* donor/gRNA was in the first position, suggesting that Csy4 processing is required on the 3' end, adjacent to the gRNA scaffold, but dispensable on the 5' end, adjacent to the msr.

Analogously to our bacterial editors, we verified that edits accumulate on genomes according to the product of the rates at each site. To this end, we compared bulk editing rates across the *ADE2* and *FAA1* sites to rates of edits in individual colonies. As in the bacterial experiments, colony sequencing matched bulk sequencing for both individual sites and for the expected frequency of double edits. We found that virtually all of the colonies sequenced contained the precise edits intended, consistent with the rates inferred from bulk Illumina amplicon sequencing (**Figures 3-6c,d**).

It is preferable to minimise the donor/gRNA unit for practical reasons of construction, just as in the prokaryotic version. Therefore, in a parallel to the prokaryotic msd array multitron, we engineered a eukaryotic msd/gRNA array multitron, transferring the msr to a distinct transcript to reduce editing unit size and avoid long direct repeats (**Extended Data Figure 3-4c**). This enabled construction of multitrons of arbitrary size using efficient one-step golden gate cloning. The msd encoding the donor remains fused to its matched gRNA, while a trans msr is able to function as a primer to create the RT-Donor internally (**Extended Data Figure 3-4d**). We tested versions of this eukaryotic arrayed msd/gRNA multitron to precisely edit two, three, or five non-adjacent sites simultaneously (**Figures 3-6e,f; Extended Data Figure 3-4e**). In each case, all targeted sites were edited, with precise edits and indels increasing over time (**Extended Data Figures 3-4f-h**).

Finally, we sought to test whether the engineered eukaryotic msd/gRNA array multitron would enable precise genome editing in human cells. We adapted an approach for multiplexing pegRNA expression[502], described initially to enable multiplexed prime and base editing, to enable the expression and processing of multiple retron msds and a single retron msr in trans. This yielded expression cassettes analogous to those developed for yeast editing, with tRNAs driving the processing of the msd/gRNA cassettes. We found that these engineered cassettes enabled the simultaneous precise edits of three non-adjacent sites in the human genome, from a single plasmid, in cultured HEK293T human cells (**Figure 3-6g**). Taken together, our data shows that the arrayed msd multitron with trans msr is a generalizable strategy for multiplexing edits within a genome.

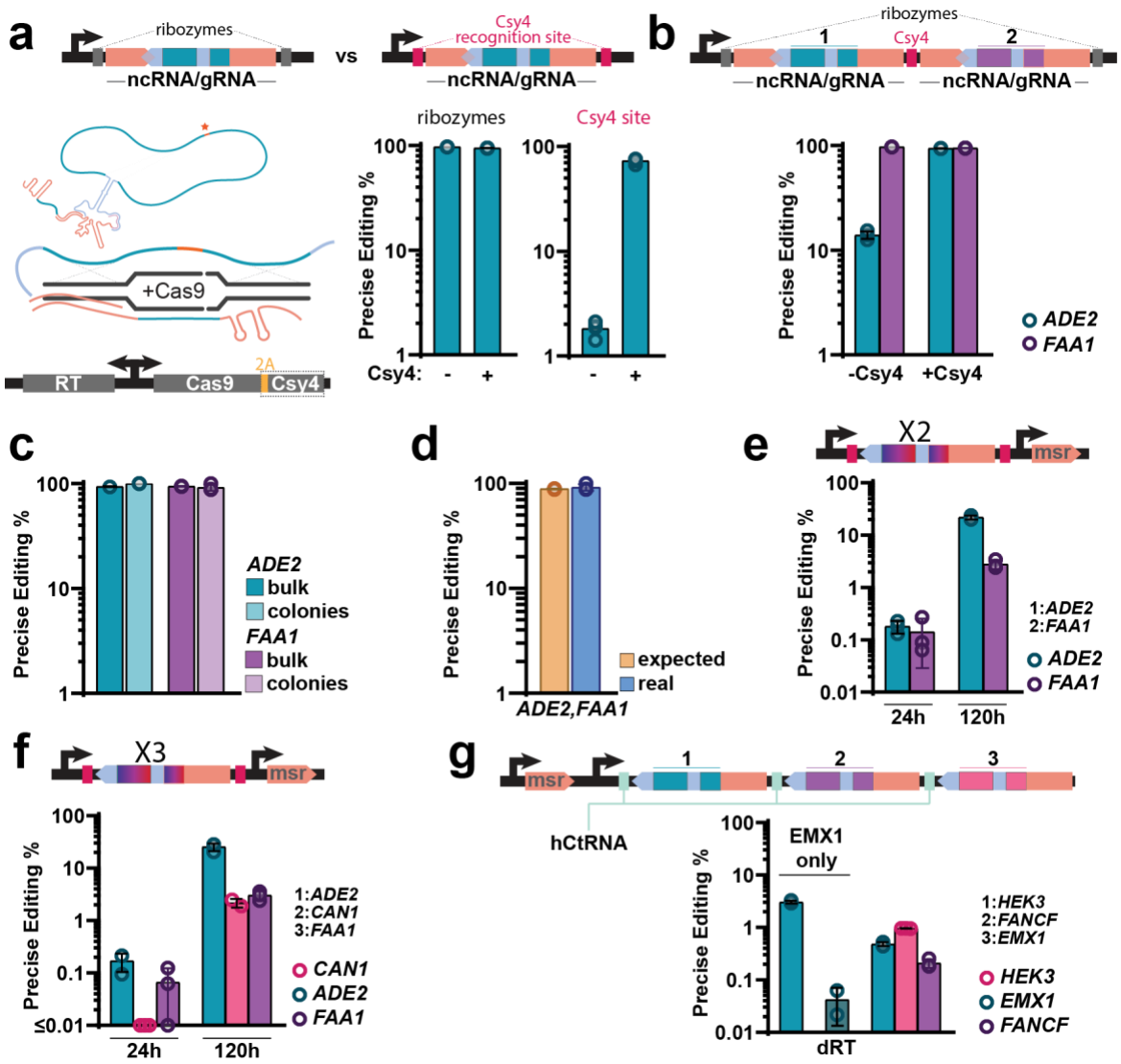


Figure 3-6: Arrayed retron msds enable multiplexed editing in eukaryotic cells.

a-g*: quantification of precise editing was determined by Illumina sequencing after 48h (yeast) and 72h (HEK293T), unless specified; circles show each of the three biological replicates, bars are mean ±SD.* ***a.*** *Top: schematic of the donor-encoding retron ncRNA/gRNA cassette, expressed from a Gal7 promoter and flanked by ribozymes or Csy4 sequences. Bottom left: schematic of a retron ncRNA/gRNA hybrid, depicted above the yeast genome-encoded protein-coding expression cassette. Bottom right: quantification of precise editing of the yeast ADE2 locus. Absence/presence of Csy4 in the protein-coding expression cassette is shown below the graph (Sidak's corrected multiple comparisons, effect of Csy4 expression, ribozyme construction P=0.2779, Csy4 construct P<0.0001).* ***b.*** *Top: schematic of an arrayed retron ncRNA/gRNA cassette, expressed from a Gal7 promoter, flanked by ribozymes, and separated by a Csy4 sequence. The editors in positions 1 and 2 target the ADE2 and FAA1 loci, respectively. Bottom: quantification of precise editing of the yeast ADE2 and FAA1 loci. Absence/presence of Csy4 in the protein-coding expression cassette is shown below the graph (Sidak's corrected multiple comparisons, effect of csy4 expression, ADE2 P<0.0001, FAA1 P=0.0012).* ***c.*** *Editing rates for each locus and time point from bulk and individual colony sequencing (bar represents mean).* ***d.*** *Quantification of expected (product of bulk rates) and real precise editing rates of double edits in individual genomes (two-way ANOVA, expected vs real, P=0. 0.4318).* ***e-f****, top: schematic of 2- and 3x arrayed retron msdRNA-gRNA cassettes, as shown in*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

***Extended Data Figure 3-4c**. Bottom: quantification of precise editing of the yeast ADE2 and FAA1 (e); and ADE2, CAN1 and FAA1 (f) loci, after 24 and 120h of editing. Two-way ANOVA, effect of expression time, e P<0.0001, f P<0.0001. **g.** Arrayed retron msds enable multiplexed editing in human cells. Top: schematic of the donor-encoding retron ncRNA/gRNA expression cassette expressed from an H1 promoter and flanked by tRNA-Cys-GCA (hCtRNA) sequences. Bottom: quantification of precise editing of the HEK293T EMX1, HEK3 and FANCF loci. Absence/presence of a catalytically active retron RT is shown below the graph. Additional statistical details in **Supplementary Table 3-1**.*

3.4 Discussion

This work demonstrates the construction, optimization, and use of multitrons for multiplexed precise editing within genomes of prokaryotic and eukaryotic cells. Final versions make use of donor-encoding retron msd arrays. Critically, we engineered the msd array format by optimizing not only for editing efficiency, but also for enabling practical cellular and molecular workflows. The compact multitron form is compatible with single-plasmid designs, one-step golden gate assembly, and plasmid removal in prokaryotic cells. These features should permit widespread adoption of the multitron editing approach. A concurrent work has shown a similar approach, providing independent validation of the utility of multiplexed retrons for recombineering[446].

We demonstrate simultaneous editing of up to five sites, with replacements of up to 8 base pairs per site, and deletions of up to 100 bases. This approach builds on previous work using oligonucleotides for MAGE by enabling efficient multisite editing without repeated transformations and by enabling a user to specify distinct combinations of donors per cell rather than relying on the random segregation of electroporated oligos. Multitrons enable a wider range of precise mutations than multiplexed base editors, and a more compact and simplified form than multiplexed prime editors.

We found that the rate of combinatorial editing on a single genome was predicted by product of rates at each individual site. As the number of editing sites increased, the rate at each site decreased. Thus, for 4+ edits, the rate of achieving all mutations on a single genome can drop well below 1:1,000. Whether his rate is high enough will depend on the application. For instance, if edited cells are to be subjected to a selective phenotyping assay, the fact that combinatorial mutants are present in the population, even

at low rates, is sufficient to enable quantification of enrichment or depletion. If however, one needs to isolate a clone without phenotypic selection, we would recommend limiting the edits per round of editing to ≤3 at this time. Further development of the technology or the addition of simultaneous counter-selection will help drive the practical number of edits up in the future.

For contextualization to other technologies, one alternative is base editing, which can also be multiplexed. On the upside, Base Editors can reach efficiencies of over 80%[501] and can be multiplexed to more than 30 loci[502]. However, it is important to note that only 2 of the 14 edits we made in bacteria and none of the edits made to yeast are suited to base editing. The deletions and RBS modifications that we made are a particularly salient example of a place where base editors fall short. MAGE is a more relevant comparison to the bacterial work and can achieve similar efficiencies, although with dramatically more hands on type to complete the multiple electroporation cycles. However, MAGE cannot be used to make the edits that we show in yeast or human cells (new to the revision) so as a technology, we would argue that the multitrons are a more universal platform.

The existing yeast genome editing toolbox is vast and spans from simple HR-based editing to more nuanced, multiplexed approaches that have enabled both trackable, genome-wide phenotypic screens and targeted, saturation mutagenesis of individual ORFs[210,313,515–519]. However, “trackable and multiplex” in this context has usually meant many changes across many genomes, with ≤1 change per genome, rather than >1 changes on an individual genome; and tools that do enable multiple changes per single genome typically do not support trackability of precise and varied edits, or require

involved and time-consuming workflows. In this sense, we believe that multitrons, in their ability to support multiple trackable and precise edits per individual genome, will naturally fit into the toolbox of yeast biologists in years to come.

We demonstrate proof-of-concept uses in molecular recording, genetic element minimization, and metabolic engineering. Future development will likely push the scale of multitrons both in the number of simultaneous mutations and the diversity of combinatorial mutations using libraries targeting two or more sites.

3.5 Methods

Biological replicates were taken from distinct samples, not the same sample measured repeatedly. Full statistics can be found in **Supplementary Table 3-1.**

Plasmid Construction

All the plasmids used in this work are listed in **Supplementary Table 3-3**. Furthermore, all the RT-donors and the oligonucleotides containing the desired mutations for the editing experiments are listed in **Supplementary Table 3-4.** All plasmids are available on Addgene: https://www.addgene.org/Seth_Shipman/

E. coli

To clone additional 70 bp donors in a single msd, pSLS.492[440] plasmid containing a *rpoB* donor was used as backbone. To clone a donor upstream of the *rpoB* donor, a 60 bp reverse oligo annealing (25bp) with the 5' region of the msd and containing 35 bp of the new donor, and a 60 bp forward oligo annealing (25bp) with the 5' end of *rpoB* donor and harbouring the other half of the new donor were used. To clone a donor downstream of the *rpoB* donor, a 60 bp forward oligo annealing (25b) with the 3' region of the msd and containing 35 bp of the new donor, and a 60 bp reverse oligo annealing (25bp) with the 3' end of *rpoB* donor and harbouring the other half of the new donor were used. After a 30 cycles PCR reaction with Q5 hot-start high-fidelity polymerase (NEB) following recommended vendor protocol, a KLD reaction (NEB) was carried out to self-ligate the plasmid encoding an additional donor.

To construct the plasmids harbouring the retron arrays in their different architectures the pCDF-DUET-1 vector (Novagen) was used as a backbone. A parental plasmid (pAGD159; **Supplementary Table 3-3**) containing a whole ncRNA with a *gyrA* donor downstream of the first T7 promoter, and Eco1-RT downstream of the second T7 promoter was constructed. To assess whole ncRNA retron arrays, the ncRNA harbouring the *rpoB* donor from pSLS.492 was amplified and cloned upstream and downstream of the *gyrA*-containing ncRNA by Gibson Assembly. To construct the plasmids containing the msd array, firstly, the msr was deleted from pAGD159 and subsequently cloned between the second T7 promoter and Eco1-RT using a Gibson Assembly approach. Finally, the msd harbouring the *rpoB* donor from pSLS.492 was amplified and cloned upstream and downstream of the *gyrA*-containing ncRNA by Gibson Assembly. To test if the msd array could act as a single transcript unit independent of the msr region, a T7 terminator was cloned between the msd array and the second T7 promoter.

To construct multitrons containing more than 2 arrayed msd a one-pot Golden Gate cloning approach was used. Firstly, a plasmid containing a sfGFP stuffer flanked by two inverted BsaI (type IIS restriction enzyme) target sites were cloned in the place of the msd Array generating pAGD236 **(see Figure 3-4c for reference).** Editing units, based on a msd with a donor were order as gBlocks (IDT) flanked by inverted BsaI target sites and compatible nucleotide overhangs to clone them in tandem. The Golden Gate protocol was carried out in 20uL reactions as follows: 1 uL pAGD236, 5uL of each gBlock (3uL for 5x msd arrays), 1.5uL BsaI (NEB), 2uL T4 DNA ligase Buffer, 0.5 uL T4 DNA ligase (NEB). The reaction consists on 30 or 60 cycles (depending on the complexity) of 5 min at 16C and 5 min at 37C and a final cycle of 10 min at 60C.

To optimise multitrons for metabolic engineering, the retron cassette (ncRNA and RT) from pSLS.492 was cloned into pORTMAGE-Ec1[480] upstream of the CspRecT gene **(Extended Data Figure 3-2a)**. RBS optimization of Eco1 RT and CspRecT genes were carried out using primers that contain the optimised RBS and self-ligating the plasmids using KLD reaction mix. Finally, recombineering operon was cloned into pKD46[489] backbone to obtain the parental temperature-sensitive multitron plasmid (pAGD248). Multitron msd array architecture with the sfGFP stuffer flanked by two inverted BsaI described previously was cloned into pAGD248 generating pAGD335. The golden gate reaction was used to clone gBlocks containing the donors into the pAGD335 backbone to generate the multitrons versions used in **Figure 3-4**.

S. cerevisiae

To assess whether Csy4 could enable the processing of editrons and retron msd/Cas9 gRNA units for genome editing, pSCL390, a derivative of pZS.157 (Addgene #114454), was generated with a yeast codon-optimised P2A-Csy4 CDS gblock (IDT) cloned downstream of the SpCas9 CDS by Gibson Assembly.

To compare the genome editing efficiencies of ribozyme-processed editrons to Csy4-processed editrons, pSCL.396, a derivative of pSCL.39 (Addgene #184973), was generated with the 5’ Hammerhead ribozyme and 3’ HDV ribozyme replaced by Csy4 recognition sites by amplification of the editron and backbone from pSCL.39 and assembled via Gibson Assembly.

To assess whether Csy4 could enable the processing of arrayed editrons, we generated pSCL.391, a derivative of pSCL.39 where a second editron, targeting the *S.*

cerevisiae FAA1 locus was added on the 3' end of the *ADE2*-targeting editron by Gibson Assembly. The cassette thus consists of two editrons, separated by a Csy4 recognition site, and flanked by a Hammerhead ribozyme and a HDV ribozyme on the 5' and 3' of the expression cassette, respectively.

To construct plasmids for the expression of retron msd arrays, first, a Golden Gate compatible entry vector, pSCL.452 was generated that carries the Gal7 promoter and terminator, alongside a cassette for expression of the retron msr from a Pol III SNR52 promoter. pSCL.452 is a derivative of a derivative of pSCL.39, generated by Gibson Assembly of the pSCL.39 backbone, amplified to replace the recombitron with inverted PaqCI sites for Golden Gate assembly, with a gblock (IDT) encoding pSNR52p-msr-SUP4t.

Next, plasmids carrying retron msd arrays for the editing of multiple loci in the yeast genome were generated by Golden Gate cloning of pre PaqCI-digested pSCL.452 with gBlocks (IDT) that encoded a PaqCI cut site, a retron msd-encoded donor and paired gRNA for editing, a Csy4 recognition sequence, and a PaqCI cut site **(Figure 3-6e)**. gBlocks were ordered with compatible nucleotide overhangs to enable random cloning of all combinations of gblocks into the entry plasmid, after PaqCI digestion. We ordered gblocks to edit the *ADE2*, *FAA1, TRP2, SGS1* and CAN1 loci. These were cloned into the PaqCI-digested pSCL.452 backbone by Golden Gate cloning, yielding plasmids pSCL.473 (editors for *ADE2, FAA1*), pSCL.475 (editors for *ADE2*, *CAN1* and *FAA1*) and pSCL.672 (editors for *ADE2, FAA1, TRP2, SGS1* and *CAN1*).

H. sapiens

All human vectors are derivatives of pSCL.273, itself a derivative of pCAGGS[520]. 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 with inserts that contain {retron RT – H1 promoter – $hCtRNA_n$_$msdRNA_n$_$gRNA_n$}, by Gibson Assembly or Golden Gate cloning (see Fig 6e for an illustration of this principle). The retron RT (or its catalytically dead counterpart) and H1 promoter fragments were synthesised through IDT, as were the $hCtRNA_n$_$msdRNA_n$_$gRNA_n$ units. Golden gate cloning of these elements alongside 3 editor units (*EMX1*, *FANCF*, and *HEK3*) yielded plasmids pSCL.757 (CAGp-Eco3RT-TYpA // H1-msr-tRNA-Cys-GCA-*EMX1*_msd-gRNA); pSCL.758 (CAGp-Eco3RT-TYpA // H1-msr-tRNA-Cys-GCA-*HEK3*_msd-gRNA-tRNA-Cys-GCA-*FANCF*_msd-gRNA-tRNA-Cys-GCA-*EMX1*_msd-gRNA) and pSCL.760 (CAGp-dEco3RT-TYpA // H1-msr-tRNA-Cys-GCA-*EMX1*_msd-gRNA).

Strains and Growth Conditions

All bacterial and yeast strains are listed in **Supplementary Table 3-5**.

Bacterial Strains

The *E. coli* strains used in this study were DH5α (New England Biolabs) for cloning purposes, bMS.346 (DE3) for retron recombineering assays. Bacteria were grown in LB medium (10 g/l tryptone, 5 g/l yeast extract, 5 g/l NaCl). Antibiotics were added as required (carbenicillin, spectinomycin, kanamycin and chloramphenicol).

Yeast Strains

All yeast strains were created by LiAc/SS carrier DNA/PEG transformation[490] of BY4742[438]. Strains for evaluating the effect of Csy4 on genome editing efficiency were created by BY4742 integration of plasmids pZS.157 (Addgene #114454) or pSCL.390. The plasmids were KpnI-linearised and inserted into the genome by homologous recombination into the *HIS3* locus. Transformants were isolated on SC-HIS plates.

Bacterial Recombineering expression and analysis

In multitron experiments edit bacterial genomes, the retron cassette encoded in a pET-21 (+) plasmid (Novagen) and the CspRecT and mutLE32K in the plasmid pORTMAGE-Ec1[480] were overexpressed using 1 mM IPTG, 1 mM m-toluic acid and 0.2% arabinose for 16 h with shaking at 37C. For the molecular recording assay **(Figure 3-1g)**, a control without m-Tol and different concentration of the inducer, ranging from 0,005 mM to 0,1 mM, were added. To engineer the lycopene metabolic pathway **(Figure 3-4)**, bMS.346 electrocompetent cells containing pAC-LYC[514] plasmid, were transformed with different multitron plasmid versions **(Supplementary Tables 3-3 and 3-4).** and growth for 16 h at 30C. Single colonies from the transformation plate were inoculated into 500uL

of LB in triplicates in 1mL deep-well plates and incubated at 30C for 24 h with vigorous shaking to prevent the cells from settling. A 1:1000 dilution of the cultures were passaged into LB 1% arabinose and incubated at 30C for 24 h with vigorous shaking. This step was repeated for a total of 72h.

After the different type of assays carried out in this study, a volume of 25 ul of culture was collected, mixed with 25 ul of water and incubated at 95C for 10 min. A volume of 1 ul of this boiled culture was used as a template in 30-ul reactions with primers flanking the edit site, which additionally contained adapters for Illumina sequencing preparation (**Supplementary Table 3-6**)**.** 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

The parental strains (–Csy4: HIS3::pZS.157; +Csy4: *HIS3*::pSCL390) were transformed with variants of the editron expression cassettes by LiAc/SS carrier DNA/PEG transformation. Single colonies from the transformation plate were inoculated into 500uL of SC-*HIS*-*URA* 2% raffinose in triplicates in 1mL deep-well plates and incubated at 30C for 24 h with vigorous shaking to prevent the cells from settling. Cultures were passaged into SC-*HIS*-*URA* 2% galactose and incubated at 30C for 24 h with vigorous shaking. This was repeated once more for experiments meant to compare the genome editing efficiencies of ribozyme-processed editrons to Csy4-processed editrons, for a total of 48h of editing; and four more times for experiments meant to assess whether arrays of retron msds could be used to edit multiple loci in the yeast genome, for a total

of 120h of editing. At each timepoint of galactose-induced editing, a 250uL aliquot of the cultures was harvested, pelleted and washed with water, and prepped for deep sequencing of the loci of interest.

To compare the bulk editing rates across sites to rates of edits in individual colonies for the Csy4-processed editrons, after 48h of editing, dilutions were plated on SC-*HIS-URA* plates. For each of 3 biological replicates, 10 colonies were grown overnight in SC-*HIS-URA* to saturation and subjected to genomic DNA extraction and targeted PCR of the *ADE2* and *FAA1* loci, as described below. Amplicons were sent for Sanger sequencing, and editing rates per biological replicated were calculated by assessing the Sanger reads for the 10 colonies per biological replicate for the expected precise edit.

Samples were prepped for deep sequencing of the edited loci as described previously[440]. Briefly, genomic DNA was extracted by (1) resuspending the cell pellets in 120uL of lysis buffer (100 mM EDTA pH 8, 50 mM Tris-HCl pH 8, 2% SDS) and heating them to 95C for 15 min; (2) cooling the lysate on ice and adding 60uL of protein precipitation buffer (7.5 M ammonium acetate), then inverting gently and placing samples at -20C for 10min; (3) centrifugation of the samples at maximum speed for 2mins (or until a clear supernatant forms) and collecting the supernatant (~100uL) in new 1.5mL tubes; (4) precipitating the nucleic acids by adding equal parts of ice-cold isopropanol to the samples, mixing the samples thoroughly and incubating the mix at -20C for 10min (or overnight for higher yield), followed by pelleting by centrifugation at maximum speed for 2min; (5) washing the pellet twice with 200 µl of ice-cold 70% ethanol, followed by air-drying it; and (6) resuspending the pellet in 40 µl of water. 0.5uL of gDNA was used as template in 20-µl PCR reactions with primers flanking the edit site in of the target locus,

which additionally contained adapters for Illumina sequencing preparation (**Supplementary Table 3-6**). Importantly, the primers do not bind to the retron msd donor sequence. These amplicons were indexed and sequenced on an Illumina MiSeq instrument and processed with custom Python software to quantify the percentage of precise edits using the retron derived RT-DNA template.

Human Cell Culture

HEK293T cells, expressing spCas9 from a piggyBac integrated, TRE3G driven, doxycycline-inducible (1 µg/ml) cassette[440], were seeded at 7 x10^5 live cells/well in coated 6-well plates and grown in DMEM +GlutaMax supplement (Thermo Fisher #10566016) overnight. Lipofectamine 3000 transfection mixes were prepared in independent triplicates and cells were transfected with 5ug of plasmid per well (3 wells per plasmid). Cells were passaged the next day and doxycycline was refreshed at passaging. Cells were grown for an additional 48h, for a total of 72h of editing. Three days after transfection, cells were collected for sequencing analysis. To prepare samples for sequencing, cell pellets were collected, and gDNA was extracted using a QIAamp DNA mini kit according to the manufacturer's instructions. DNA was eluted in 150 µl of ultra-pure, nuclease-free water. 0.5uL of gDNA was used as template in 20-µl PCR reactions with primers flanking the edit site in of the target locus, which additionally contained adapters for Illumina sequencing preparation (**Supplementary Table 3-6**). Importantly, the primers do not bind to the retron msd donor sequence. These amplicons were indexed and sequenced on an Illumina MiSeq instrument and processed with custom Python

software to quantify the percentage of precise edits using the retron derived RT-DNA template.

Whole-Genome Sequencing to Measure Off-Target Mutagenesis

A total of 7 genomes were sequenced using a shot-gun approach: *E. coli* bMS.346 parental strain, 3 individual colonies after one recombineering round using a wild-type Eco1 RT and 3 individual colonies after one recombineering round using a dead Eco 1 RT. Prior to sequencing, 3 ml LB liquid culture of each isolate was grown for 16h at 37C. The gDNA was isolated by using the Quick-DNA/RNA™ Miniprep Plus Kit (Zymo Research). Extracted gDNA was measured using QubitTM 1X dsDNA High Sensitive (HS; Thermo Scientific). gDNA was tagmented using Tn5 transposase using the following reaction (50uL): 25 uL 2x TD Buffer (20 mM Tris-HCl pH 7.6, 10 mM MgCl2 and 20% dimethyl formamide), 2.5uL Tn5 (in-house prepared) and 50 ng gDNA. The reaction was incubated for 1.5h at 37C. The gDNA was cleaned-up and eluted in 15uL using the DNA Clean & Concentrator (Zymo Research). Tagmented gDNAs were indexed and sequenced on an Illumina MiSeq instrument. *E. coli* strain bMS.346 whole genome variants were called against *E. coli* K12 sbstr. MG1655 genome (accession no. NC_000913) using Geneious Prime 2023.2.1 software alignment tools. Variants appearing in the genome of the wild-type and dead RT isolates were called against the bMS.346 parental strain.

Colorimetric screen and assay for lycopene production

After cycle 3 (72h) of the metabolic engineering assay, cells from the edited bMS.346 populations using different multitrons were plated on LB-chloramphenicol agar plates and grown for 1 day at 30C and 2 days more in darkness and at room temperature to produce red colonies. Per edited population with a multitron, plates containing around 10^3 colonies were screened by visual inspection searching for increased red colour intensity. A total of 84 colonies (12 isolates from each multitron version and 12 from the control) were selected for lycopene quantification. These isolated colonies were grown into 1 mL LB-chloramphenicol in 1 mL deep-well plates for 24h at 37C to cure multitron plasmid. For lycopene extraction, 1 ml of cells were centrifuged at 16,000g for 30s, the supernatant was removed and the cell pellet was resuspended with 1 mL water. Cells were re-centrifuged at 16,000g for 30 s, the supernatant was removed and the cells were resuspended in 200 ml acetone and incubated in the dark for 15 min at 55C with intermittent vortexing. The mixture was centrifuged at 16,000g for 1 min and the supernatant containing the lycopene was transferred to 96 white/clear bottom plate. Absorbance at 470 nm of the extracted lycopene solution was measured using a spectrophotometer to determine the lycopene content. Lycopene yield of the different colonies from each was calculated by normalizing the times of lycopene production against the control. Cells coming from different clusters of lycopene production were re-streaked on LB-chloramphenicol agar plates grown for 24h at 30C and for another 48h at room temperature. Between 3 and 8 colonies from each re-striking were selected to quantify the lycopene production following the described protocol and for Sanger sequencing across the *dxs/idi* targets.

Assessment of plasmid stability

E. coli

Recombineering plasmid was transformed into *E. coli* strain bMS.346, followed by 5 days of growing and diluting in the presence or absence of the arabinose. A dilution of the final culture was diluted and plated. Finally, the msd Array of 10 individual colonies per replicate (n=3) were amplified and sequenced to assess genetic stability of the multitron approach (see **Extended Data Figure 3-2f** for reference).

S. cerevisiae

Three individual colonies of yeast carrying 2, 3 or 5 donor arrayed retron msdRNA-Cas9 gRNA expression cassettes were inoculated in SC-*URA-HIS* 2% Raffinose media, and passaged 5 times overnight in SC-*URA-HIS* with 2% Galactose, for a total of 120h of editing at 30C. After 120h of editing, dilutions were plated on SC-*HIS-URA* plates and 10 colonies for each biological replicates were subjected to plasmid extraction. Plasmids were sent for whole-plasmid sequencing and consensus reads were aligned to the reference plasmid.

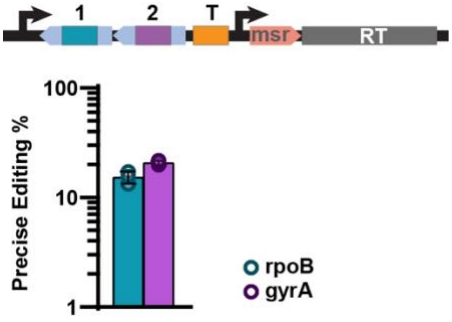
Data availability

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 is available on NCBI SRA as BioProject ID PRJNA1107632.

Code availability

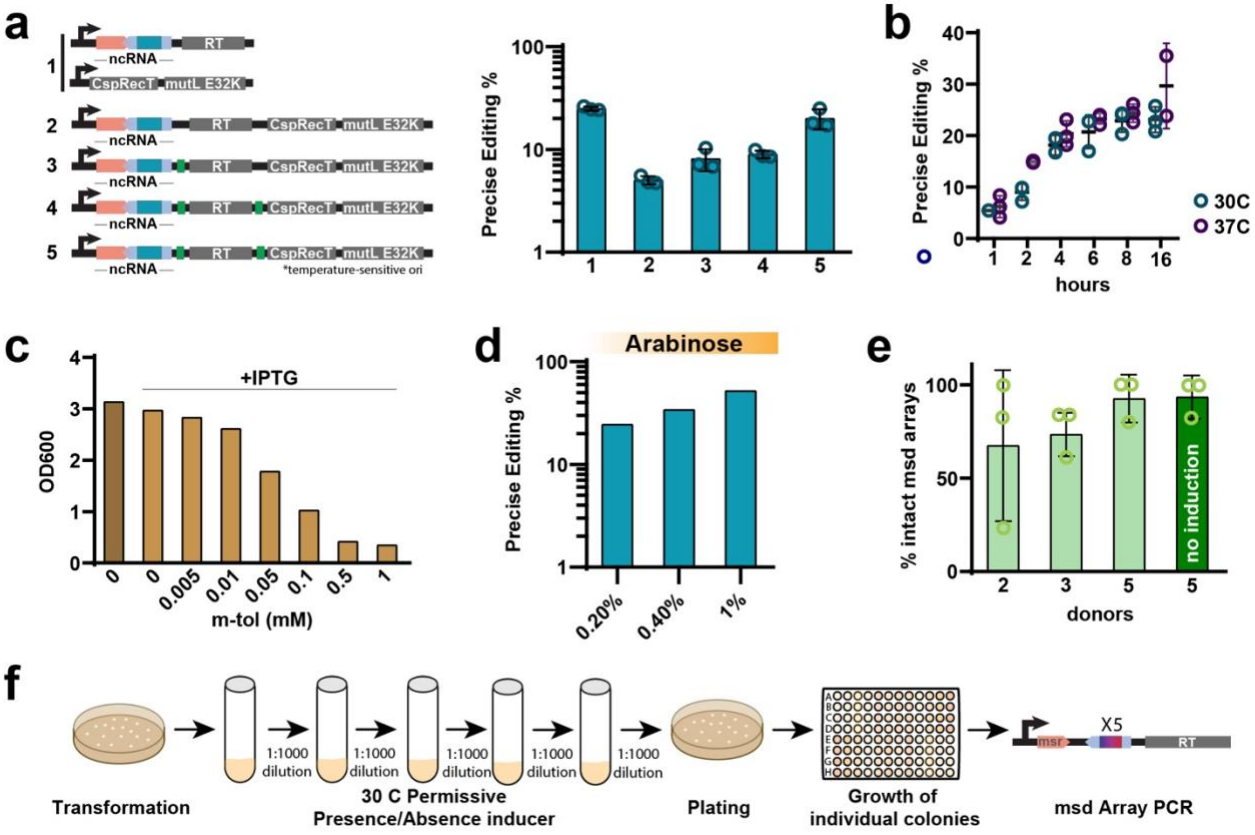
Custom code to process or analyse data from this study is available on GitHub: https://github.com/Shipman-Lab/multitrons (DOI: 10.5281/zenodo.11289190)

3.6 Supplementary Information



Extended Data Figure 3-7: Trans msr multitron architecture enables precise genome editing

Top: Schematic of retron recombineering using an msd array with a single msr sequence in trans including a terminator (T) between the msd array and msr. Bottom: quantification of precise editing rates for precise editing of rpoB or gyrA simultaneously by Illumina sequencing after 24h of editing. Circles show each of the three biological replicates, bars are mean ±SD.

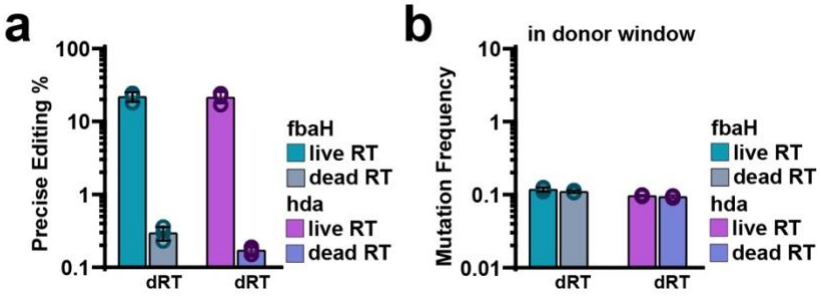


Extended Data Figure 3-8: Optimization of retron recombineering using a single plasmid.

(Figure caption continued on the next page)

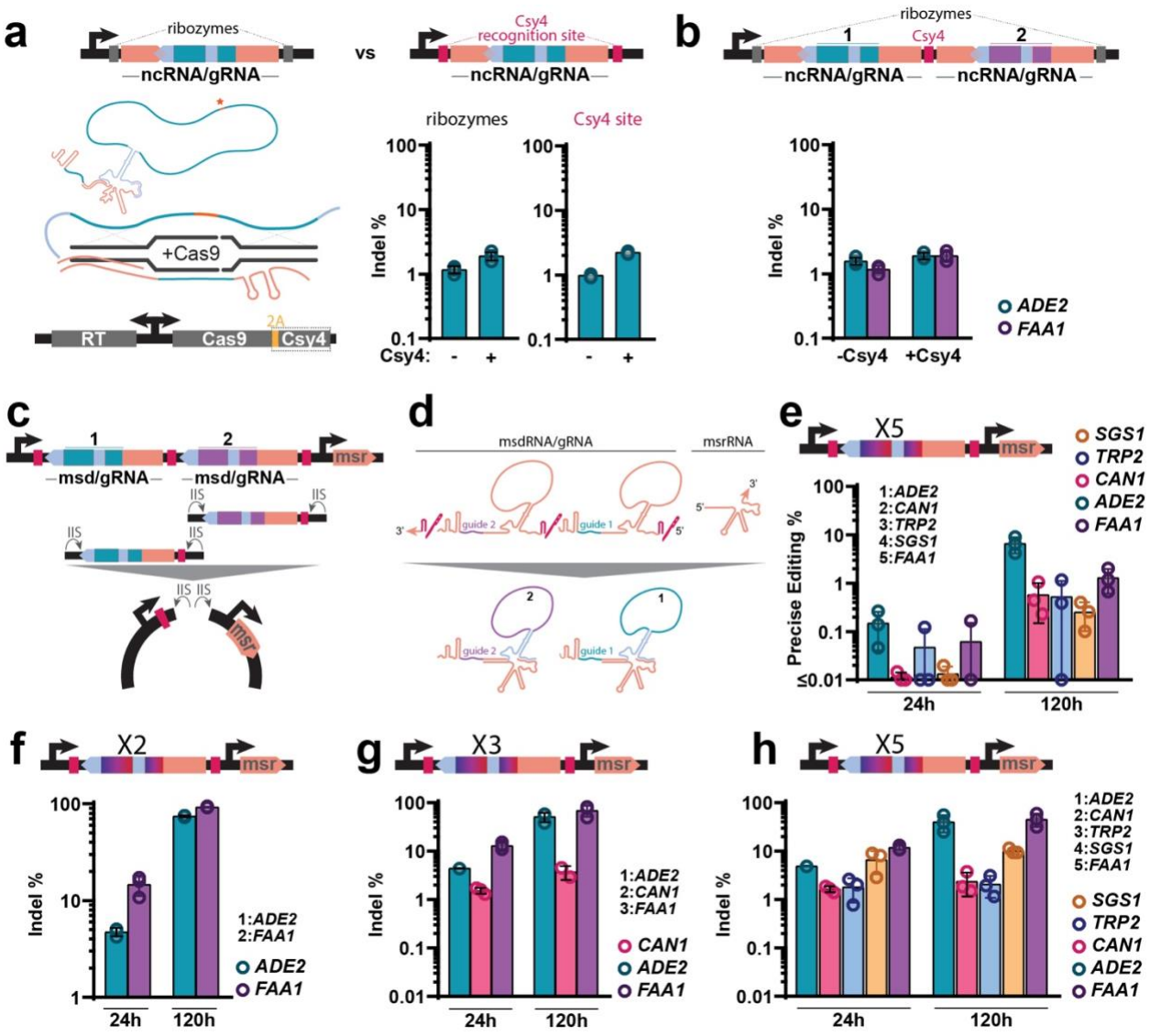
(Figure caption continued from the next page)

a. *Left: schematic of the different retron operon architectures tested. ncRNA with donor (orange and blue), genes required (grey) and optimised ribosome binding sites (RBS) regions (green) are indicated Right: quantification of rates for precise rpoB editing, circles show each of the three biological replicates, bars are mean ±SD.* ***b.*** *Quantification of precise editing rates for rpoB target site at 30 and 37ºC, circles show each of the three biological replicates, lines are mean ±SD.* ***c****. Quantification of* OD_{600} *using increasing concentrations of m-toluic acid after 16h of bacterial growing (n=1).* ***d****. Quantification of precise editing rates for rpoB using different concentrations of arabinose (n=1).* ***e.*** *Quantification of colonies with intact msd arrays. A total of 30 colonies coming from 3 different replicates were sequenced, bars are mean ±SD All precise editing rates were quantified using Illumina MiSeq after 24h of editing.* ***f****. Scheme of the protocol used to analyse genetic stability of the retron arrays. Briefly, recombineering plasmid was transformed into E. coli strain bMS.346, followed by 5 days of growing and diluting in the presence or absence of the arabinose. A dilution of the final culture was diluted and plated. Finally, the msd Array of 10 individual colonies per replicate (n=3) were amplified and sequenced to assess genetic stability of the multitron approach.*



Extended Data Figure 3-9: Local off-target mutations.

a*. Quantification of precise editing rates for fbaH and hda genes using a live or dead version of Eco1 RT, circles show each of the three biological replicates, bars are mean ±SD.* ***b****. Local off-target mutation frequency in the 70 bp region of the chromosome homologous to fbaH and hda editing donors using a live of dead version of Eco1 RT circles show each of the three biological replicates, bars are mean ±SD. All data was quantified using Illumina MiSeq after 24h of editing.*



Extended Data Figure 3-10: Intended and undesired on-target mutation rates caused by arrayed retron multiplexed editing in yeast cells.

a. *Top: Schematic of the donor encoding retron ncRNA/gRNA expression cassette expressed from a Gal7 Pol II promoter and flanked by ribozymes versus a new construction replacing ribozymes with Csy4 sequences. Bottom left: schematic of a retron ncRNA-Cas9 gRNA hybrid for genome editing in yeast, depicted above the protein-coding expression cassette which is inserted into the yeast genome. Bottom right: quantification of indel rates of the ADE2 locus in yeast by Illumina sequencing after 48h of editing. Circles show each of the three biological replicates, bars are mean ±SD; absence/presence of Csy4 in the protein-coding expression cassette is shown below the graph.* ***b.*** *Top: schematic of an arrayed retron ncRNA-Cas9 gRNA expression cassette, expressed from a Gal7 Pol II promoter, flanked by ribozymes, and separated by a Csy4 sequence. The retron editors in positions 1 and 2 target the ADE2 and FAA1 locus, respectively. Bottom: quantification of indel rates of the ADE2 and FAA1 loci in yeast by Illumina sequencing after 48h of editing. Circles show each of the three biological replicates, bars are mean ±SD; absence/presence of Csy4 in the protein-coding expression cassette is shown below the graph.* ***c.*** *Top: schematic of an arrayed retron msdRNA-Cas9 gRNA expression cassette, expressed from a Gal7 Pol II promoter, flanked and separated by a Csy4 sequence; the msrRNA is expressed in trans from a SNR52 Pol III promoter. Bottom: assembly schematic for one-pot Golden Gate cloning of multiple msdRNA-sgRNA*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

*editors. **d**. Schematic showing the presumed processing, annealing and reverse-transcription involved in the generation of editing donors from arrayed retron msdRNA-Cas9 gRNA cassettes. **e.** top: schematic of 5x arrayed retron msdRNA-Cas9 gRNA expression cassettes, as shown in **c**. Bottom: quantification of precise editing of the various yeast loci targeted by the retron editors shown above, by Illumina sequencing, after 24 and 120h of editing. The editors target ADE2, CAN1, TRP2, SGS1 and FAA1. Two-way ANOVA, effect of expression time, P=0.0038. Circles show each 3 biological replicates, bars are mean ±SD. **f-h**, top: schematic of 2x, 3x or 5x arrayed retron msdRNA-Cas9 gRNA expression cassettes. Bottom: quantification of indel rates of the various yeast loci targeted by the retron editors shown above, by Illumina sequencing, after 24 and 120h of editing. Individual open circles show each of three biological replicates per condition, bars are mean ±SD The editors target ADE2 and FAA1 (f); ADE2, CAN1 and FAA1 (g); and ADE2, CAN1, TRP2, SGS1 and FAA1 (h).*

3.7 Supplemental Files

Supplementary Information_Chapter3.pdf

This PDF file contains:

- Supplementary Table 3-1: Statistical analysis
- Supplementary Table 3-2: analysis of off-target mutations using Multitrons
- Supplementary Table 3-3: Plasmids used in this study
- Supplementary Table 3-4: Donors used in this study
- Supplementary Table 3-5: Strains used in this study
- Supplementary Table 3-6: Primers used in this study

Chapter 4 SspA is a transcriptional regulator of CRISPR adaptation in *E. coli*

4.1 Abstract

The CRISPR integrases Cas1-Cas2 create immunological memories of viral infection by storing phage-derived DNA in CRISPR arrays, a process known as CRISPR adaptation. A number of host factors have been shown to influence adaptation, but the full pathway from infection to a fully integrated, phage-derived sequences in the array remains incomplete. Here, we deploy a new CRISPRi-based screen to identify putative host factors that participate in CRISPR adaptation in the *E. coli* Type I-E system. Our screen uncovers a novel host factor, SspA, which transcriptionally regulates CRISPR adaptation. One target of SspA is H-NS, a known repressor of CRISPR interference proteins, but we find that the role of SspA on adaptation is not H-NS-dependent. We propose a new model of CRISPR-Cas defence that includes independent cellular control of adaptation and interference by SspA.

4.2 Introduction

CRISPR-Cas is an adaptive immune system found in archaea and bacteria, used to defend the host from foreign invaders, such as viruses or mobile genetic elements[101,113,114,122,521]. This defence is mediated by Cas (CRISPR associated) proteins, which are capable of creating immune memories of invading nucleic acids and using those memories to mount RNA-guided degradation of invaders in the event of a future encounter[105,522–524]. This process of storing immunological memory is known as CRISPR adaptation, and is mediated by a phylogenetically-conserved duo of proteins, Cas1 and Cas2, that form an integrase complex capable of inserting new DNA fragments (prespacers) into the cell's CRISPR array[98,107,175,525].

Studies spanning the past two decades have uncovered substantial mechanistic understanding of how CRISPR adaptation works and some of the key host factors that assist the CRISPR Cas1-Cas2 integrase complex in creating immune memories[108,149,152–160]. Double-stranded DNA fragments are the preferred substrate for the CRISPR Cas1-Cas2 integrases and can arise from a variety of sources, such as foreign DNA degradation by helicase-nuclease enzymatic complexes like the RecBCD complex[161] or AddAB[162] as well as from the replicating bacterial and phage genomes[165]. Fragments captured by the CRISPR Cas1-Cas2 integrases can then undergo trimming by Cas4[166], DnaQ[167] or other host exonucleases[168], generating free 3' OH groups required as substrates for spacer integration[149,167]. Cas1-Cas2 integrase docking at the Leader-Repeat junction of the CRISPR array requires the Integration Host Factor (IHF)[157,170–172], which generates a bend in the Leader sequence that accommodates the integrase complex and allows it to form stabilising contacts with the DNA[172]. Docking enables the CRISPR integrase

complex to catalyse a series of two nucleophilic attacks and add a new spacer at the Leader-Repeat junction[158,172,175].

Spacer integration creates staggered double strand breaks at either end of the duplicated Repeat. Recent *in vitro* evidence suggests that host polymerases, in coordination with genome replication or transcription, could aid in repairing the CRISPR array[158] (**Figure 4-1a**). The expanded and repaired CRISPR array is capable of supporting further rounds of spacer acquisition.

Despite this knowledge, several open questions remain, including what host factors are responsible for regulation of CRISPR-Cas activity[90,526–529], and repair of the CRISPR array post-spacer integration[158,169]. Furthermore, in contrast to noteworthy successes in heterologous reconstitution and harnessing of the CRISPR interference machinery across the tree of life, most notably CRISPR-Cas9, there are conspicuously few reports of successful heterologous expression of a CRISPR adaptation system outside of its native host[147,412], and no reports in eukaryotic systems. We, therefore, set out to discover additional host factors required for CRISPR adaptation.

Here, we develop and use a CRISPRi-based genetic screen to identify new host factors that participate in CRISPR adaptation in the Type I-E *E. coli* system. We report that a novel host factor, SspA, acts as a transcriptional-level regulator of CRISPR adaptation. We further find that SspA regulation of CRISPR adaptation does not function via H-NS, a known regulator of CRISPR interference and a member of the SspA regulon. Our data supports independent pathways for regulating the adaptation and interference components of CRISPR immunity, both downstream of SspA.

4.3 Results

CRISPRi screen identifies adaptation host factors

We designed a genome-wide CRISPRi screen to identify potential host factors that participate in Type I-E CRISPR adaptation (**Figure 4-1a**). This screen utilises a library of 92,919 gRNAs that are distributed across a population of *E. coli*, each of which direct a catalytically dead Cas9 (dCas9) to knock down transcription at a single locus, with multiple redundant gRNAs per gene[303,304]. We utilised this CRISPRi library in a negative selection scheme designed to deplete adaptation-competent cells. Specifically, we electroporated oligonucleotide prespacers that matched an essential gene into *E. coli* expressing a Type I-E CRISPR system. Integration of this prespacer into the CRISPR array would lead to the generation of a self-targeting crRNA and ultimately death of the adaptation-competent library members. CRISPRi knockdown of host factors involved in CRISPR adaptation would reduce adaptation and subsequent self-targeting, leading to enrichment of host factor gRNAs in the population following selection (**Figure 4-1b**).

For this screen, we used *E. coli* LC-E75, a derivative of K-12 MG1655, which encodes a Tetracycline-inducible dCas9 cassette integrated at the Phage 186 attB site[304]. We built an LC-E75 strain that carried a plasmid-encoded, IPTG-inducible Cas1-Cas2 cassette (plasmid hereon referred to as pSCL565). We then electroporated this strain and *E. coli* K-12 MG1655 (parental strain serving as a control) with a library of 92,919 plasmid-encoded sgRNAs, which target both coding and non-coding regions across the *E. coli* genome[304]. The libraries were grown overnight, and subsequently passaged and grown to mid-log phase (~3h) with dCas9 induction; the remainder of the overnight library cultures were harvested for sgRNA library sequencing (pre-screen library). Then, cells

were co-electroporated with (1) a plasmid encoding an m-Toluic acid-inducible *E. coli* Cas3-Cascade, the effector of the Type I-E CRISPR interference system, and (2) a 35bp dsDNA spacer targeting the essential gene *murA*. Cells were rescued in media containing inducers for Cas3-Cascade and dCas9 and antibiotics to select for their respective plasmids, cultured overnight, and harvested for sgRNA library sequencing (post-screen library). We extracted the sgRNA plasmid libraries and prepared samples for sequencing by amplifying the sgRNAs using a primer pool targeting the region upstream of the sgRNA promoter and downstream of the tracrRNA. The primers contained Illumina adapters to make the amplicons compatible with our downstream sequencing prep. Sequencing of the sgRNAs libraries yielded sgRNA counts for the dCas9-expressing LC-E75 and control dCas9-less parental strains, which allowed the calculation of the binned enrichment/depletion of sgRNAs across the *E. coli* genome (**Figure 4-1c**).

We found peaks of sgRNA enrichment that were distributed across the *E. coli* genome and did not cluster around the *murA* locus. Additionally, we identified *polA*, *priA* and *gyrA*, essential genes previously suggested to play a role in the CRISPR adaptation process. This highlights the advantage of a knock-down approach over transposon-based knock-out approaches, where essential genes would have been lost from the library altogether. We found several other regions of the *E. coli* genome where sgRNAs were strongly enriched, suggesting additional host factors.

We quantified differentially enriched or depleted sgRNAs from their cumulative sgRNA counts (sum of all sgRNAs per gene), by comparing each experimental sample (+dCas9) to its paired control (–dCas9) using PyDESeq2 package[303,304,530]. We filtered out genes with less than 10 cumulative reads, and controlled for variation in relative

sgRNA library composition by including pre-screen sgRNA counts as an interaction factor in the model. We found 571 differentially enriched/depleted genes and gene-adjacent regions, out of a total of 12,809 gene/gene-adjacent regions considered in our analysis (**Figure 4-1d**). Interestingly, a subset of the differentially enriched genes (i.e., CRISPR adaptation deficient when knocked-down) also had their gene-adjacent regions differentially enriched (shown with asterisked gene names).

We selected the top 8 gene regions with highest $\log_2$ fold changes for individual validation using knockout mutants from the Keio collection[531] in a naive adaptation assay. One additional gene, *gyrA*, is essential and could not be validated with a knockout. Although polA knockouts are non-viable, a *polA* Klenow fragment deletion mutant is viable[532], and was thus used in validation assays alongside the other non-essential genes.

We electroporated wild-type and knockout strains with pSCL565 and grew them in liquid culture for 48h without inducers for Cas1-Cas2 to achieve a moderate level of expression from transcriptional leak. We then sequenced the CRISPR II array of these cells (i.e., endogenous CRISPR array flanked by the *ygcE* and *ygcF* genes[533], hereon referred to as CRISPR-II) and quantified the rate of CRISPR adaptation as the fraction of sequenced arrays that had acquired new spacers. Biological replicates run on different days were normalised to the CRISPR adaptation rate of the wild-type parental Keio strain (**Figure 4-1e**). We found that 3 mutants showed significantly decreased rates of CRISPR adaptation compared to the wild-type strain: *pcnB*, *sspA*, and *polA Δ*Klenow.

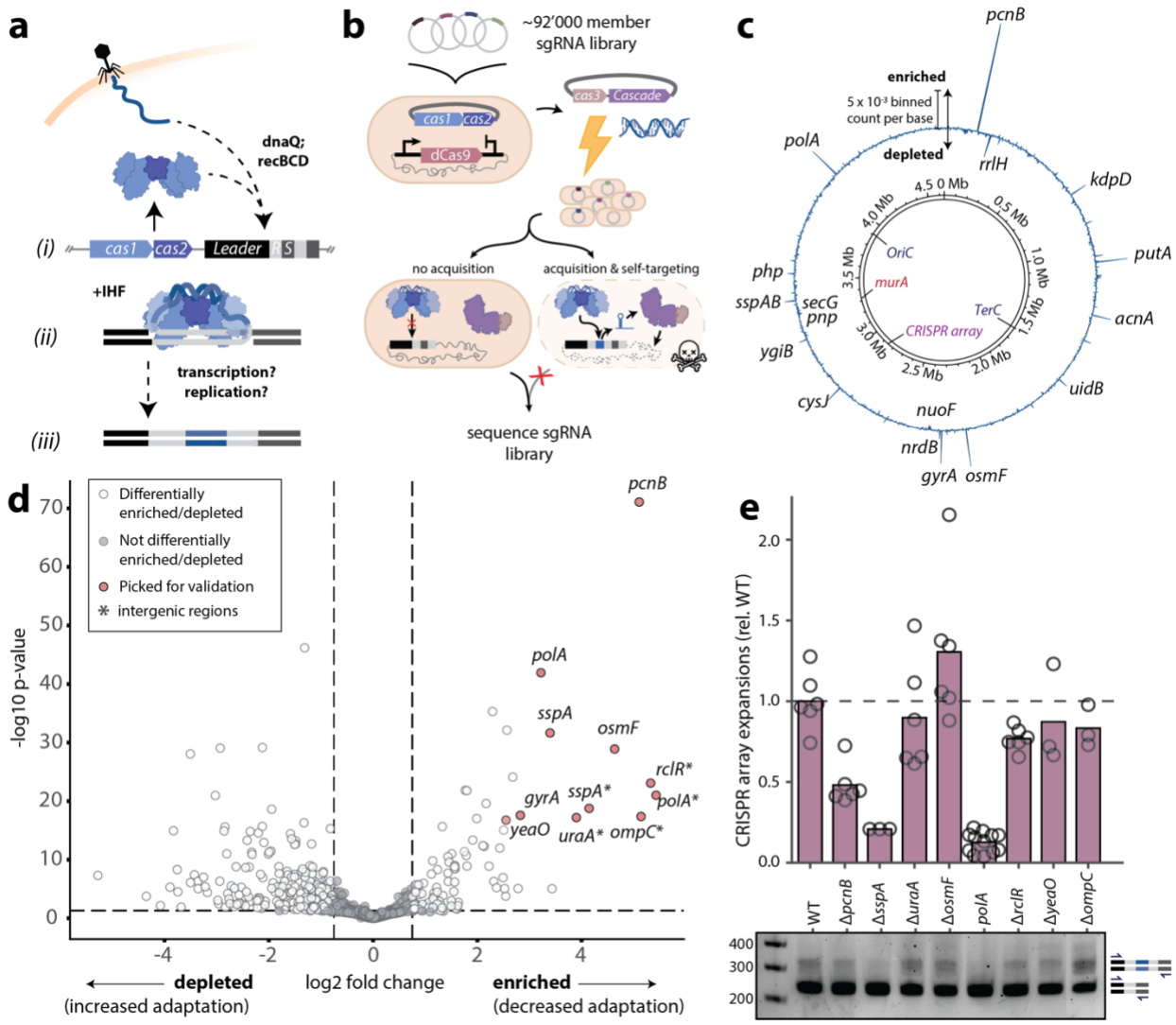


Figure 4-1: CRISPRi screen identifies adaptation host factors.

a. *Overview of the CRISPR adaptation process, highlighting key known host factors.* ***b****. Schematic of the CRISPRi adaptation host factor screen.* ***c****. Binned coverage plot of sgRNAs across the E. coli genome. sgRNA occupancy was calculated as the difference between the normalised (post/pre-screen) binned sgRNA counts per base of the experimental (+dCas9) and paired control (–dCas9) conditions. Regions of the genome with high ("enriched") sgRNA coverage are interpreted to be genomic loci that positively regulate CRISPR adaptation; regions of the genome with low (or negative, i.e., "depleted") sgRNA coverage are interpreted to be genomic loci that negatively regulate CRISPR adaptation. The highest-ranking regions with attributable genes are labelled; other labelled loci are the Ori and Ter regions, the murA gene, and the CRISPR-II array. n = 9 biological replicates.* ***d****. Volcano plot showing log_2 fold change for each sgRNA versus adjusted $–log_{10}$ p-values (n = 9 biological replicates). The horizontal dashed line represents an adjusted p-value of 0.05; the vertical lines represent log_2 fold changes of –0.75 and 0.75. Genes targeted by sgRNAs differentially enriched that were selected for individual validation are coloured in pink.* ***e****. Top: deep-sequencing based measurement of the rates of new spacer acquisition in Keio knockouts harbouring pSCL565, after growth for 48h in liquid culture without induction of Cas1-Cas2 expression. Acquisition rates are shown relative to the wild-type parental strain. Open circles represent biological replicates (n ≥ 3), bars are the mean (one-way ANOVA effect of strain P<0.0001; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ΔpcnB P=0.00217, ΔsspA P=0.000102, polA ΔKlenow P<0.0001; others ns). Bottom:*

(Figure caption continued on the next page)

(Figure caption continued from the next page)
representative agarose gel for the data shown. Expansions of the CRISPR array can be seen as higher sized bands above the parental array length. Additional statistical details in ***Supplementary Table 4-1****.*

Features of spacers acquired in Knockout Strains

Spacers captured by the CRISPR Cas1-Cas2 integrases come from a variety of sources. Defence associated sources include mobile genetic elements and phages. However, in the absence of interference machinery, spacers derived from the bacterial genome and plasmids accumulate (**Figure 4-2a**). We next tested whether any of the hits that we chose for validation modified the source of new spacers. We found that, consistent with previous findings[152,403], the majority of new spacers acquired in the wild-type strain were plasmid-derived (**Figure 4-2b**). This finding held for all mutants except *polA* ΔKlenow, which acquired spacers solely from the genome. The breakdown of spacer origin as a percent of all newly acquired spacers starkly illustrates this finding (**Figure 4-2c**).

We next sought to determine whether the differences in new spacer acquisition could be explained by a change in PAM preference or other motifs up- or downstream of the spacer. We searched 15bp up- and downstream of the newly acquired spacer in its source location, and found that all mutants showed similar PAM preferences to the wild-type strain, consistent with previous reports[109]. Similarly, all mutants except the *yeaO* deletion mutant showed no additional up- and downstream motif preferences of preference, beyond the AAG PAM.

The *yeaO* mutant displayed strong motif preferences up- and downstream of the genome-derived spacers (**Figure 4-2d)**, which prompted us to map all newly acquired

spacers for each mutant to their respective source on either the *E. coli* genome or pSCL565 plasmid (**Figures 4-2e-h;** see **Extended Data Figure 1** for an expanded view of these figures). We found that the distribution of new spacers from both sources were mostly consistent between the wild-type and the mutants tested (**Extended Data Figures 4-1a-i)** with two exceptions: the *yeaO* and *polA* ΔKlenow mutants.

We found that, as suggested by the prespacer neighbourhood motif analysis (**Figure 4-2d**), the *yeaO* mutant acquired spacers almost uniquely from one location in the genome, which maps to the gene *insG*, encoding an IS4 transposase (**Figure 4-2g**; **Extended Data Figure 4-1h**). It is possible that the high number of *insG*-derived spacers were a product of an early acquisition event from which most of the sequenced arrays descended, or if there had been multiple independent acquisition events leading to *insG*-derived spacers across CRISPR arrays. We can distinguish the two by looking at multiply expanded CRISPR arrays, or arrays that had acquired two or more new spacers, due to the CRISPR Cas1-Cas2 integrases' preference for Leader-proximal spacer insertion: this feature makes the CRISPR arrays temporally ordered, with spacers acquired at a later stage being closer to the Leader sequence than spacers acquired earlier, or than vestigial spacers[403,410–412]. We found *insG*-derived spacers in multiple different positions with respect to the Leader and found *insG*-derived spacers in arrays with distinct additional new Leader-distal spacers across three biological replicates (**Extended Data Figure 4-2**), suggesting that *insG*-derived spacers represent more than one acquisition event in parallel lineages.

Though there is no literature describing *yeaO*, it is predicted to contain a DUF488-like domain, which is ubiquitously distributed across prokaryotes and some viruses, and

has been found in genomic neighbourhoods linked to prophages and defence islands[534]. Structural searches using DALI[535] and Foldseek[536] suggested that YeaO is structurally similar to putative transcriptional regulators; in turn, *insG* is predicted to encode a transposase in a IS4 transposable element, previously been reported to be non-mobilised in *E. coli*[537]. We suggest that YeaO could be regulating the InsG-mediated mobilisation of IS4; the transposition events could cause the generation of DNA fragments that could serve as prespacers for the Cas1-Cas2 integrases, decoying these away from the electroporated *murA* prespacer and phenocopying a decrease in CRISPR adaptation, thus explaining the detection of YeaO in our screen.

The *polA* ΔKlenow mutant had a similar distribution of prespacers originating from the genome when compared to the wild-type, but we were unable to map any prespacers to the plasmid, suggesting that the plasmid was unable to serve as a source of prespacers in this mutant (**Figure 4-2h**; **Extended Data Figure 4-1f**). Given the loss of CRISPR adaptation from the pSCL565 plasmid but wild-type levels of CRISPR adaptation from the genome (**Figure 4-2b**), we hypothesised that the *polA* ΔKlenow mutant could be deficient in plasmid replication[538]. To test this, we measured the relative number of copies of the pSCL565 *Ori* and *cas1* sequences (the latter also found in the genome) in the wild-type and *polA* ΔKlenow mutant. We found a nearly 80-fold difference in the relative number of copies of pSCL565 that the *polA* ΔKlenow mutant contains compared to the wild-type strain, which would explain this strain's decreased ability to acquire new spacers from the plasmid (significantly decreased number of plasmid copies per cell) and also why it was identified as a hit in the initial screen (**Figure 4-2i**).

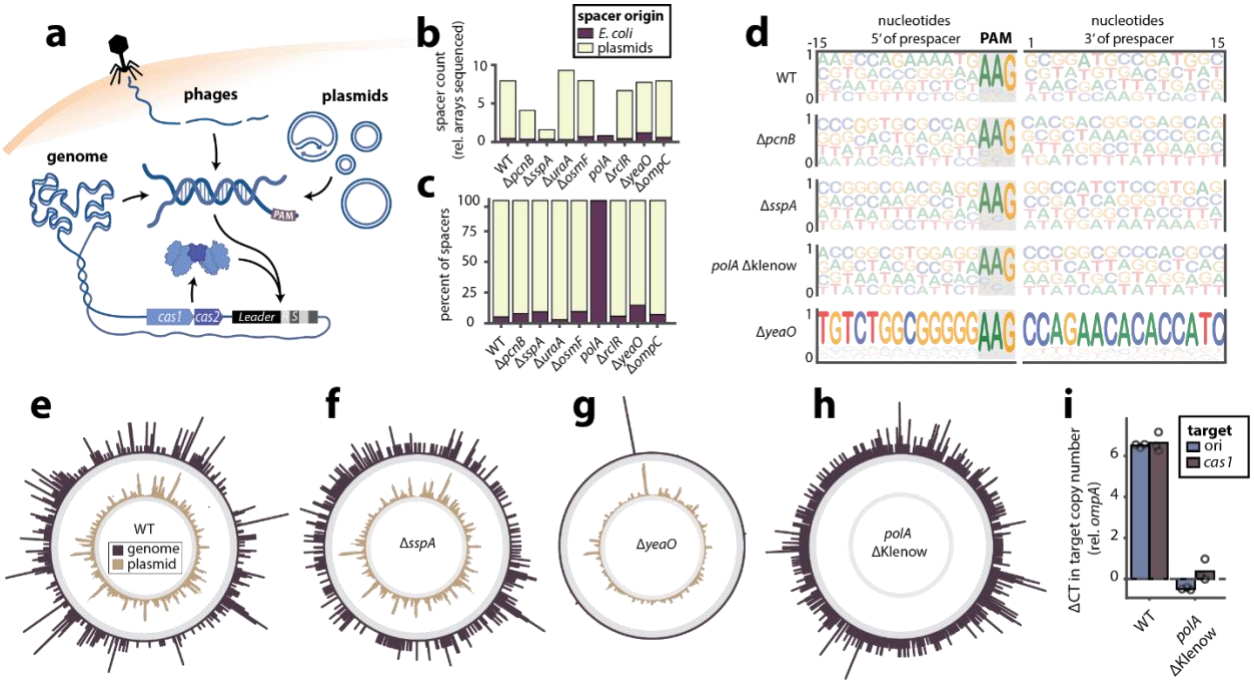


Figure 4-2: Features of spacers acquired in knockout strains.

a. *Prespacer substrates for CRISPR adaptation arise from a variety of sources.* ***b.*** *Breakdown of normalised spacer count (total number of new spacers / number of CRISPR arrays sequenced) according to spacer origin (E. coli or plasmid) and strain of interest.* ***c.*** *Breakdown of percent of spacer attributable to each spacer origin (E. coli or plasmid) and strain of interest.* ***d.*** *Motifs in the 15bp up- and downstream of the newly acquired spacer in its source location.* ***e-f:*** *Binned coverage plot of newly acquired spacer across the E. coli genome (outer, purple) and pSCL565 plasmid (inner, tan) for the wild-type strain (**e**) and derivatives (**f-h**). See* ***Extended Data figure 4-1*** *for the full set.* ***i.*** *qPCR-based measurement of the relative copy number of pSCL565 Ori and cas1 sequences in the wild-type and polA ΔKlenow mutant. Open circles represent biological replicates (n ≥ 3), bars are the mean (one-way ANOVA effect of strain and target P<0.0001; Sidak's corrected multiple comparisons for wild-type vs. ΔsspA, CDF ori copy number P<0.0001, cas1 copy number P<0.0001). Additional statistical details in* ***Supplementary Table 4-1****.*

SspA is a transcriptional regulator of CRISPR adaptation

Our CRISPR adaptation assays and downstream analysis of acquired spacers revealed that Δ*sspA* was consistently and significantly defective in naïve CRISPR adaptation, despite no other noticeable differences in the features of its acquired spacers when compared to the wild-type parental strain (**Figures 4-2b, c, e,** and **f**). Additionally, we found that the decrease in CRISPR adaptation in the Δ*sspA* background was not due

to decreased protein expression levels from pSCL565 (**Extended Data Figure 4-3**). We thus selected *sspA* for further mechanistic characterisation.

E. coli SspA was discovered four decades ago during a screen for proteins induced by the stringent response[539]. Over the years, its reported cellular functions have increased, and SspA has become particularly linked to global stress response[540,541] through its action as an RNA polymerase (RNAP)-associated protein[542,543]. Crystal structures of *E. coli* RNAP-promoter open complex with SspA have revealed that SspA inhibits σ^{70} promoter escape through contacts with both RNAP and σ^{70} through a conserved PHP motif[540,543–545]. This promoter escape inhibition induces a rewiring of the cellular transcriptomic landscape towards expression of σ^{S} genes, with implications on stress tolerance, motility and virulence[540,543–545]. The *sspA* gene is encoded in a two-member operon, upstream of *sspB*. SspB acts as a specificity-enhancing factor for the ClpXP protease[546]. It helps maintain protein homeostasis by escorting SsrA-tagged peptides, resulting from stalled ribosomes, to the ClpXP protease and promoting their degradation (**Figure 4-3a**), thus simultaneously freeing ribosomes and replenishing the pool of amino-acids that can become a precious resource in conditions of starvation.

Though we found *sspA* and not *sspB* as a significant hit in our screen, we sought to confirm that the defects in CRISPR adaptation observed in the Δ*sspA* mutant were due strictly to the lack of SspA, and not due to polar effects of this mutation on the downstream *sspB* gene. We compared the rates of CRISPR adaptation in wild-type strains to those in Δ*sspA::kan*^{R} and Δ*sspB::kan*^{R} mutants carrying pSCL565 (**Figure 4-3b**). We found that the Δ*sspA* mutant was deficient at new spacer acquisition, but the Δ*sspB* mutant acquired spacers at rates indistinguishable from the wild-type strain (**Figure 4-3c**). We attempted

to deliver an *sspA* rescue plasmid into the Δ*sspA*::*kan*^{R} strain, but this yielded no transformants over multiple attempts. However, we found that we could rescue the CRISPR adaptation phenotype to wild-type levels when Δ*sspA::kan*^{R} carrying pSCL565 were additionally electroporated with an *sspAB* cassette, encoding the SspA and SspB proteins under control of their native promoter and on a low-copy (~5) plasmid. This suggests that lack of *sspA* alone is sufficient to cause the loss of adaptation phenotype, and that this can be rescued by supplying a copy of the *sspA* gene in *trans*, under its native regulation.

Given these findings, we next sought to determine which part of the SspA protein was responsible for the loss of adaptation phenotype. We were particularly interested in the SspA PHP$^{84\text{-}86}$ motif, which has been reported to be indispensable for stabilisation of interactions between SspA, σ^{70} and the RNAP complex. Via this interaction, SspA acts as a transcriptional repressor of σ^{70} promoters by inhibiting promoter escape[543]. Triple-Alanine substitutions in this motif cause pleiotropic cellular effects such as increased swarming and defects in acid-resistance and phage P1 growth[541,547]. Thus, given SspA's role as a transcriptional rewiring agent, we decided to test whether an SspA PHP$^{84\text{-}86}$>AAA$^{84\text{-}86}$ mutant, deficient in σ^{70}-RNAP binding, would also phenocopy the Δ*sspA* mutant in terms of loss of CRISPR adaptation. To do this, we designed rescue plasmids, encoding variants of the *sspAB* operon under endogenous regulation, on low-copy (~5) plasmids (**Figure 4-3d**).

We found that rescue plasmids encoding the full *sspAB* operon or an early frameshifted *sspB* could rescue CRISPR adaptation to levels comparable to wild-type. However, rescue plasmids encoding SspB, an early frameshifted *sspA*, early frameshifted

sspA and *sspB*, and crucially, the SspA PHP$^{84\text{-}86}$>AAA$^{84\text{-}86}$-SspB RNAP binding mutant were all deficient in CRISPR adaptation (**Figure 4-3e**). Taken together, our data is consistent with a model in which SspA's role as RNAP-σ^{70} interactor and transcriptional rewiring agent is required for functional CRISPR adaptation.

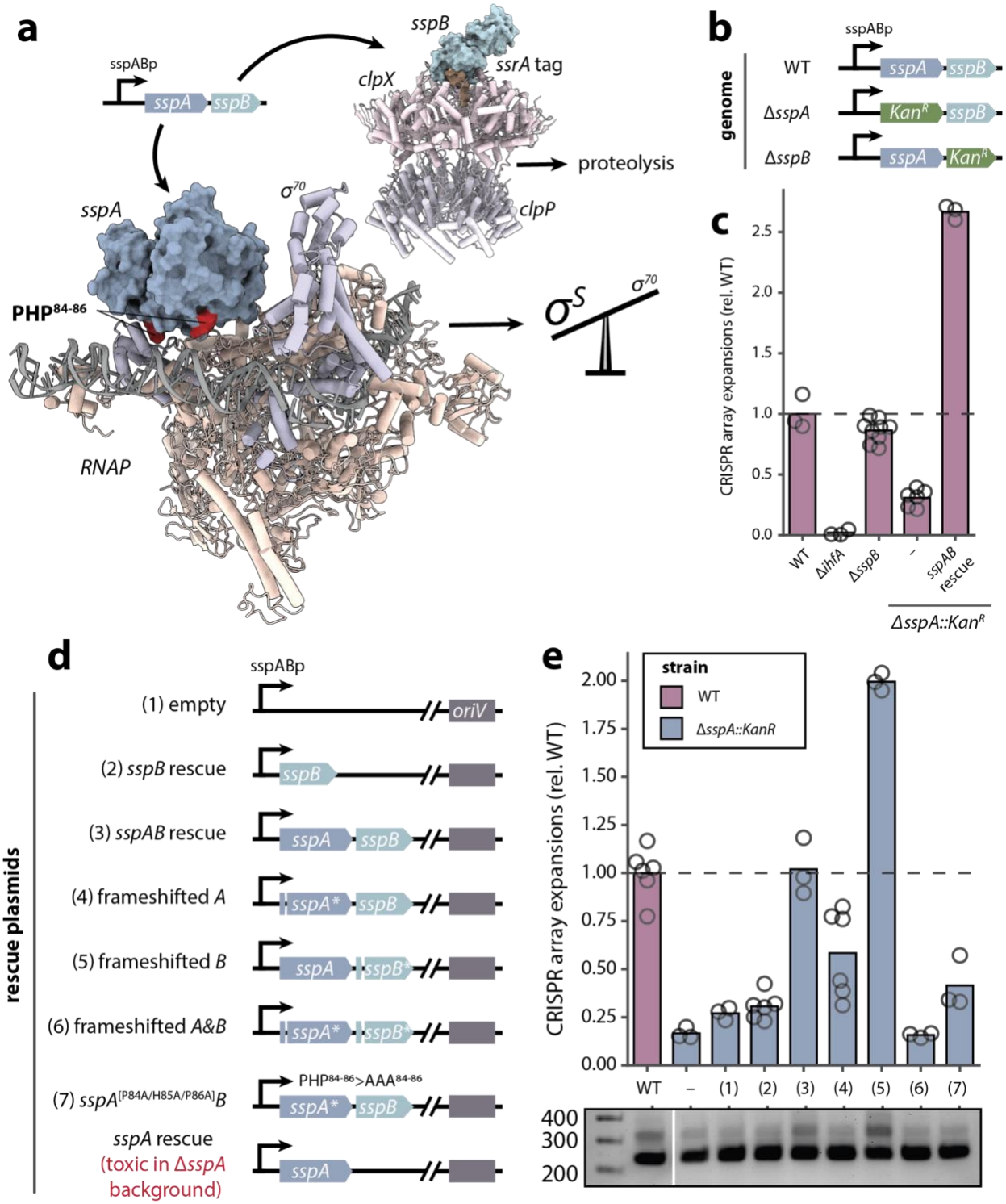


Figure 4-3: SspA is a transcriptional regulator of CRISPR adaptation.

a. *sspAB operon, proteins and function. Bottom left: crystal structure of an SspA dimer (blue) in complex with E. coli RNAP-promoter open complex, showing the conserved SspA PHP$^{84-86}$ residues (red) interacting with RNAP (pink) and σ^{70} (purple) (PDB 7DY6[543]). Top right: crystal structure of SspB escorting an SsrA-tagged substrate being delivered to the ClpXP protease complex (PDB 8ET3[546]).* ***b****. Schematic of the sspAB*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

operon of WT, ΔsspA::kan^R and ΔsspB::kan^R strains. kan^R: kanamycin resistance cassette. ***c****. Deep-sequencing based measurement of the rates of new spacer acquisition in strains harbouring pSCL565 and, in the case of the ΔsspA::kan^R, either an empty plasmid or a low (~5) copy plasmid encoding the sspAB operon, after growth for 48h in liquid culture. Adaptation rates are shown relative to the wild-type parental strain. Open circles represent biological replicates (n ≥ 3), bars are the mean. Horizontal dashed line represents the mean rate of spacer acquisition in the wild-type strain (one-way ANOVA effect of strain P<0.0001; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ΔsspA P<0.0001, ΔsspB P=0.109807; ΔsspA vs. ΔsspB P<0.0001).* ***d****. Schematic of the sspAB operon variant rescue plasmids. All plasmids are low (~5) copy, and encode variants of the sspAB operon under its native regulation. Frameshift mutants of SspA (AN$^{5-6}$>AQ$^{5-6}$ GCC|AAC>GC**T**|CAA|C) and SspB (PR$^{9-10}$>PS$^{9-10}$ CCA|CGT>CCA|**T**CG|T) encode sequences with single base insertions to cause protein translation to terminate early. The SspA PHP$^{84-86}$>AAA$^{84-86}$ mutant is RNAP-binding deficient and thus does not enable the shift in promoter use (σ^{70} → σ^{S})[543]. A single sspA rescue plasmid yielded no transformants into the ΔsspA::kan^R strain over multiple attempts.* ***e****. Top: deep-sequencing based measurement of the rates of new spacer acquisition in strains harbouring pSCL565 and, in the case of the ΔsspA::kan^R, either an empty plasmid or a low (~5) copy plasmid encoding variants of the sspAB operon as described in* ***d****., after growth for 48h in liquid culture. Adaptation rates are shown relative to the wild-type parental strain. Open circles represent biological replicates (n ≥ 3), bars are the mean. Horizontal dashed line represents the mean rate of spacer acquisition in the wild-type strain (one-way ANOVA effect of strain P<0.0001; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ΔsspA P<0.0001, ΔsspA + empty plasmid P<0.0001, ΔsspA + sspAB rescue P = 1, ΔsspA + sspA* (PHP84-86>AAA84-86) & sspB rescue P<0.0001; ΔsspA vs. rescues, ΔsspA + empty vector P=0.997758, ΔsspA + sspA* (PHP84-86>AAA84-86) & sspB P=0.334315, ΔsspA + sspAB P<0.0001, ΔsspA + sspB P=0.892991, ΔsspA + sspA* & sspB* (frameshifted) P=1). Bottom: representative agarose gel for the data shown. Expansions of the CRISPR array can be seen as higher sized bands above the parental array length. Additional statistical details in* ***Supplementary Table 4-1.***

H-NS regulates CRISPR interference downstream of SspA

Having established SspA's role as a regulator of CRISPR adaptation, we sought to test whether it could also play a role in regulating CRISPR interference. As SspA has been shown to be rapidly and highly upregulated following lambda infection[548], it could serve as a link between phage infection and CRISPR defence generally. Given previous reports of the role of SspA in downregulating levels of H-NS[541,549], a repressor of the CRISPR interference machinery, we hypothesised that SspA could be acting on the CRISPR-Cas system via H-NS[550–552] (**Figure 4-4a**).

To assess the effects of SspA on CRISPR mediated anti-phage defence and the potential interactions between SspA and H-NS in regulating this defence, we constructed Δ*sspA::FRT*, Δ*hns::FRT* and Δ*sspA::FRT* Δ*hns::FRT E. coli* strains (**Figure 4-4b**).

Previous studies have shown that H-NS is a strong repressor of CRISPR-Cas gene expression, but that this repression can be relieved by knocking out H-NS[550,551]. This de-repression can result in defence against bacteriophages, provided that these cells' CRISPR arrays encode one or more spacers targeting the phage genome (hereinafter referred to as "pre-immunised *E. coli*")[101,550,551].

We electroporated our mutant strains with a plasmid carrying a CRISPR array encoding a first spacer complementary to the lambda genome (T: target[101,551]) or a control CRISPR array with a non-target first spacer (NT: non-target). Then, we infected these pre-immunised strains with varying titres of λ_{vir} and quantified phage defence (**Figure 4-4c**). Because of the pre-immunisation, this assay measures the ability of mutants to mount anti-phage defence via CRISPR interference and should be CRISPR adaptation-independent.

Plaque assays revealed that a wild-type strain was unable to mount defence against new rounds of infection even when pre-immunised with an anti-λ spacer, as reported previously[101,550,551] (**Figure 4-4d**). Pre-immunised Δ*sspA* mutants were similarly unable to defend against λ_{vir}. However, pre-immunised Δ*hns* mutants were capable of mounting considerable defence against λ_{vir}. Interestingly, we saw no differences in anti-λ_{vir} defence between the Δ*hns* and Δ*hns* Δ*sspA* pre-immunised mutants, suggesting that the CRISPR-Cas mediated anti-phage defence observed in the Δ*hns* Δ*sspA* mutants was determined solely by the lack of CRISPR interference repression by H-NS, and that Δ*sspA* has no additive effect on CRISPR interference-mediated anti-phage defence on the Δ*hns* background. Quantification of efficiency of plating confirmed these findings (**Figure 4-4e**), as did additional experiments measuring anti-phage defence in overnight liquid culture

growth assays (**Figure 4-4f**). Together, these results suggest that H-NS and SspA are epistatic for CRISPR interference, with H-NS acting downstream of SspA on the regulation of CRISPR interference-mediated anti-phage defence.

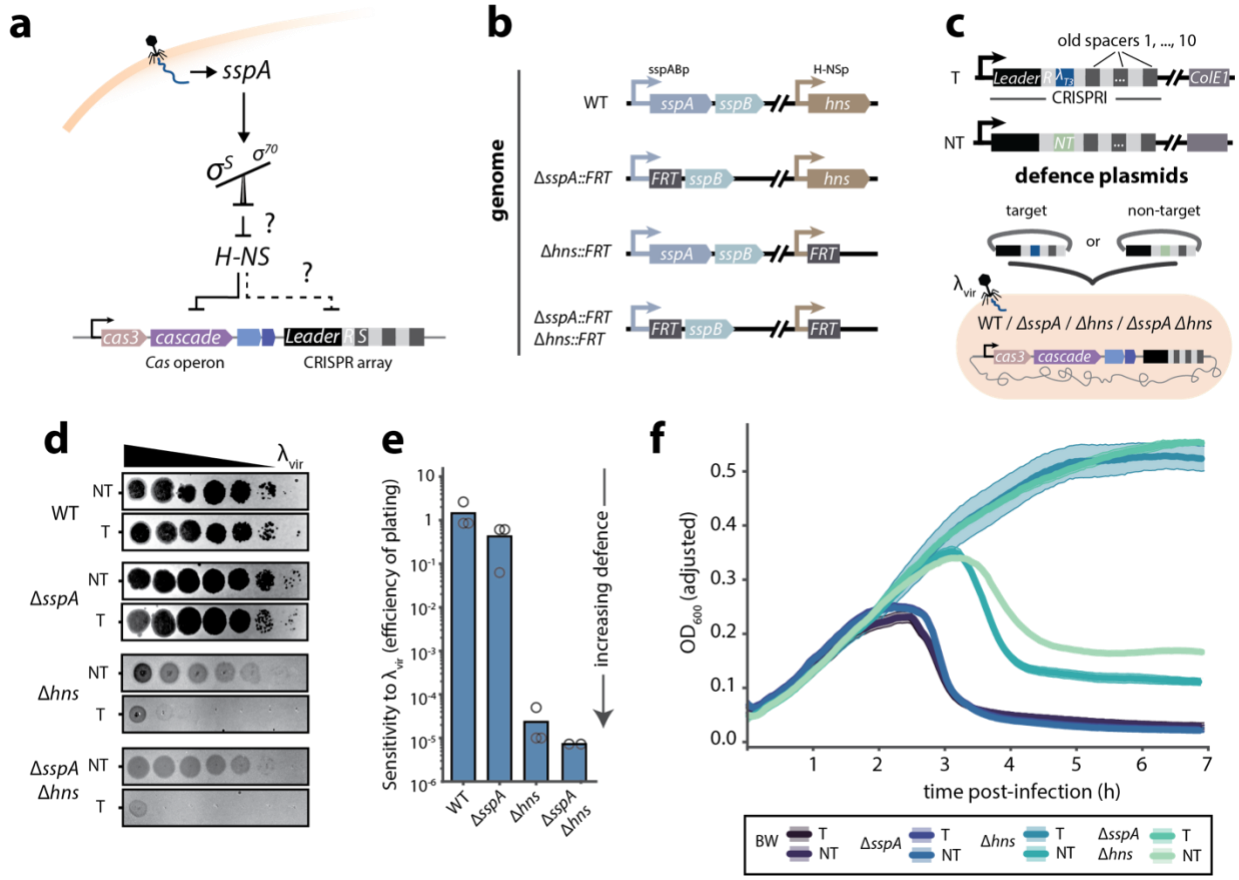


Figure 4-4: H-NS regulates CRISPR interference downstream of SspA.

***a**. Model for SspA-mediated regulation of CRISPR-Cas defence. Phage infection triggers upregulation of SspA[548], which in turn induces a global transcriptional shift towards σ^S-regulated promoters. This results in H-NS downregulation[541,549], induction of CRISPR-Cas mediated defence through de-repression Cas gene expression[550,551], leading to increased rates of CRISPR adaptation and interference. **b**. Schematic of the sspAB and hns operons of WT, ΔsspA::FRT, Δhns::FRT and ΔsspA::FRT Δhns::FRT strains. FRT: flippase recognition target, a scar left after the removal of resistance cassettes. **c**. Schematic of the CRISPR interference-mediated defence assays in pre-immunised E. coli strains. Top: schematic of the CRISPR-I immunisation (defence) plasmids. All plasmids are low (~5) copy, and encode an E. coli CRISPR-I array with a first spacer encoding either a Target (complementary to the λ genome[101,551]), or a Non-Target (NT) spacer. Bottom: The experimental strains were electroporated with either the T or NT plasmid, and infected to varying titres of λ_{vir}. Note that the strains encode a complete endogenous E. coli Type I-E CRISPR-Cas system. **d**. Representative plaque assays of λ_{vir} on experimental strains (described above) pre-immunised with either T or NT defence plasmids. Strains were infected with λ_{vir} and grown on plates at 30ºC for 16h. Full plaque assay plates for n = 3 biological replicates in **Extended Data Figure 4-4**. **e**. Efficiency of plating of λ_{vir} on experimental strains. Open circles represent biological replicates (n ≥ 3) of individual plaque*

(Figure caption continued on the next page)

(Figure caption continued from the next page)
assays, bars are the mean (one-way ANOVA effect of strain P=0.033454; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ΔsspA P=0.181757, Δhns P=0.043319, ΔsspA Δhns P = 0.043316; for Δhns vs. ΔsspA Δhns P=1). ***f.*** *Anti-phage defence and growth in overnight liquid culture of experimental strains, post λ_{vir} infection (MOI: 0.1). Hue around solid line (mean) represents the standard deviation across 3 biological replicates.*

SspA regulates CRISPR adaptation independently of H-NS

Given that SspA may regulate CRISPR interference via H-NS, we sought to determine whether SspA, in turn, regulates CRISPR adaptation via H-NS as well. To do so, we performed deep sequencing of the CRISPR arrays from samples harvested 3h post λ_{vir} infection in liquid cultures of wild-type, Δ*hns*, Δ*sspA*, and Δ*hns* Δ*sspA* mutants harbouring either T or NT plasmids. We found no differences in the rates of CRISPR adaptation across conditions, except in the Δ*hns* + T cultures, which substantially increased rates of CRISPR adaptation (**Figure 4-5a**). These new spacers were primarily λ_{vir} derived (**Figure 4-5b**), and that the majority of the acquired spacers are found immediately downstream and on same strand as the immunising spacer, consistent with primed CRISPR adaptation (**Figures 4-5c-d, Extended Data Figure 4-5a**). Interestingly, we saw a substantial decrease in λ_{vir} derived spacers in the Δ*hns* Δ*sspA* + T conditions (**Extended Data Figure 4-5b**). Although the rates of CRISPR adaptation in the Δ*hns* + T condition were low (0.5% of CRISPR arrays expanded, i.e., 5 cells per thousand with a newly expanded array) and could not explain the defence demonstrated by the Δ*hns* + T cultures at the time of sample collection (**Figure 4-4f**), our results underscore the requirement for SspA for adequate primed CRISPR acquisition, in a closer-to-natural and defence-relevant setting.

We next sought to determine whether SspA modulates naïve CRISPR adaptation via H-NS. For this, we used the Δ*sspA::FRT*, Δ*hns::FRT* and Δ*sspA::FRT* Δ*hns::FRT E. coli* strains (**Figure 4-5e**), and assessed the mutants' ability to acquire new spacers after co-electroporation of pSCL565 alongside a low (~5) copy rescue plasmid encoding the *sspAB* operon, *hns* operon, or both, under their native genomic contexts and regulation (**Figure 4-5f**). We found that Δ*sspA*, Δ*hns*, and Δ*sspA* Δ*hns* mutant strains all showed defects in CRISPR adaptation, with the double Δ*sspA* Δ*hns* mutant showing the strongest defect (**Figure 4-5g**). Complementation of the knockout strains with their respective rescue plasmids restored CRISPR adaptation to levels comparable to wild-type.

Since H-NS deletion de-represses CRISPR interference (**Figures 4-4d-f**), we hypothesised that its effect on CRISPR adaptation could be indirect, through the removal of cells that acquired genome-derived spacers via CRISPR interference-mediated self-targeting. To remove the confounding effect of increased self-targeting in the Δ*hns* background, we built Δ*cas3-cascade::cm*^{R} knockouts on top of the Δ*sspA* and Δ*hns* genetic backgrounds, and assessed the mutants' ability to acquire new spacers after electroporation with pSCL565. We found that although the Δ*sspA* Δ*cas3-cascade::cm*^{R} mutant still remained substantially CRISPR adaptation deficient, the Δ*hns* Δ*cas3-cascade::cm*^{R} mutant recovered CRISPR adaptation to levels comparable to wild-type (**Figure 4-5h**). This confirmed that the apparent CRISPR adaptation deficiency of the Δ*hns* mutant was caused by self-targeting through de-repression of CRISPR interference, and not additional effects on CRISPR adaptation. Taken together, our data supports a role for SspA in CRISPR adaptation that is independent of H-NS.

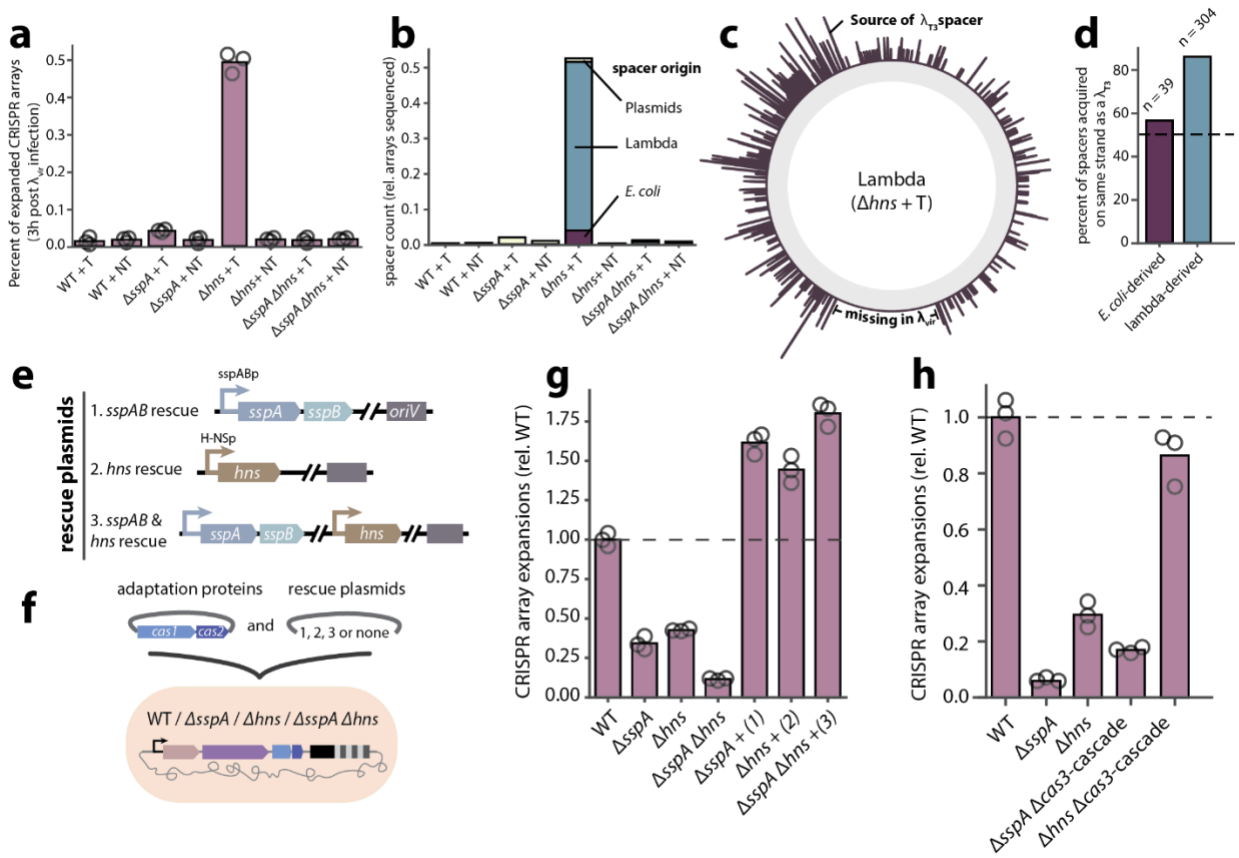


Figure 4-5: SspA regulates CRISPR adaptation independently of H-NS.

***a**. Deep-sequencing based measurement of the rates of new spacer acquisition in strains pre-immunised with either a T or NT defence plasmid, harvested 3h post λ_{vir} infection in liquid culture and growth at 30ºC. Open circles represent biological replicates (n ≥ 3), bars are the mean (one-way ANOVA effect of strain P< <0.0001; Sidak's corrected multiple comparisons for wild-type +T vs. knockouts +T, ΔsspA P=082553, Δhns P<0.0001, ΔsspA Δhns P=0.999999; ΔsspA +T vs. knockouts +T, Δhns P<0.0001, ΔsspA Δhns P=0.154762; Δhns +T vs. Δhns +NT P<0.0001; Δhns +T vs. ΔsspA Δhns +T P<0.0001). **b.** Breakdown of normalised spacer count (total number of new spacers / number of CRISPR arrays sequenced) according to spacer origin (E. coli, lambda or plasmid) and strain of interest. **c.** Binned coverage plot of Δhns + T newly acquired spacers across the lambda genome (outer, purple). The location of the T immunisation spacer is shown on the lambda genome; "missing in λ_{vir}" indicates a genomic region missing in our strain of λ_{vir}. **d**. Percent of spacers acquired that are on the same strand as the T immunisation spacer, according to the spacer source (E. coli or lambda). **e**. Schematic of the sspAB and hns operonic rescue plasmids. All plasmids are low (~5) copy, and encode either 1. The sspAB operon, 2. The hns operon, or 3. both, under their native regulation. **f**. Schematic of the CRISPR adaptation assays in wild-type, sspA and/or hns mutant strains. Strains were electroporated with pSCL565 and rescue plasmids 1., 2., or 3. (see **e**.), and assessed for their ability to acquire new spacers into the endogenous CRISPR I array. **g**. PCR-based detection of new spacer acquisition into the CRISPR I array of wild-type, of WT, ΔsspA::FRT, Δhns::FRT and ΔsspA::FRT Δhns::FRT strains harbouring pSCL565 and rescue plasmids 1., 2., or 3. (see **e**.), after growth for 48h in liquid culture. Open circles represent biological replicates (n ≥ 3), bars are the mean. Horizontal dashed line represents the mean rate of spacer acquisition in the wild-type strain (one-way ANOVA effect of strain P< <0.0001; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ΔsspA P<0.0001, Δhns P<0.0001, ΔsspA Δhns P<0.0001; ΔsspA vs. knockouts, Δhns P=0.714182, ΔsspA Δhns P=0.002269, ΔsspA + sspAB rescue P<0.0001; Δhns vs. knockouts, ΔsspA Δhns P<0.0001, Δhns + hns rescue P<0.0001; ΔsspA Δhns vs. ΔsspA Δhns + sspA & hns rescues P<0.0001). **h**. PCR-based detection of new*

(Figure caption continued on the next page)

(Figure caption continued from the next page)

spacer acquisition into the CRISPR I array of WT, ∆sspA::FRT, ∆hns::FRT, ∆sspA::FRT ∆cas3-Cascade::Cm^R or ∆hns::FRT ∆cas3-Cascade::Cm^R strains harbouring pSCL565 after growth for 48h in liquid culture. Open circles represent biological replicates (n ≥ 3), bars are the mean (one-way ANOVA effect of strain P<0.0001; Sidak's corrected multiple comparisons for wild-type vs. knockouts, ∆sspA P<0.0001, ∆hns P<0.0001, ∆sspA ∆cas3-cascade P<0.0001, ∆hns ∆cas3-cascade P=0.125466; ∆sspA vs. ∆hns P=0.004161; ∆sspA vs. ∆sspA ∆cas3-cascade P=0.310715; ∆hns vs. ∆hns ∆cas3-cascade P<0.0001; ∆sspA ∆cas3-cascade vs. ∆hns ∆cas3-cascade P<0.0001). Horizontal dashed line represents the mean rate of spacer acquisition in the wild-type strain. Additional statistical details in ***Supplementary Table 4-1****.*

4.4 Discussion

We developed a novel negative selection CRISPRi screen, designed around the concept of stimulated CRISPR self-immunity, to identify potential host factors that participate in CRISPR adaptation in *E. coli*. We identified a new host factor in our screen, SspA. In validation experiments, adaptation assays and downstream analysis of newly acquired spacers revealed that a *sspA* knockout mutant is consistently and significantly defective in naïve CRISPR adaptation, despite no other noticeable differences in the features of its acquired spacers when compared to the wild-type parental strain. Further, we found that mutations that abolish SspA's ability to bind to the RNA Polymerase complex cause a loss-of-adaptation phenotype, suggesting that SspA acts as a transcriptional-level regulator of CRISPR adaptation. A series of phage sensitivity and CRISPR adaptation assays revealed that SspA regulates CRISPR adaptation independently of H-NS, a known regulator of CRISPR interference-mediated anti-phage defence and a member of the SspA regulon. Taken together, our data support independent control of CRISPR adaptation and interference downstream of SspA.

We find that our data is consistent with a model where the immunisation and interference steps could occur separately, perhaps even temporally so. We speculate that phage infection could trigger the rapid accumulation of SspA[548], opening a window for the acquisition of new spacers; this window may close rapidly as the levels of SspA decline, but this sudden SspA accumulation may be enough to cause downregulation of H-NS[541], thus opening a second window for CRISPR interference to occur (**Figure 4-6**). However, more studies are required to determine whether the sudden accumulation of SspA in response to phage infection is a ubiquitous response beyond lambda, what phage

element or phage-induced signal triggers this sudden spike in SspA levels, and whether this spike is indeed sufficient to significantly deplete levels of H-NS and open a window for CRISPR interference to occur. Further, though our data strongly suggests that SspA acts on CRISPR adaptation at a transcriptional level, additional work is needed to discover the target(s) of the SspA-mediated transcriptional rewiring.

Though our screen revealed SspA as a novel regulator of CRISPR adaptation, we did not identify host factors involved in the repair of the CRISPR array. Although *polA* was a promising hit, with *in vitro* evidence that its Klenow fragment is capable of repairing CRISPR arrays that have been cleared of the Cas1-Cas2 integrases[158], our results do not support this role. Though we found that CRISPR adaptation levels were significantly diminished in the *polA* ΔKlenow mutant, this decrease was attributable to the loss of acquisition of plasmid-derived spacers; The loss-of-adaptation phenotype seen in the Δ*pcnB* mutant is likely due to a similar effect, as *pcnB* has been shown to be required for copy number maintenance of ColE1 and other plasmids[553]. We cannot, however, rule out a role for *polA* in array repair, though it is conceivable that there is redundancy in host factors capable of this task. Indeed, functional redundancy of host factors is a possible explanation for not capturing the comprehensive set of these proteins. We anticipate that more complex combinatorial knockdown and activation screens could be used to tackle this problem. Furthermore, we believe that pairing genetic screens such as our CRISPRi screen with orthogonal physical screens, such as proximity labelling and pull-down assays[554], will yield rich and informative datasets, which are likely to uncover a more comprehensive set of host factors required for CRISPR adaptation.

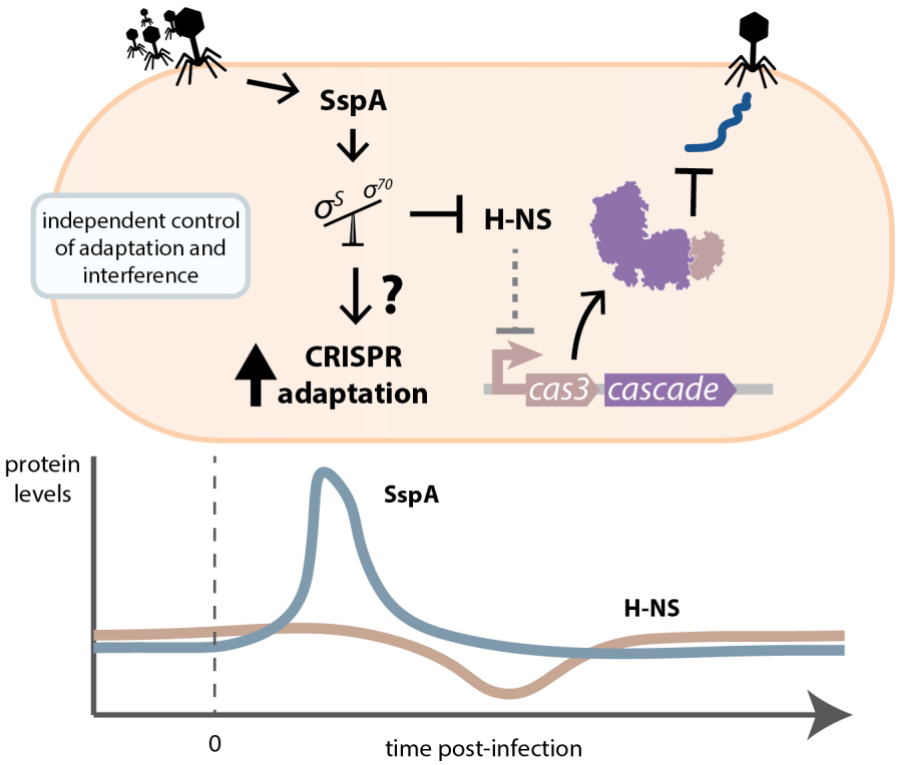


Figure 4-6: Proposed model for the independent control of CRISPR adaptation and interference.
In both cases, the regulation of CRISPR immunity happens downstream of SspA, through its role as a global transcriptional rewiring agent.

4.5 Methods

Biological replicates were taken from distinct samples, not the same sample measured repeatedly. Full statistics can be found in **Supplementary Table 4-1.**

Bacterial strains and culturing

All strains used in this study can be found in **Supplementary Table 4-2**. Wild-type *E. coli* K-12 W3110 (BW25113) strain, generously provided by Joseph Bondy-Denomy, was used for all experiments in this study, unless specified. *E. coli* K-12 MG1655 and LC-E75[304] (derivative of MG1655, Addgene #115925) were used for the CRISPRi screen. *E. coli* NEB-5-alpha (NEB C2987) was used for plasmid cloning. Keio collection[531] single-gene knock-out (KO) mutants, derivatives of BW25113, were generously provided by Carol Gross.

Additional deletions on Keio single-gene KO backgrounds were generated by λ_{Red} recombinase-mediated insertion of an FRT-flanked chloramphenicol (Cm^R) resistance cassette[489]. This cassette was amplified from pKD3[489] (Addgene #45604) with homology arms (50bp each) corresponding to the genomic sequences immediately up- and downstream of the intended deletion site. This amplicon was electroporated into the Keio strains expressing the λ_{Red} recombinase from pKD46[489]. Clones were isolated by selection on LB + chloramphenicol (10 μg/mL) plates. After PCR genotyping and sequencing to confirm locus-specific insertion, the chloramphenicol and pre-existing kanamycin cassettes was excised by transient expression of FLP recombinase from pE-FLP[555] (Addgene #45978) to leave a single FRT scar, whenever specified in the text (i.e., Δ*gene*::*FRT*).

The *polA* ΔKlenow mutant was generated by λ_{Red} recombinase-mediated insertion of an FRT-flanked Cm^R resistance cassette into the Klenow fragment of *E. coli* BW25113 Polymerase I. This cassette was amplified from pKD3 with homology arms (50bp each), corresponding to the genomic regions flanking the Klenow fragment, as reported previously[532]. This amplicon was electroporated into BW25113 expressing the λ_{Red} recombinase from pKD46. Clones were isolated by selection on LB + chloramphenicol (10 μg/mL) plates. PCR genotyping and sequencing confirmed the locus-specific insertion.

For the CRISPRi screen and CRISPR-Cas adaptation experiments, LB containing 1.5% w:v agar was used to grow strains on plates (growth at 37C until single colonies became visible, usually ~16h). Strains were subsequently grown in LB broth at 37C with 250 r.p.m. shaking, with appropriate inducers and antibiotics as described below.

For CRISPR-Cas defence experiments, strains were grown in LB broth supplemented with 10 mM $MgSO_4$ and 0.2% maltose at 30C with 250 r.p.m. shaking, with appropriate inducers and antibiotics as described below. For plaque assays, cells were mixed with top agar (0.5% w:v LB agar, supplemented with 10 mM $MgSO_4$ and 0.2% maltose and the appropriate antibiotics) poured over LB plates supplemented with the appropriate antibiotics, and grown at 30C overnight.

Inducers and antibiotics were used at the following working concentrations: 2 mg/mL L-Arabinose (GoldBio A-300), 1 mM IPTG (GoldBio I2481C), 1mM m-Toluic acid, 1 ug/mL anhydrotetracycline, 35 μg/mL kanamycin (GoldBio K-120), 25 μg/mL spectinomycin (GoldBio S-140), 100 μg/mL carbenicillin (GoldBio C-103), 25 μg/mL chloramphenicol (GoldBio C-105).

Phage strains and culturing

A virulent variant of phage Lambda (λ_{vir})[556], generously provided by Luciano Marraffini, was used throughout this study. λ_{vir} was propagated on BW25113 grown in LB at 30C, based on previous studies[557]. Briefly, overnights of *E. coli* BW25113 were grown at 30C in 5mL LB + 10 mM $MgSO_4$ and 0.2% maltose. The next day, 300uL of bacterial culture was infected with 10uL of serial dilutions of λ_{vir} in LB + 10 mM $MgSO_4$ and 0.2% maltose, incubated at 30C for 15min, and added to 5mL top agar, mixed gently and poured over LB agar plates. Plates were grown overnight at 30C. Plates from the dilution series that showed evidence of confluent lysing of *E. coli* were covered in 5mL LB supplemented with 10 mM $MgSO_4$ and 0.2% maltose, placed on a shaker to agitate gently at room temperature for 2h. Then, the lysate was transferred to a 15mL conical tube, centrifuged at 4500g x 15min to remove the bacterial debris, and filtered through a 0.2um filter. Phage titres were determined by preparing 1:10 dilutions of λ_{vir} in LB supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and spotting 2.5uL of the dilutions over top agar lawns of BW25113, which had been previously prepared by mixing 100uL of the overnight culture with 5mL of top agar (0.5% w:v LB agar, supplemented with 10 mM $MgSO_4$ and 0.2% maltose) and poured over LB agar plates. Serial dilutions of λ_{vir} were prepared in LB supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and 2.5uL of each dilution was spotted on the top agar using a multichannel pipette. Plates were tilted to allow phage spots to drip down the plate for easier quantification, and left to dry completely at room temperature. Plates were incubated at 30C overnight.

Plasmids

Plasmid information can be found in **Supplementary Table 4-3**. Plasmid pSCL565, encoding an IPTG-inducible *E. coli* Cas1-Cas2 cassette, spectinomycin resistance cassette and a pCDF ori, was constructed by PCR amplification of pCas1+2[152] (Addgene #72676) to replace the T7 promoter by an IPTG-inducible Lac promoter.

Plasmid pSCL563, encoding an m-Tol-inducible *E. coli* Cas3-Cascade operon, carbenicillin resistance cassette and a pRSF ori was constructed by Gibson cloning.

The *sspAB* rescue set of plasmids, designed to rescue the loss of CRISPR adaptation phenotype of the Δ*sspA* mutant, were constructed by first Gibson cloning the *sspAB* operon (including 236bp upstream of *sspA,* containing the predicted promoter[558] between *rpsI* and *sspA*) into a low copy plasmid backbone (pSC101 ori) containing a carbenicillin resistance cassette. This yielded pSCL735 (*sspAB* rescue). Variants of the *sspAB* operon were generated by targeted PCRs to yield pSCL747 (*sspA* rescue, not tested because toxic in Δ*sspA* background); pSCL748 (*sspB* rescue); pSCL751 (*sspA* frameshifted AN^{5-6}>AQ^{5-6} GCC|AAC>GC**T**|*CAA*|C + *sspB*); pSCL752 (*sspA* + *sspB* frameshifted PR^{9-10}>PS^{9-10} CCA|CGT>CCA|***T****CG*|T); pSCL753 (*sspA* frameshifted AN^{5-6}>AQ^{5-6} GCC|AAC>GC**T**|*CAA*|C + *sspB* frameshifted PR^{9-10}>PS^{9-10} CCA|CGT>CCA|***T****CG*|T); and pSCL770 (*sspA* PHP^{84-86}>AAA^{84-86} + *sspB*).

The CRISPR defence set of plasmids, designed to pre-immunise *E. coli* strains against λ_{vir} by expressing an *E. coli* CRISPR-I array with a first spacer encoding either a Target (complementary to the λ genome[101,551] or a Non-Target (NT) spacer, were constructed by cloning a $spacer_1$-swapped *E. coli* CRISPR-I array (Cas-adjacent array in K-12 *E. coli*) into a high copy plasmid backbone (ColE1) containing a kanamycin

resistance cassette. This yielded pSCL787 (Target, spacer$_1$ complementary to the λ_{vir} *R* gene[101] and pSCL788 (Non-Target, spacer$_1$ complementary to the *S. cerevisiae ade2* gene).

The *sspAB-hns* rescue set of plasmids, designed to rescue the loss of CRISPR adaptation phenotype of the Δ*hns* and Δ*sspA* Δ*hns* mutants, were constructed by Gibson cloning the *hns* operon (including 419bp upstream and 122bp downstream of *hns,* containing the predicted promoter[558], regulatory and terminator regions contained between *tdk-hns-galU*, respectively) and/or *sspAB* operons (as above) into a low copy plasmid backbone (pSC101 ori) containing a carbenicillin resistance cassette. This yielded pSCL785 (*hns* rescue) and pSCL786 (*hns // sspAB* rescue). pSCL832 was constructed from pSCL565 by swapping the *E. coli* Cas1-Cas2 CDS with an eGFP CDS via Gibson Assembly.

CRISPRi adaptation host factor screen

LC-E75[304], a derivative of MG1655 *E. coli* encoding a Tetracycline-inducible dCas9 cassette integrated at the Phage 186 *attB* site, and *E. coli* MG1655 were electroporated with pSCL565, and transformants were isolated on LB + spectinomycin after overnight growth at 37C. Single colonies were inoculated into 5mL LB + spectinomycin, and grown overnight. Each experiment was repeated 3 times in triplicates, for a total of 9 paired LC-E75 (experiment) – MG1655 (control) screens.

The next day, cultures were electroporated with a library of 92,919 sgRNAs (psgRNA[304] Pooled Library #115927, Addgene), targeting coding and non-coding regions across the *E. coli* genome, as described in[303,304]. Briefly, 4mL of the overnight cultures

were diluted into 400mL LB + spectinomycin, and grown for 2h at 37C with shaking (250 r.p.m.). Cells were then subjected to an electroporation prep: cultures were split into 50mL falcon tubes, chilled on ice for 10min, and pelleted at 4000g for 15min at 4C. The supernatants were discarded, and cells were washed with 30mL of ice-cold ultra-pure, DNAse/RNAse free, pyrogen free H_2O (updH_2O). The resuspended cultures were chilled on ice for another 10min, then pelleted at 4000g for 15 min at 4C. These wash steps were repeated twice, for 3 total washes. After the last wash, cells were resuspended in 600uL of 10% glycerol in updH_2O (~800uL final volume).

Then, 180uL of cells were added to 0.2cm gap electroporation cuvettes (BioRad #1652086), and ~1ug of the sgRNA library was mixed with the cells (total volume in electroporation cuvette < 200uL). Cells were electroporated with the following settings: 2.5 kV, 25uF, 200Ω. After the pulse, cells were quickly recovered in 25mL of pre-warmed LB + spectinomycin, and placed in a shaking incubator for 1h at 37C. The cultures were then transferred into 75mL of pre-warmed LB + spectinomycin + kanamycin, and dilutions were plated on LB + spectinomycin + kanamycin to estimate CFUs.

The next day, CFUs were estimated, and the experiments were continued only if the library coverage was estimated to be >1000x. If so, 20mL of the overnight cultures were diluted in 1L warmed LB + spectinomycin + kanamycin + 1uM anhydrotetracycline (aTc); the remainder of the overnight cultures was collected by centrifugation for pre-experiment library quantification.

Cultures were grown for 3h, after which the electroporation prep was performed as described above. After the last centrifugation step, each pellet was resuspended in 150uL of a mix of *murA* targeting pre-spacer oligonucleotides and ~1ug of pSCL563 in updH_2O.

The *murA* targeting prespacer mix was prepared by combining and annealing complementary single-stranded oligos that encode a prespacer targeting the essential gene *murA* (F and R sequences: AGGTTATGGCAACCGATCTGCGTGCATCAGCAAGC; GCTTGCTGATGCACGCAGATCGGTTGCCATAACCT), to a final concentration of 3.125 uM per oligo. After electroporation, cells were rescued with 5mL of pre-warmed LB + carbenicillin + kanamycin + 1 mM m-Toluic acid + 1uM aTc, and placed in a shaking incubator for 1h at 37C. Then, these cultures were then transferred into 20 mL of pre-warmed LB + carbenicillin + kanamycin + 1 mM m-Toluic acid + 1uM aTc, and placed in a shaking incubator overnight at 37C. Cultures (post-experiment library samples) were harvested the next day by centrifugation, 4000g x 30min, followed by plasmid extraction using the Qiagen Plasmid Plus Midi kit (cat. no. 12143).

Sequencing of the sgRNA libraries was performed as follows. 1uL of the plasmid extractions were used as template in 50uL PCR reactions, using 37uL of updH_2O, 10uL 5X Q5 reaction buffer, 1uL 10mM dNTPs, 1uL Q5 Hot Start HiFi DNA polymerase and 0.25uL 100uM Forward and Reverse primers. The primers used contained Illumina adapters to make the amplicons compatible with our downstream sequencing prep, as well as 1-5 random nucleotides between the Illumina adapter and the annealing sequence to introduce diversity into the sequencing library. The PCR reaction was run using the standard recommended Q5 cycling conditions: 98C initial denaturation x 30s; 30 cycles of 98C x 10s, 62C x 30s, 72C x 30s; final extension of 2min at 72C. Amplicons were then cleaned up using AMPure XP beads (A63880), indexed using custom indexing oligos, and sequenced on an Illumina NextSeq instrument with ~2million reads per biological replicate. A list of primers can be found in **Supplementary Table 4-4**.

Fluorescence-based monitoring of the Lac promoter activity

E. coli BW25113 (control) and Δ*sspA* strains were transformed with pSCL832 by electroporation, and transformants were isolated on LB + spectinomycin after overnight growth at 37C. Single colonies ($n \geq 3$) were inoculated into 3mL of LB + spectinomycin and grown overnight with 250 r.p.m shaking at 37C. The next day, cultures were diluted 1:100 in 3mL of LB + spectinomycin and grown to log phase (~4h). Subsequently, OD_{600} of the cultures was measured on a Spectramax i3 plate reader, and cultures were normalised to an OD_{600} = 0.05. 200uL of cultures were placed on clear-bottom plate and incubated at 37C on a Spectramax i3 plate reader, with fluorescence readings (wavelength = 508nm) every 30s for a total of 7.5h.

qPCR

E. coli BW25113 (control) and Δ*sspA* strains were transformed with pSCL565 by electroporation, and transformants were isolated on LB + spectinomycin after overnight growth at 37C. Single colonies ($n \geq 3$) were inoculated into 3mL of LB + spectinomycin and grown overnight with 250 r.p.m shaking at 37C. The next day, cultures were diluted 1:100 in 3mL of LB + spectinomycin and grown to log phase (~4h). Then, 1mL of cultures was harvested by centrifugation (21,000 g x 1min), then resuspended in 250uL of updH_2O. These samples were heated to 95C for 15min, then placed on ice to cool. Then, lysates were treated with 2 units of Proteinase K (NEB) for 30min, followed by Proteinase K inactivation by incubation at 95C for 10min. Lastly, lysates were centrifuged at 21,000 g for 2min, and supernatants were diluted 1:500 in updH_2O. 5uL of the diluted supernatant was used in 20uL qPCR reactions, set up using the NEB Luna Universal qPCR Master

Mix following the manufacturer's instructions. qPCR Primers were designed to target pSCL565's CDF ori and *cas1* regions, using the genomic *ompA* as a reference. Primers are listed in **Supplementary Table 4-4**.

Naïve CRISPR-Cas adaptation

E. coli BW25113 (control) and strains of interest were transformed with pSCL565 by electroporation, and transformants were isolated on LB + spectinomycin after overnight growth at 37C. In the case of "plasmid rescue" experiments, strains of interest we co-transformed with pSCL565 and the rescue plasmid by electroporation, and transformants were isolated on LB + spectinomycin + carbenicillin after overnight growth at 37C.

Single colonies ($n \geq 3$) were inoculated into individual wells of a 96-well deep well plate containing 500uL of LB + spectinomycin (and carbenicillin, if needed), and grown for 48h with 1000 r.p.m shaking at 37C. After 48h of growth, 75uL of the cultures were mixed with 75uL of updH_2O, heated to 95C for 10min, and spun-down. 0.5uL of the supernatant was used as template for 25uL PCR reactions (same recipe and cycling protocol as above). We designed primers to amplify a region of the *E. coli* CRISPR-II array, contained between the end of the Leader sequence and the second pre-existing spacer. To reduce the number of indices needed per sample, we designed 3 barcoded F primers (one per biological replicate) to amplify the CRISPR arrays – these would enable us to pool the samples post-CRISPR array amplification, and de-multiplex the biological replicates during data analysis. A list of primers can be found in Supplementary Table 4.

In some cases, CRISPR array expansions are visible on an agarose gel as laddering caused by larger arrays (expanded) migrating slower than the shorter parental arrays. We visualised this by running 5uL of the pooled PCR products on Invitrogen 2% Agarose SYBR safe E-Gels (A42135). Gels were re-stained with SYBR Gold before imaging.

Phage plaque assays

BW25113 (control), Δ*sspA*::FRT, Δ*hns*::FRT, and Δ*sspA*::FRT Δ*hns*::FRT strains were transformed with plasmids encoding either Target or Non-Target CRISPR-I arrays (pSCL787 and pSCL788, respectively), and transformants were isolated on LB + kanamycin after overnight growth at 37C. Single colonies ($n \geq 3$) were inoculated into 3mL of LB + kanamycin supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and grown overnight with 250 r.p.m shaking at 30C. The next day, top agar lawns of each bacterial culture were prepared by mixing 100uL of overnight cultures with 5mL of top agar (0.5% w:v LB agar, supplemented with 10 mM $MgSO_4$ and 0.2% maltose and kanamycin). Top agar mixtures were poured over LB agar + kanamycin plates and left to dry at room temperature, partially open by a sterilizing flame. Serial dilutions of λ_{vir} were prepared in LB supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and 2.5uL of each dilution was spotted on the top agar using a multichannel pipette, and left to dry completely at room temperature. Plates were incubated at 30C overnight.

Efficiency of plating was calculated as the number of plaques formed by λ_{vir} on lawns of a strain harbouring pSCL787 (Target) divided by the plaques formed by λ_{vir} on

lawns of a strain harbouring pSCL788 (Non-Target). Full plaque assay plates for all *n* = 3 biological replicates in **Extended Data Figure 4-5**.

Phage resistance infection growth curves

BW25113 (control), Δ*sspA*::FRT, Δ*hns*::FRT, and Δ*sspA*::FRT Δ*hns*::FRT strains were transformed with plasmids encoding either Target or Non-Target CRISPR-I arrays (pSCL787 and pSCL788, respectively), and transformants were isolated on LB + kanamycin after overnight growth at 37C. Single colonies ($n \geq 3$) were inoculated into 3mL of LB + kanamycin supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and grown overnight with 250 r.p.m shaking at 30C. The next day, cultures were diluted 1:100 in 3mL of LB + kanamycin supplemented with 10 mM $MgSO_4$ and 0.2% maltose and grown to log phase (~4h). Subsequently, OD_{600} of the cultures was measured on a Spectramax i3 plate reader, and cultures were normalised to an OD_{600} = 0.05. 200uL of cultures was infected with a range of MOIs (10 → 10^{-8}), using serial dilutions of λ_{vir} prepared in LB supplemented with 10 mM $MgSO_4$ and 0.2% maltose. Cultures were loaded on clear-bottom plate and incubated at 30C on a Spectramax i3 plate reader, with OD_{600} readings every 2.5mins for a total of 16h.

CRISPR-Cas primed adaptation after phage infection

BW25113 (control), Δ*sspA*::FRT, Δ*hns*::FRT, and Δ*sspA*::FRT Δ*hns*::FRT strains were transformed with plasmids encoding either Target or Non-Target CRISPR-I arrays (pSCL787 and pSCL788, respectively), and transformants were isolated on LB + kanamycin after overnight growth at 37C. Single colonies ($n \geq 3$) were inoculated into 3mL

of LB + kanamycin supplemented with 10 mM $MgSO_4$ and 0.2% maltose, and grown overnight with 250 r.p.m shaking at 30C. The next day, cultures were diluted 1:100 in 3mL of LB + kanamycin supplemented with 10 mM $MgSO_4$ and 0.2% maltose and grown to log phase (~4h). Subsequently, OD_{600} of the cultures was measured on a Spectramax i3 plate reader, and cultures were normalised to an OD_{600} = 0.05. 200uL of cultures was infected with λ_{vir} at an MOI of 0.1, and cultures were loaded on clear-bottom plate and incubated at 30C on a Spectramax i3 plate reader, with OD_{600} readings every 1.5mins for a total of 3h.

After 3h, the cultures were harvested by centrifugation, resuspended in 100uL of updH_2O, heated to 95C for 10min, and spun-down. 1uL of the supernatant was used as template for 25uL PCR reactions (same recipe and cycling protocol as above). We designed primers to amplify a region of the *E. coli* CRISPR-II array, contained between the end of the Leader sequence and the second pre-existing spacer. The barcoded F primer approach, described above, was used to pool PCRs and de-multiplex biological replicates during data analysis.

Protein model structures

Protein model coordinates were retrieved from the RSCB Protein Data Bank (codes 7DY6 and 8ET3). Figures were prepared using UCSF ChimeraX[559].

Data analysis

The data analysis for this project can be broken down into 5 modules: (1) processing of the sequencing reads to extract, count, and group sgRNAs by gene/gene-adjacent regions; (2) generate binned coverage plots of sgRNAs across the *E. coli* genome; (3) identify the statistically enriched/depleted sgRNAs, using PyDESeq2[530], a Python implementation of DEseq2[560]; (4) quantify the rates of CRISPR adaptation; and (5) extract new spacers perform spacer analysis. All data analysis was performed in Jupyter Lab[561], and all code to replicate this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen.

Sequencing data processing: from reads to sgRNA counts

First, fastq reads were trimmed using sickle-trim[562]. For each fastq, a counter of sgRNAs was generated by extracting the sgRNA from each read, provided that this sgRNA could be found in the original synthesised psgRNA library[304]. Then, the sgRNAs were BLASTed[563] against the *E. coli* MG1655 genome and the top hit was saved. For each sample, a DataFrame of genomic_location–sgRNA–count was generated, and used for downstream analysis. All data corresponding to the screens can be found in **Supplementary Tables 4-5** and **4-6**.

The Jupyter Notebook for this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/blast_screen_hits_clean.ipynb

Binned coverage plot of sgRNAs across the E. coli genome

We generated occupancy arrays for each sample, using counts generated above. These arrays contain cumulative counts of sgRNAs per base, i.e., occupancy O = [c_1, c_2, …, c_n], where n is the size of the *E. coli* MG1655 genome and c_i are the total sgRNA counts at that position. We then normalised the counts to the total sgRNA count in that sample, i.e., O = [c_1/sum_sgRNAs, c_2/sum_sgRNAs, …, c_n/sum_sgRNAs], where sum_sgRNAs is total sgRNA count. Next, we calculated the mean occupancy for the experimental and control conditions, i.e., $O_{LC\text{-}E75}$ = ($O_{biorep1}$ + $O_{biorep2}$ + … +$O_{biorep9}$) / 9, where $O_{LC\text{-}E75}$ is the mean occupancy for the experimental condition, and O_{biorep_i} are the normalised counts for each biological replicate of the screen run in the experimental condition. Lastly, we calculated the delta occupancy, or difference between the mean sgRNA occupancies of the experimental and control conditions, and posteriorly calculated the mean delta sgRNA occupancy in a sliding window, in the interest of interpretability. We used pyCirclize[564] to generate the final occupancy plot.

The Jupyter Notebook for this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/plot_genome_coverage_clean.ipynb

Identification of enriched/depleted sgRNAs

We performed statistical testing for enriched/depleted sgRNAs from binned sgRNA (sum of all sgRNAs per gene) count data generated in (1) using the PyDESeq2 package[530], and compared each experimental sample to its paired control, and controlled for pre-experimental variation in the relative sgRNA library composition by including the sgRNA counts from the pre-experiment library as an interactor factor (i.e., sgRNA_counts

~ input_lib_counts + knockdown (yes/no)). Genes that have less than a total of 10 reads for all of their sgRNAs in the dataset were removed from the analysis. The log_2FoldChange (log_2FC) value represents the enrichment or depletion of each gene. The lists of all genes, log_2FC and adjusted p-values can be found in **Supplementary Table 4-6**. The Jupyter Notebook for this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/deseq2_volcano_allhits_clean.ipynb

Quantification of the rates of CRISPR adaptation

First, fastq reads were trimmed using sickle-trim. For each fastq, we filtered for reads containing the Leader-repeat junction of the *E. coli* CRISPR-II array. We then identified newly acquired spacers from the array sequences by recursive identification of CRISPR repeats and comparison of putative new spacers to pre-existing spacers in the array, using a lenient search algorithm allowing for a maximum of 3bp mismatches. We generated sums of new expansions in CRISPR arrays per condition, and used these to calculate the rate of CRISPR adaptation (100 * number of newly expanded CRISPR arrays / total number of arrays sequenced). Lastly, we normalised the rate of CRISPR adaptation for each condition by the wild-type rate CRISPR adaptation, so as to make inter-experiment comparisons feasible and more interpretable. All normalised rates corresponding to the CRISPR adaptation experiments, as well as the “run” label (i.e., batch in which the experiment were run and sequenced) can be found in **Supplementary Table 4-7**.

The Jupyter Notebook for this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/spacer_fishing_clean.ipynb

Newly-acquired spacer analysis

Spacer analysis involves in several steps:

Extraction of new spacers

We began by using the same recursive new spacer search algorithm described above to extract new spacers. In parallel to extracting spacers, we also stored information regarding the total number of arrays sequenced and the fraction of those that were expanded, to use as normalisation for comparisons across samples that might vary in sequencing depth or quality.

Identification of spacer origin

Next, we generated a counter of newly acquired spacers and their frequencies. We used this to generate FASTA files of new spacers and their counts, which were subsequently BLASTed to two databases: the *E. coli* K-12 genome (taxid 511145) and pSCL565, to capture spacers derived from the Cas1-Cas2 expression plasmid. To identify the source of acquired spacers during CRISPR-Cas primed adaptation amidst phage infection experiments, unique spacers extracted in steps described above were BLASTed to four databases: the *E. coli* K-12 genome (taxid 511145); the bacteriophage lambda genome (taxid 2681611); and pSCL787 or pSCL788, to capture spacers derived from the defence plasmids. In both cases, BLAST searches were performed with high stringency (≥90% identity, i.e., 30/33bp match between spacer and reference query) to obtain unique matches to the reference maps. We then parsed the BLAST results and filtered the

genome-matching spacers for LacI, Cas1 and Cas2, as we assumed that spacers from these sources were most likely plasmid-derived.

Mapping spacers to reference genomes

Using the spacer genomic (lambda or *E. coli* K-12) or plasmidic (pSCL565, pSCL787 and pSCL788) location, target locus and counts, we generated coverage maps of the different genomes and plasmids where the spacers could have been sourced from, as well as spacer counts per location (i.e., counts of how many of the new spacers were *E. coli*, lambda, or plasmid derived). Briefly, for each BLAST record, we first assessed whether the BLAST record mapped to any of our reference genomes, and if so, added counts to spacer origin and occupancy counters. The occupancy array is generated analogously to those used to estimate sgRNA coverage (see above), and is genome-size aware (i.e., accounts for start-end junctions).

Spacer neighbourhood analysis

We also used the spacer ←→ genome information to look into the 15bp up and downstream of the genomic origin of the new spacer, in the hopes of capturing information regarding the PAM (canonically, AAG for this CRISPR adaptation system) and any other discernible motifs. This was done by mapping the spacer back to its reference genome, using the BLAST results, and extracting 15 bases upstream and downstream of the spacer. These sequences were compiled and Logomaker[565] was used to generate sequence logos for the up and downstream region. This yielded **Figure 4-2d**.

Spacer origin distribution

Next, we used the spacer origin counters to obtain information about the breakdown of spacers by their origin (*E. coli*, lambda, or plasmids). To do so, we first normalised the spacer count per location to the number of arrays sequenced. In parallel, we also normalised the spacer count per location to the total number of new spacers identified, converting this metric to the percent of spacers mapping to each location. This allowed us to then plot the new spacer count with respect to spacer origin and strain of interest (**Figure 4-2b** and **Figure 4-5b**), in addition to the percent of new spacers belonging to each spacer origin and strain of interest (**Figure 4-2c**).

Coverage plots

Lastly, we generated genome coverage plots for *E. coli*, lambda, and the plasmids, as described above. For the *E. coli* and lambda genomes, we generated binned coverage plots by calculating the coverage as a sliding mean, or binned coverage. The spacer coverage for plasmids was generated without binning spacer occupancy. This analysis yielded **Figures 4-2e-h**, **Figure 4-5c**, **Extended Data figure 4-1** and **Extended Data figure 4-5**.

The Jupyter Notebook for this analysis can be found here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/map_new_spacers_clean.ipynb

Biological replicates

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

Data availability

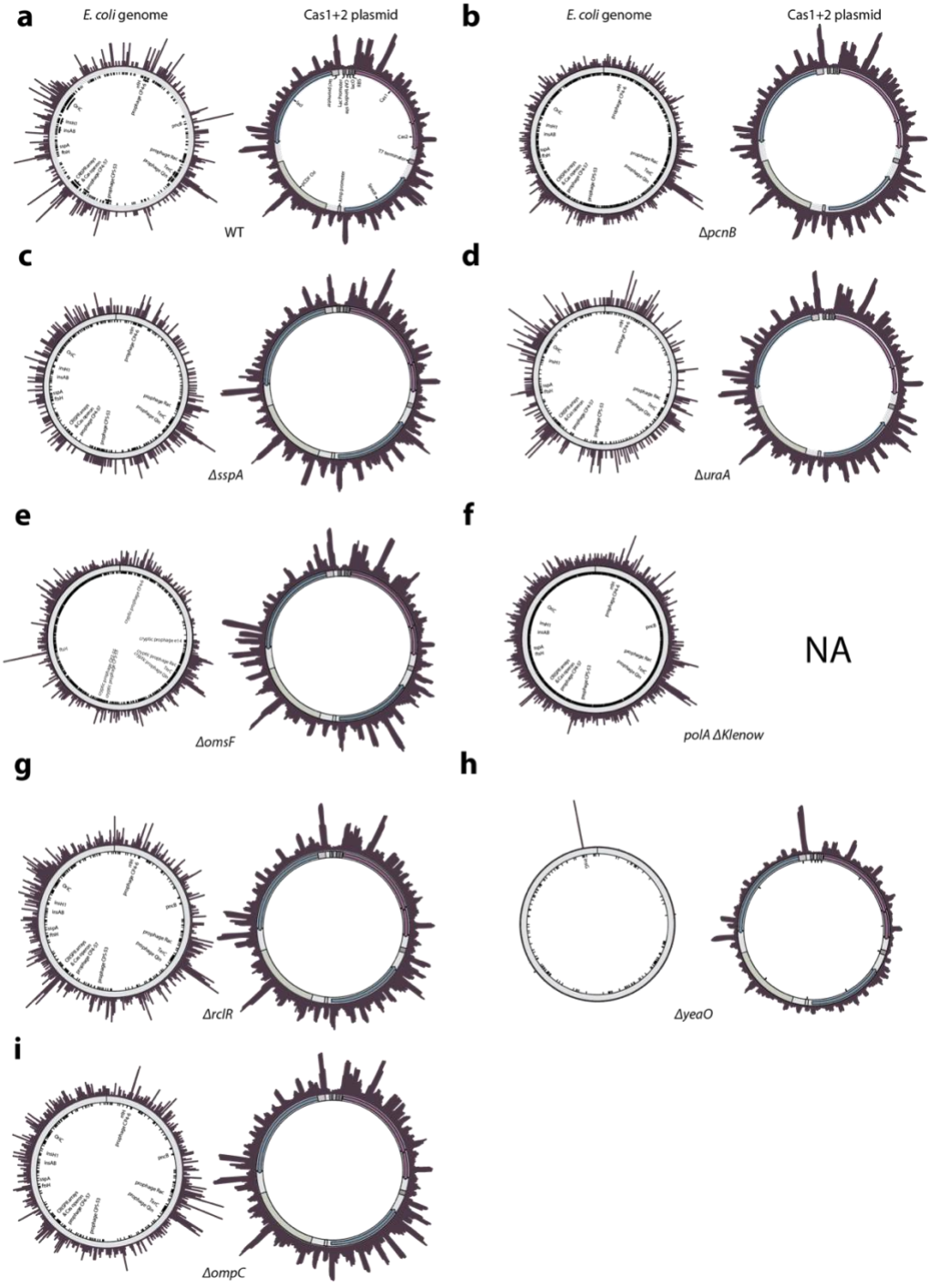
All data supporting the findings of this study are available within the article and its supplementary information.

Data used to generate all figures and perform statistical analysis, alongside a Jupyter Notebook to recreate our figures is available on GitHub: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen/blob/main/plot_run_stats_clean.ipynb. All sequencing data associated with this study is available on NCBI SRA (PRJNA1109382).

Code availability

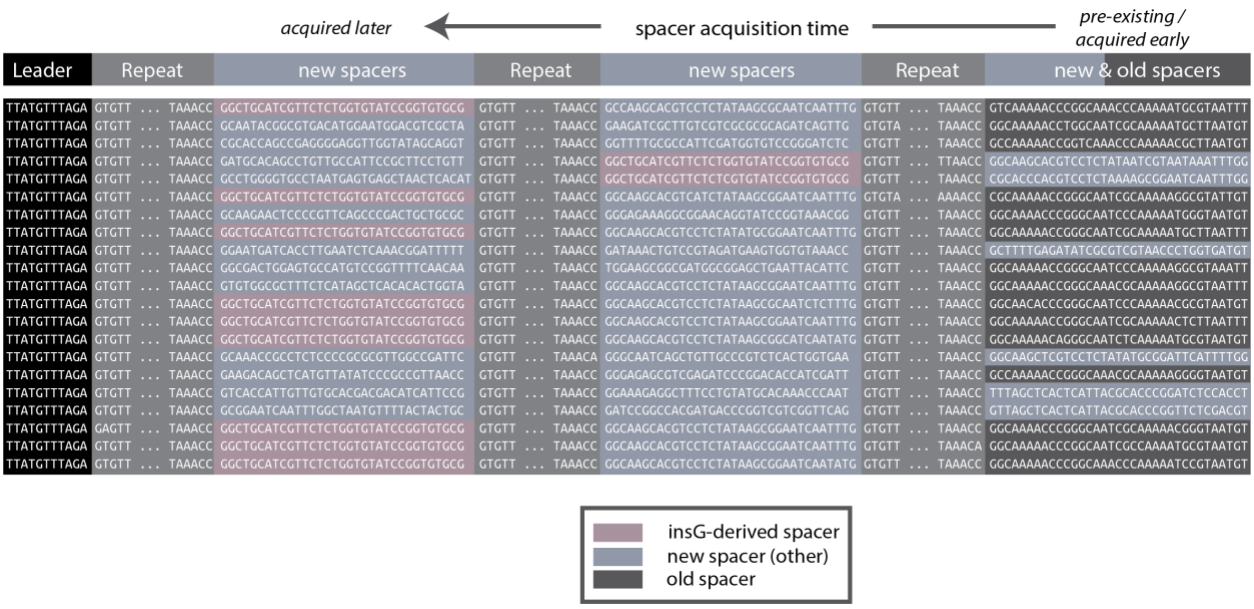
All code used to process or analyse data from this study is available on GitHub here: https://github.com/Shipman-Lab/CRISPRi_host_factor_screen.

4.6 Supplementary Information



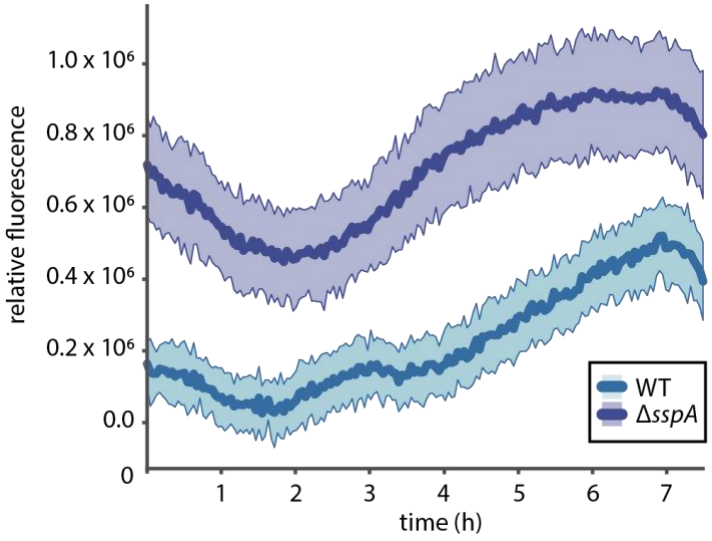
Extended Data Figure 4-11: Binned coverage plot of newly acquired spacer across the E. coli genome (left) and pSCL565 plasmid (right) for strains selected for individual validation.

a-i*: wild-type, ΔpcnB, ΔsspA, ΔuraA, ΔomsF, polA ΔKlenow, ΔrclR, ΔyeaO and ΔompC. Wild-type is E. coli BW25113, parental strain to the Keio collection; all other strains besides polA ΔKlenow are from the Keio collection. polA ΔKlenow was constructed as described previously*[532]*.*



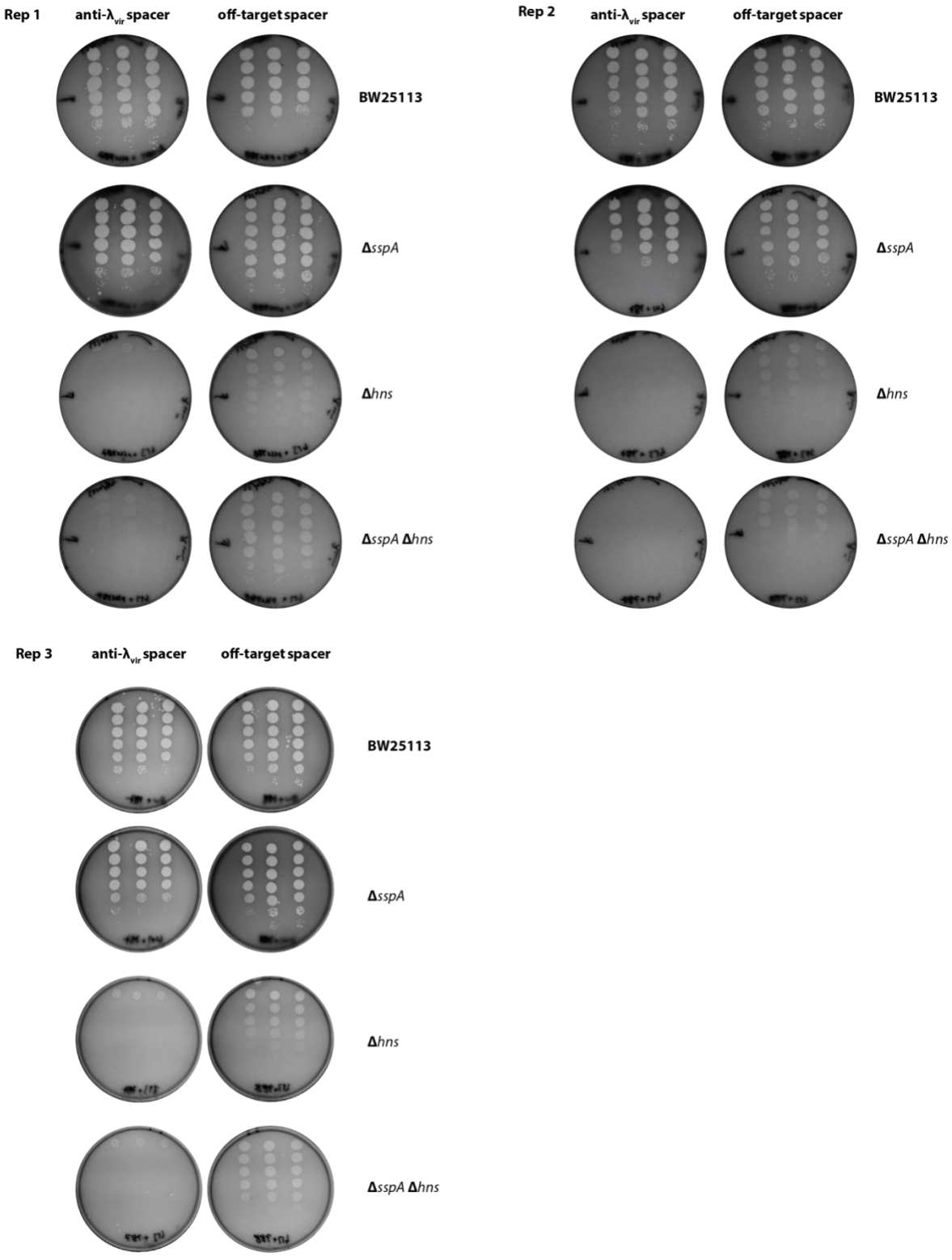
Extended Data Figure 4-12: Alignment of sequencing reads corresponding to doubly and triply expanded CRISPR array from the ΔyeaO mutant strain, highlighting insG-derived spacers, new spacers and old or pre-existing spacers.

Spacers acquired at a later stage are closer to the Leader sequence than spacers acquired earlier, or than pre-existing spacers[403,410,411].



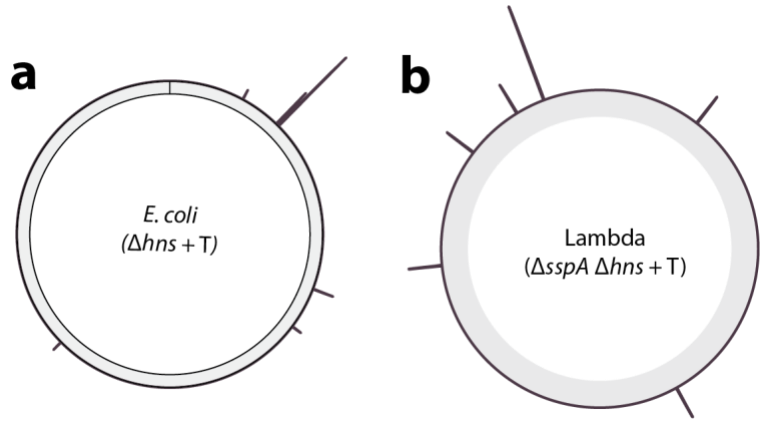
Extended Data Figure 4-13: Fluorescence-based monitoring of the Lac promoter activity, used to express the Cas1-Cas2 integrases on pSCL565, in wild-type and ΔsspA cells, over the course of 7h of liquid culture.

Hue around solid line (mean) represents the standard deviation across 3 biological replicates.



Extended Data Figure 4-14: Full plates and 3 biological replicates of plaque assays.

Plaques of λ_{vir} on WT, ΔsspA::FRT, Δhns::FRT and ΔsspA::FRT Δhns::FRT strains, pre-immunised with either T or NT defence plasmids, corresponding to ***Figure 4-4d.*** *Strains were infected with λ_{vir} and grown on plates at 30ºC for 16h.*



Extended Data Figure 4-15: Distribution of newly acquired spacers in ∆hns +T and ∆sspA ∆hns +T strains upon lambda infection.

a*. Binned coverage plot of ∆hns + T newly acquired spacers across the E. coli genome (outer, purple).* ***b****. Binned coverage plot of ∆sspA ∆hns + T newly acquired spacers across the lambda genome (outer, purple).*

4.7 Supplemental Files

Supplementary_Tables_Chapter4.xlsx

This PDF file contains:

- Supplementary Table 4-1: Statistical analysis
- Supplementary Table 4-2: Strains used in this study
- Supplementary Table 4-3: Plasmids used in this study
- Supplementary Table 4-4: Oligos used in this study
- Supplementary Table 4-5: Summary of library screen samples
- Supplementary Table 4-6: Differential analysis results (DEseq2 analysis)
- Supplementary Table 4-7: Rates of naïve CRISPR adaptation for all experiments reported in this work
- Supplementary Table 4-8: qPCR data and analysis
- Supplementary Table 4-9: Efficiency of plating and analysis

References

1. Engels, F., Dutt, C. P. & Haldane, J. B. S. *Dialectics of Nature*. (International Publishers, New York, 1976).
2. Wilczynski, J. *An Encyclopedic Dictionary of Marxism, Socialism and Communism: Economic, Philosophical, Political and Sociological Theories, Concepts, Institutions and Practices - Classical and Modern, East-West Relations Included*. (Macmillan, London, 1984).
3. Gould, S. J. *An Urchin in the Storm: Essays about Books and Ideas*. (W.W. Norton, New York, 1988).
4. Chevallereau, A., Pons, B. J., Van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
5. Marx, K. & Engels, F. *Collected Works. 25: Frederick Engels: Anti-Dühring. Dialectics of Nature*. (International Publ, New York, 1987).
6. Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.-L. & Brüssow, H. Phage-Host Interaction: an Ecological Perspective. *J. Bacteriol.* **186**, 3677–3686 (2004).
7. Stern, A. & Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *BioEssays* **33**, 43–51 (2011).
8. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
9. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).

10. Bertozzi Silva, J., Storms, Z. & Sauvageau, D. Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**, fnw002 (2016).

11. Seed, K. D. *et al.* Phase Variable O Antigen Biosynthetic Genes Control Expression of the Major Protective Antigen and Bacteriophage Receptor in Vibrio cholerae O1. *PLoS Pathog.* **8**, e1002917 (2012).

12. Van Der Woude, M. W. & Bäumler, A. J. Phase and Antigenic Variation in Bacteria. *Clin. Microbiol. Rev.* **17**, 581–611 (2004).

13. Tal, N. *et al.* Bacteria deplete deoxynucleotides to defend against bacteriophage infection. *Nat. Microbiol.* **7**, 1200–1209 (2022).

14. Huiting, E. & Bondy-Denomy, J. Defining the expanding mechanisms of phage-mediated activation of bacterial immunity. *Curr. Opin. Microbiol.* **74**, 102325 (2023).

15. Arber, W. Restriction Endonucleases. *Angew. Chem. Int. Ed. Engl.* **17**, 73–79 (1978).

16. Lopatina, A., Tal, N. & Sorek, R. Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annu. Rev. Virol.* **7**, 371–384 (2020).

17. Luria, S. E. & Delbrück, M. MUTATIONS OF BACTERIA FROM VIRUS SENSITIVITY TO VIRUS RESISTANCE. *Genetics* **28**, 491–511 (1943).

18. Burnet, F. M. “Smooth-rough” variation in bacteria in its relation to bacteriophage. *J. Pathol. Bacteriol.* **32**, 15–42 (1929).

19. Hyman, P. & Abedon, S. T. Bacteriophage Host Range and Bacterial Resistance. in *Advances in Applied Microbiology* vol. 70 217–248 (Elsevier, 2010).

20. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).

21. Clément, J. M., Lepouce, E., Marchal, C. & Hofnung, M. Genetic study of a membrane protein: DNA sequence alterations due to 17 lamB point mutations affecting adsorption of phage lambda. *EMBO J.* **2**, 77–80 (1983).

22. Hofnung, M., Jezierska, A. & Braun-Breton, C. lamB mutations in E. coli K12: Growth of λ host range mutants and effect of nonsense suppressors. *Mol. Gen. Genet. MGG* **145**, 207–213 (1976).

23. Werts, C., Michel, V., Hofnung, M. & Charbit, A. Adsorption of bacteriophage lambda on the LamB protein of Escherichia coli K-12: point mutations in gene J of lambda responsible for extended host range. *J. Bacteriol.* **176**, 941–947 (1994).

24. Samuel, A. D. T. *et al.* Flagellar determinants of bacterial sensitivity to χ-phage. *Proc. Natl. Acad. Sci.* **96**, 9863–9866 (1999).

25. Komano, T., Kubo, A. & Nisioka, T. Shufflon: multi-inversion of four contiguous DNA segments of plasmid R64 creates seven different open reading frames. *Nucleic Acids Res.* **15**, 1165–1172 (1987).

26. Raimondo, L. M., Lundh, N. P. & Martinez, R. J. Primary Adsorption Site of Phage PBS1: the Flagellum of *Bacillus subtilis*. *J. Virol.* **2**, 256–264 (1968).

27. Valentine, R. C. & Strand, M. Complexes of F-Pili and RNA Bacteriophage. *Science* **148**, 511–513 (1965).

28. Abraham, J. M., Freitag, C. S., Clements, J. R. & Eisenstein, B. I. An invertible element of DNA controls phase variation of type 1 fimbriae of Escherichia coli. *Proc. Natl. Acad. Sci.* **82**, 5724–5727 (1985).

29. Segal, E. Role of chromosomal rearrangement in N. gonorrhoeae pilus phase variation. *Cell* **40**, 293–300 (1985).

30. Zieg, J., Silverman, M., Hilmen, M. & Simon, M. Recombinational Switch for Gene Expression. *Science* **196**, 170–172 (1977).

31. Hood, D. W. *et al.* DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc. Natl. Acad. Sci.* **93**, 11121–11125 (1996).

32. Bayliss, C. D., Sweetman, W. A. & Moxon, E. R. Mutations in *Haemophilus influenzae* Mismatch Repair Genes Increase Mutation Rates of Dinucleotide Repeat Tracts but Not Dinucleotide Repeat-Driven Pilin Phase Variation Rates. *J. Bacteriol.* **186**, 2928–2935 (2004).

33. Bikard, D. & Marraffini, L. A. Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. *Curr. Opin. Immunol.* **24**, 15–20 (2012).

34. Luria, S. E. HOST-INDUCED MODIFICATIONS OF VIRUSES. *Cold Spring Harb. Symp. Quant. Biol.* **18**, 237–244 (1953).

35. Arber, W. Host-Controlled Modification of Bacteriophage. *Annu. Rev. Microbiol.* **19**, 365–378 (1965).

36. Bertani, G. & Weigle, J. J. HOST CONTROLLED VARIATION IN BACTERIAL VIRUSES. *J. Bacteriol.* **65**, 113–121 (1953).

37. Arber, W. & Dussoix, D. Host specificity of DNA produced by Escherichia coli. *J. Mol. Biol.* **5**, 18–36 (1962).

38. Dussoix, D. & Arber, W. Host specificity of DNA produced by Escherichia coli. *J. Mol. Biol.* **5**, 37–49 (1962).

39. Arber, W. Host-Controlled Modification of DNA. in *Molecular Genetics* (eds. Wittmann, H. G. & Schuster, H.) 155–160 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1968). doi:10.1007/978-3-642-87534-2_14.

40. Arber, W. Host specificity of DNA produced by Escherichia coli. *J. Mol. Biol.* **11**, 247–256 (1965).

41. Smith, J. D., Arber, W. & Kühnlein, U. Host specificity of DNA produced by Escherichia coli. *J. Mol. Biol.* **63**, 1–8 (1972).

42. Korona, R. & Levin, B. R. PHAGE-MEDIATED SELECTION AND THE EVOLUTION AND MAINTENANCE OF RESTRICTION-MODIFICATION. *Evolution* **47**, 556–575 (1993).

43. Temin, H. M. Retrons in bacteria. *Nature* **339**, 254–255 (1989).

44. Yee, T., Furuichi, T., Inouye, S. & Inouye, M. Multicopy single-stranded DNA isolated from a gram-negative bacterium, Myxococcus xanthus. *Cell* **38**, 203–209 (1984).

45. Furuichi, T., Dhundale, A., Inouye, M. & Inouye, S. Branched RNA covalently linked to the 5′ end of a single-stranded DNA in Stigmatella aurantiaca: Structure of msDNA. *Cell* **48**, 47–53 (1987).

46. Furuichi, T., Inouye, S. & Inouye, M. Biosynthesis and structure of stable branched RNA covalently linked to the 5′ end of multicopy single-stranded DNA of Stigmatella aurantiaca. *Cell* **48**, 55–62 (1987).

47. 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).

48. Lampson, B. C. *et al.* Reverse Transcriptase in a Clinical Strain of *Escherichia coli* : Production of Branched RNA-Linked msDNA. *Science* **243**, 1033–1038 (1989).

49. Inouye, S., Hsu, M.-Y., Eagle, S. & Inouye, M. Reverse transcriptase associated with the biosynthesis of the branched RNA-linked msDNA in Myxococcus xanthus. *Cell* **56**, 709–717 (1989).

50. 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).

51. Gao, L. *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077–1084 (2020).

52. Millman, A. *et al.* Bacterial Retrons Function In Anti-Phage Defense. *Cell* **183**, 1551-1561.e12 (2020).

53. Bobonis, J. *et al.* Bacterial retrons encode phage-defending tripartite toxin–antitoxin systems. *Nature* **609**, 144–150 (2022).

54. Wang, Y. *et al.* Cryo-EM structures of Escherichia coli Ec86 retron complexes reveal architecture and defence mechanism. *Nat. Microbiol.* **7**, 1480–1489 (2022).

55. Carabias, A. *et al.* Retron-Eco1 assembles NAD+-hydrolyzing filaments that provide immunity against bacteriophages. *Mol. Cell* **84**, 2185-2202.e12 (2024).

56. Mayo-Muñoz, D., Pinilla-Redondo, R., Birkholz, N. & Fineran, P. C. A host of armor: Prokaryotic immune strategies against mobile genetic elements. *Cell Rep.* **42**, 112672 (2023).

57. Snyder, L. Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? *Mol. Microbiol.* **15**, 415–420 (1995).

58. Niewoehner, O. *et al.* Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543–548 (2017).

59. Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605–609 (2017).

60. Durmaz, E. & Klaenhammer, T. R. Abortive Phage Resistance Mechanism AbiZ Speeds the Lysis Clock To Cause Premature Lysis of Phage-Infected *Lactococcus lactis*. *J. Bacteriol.* **189**, 1417–1425 (2007).

61. Bingham, R., Ekunwe, S. I. N., Falk, S., Snyder, L. & Kleanthous, C. The Major Head Protein of Bacteriophage T4 Binds Specifically to Elongation Factor Tu. *J. Biol. Chem.* **275**, 23219–23226 (2000).

62. Schmitt, C. K. & Molineux, I. J. Expression of gene 1.2 and gene 10 of bacteriophage T7 is lethal to F plasmid-containing Escherichia coli. *J. Bacteriol.* **173**, 1536–1543 (1991).

63. Tamulaitiene, G. *et al.* Activation of Thoeris antiviral system via SIR2 effector filament assembly. *Nature* **627**, 431–436 (2024).

64. Morehouse, B. R. *et al.* STING cyclic dinucleotide sensing originated in bacteria. *Nature* **586**, 429–433 (2020).

65. LeRoux, M. *et al.* The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat. Microbiol.* **7**, 1028–1040 (2022).

66. Banh, D. V. *et al.* Bacterial cGAS senses a viral RNA to initiate immunity. *Nature* **623**, 1001–1008 (2023).

67. Cohen, D. *et al.* Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature* **574**, 691–695 (2019).

68. Parma, D. H. *et al.* The Rex system of bacteriophage lambda: tolerance and altruistic cell death. *Genes Dev.* **6**, 497–510 (1992).

69. Lau, R. K. *et al.* Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity. *Mol. Cell* **77**, 723-733.e6 (2020).

70. Levitz, R. *et al.* The optional E. coli prr locus encodes a latent form of phage T4-induced anticodon nuclease. *EMBO J.* **9**, 1383–1389 (1990).

71. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *J. Bacteriol.* **193**, 6039–6056 (2011).

72. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).

73. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).

74. Hochhauser, D., Millman, A. & Sorek, R. The defense island repertoire of the Escherichia coli pan-genome. *PLOS Genet.* **19**, e1010694 (2023).

75. Millman, A. *et al.* An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556-1569.e5 (2022).

76. Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. *Nat. Rev. Microbiol.* **21**, 686–700 (2023).

77. Ojima, S. *et al.* Systematic Discovery of Phage Genes that Inactivate Bacterial Immune Systems. Preprint at https://doi.org/10.1101/2024.04.14.589459 (2024).

78. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome. *Nat. Microbiol.* **7**, 1568–1579 (2022).

79. Rousset, F. *et al.* Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe* **30**, 740-753.e5 (2022).

80. Van Houte, S., Buckling, A. & Westra, E. R. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiol. Mol. Biol. Rev.* **80**, 745–763 (2016).

81. Rostøl, J. T. & Marraffini, L. (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host Microbe* **25**, 184–194 (2019).

82. Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. Remarkable Mechanisms in Microbes to Resist Phage Infections. *Annu. Rev. Virol.* **1**, 307–331 (2014).

83. Payne, L. J. *et al.* PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res.* **50**, W541–W550 (2022).

84. Shomar, H. *et al.* Viperin immunity evolved across the tree of life through serial innovations on a conserved scaffold. *Nat. Ecol. Evol.* (2024) doi:10.1038/s41559-024-02463-z.

85. Bernheim, A., Cury, J. & Poirier, E. Z. The immune modules conserved across the tree of life: Towards a definition of ancestral immunity. *PLOS Biol.* **22**, e3002717 (2024).

86. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).

87. Rousset, F. & Sorek, R. The evolutionary success of regulated cell death in bacterial immunity. *Curr. Opin. Microbiol.* **74**, 102312 (2023).

88. Tesson, F. & Bernheim, A. Synergy and regulation of antiphage systems: toward the existence of a bacterial immune system? *Curr. Opin. Microbiol.* **71**, 102238 (2023).

89. Wu, Y. *et al.* Bacterial defense systems exhibit synergistic anti-phage activity. *Cell Host Microbe* **32**, 557-572.e6 (2024).

90. Patterson, A. G., Yevstigneyeva, M. S. & Fineran, P. C. Regulation of CRISPR–Cas adaptive immune systems. *Curr. Opin. Microbiol.* **37**, 1–7 (2017).

91. Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nat. Commun.* **13**, 5731 (2022).

92. Tang, S. *et al.* De novo gene synthesis by an antiviral reverse transcriptase. Preprint at https://doi.org/10.1101/2024.05.08.593200 (2024).

93. 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).

94. Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications. *Nat. Microbiol.* **2**, 17092 (2017).

95. Mojica, F. J. M. & Rodriguez-Valera, F. The discovery of CRISPR in archaea and bacteria. *FEBS J.* **283**, 3162–3169 (2016).

96. Ishino, Y., Krupovic, M. & Forterre, P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J. Bacteriol.* **200**, (2018).

97. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

98. Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, eaal5056 (2017).

99. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).

100. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).

101. Brouns, S. J. J. *et al.* Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* **321**, 960–964 (2008).

102. 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).

103. 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).

104. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).

105. Marraffini, L. A. & Sontheimer, E. J. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science* **322**, 1843–1845 (2008).

106. Horvath, P. *et al.* Diversity, Activity, and Evolution of CRISPR Loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401–1412 (2008).

107. Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR-Cas Systems. *Mol. Cell* **61**, 797–808 (2016).

108. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).

109. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).

110. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61 (2007).

111. Deveau, H. *et al.* Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).

112. Hille, F. & Charpentier, E. CRISPR-Cas: biology, mechanisms and relevance. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150496 (2016).

113. Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* **327**, 167–170 (2010).

114. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

115. Van Der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).

116. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6**, 181–186 (2008).

117. 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).

118. Makarova, K. S. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* **30**, 482–496 (2002).

119. Horvath, P. *et al.* Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.* **131**, 62–70 (2009).

120. Godde, J. S. & Bickerton, A. The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes. *J. Mol. Evol.* **62**, 718–729 (2006).

121. Andersson, A. F. & Banfield, J. F. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science* **320**, 1047–1050 (2008).

122. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).

123. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).

124. Hale, C. R. *et al.* RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **139**, 945–956 (2009).

125. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* **10**, 200–207 (2008).

126. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).

127. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

128. Tamulaitis, G., Venclovas, Č. & Siksnys, V. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends Microbiol.* **25**, 49–61 (2017).

129. McMahon, S. A. *et al.* Structure and mechanism of a Type III CRISPR defence DNA nuclease activated by cyclic oligoadenylate. *Nat. Commun.* **11**, 500 (2020).

130. Kolesnik, M. V., Fedorova, I., Karneyeva, K. A., Artamonova, D. N. & Severinov, K. V. Type III CRISPR-Cas Systems: Deciphering the Most Complex Prokaryotic Immune System. *Biochem. Mosc.* **86**, 1301–1314 (2021).

131. Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019).

132. Abudayyeh, O. O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

133. Swarts, D. C. & Jinek, M. Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol. Cell* **73**, 589-600.e4 (2019).

134. Zhang, L. *et al.* Cas1 mediates the interference stage in a phage-encoded CRISPR-Cas system. Preprint at https://doi.org/10.1101/2024.03.09.584257 (2024).

135. An alternative mechanism for recruiting Cas2/3 in a phage-encoded CRISPR–Cas system. *Nat. Chem. Biol.* (2024) doi:10.1038/s41589-024-01667-5.

136. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).

137. Mohanraju, P. *et al.* Alternative functions of CRISPR–Cas systems in the evolutionary arms race. *Nat. Rev. Microbiol.* **20**, 351–364 (2022).

138. Faure, G., Makarova, K. S. & Koonin, E. V. CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. *J. Mol. Biol.* **431**, 3–20 (2019).

139. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

140. Klompe, S. E. *et al.* Evolutionary and mechanistic diversity of Type I-F CRISPR-associated transposons. *Mol. Cell* **82**, 616-628.e5 (2022).

141. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl. Acad. Sci.* **114**, (2017).

142. Gao, N. J. *et al.* Functional and Proteomic Analysis of Streptococcus pyogenes Virulence Upon Loss of Its Native Cas9 Nuclease. *Front. Microbiol.* **10**, 1967 (2019).

143. Shabbir, M. A. B. *et al.* The Involvement of the Cas9 Gene in Virulence of Campylobacter jejuni. *Front. Cell. Infect. Microbiol.* **8**, 285 (2018).

144. Spencer, B. L. *et al.* Cas9 Contributes to Group B Streptococcal Colonization and Disease. *Front. Microbiol.* **10**, 1930 (2019).

145. Wang, J. Y. *et al.* Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex. *Nat. Commun.* **12**, 2571 (2021).

146. Mohr, G. *et al.* A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition. *Mol. Cell* **72**, 700-714.e8 (2018).

147. González-Delgado, A., Mestre, M. R., Martínez-Abarca, F. & Toro, N. Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR–Cas system in Vibrio vulnificus. *Nucleic Acids Res.* **47**, 10202–10211 (2019).

148. Silas, S. *et al.* Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016).

149. Wang, J. Y. *et al.* Genome expansion by a CRISPR trimmer-integrase. *Nature* **618**, 855–861 (2023).

150. Drabavicius, G. *et al.* DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep.* **19**, e45543 (2018).

151. Altae-Tran, H. *et al.* Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382**, eadi1910 (2023).

152. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569–5576 (2012).

153. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).

154. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535–538 (2015).

155. Ivančić-Baće, I., Cass, S. D., Wearne, S. J. & Bolt, E. L. Different genome stability proteins underpin primed and naïve adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic Acids Res.* **43**, 10821–10830 (2015).

156. Killelea, T. *et al.* Cas1–Cas2 physically and functionally interacts with DnaK to modulate CRISPR Adaptation. *Nucleic Acids Res.* **51**, 6914–6926 (2023).

157. Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Curr. Biol.* **31**, 3515-3524.e6 (2021).

158. Budhathoki, J. B. *et al.* Real-time observation of CRISPR spacer acquisition by Cas1–Cas2 integrase. *Nat. Struct. Mol. Biol.* **27**, 489–499 (2020).

159. Malone, L. M., Hampton, H. G., Morgan, X. C. & Fineran, P. C. Type I CRISPR-Cas provides robust immunity but incomplete attenuation of phage-induced cellular stress. *Nucleic Acids Res.* **50**, 160–174 (2022).

160. Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *Proc. Natl. Acad. Sci.* **114**, (2017).

161. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).

162. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544**, 101–104 (2017).

163. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).

164. Swarts, D. C., Mosterd, C., Van Passel, M. W. J. & Brouns, S. J. J. CRISPR Interference Directs Strand Specific Spacer Acquisition. *PLoS ONE* **7**, e35888 (2012).

165. Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. *Trends Biochem. Sci.* **47**, 464–476 (2022).

166. Hu, C. *et al.* Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* **598**, 515–520 (2021).

167. Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141–145 (2020).

168. Shiriaeva, A. A. *et al.* Host nucleases generate prespacers for primed adaptation in the *E. coli* type I-E CRISPR-Cas system. *Sci. Adv.* **8**, eabn8650 (2022).

169. Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23**, 876–883 (2016).

170. Yoganand, K. N. R., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res.* **45**, 367–381 (2017).

171. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* **62**, 824–833 (2016).

172. Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).

173. Santiago-Frangos, A. *et al.* Structure reveals why genome folding is necessary for site-specific integration of foreign DNA into CRISPR arrays. *Nat. Struct. Mol. Biol.* **30**, 1675–1685 (2023).

174. McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol. Cell* **64**, 616–623 (2016).

175. Sasnauskas, G. & Siksnys, V. CRISPR adaptation from a structural perspective. *Curr. Opin. Struct. Biol.* **65**, 17–25 (2020).

176. Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J. & Mojica, F. J. M. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli . *RNA Biol.* **10**, 792–802 (2013).

177. Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res.* **43**, 1749–1758 (2015).

178. Nussenzweig, P. M. & Marraffini, L. A. Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria. *Annu. Rev. Genet.* **54**, 93–120 (2020).

179. Leenay, R. T. & Beisel, C. L. Deciphering, Communicating, and Engineering the CRISPR PAM. *J. Mol. Biol.* **429**, 177–191 (2017).

180. Dhingra, Y., Suresh, S. K., Juneja, P. & Sashital, D. G. PAM binding ensures orientational integration during Cas4-Cas1-Cas2-mediated CRISPR adaptation. *Mol. Cell* **82**, 4353-4367.e6 (2022).

181. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109**, (2012).

182. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).

183. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).

184. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823–826 (2013).

185. Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 230–232 (2013).

186. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl. Acad. Sci.* **110**, 15644–15649 (2013).

187. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154**, 1380–1389 (2013).

188. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).

189. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

190. Schwank, G. *et al.* Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. *Cell Stem Cell* **13**, 653–658 (2013).

191. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80–84 (2014).

192. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**, 84–87 (2014).

193. Chen, C., Fenk, L. A. & De Bono, M. Efficient genome editing in Caenorhabditis elegans by CRISPR-targeted homologous recombination. *Nucleic Acids Res.* **41**, e193–e193 (2013).

194. Katic, I. & Großhans, H. Targeted Heritable Mutation and Gene Conversion by Cas9-CRISPR in *Caenorhabditis elegans*. *Genetics* **195**, 1173–1176 (2013).

195. Friedland, A. E. *et al.* Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nat. Methods* **10**, 741–743 (2013).

196. Sebo, Z. L., Lee, H. B., Peng, Y. & Guo, Y. A simplified and efficient germline-specific CRISPR/Cas9 system for Drosophila genomic engineering. *Fly (Austin)* **8**, 52–57 (2014).

197. Gratz, S. J., Wildonger, J., Harrison, M. M. & O'Connor-Giles, K. M. CRISPR/Cas9-mediated genome engineering and the promise of designer flies on demand. *Fly (Austin)* **7**, 249–255 (2013).

198. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).

199. Jao, L.-E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci.* **110**, 13904–13909 (2013).

200. Xiao, A. *et al.* Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. *Nucleic Acids Res.* **41**, e141–e141 (2013).

201. Blitz, I. L., Biesinger, J., Xie, X. & Cho, K. W. Y. Biallelic genome modification in F0 Xenopus tropicalis embryos using the CRISPR/Cas system. *genesis* **51**, 827–834 (2013).

202. Nakayama, T. *et al.* Simple and efficient CRISPR/Cas9-mediated targeted mutagenesis in Xenopus tropicalis. *genesis* **51**, 835–843 (2013).

203. Wang, H. *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* **153**, 910–918 (2013).

204. Li, D. *et al.* Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 681–683 (2013).

205. Li, W., Teng, F., Li, T. & Zhou, Q. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 684–686 (2013).

206. Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J. D. G. & Kamoun, S. Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 691–693 (2013).

207. Jiang, W. *et al.* Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Res.* **41**, e188–e188 (2013).

208. Shan, Q. *et al.* Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 686–688 (2013).

209. Li, J.-F. *et al.* Multiplex and homologous recombination–mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat. Biotechnol.* **31**, 688–691 (2013).

210. DiCarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).

211. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

212. Koonin, E. V., Gootenberg, J. S. & Abudayyeh, O. O. Discovery of Diverse CRISPR-Cas Systems and Expansion of the Genome Engineering Toolbox. *Biochemistry* **62**, 3465–3487 (2023).

213. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).

214. Karvelis, T., Young, J. K. & Siksnys, V. A pipeline for characterization of novel Cas9 orthologs. in *Methods in Enzymology* vol. 616 219–240 (Elsevier, 2019).

215. Gasiunas, G. *et al.* A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat. Commun.* **11**, 5512 (2020).

216. Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).

217. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).

218. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).

219. Schmid-Burgk, J. L. *et al.* Highly Parallel Profiling of Cas9 Variant Specificity. *Mol. Cell* **78**, 794-800.e8 (2020).

220. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).

221. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).

222. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

223. Huang, T. P. *et al.* High-throughput continuous evolution of compact Cas9 variants targeting single-nucleotide-pyrimidine PAMs. *Nat. Biotechnol.* **41**, 96–107 (2023).

224. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

225. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

226. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79**, 181–211 (2010).

227. Yeh, C. D., Richardson, C. D. & Corn, J. E. Advances in genome editing through control of DNA repair pathways. *Nat. Cell Biol.* **21**, 1468–1478 (2019).

228. Nambiar, T. S. *et al.* Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nat. Commun.* **10**, 3395 (2019).

229. Charpentier, M. *et al.* CtIP fusion to Cas9 enhances transgene integration by homology-dependent repair. *Nat. Commun.* **9**, 1133 (2018).

230. Canny, M. D. *et al.* Inhibition of 53BP1 favors homology-dependent DNA repair and increases CRISPR-Cas9 genome-editing efficiency. *Nat. Biotechnol.* **36**, 95–102 (2018).

231. Pinder, J., Salsman, J. & Dellaire, G. Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Res.* **43**, 9379–9392 (2015).

232. Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat. Biotechnol.* **33**, 538–542 (2015).

233. Yu, C. *et al.* Small molecules enhance CRISPR genome editing in pluripotent stem cells. *Cell Stem Cell* **16**, 142–147 (2015).

234. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

235. Robert, F., Barbeau, M., Éthier, S., Dostie, J. & Pelletier, J. Pharmacological inhibition of DNA-PK stimulates Cas9-mediated genome editing. *Genome Med.* **7**, 93 (2015).

236. Wang, C. *et al.* Microbial single-strand annealing proteins enable CRISPR gene-editing tools with improved knock-in efficiencies and reduced off-target effects. *Nucleic Acids Res.* **49**, e36 (2021).

237. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).

238. Cullot, G. *et al.* CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**, 1136 (2019).

239. 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).

240. Ihry, R. J. *et al.* p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018).

241. 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).

242. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).

243. Renaud, J.-B. *et al.* Improved Genome Editing Efficiency and Flexibility Using Modified Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases. *Cell Rep.* **14**, 2263–2272 (2016).

244. Savic, N. *et al.* Covalent linkage of the DNA repair template to the CRISPR-Cas9 nuclease enhances homology-directed repair. *eLife* **7**, e33761 (2018).

245. Aird, E. J., Lovendahl, K. N., St. Martin, A., Harris, R. S. & Gordon, W. R. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Commun. Biol.* **1**, 1–6 (2018).

246. Carlson-Stevermer, J. *et al.* Assembly of CRISPR ribonucleoproteins with biotinylated oligonucleotides via an RNA aptamer for precise gene editing. *Nat. Commun.* **8**, 1711 (2017).

247. 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**, e04766 (2014).

248. Nishiyama, J., Mikuni, T. & Yasuda, R. Virus-Mediated Genome Editing via Homology-Directed Repair in Mitotic and Postmitotic Cells in Mammalian Brain. *Neuron* **96**, 755-768.e5 (2017).

249. Han, W. *et al.* Efficient precise integration of large DNA sequences with 3′-overhang dsDNA donors using CRISPR/Cas9. *Proc. Natl. Acad. Sci.* **120**, e2221127120 (2023).

250. Lee, K. *et al.* Synthetically modified guide RNA and donor DNA are a versatile platform for CRISPR-Cas9 engineering. *eLife* **6**, e25312 (2017).

251. Yu, Y. *et al.* An efficient gene knock-in strategy using 5'-modified double-stranded DNA donors with short homology arms. *Nat. Chem. Biol.* **16**, 387–390 (2020).

252. Nguyen, D. N. *et al.* Polymer-stabilized Cas9 nanoparticles and modified repair templates increase genome editing efficiency. *Nat. Biotechnol.* **38**, 44–49 (2020).

253. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* **23**, 11–22 (2021).

254. Okada, M., Kanamori, M., Someya, K., Nakatsukasa, H. & Yoshimura, A. Stabilization of Foxp3 expression by CRISPR-dCas9-based epigenome editing in mouse primary T cells. *Epigenetics Chromatin* **10**, 24 (2017).

255. Wang, H. *et al.* Epigenetic Targeting of Granulin in Hepatoma Cells by Synthetic CRISPR dCas9 Epi-suppressors. *Mol. Ther. - Nucleic Acids* **11**, 23–33 (2018).

256. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9–histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).

257. Nuñez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* **184**, 2503-2519.e17 (2021).

258. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173–1183 (2013).

259. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**, 1116–1121 (2013).

260. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nat. Methods* **10**, 973–976 (2013).

261. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nat. Methods* **10**, 977–979 (2013).

262. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* **12**, 326–328 (2015).

263. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

264. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).

265. Farzadfard, F., Perli, S. D. & Lu, T. K. Tunable and Multifunctional Eukaryotic Transcription Factors Based on CRISPR/Cas. *ACS Synth. Biol.* **2**, 604–613 (2013).

266. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).

267. Przybyla, L. & Gilbert, L. A. A new era in functional genomics screens. *Nat. Rev. Genet.* **23**, 89–103 (2022).

268. Kampmann, M. CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine. *ACS Chem. Biol.* **13**, 406–416 (2018).

269. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, e19760 (2016).

270. Mandegar, M. A. *et al.* CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell* **18**, 541–553 (2016).

271. Dräger, N. M. *et al.* A CRISPRi/a platform in human iPSC-derived microglia uncovers regulators of disease states. *Nat. Neurosci.* **25**, 1149–1162 (2022).

272. Tian, R. *et al.* Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nat. Neurosci.* **24**, 1020–1034 (2021).

273. Jensen, T. I. *et al.* Targeted regulation of transcription in primary cells using CRISPRa and CRISPRi. *Genome Res.* **31**, 2120–2130 (2021).

274. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, eaah7111 (2017).

275. Jost, M. *et al.* Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. *Mol. Cell* **68**, 210-223.e6 (2017).

276. Kampmann, M. Elucidating drug targets and mechanisms of action by genetic screens in mammalian cells. *Chem. Commun.* **53**, 7162–7167 (2017).

277. Acosta-Alvear, D. *et al.* Paradoxical resistance of multiple myeloma to proteasome inhibitors by decreased levels of 19S proteasomal subunits. *eLife* **4**, e08153 (2015).

278. Kruth, K. A. *et al.* Suppression of B-cell development genes is key to glucocorticoid efficacy in treatment of acute lymphoblastic leukemia. *Blood* **129**, 3000–3008 (2017).

279. Jost, M. & Weissman, J. S. CRISPR Approaches to Small Molecule Target Identification. *ACS Chem. Biol.* **13**, 366–375 (2018).

280. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

281. Zalatan, J. G. *et al.* Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell* **160**, 339–350 (2015).

282. Bassik, M. C. *et al.* A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility. *Cell* **152**, 909–922 (2013).

283. Du, D. *et al.* Genetic interaction mapping in mammalian cells using CRISPR interference. *Nat. Methods* **14**, 577–580 (2017).

284. Shen, J. P. *et al.* Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* **14**, 573–576 (2017).

285. Han, K. *et al.* Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.* **35**, 463–474 (2017).

286. Replogle, J. M. *et al.* Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors. *eLife* **11**, e81856 (2022).

287. Replogle, J. M. *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).

288. Kampmann, M., Bassik, M. C. & Weissman, J. S. Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. *Proc. Natl. Acad. Sci.* **110**, E2317–E2326 (2013).

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

290. Yan, J. *et al.* Improving prime editing with an endogenous small RNA-binding protein. *Nature* **628**, 639–647 (2024).

291. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

292. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

293. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).

294. Yao, D. *et al.* Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nat. Biotechnol.* 1–14 (2023) doi:10.1038/s41587-023-01964-9.

295. Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, eaaz6063 (2020).

296. Ursu, O. *et al.* Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat. Biotechnol.* **40**, 896–905 (2022).

297. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

298. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).

299. Chardon, F. M. *et al.* Multiplex, single-cell CRISPRa screening for cell type specific regulatory elements. 2023.03.28.534017 Preprint at https://doi.org/10.1101/2023.03.28.534017 (2024).

300. Jost, M. *et al.* Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* **38**, 355–364 (2020).

301. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).

302. Chavez, A. *et al.* Comparison of Cas9 activators in multiple species. *Nat. Methods* **13**, 563–567 (2016).

303. Rousset, F. *et al.* Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLOS Genet.* **14**, e1007749 (2018).

304. Cui, L. *et al.* A CRISPRi screen in E. coli reveals sequence-specific toxicity of dCas9. *Nat. Commun.* **9**, 1912 (2018).

305. Yao, L., Cengic, I., Anfelt, J. & Hudson, E. P. Multiple Gene Repression in Cyanobacteria Using CRISPRi. *ACS Synth. Biol.* **5**, 207–212 (2016).

306. Stachler, A.-E., Schwarz, T. S., Schreiber, S. & Marchfelder, A. CRISPRi as an efficient tool for gene repression in archaea. *Methods* **172**, 76–85 (2020).

307. Depardieu, F. & Bikard, D. Gene silencing with CRISPRi in bacteria and optimization of dCas9 expression levels. *Methods* **172**, 61–75 (2020).

308. Ho, H., Fang, J. R., Cheung, J. & Wang, H. H. Programmable CRISPR-Cas transcriptional activation in bacteria. *Mol. Syst. Biol.* **16**, e9427 (2020).

309. Zetsche, B. *et al.* Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2017).

310. Port, F., Starostecka, M. & Boutros, M. Multiplexed conditional genome editing with Cas12a in *Drosophila*. *Proc. Natl. Acad. Sci.* **117**, 22890–22899 (2020).

311. Hsiung, C. C.-S. *et al.* Engineered CRISPR-Cas12a for higher-order combinatorial chromatin perturbations. *Nat. Biotechnol.* 1–15 (2024) doi:10.1038/s41587-024-02224-0.

312. Campa, C. C., Weisbach, N. R., Santinha, A. J., Incarnato, D. & Platt, R. J. Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. *Nat. Methods* **16**, 887–893 (2019).

313. Świat, M. A. *et al.* FnCpf1: a novel and efficient genome editing tool for Saccharomyces cerevisiae. *Nucleic Acids Res.* **45**, 12585–12598 (2017).

314. Ao, X. *et al.* A Multiplex Genome Editing Method for Escherichia coli Based on CRISPR-Cas12a. *Front. Microbiol.* **9**, (2018).

315. Deaner, M., Mejia, J. & Alper, H. S. Enabling Graded and Large-Scale Multiplex of Desired Genes Using a Dual-Mode dCas9 Activator in Saccharomyces cerevisiae. *ACS Synth. Biol.* **6**, 1931–1943 (2017).

316. Tak, Y. E. *et al.* Inducible and multiplex gene regulation using CRISPR–Cpf1-based transcription factors. *Nat. Methods* **14**, 1163–1166 (2017).

317. Yang, Z., Edwards, H. & Xu, P. CRISPR-Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in *Yarrowia lipolytica*. *Metab. Eng. Commun.* **10**, e00112 (2020).

318. Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).

319. Karvelis, T. *et al.* PAM recognition by miniature CRISPR–Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic Acids Res.* **48**, 5016–5023 (2020).

320. Wu, Z. *et al.* Programmed genome editing by a miniature CRISPR-Cas12f nuclease. *Nat. Chem. Biol.* **17**, 1132–1138 (2021).

321. Kim, D. Y. *et al.* Efficient CRISPR editing with a hypercompact Cas12f1 and engineered guide RNAs delivered by adeno-associated virus. *Nat. Biotechnol.* **40**, 94–102 (2022).

322. Xu, X. *et al.* Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol. Cell* **81**, 4333-4345.e4 (2021).

323. Wang, Y. *et al.* Guide RNA engineering enables efficient CRISPR editing with a miniature Syntrophomonas palmitatica Cas12f1 nuclease. *Cell Rep.* **40**, 111418 (2022).

324. McGaw, C. *et al.* Engineered Cas12i2 is a versatile high-efficiency platform for therapeutic genome editing. *Nat. Commun.* **13**, 2833 (2022).

325. Zhang, H., Li, Z., Xiao, R. & Chang, L. Mechanisms for target recognition and cleavage by the Cas12i RNA-guided endonuclease. *Nat. Struct. Mol. Biol.* **27**, 1069–1076 (2020).

326. Pausch, P. *et al.* DNA interference states of the hypercompact CRISPR–CasΦ effector. *Nat. Struct. Mol. Biol.* **28**, 652–661 (2021).

327. East-Seletsky, A. *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270–273 (2016).

328. Abudayyeh, O. O. *et al.* RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).

329. Singsuksawat, E. *et al.* Potent programmable antiviral against dengue virus in primary human cells by Cas13b RNP with short spacer and delivery by VLP. *Mol. Ther. - Methods Clin. Dev.* **21**, 729–740 (2021).

330. Fareh, M. *et al.* Reprogrammed CRISPR-Cas13b suppresses SARS-CoV-2 replication and circumvents its mutational escape through mismatch tolerance. *Nat. Commun.* **12**, 4270 (2021).

331. Yan, W. X. *et al.* Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Mol. Cell* **70**, 327-339.e5 (2018).

332. Aman, R. *et al.* RNA virus interference via CRISPR/Cas13a system in plants. *Genome Biol.* **19**, 1 (2018).

333. Kushawah, G. *et al.* CRISPR-Cas13d Induces Efficient mRNA Knockdown in Animal Embryos. *Dev. Cell* **54**, 805-817.e7 (2020).

334. Kannan, S. *et al.* Compact RNA editors with small Cas13 proteins. *Nat. Biotechnol.* **40**, 194–197 (2022).

335. Adler, B. A. *et al.* Broad-spectrum CRISPR-Cas13a enables efficient phage genome editing. *Nat. Microbiol.* **7**, 1967–1979 (2022).

336. Wessels, H.-H. *et al.* Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. *Nat. Methods* **20**, 86–94 (2023).

337. Wessels, H.-H. *et al.* Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.* **38**, 722–727 (2020).

338. Li, S. *et al.* Screening for functional circular RNAs using the CRISPR–Cas13 system. *Nat. Methods* **18**, 51–59 (2021).

339. Li, S., Wu, H. & Chen, L.-L. Screening circular RNAs with functional potential using the RfxCas13d/BSJ-gRNA system. *Nat. Protoc.* **17**, 2085–2107 (2022).

340. Özcan, A. *et al.* Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. *Nature* **597**, 720–725 (2021).

341. Wang, H. *et al.* CRISPR-mediated live imaging of genome editing and transcription. *Science* **365**, 1301–1305 (2019).

342. Yang, L.-Z. *et al.* Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems. *Mol. Cell* **76**, 981-997.e7 (2019).

343. Wilson, C., Chen, P. J., Miao, Z. & Liu, D. R. Programmable m6A modification of cellular RNAs with a Cas13-directed methyltransferase. *Nat. Biotechnol.* **38**, 1431–1440 (2020).

344. Cox, D. B. T. *et al.* RNA editing with CRISPR-Cas13. *Science* **358**, 1019–1027 (2017).

345. Konermann, S. *et al.* Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* **173**, 665-676.e14 (2018).

346. Van Beljouw, S. P. B. *et al.* The gRAMP CRISPR-Cas effector is an RNA endonuclease complexed with a caspase-like peptidase. *Science* **373**, 1349–1353 (2021).

347. Kato, K. *et al.* Structure and engineering of the type III-E CRISPR-Cas7-11 effector complex. *Cell* **185**, 2324-2337.e16 (2022).

348. Mahler, M., Costa, A. R., Beljouw, S. P. B. van, Fineran, P. C. & Brouns, S. J. J. Approaches for bacteriophage genome engineering. *Trends Biotechnol.* **41**, 669–685 (2023).

349. Ramirez-Chamorro, L., Boulanger, P. & Rossier, O. Strategies for Bacteriophage T5 Mutagenesis: Expanding the Toolbox for Phage Genome Engineering. *Front. Microbiol.* **12**, 667332 (2021).

350. Hupfeld, M. *et al.* A functional type II-A CRISPR-Cas system from Listeria enables efficient genome editing of large non-integrating bacteriophage. *Nucleic Acids Res.* **46**, 6920–6933 (2018).

351. 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).

352. 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).

353. 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).

354. Mayo-Muñoz, D. *et al.* Anti-CRISPR-Based and CRISPR-Based Genome Editing of Sulfolobus islandicus Rod-Shaped Virus 2. *Viruses* **10**, 695 (2018).

355. Pyenson, N. C., Gayvert, K., Varble, A., Elemento, O. & Marraffini, L. A. Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. *Cell Host Microbe* **22**, 343-353.e3 (2017).

356. Guan, J. *et al.* RNA targeting with CRISPR-Cas13a facilitates bacteriophage genome engineering. 2022.02.14.480438 Preprint at https://doi.org/10.1101/2022.02.14.480438 (2022).

357. Lemay, M.-L., Tremblay, D. M. & Moineau, S. Genome Engineering of Virulent Lactococcal Phages Using CRISPR-Cas9. *ACS Synth. Biol.* **6**, 1351–1358 (2017).

358. Shen, J., Zhou, J., Chen, G.-Q. & Xiu, Z.-L. Efficient Genome Engineering of a Virulent Klebsiella Bacteriophage Using CRISPR-Cas9. *J. Virol.* **92**, 10.1128/jvi.00534-18 (2018).

359. Schilling, T., Dietrich, S., Hoppert, M. & Hertel, R. A CRISPR-Cas9-Based Toolkit for Fast and Precise In Vivo Genetic Engineering of Bacillus subtilis Phages. *Viruses* **10**, 241 (2018).

360. Tao, P., Wu, X., Tang, W.-C., Zhu, J. & Rao, V. Engineering of Bacteriophage T4 Genome Using CRISPR-Cas9. *ACS Synth. Biol.* **6**, 1952–1961 (2017).

361. Qin, W., Cho, K. F., Cavanagh, P. E. & Ting, A. Y. Deciphering molecular interactions by proximity labeling. *Nat. Methods* **18**, 133–143 (2021).

362. Kang, M.-G. & Rhee, H.-W. Molecular Spatiomics by Proximity Labeling. *Acc. Chem. Res.* **55**, 1411–1422 (2022).

363. Guo, J. *et al.* The development of proximity labeling technology and its applications in mammals, plants, and microorganisms. *Cell Commun. Signal.* **21**, 269 (2023).

364. Gao, X. D., Rodríguez, T. C. & Sontheimer, E. J. Chapter Sixteen - Adapting dCas9-APEX2 for subnuclear proteomic profiling. in *Methods in Enzymology* (ed. Bailey, S.) vol. 616 365–383 (Academic Press, 2019).

365. Escobar, T. M. *et al.* Active and Repressed Chromatin Domains Exhibit Distinct Nucleosome Segregation during DNA Replication. *Cell* **179**, 953-963.e11 (2019).

366. Tsui, C. *et al.* dCas9-targeted locus-specific protein isolation method identifies histone gene regulators. *Proc. Natl. Acad. Sci.* **115**, E2734–E2741 (2018).

367. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170**, 1028-1043.e19 (2017).

368. Fujita, T. & Fujii, H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem. Biophys. Res. Commun.* **439**, 132–136 (2013).

369. Fujita, T., Yuno, M. & Fujii, H. Efficient sequence-specific isolation of DNA fragments and chromatin by in vitro enChIP technology using recombinant CRISPR ribonucleoproteins. *Genes Cells* **21**, 370–377 (2016).

370. Schmidtmann, E., Anton, T., Rombaut, P., Herzog, F. & Leonhardt, H. Determination of local chromatin composition by CasID. *Nucleus* **7**, 476–484 (2016).

371. Li, P. *et al.* Nuclear localization of Desmoplakin and its involvement in telomere maintenance. *Int. J. Biol. Sci.* **15**, 2350–2362 (2019).

372. Gao, X. D. *et al.* C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9–APEX2. *Nat. Methods* **15**, 433–436 (2018).

373. Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9–APEX-mediated proximity labeling. *Nat. Methods* **15**, 437–439 (2018).

374. Torres, M. & Kramer, A. Proteomic Insights into Circadian Transcription Regulation: Novel E-box Interactors Revealed by Proximity Labeling. 2024.04.18.590107 Preprint at https://doi.org/10.1101/2024.04.18.590107 (2024).

375. Han, S. *et al.* RNA-protein interaction mapping via MS2- or Cas13-based APEX targeting. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22068–22079 (2020).

376. Lin, X., Fonseca, M. A. S., Breunig, J. J., Corona, R. I. & Lawrenson, K. In vivo discovery of RNA proximal proteins via proximity-dependent biotinylation. *RNA Biol.* **18**, 2203–2217 (2021).

377. Yi, W. *et al.* CRISPR-assisted detection of RNA–protein interactions in living cells. *Nat. Methods* **17**, 685–688 (2020).

378. Zhang, Z. *et al.* Capturing RNA–protein interaction via CRUIS. *Nucleic Acids Res.* **48**, e52 (2020).

379. Li, Y. *et al.* CBRPP: a new RNA-centric method to study RNA–protein interactions. *RNA Biol.* **18**, 1608–1621 (2021).

380. Qiu, W. *et al.* Determination of local chromatin interactions using a combined CRISPR and peroxidase APEX2 system. *Nucleic Acids Res.* **47**, e52 (2019).

381. Gräwe, C., Stelloo, S., Hout, F. A. H. van & Vermeulen, M. RNA-Centric Methods: Toward the Interactome of Specific RNA Transcripts. *Trends Biotechnol.* **39**, 890–900 (2021).

382. Lu, M. & Tokuyasu, T. A. CRISPR-Cas13-Based RNA-Interacting Protein Detection in Living Cells. *Biochemistry* **59**, 1791–1792 (2020).

383. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

384. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).

385. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, aag0511 (2016).

386. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).

387. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).

388. Zafar, H., Lin, C. & Bar-Joseph, Z. Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. *Nat. Commun.* **11**, 3055 (2020).

389. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

390. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).

391. Saunders, L. M. *et al.* Embryo-scale reverse genetics at single-cell resolution. *Nature* **623**, 782–791 (2023).

392. Guernet, A. *et al.* CRISPR-Barcoding for Intratumor Genetic Heterogeneity Modeling and Functional Analysis of Oncogenic Driver Mutations. *Mol. Cell* **63**, 526–538 (2016).

393. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

394. Raj, B., Gagnon, J. A. & Schier, A. F. Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR–Cas9 barcodes by scGESTALT. *Nat. Protoc.* **13**, 2685–2713 (2018).

395. Junker, J. P. *et al.* Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. 056499 Preprint at https://doi.org/10.1101/056499 (2017).

396. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synth. Biol.* **6**, 936–942 (2017).

397. Pan, X., Li, H., Putta, P. & Zhang, X. LinRace: cell division history reconstruction of single cells using paired lineage barcode and gene expression data. *Nat. Commun.* **14**, 8388 (2023).

398. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).

399. Shlyakhtina, Y., Bloechl, B. & Portal, M. M. BdLT-Seq as a barcode decay-based method to unravel lineage-linked transcriptome plasticity. *Nat. Commun.* **14**, 1085 (2023).

400. Baron, C. S. & van Oudenaarden, A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* **20**, 753–765 (2019).

401. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).

402. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).

403. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).

404. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).

405. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).

406. Munck, C., Sheth, R. U., Freedberg, D. E. & Wang, H. H. Real-time capture of horizontal gene transfers from gut microbiota by engineered CRISPR-Cas acquisition. 492751 Preprint at https://doi.org/10.1101/492751 (2018).

407. Munck, C., Sheth, R. U., Freedberg, D. E. & Wang, H. H. Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. *Nat. Commun.* **11**, 95 (2020).

408. Yim, S. S. *et al.* Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.* **17**, 246–253 (2021).

409. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* **19**, 718–732 (2018).

410. Bhattarai-Kline, S. *et al.* Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* **608**, 217–225 (2022).

411. Lear, S. K., Lopez, S. C., González-Delgado, A., Bhattarai-Kline, S. & Shipman, S. L. Temporally resolved transcriptional recording in E. coli DNA using a Retro-Cascorder. *Nat. Protoc.* **18**, 1866–1892 (2023).

412. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).

413. Schmidt, F. *et al.* Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science* **376**, eabm6038 (2022).

414. Tanna, T., Schmidt, F., Cherepkova, M. Y., Okoniewski, M. & Platt, R. J. Recording transcriptional histories using Record-seq. *Nat. Protoc.* **15**, 513–539 (2020).

415. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

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

417. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).

418. Molla, K. A. & Yang, Y. CRISPR/Cas-Mediated Base Editing: Technical Considerations and Practical Applications. *Trends Biotechnol.* **37**, 1121–1142 (2019).

419. Yang, B., Yang, L. & Chen, J. Development and Application of Base Editors. *CRISPR J.* **2**, 91–104 (2019).

420. Hess, G. T., Tycko, J., Yao, D. & Bassik, M. C. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol. Cell* **68**, 26–43 (2017).

421. 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).

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

423. Bigelyte, G. *et al.* Miniature type V-F CRISPR-Cas nucleases enable targeted DNA modification in cells. *Nat. Commun.* **12**, 6191 (2021).

424. Levy, J. M. *et al.* Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat. Biomed. Eng.* **4**, 97–110 (2020).

425. Porto, E. M. & Komor, A. C. In the business of base editors: Evolution from bench to bedside. *PLOS Biol.* **21**, e3002071 (2023).

426. Chen, L. *et al.* Engineering a precise adenine base editor with minimal bystander editing. *Nat. Chem. Biol.* **19**, 101–110 (2023).

427. Yan, J., Cirincione, A. & Adamson, B. Prime Editing: Precision Genome Editing by Reverse Transcription. *Mol. Cell* **77**, 210–212 (2020).

428. Lin, Q. *et al.* Prime genome editing in rice and wheat. *Nat. Biotechnol.* **38**, 582–585 (2020).

429. Liu, Y. *et al.* Efficient generation of mouse models with the prime editing system. *Cell Discov.* **6**, 1–4 (2020).

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

431. Aida, T. *et al.* Prime editing primarily induces undesired outcomes in mice. 2020.08.06.239723 Preprint at https://doi.org/10.1101/2020.08.06.239723 (2020).

432. Huang, M. E. *et al.* C-to-G editing generates double-strand breaks causing deletion, transversion and translocation. *Nat. Cell Biol.* **26**, 294–304 (2024).

433. Shimamoto, T., Hsu, M. Y., Inouye, S. & Inouye, M. Reverse transcriptases from bacterial retrons require specific secondary structures at the 5'-end of the template for the cDNA priming reaction. *J. Biol. Chem.* **268**, 2684–2692 (1993).

434. 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).

435. 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).

436. Mao, J. R., Shimada, M., Inouye, S. & Inouye, M. Gene regulation by antisense DNA produced in vivo. *J. Biol. Chem.* **270**, 19684–19687 (1995).

437. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).

438. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).

439. 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).

440. 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).

441. Buffington, J. D. *et al.* Discovery and Engineering of Retrons for Precise Genome Editing.

442. Khan, A. G. *et al.* An experimental census of retrons for DNA production and genome editing. *bioRxiv* 2024.01.25.577267 (2024) doi:10.1101/2024.01.25.577267.

443. Crawford, K. D., Khan, A. G., Lopez, S. C., Goodarzi, H. & Shipman, S. L. High Throughput Variant Libraries and Machine Learning Yield Design Rules for Retron Gene Editors. 2024.07.08.602561 Preprint at https://doi.org/10.1101/2024.07.08.602561 (2024).

444. 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. *Nat. Chem. Biol.* 1–11 (2024) doi:10.1038/s41589-024-01665-7.

445. Fishman, C. B. *et al.* Continuous Multiplexed Phage Genome Editing Using Recombitrons. Preprint at https://doi.org/10.1101/2023.03.24.534024 (2023).

446. Liu, W. *et al.* Retron-mediated multiplex genome editing and continuous evolution in *Escherichia coli*. *Nucleic Acids Res.* **51**, 8293–8307 (2023).

447. Schubert, M. G. *et al.* High throughput functional variant screens via in-vivo production of single-stranded DNA. Preprint at https://doi.org/10.1101/2020.03.05.975441 (2020).

448. 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).

449. Vibhute, M. A. *et al.* Intracellular Expression of a Fluorogenic DNA Aptamer Using Retron Eco2. 2024.05.21.595248 Preprint at https://doi.org/10.1101/2024.05.21.595248 (2024).

450. Kaur, N. & Pati, P. K. Retron Library Recombineering: Next Powerful Tool for Genome Editing after CRISPR/Cas. *ACS Synth. Biol.* **13**, 1019–1025 (2024).

451. Liu, J. *et al.* Generation of DNAzyme in Bacterial Cells by a Bacterial Retron System. *ACS Synth. Biol.* **13**, 300–309 (2024).

452. Ni, Y. *et al.* Reducing competition between msd and genomic DNA significantly improved the editing efficiency of the retron editing system. 2024.06.04.597346 Preprint at https://doi.org/10.1101/2024.06.04.597346 (2024).

453. Ellington, A. J. & Reisch, C. R. Efficient and iterative retron-mediated in vivo recombineering in Escherichia coli. *Synth. Biol.* **7**, ysac007 (2022).

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

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

456. Koonin, E. V. & Makarova, K. S. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).

457. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).

458. Faure, G. *et al.* CRISPR-Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).

459. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR–Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).

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

461. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).

462. Lampe, G. D. *et al.* Targeted DNA integration in human cells without double-strand breaks using CRISPR-associated transposases. *Nat. Biotechnol.* **42**, 87–98 (2024).

463. Vo, P. L. H. *et al.* CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.* **39**, 480–489 (2021).

464. George, J. T. *et al.* Mechanism of target site selection by type V-K CRISPR-associated transposases. *Science* **382**, eadj8543 (2023).

465. Walker, M. W. G., Klompe, S. E., Zhang, D. J. & Sternberg, S. H. Novel molecular requirements for CRISPR RNA-guided transposition. *Nucleic Acids Res.* **51**, 4519–4535 (2023).

466. Gelsinger, D. R. *et al.* Bacterial genome engineering using CRISPR-associated transposases. *Nat. Protoc.* **19**, 752–790 (2024).

467. Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernández, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. *Nature* **577**, 271–274 (2020).

468. Hsieh, S.-C. & Peters, J. E. Discovery and characterization of novel type I-D CRISPR-guided transposons identified among diverse Tn7-like elements in cyanobacteria. *Nucleic Acids Res.* **51**, 765–782 (2023).

469. Rybarski, J. R., Hu, K., Hill, A. M., Wilke, C. O. & Finkelstein, I. J. Metagenomic discovery of CRISPR-associated transposons. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2112279118 (2021).

470. Yang, S. *et al.* Orthogonal CRISPR-associated transposases for parallel and multiplexed chromosomal integration. *Nucleic Acids Res.* **49**, 10192–10202 (2021).

471. Rubin, B. E. *et al.* Species- and site-specific genome editing in complex bacterial communities. *Nat. Microbiol.* **7**, 34–47 (2022).

472. Chen, W. *et al.* Targeted genetic screening in bacteria with a Cas12k-guided transposase. *Cell Rep.* **36**, 109635 (2021).

473. Voltaire & Havens, G. R. *Candide: Ou, L'optimisme*. (Holt, Rinehart and Winston, New York, 1969).

474. Lopez, S. C., Lee, Y., Zhang, K. & Shipman, S. L. SspA is a transcriptional regulator of CRISPR adaptation in E. coli. 2024.05.24.595836 Preprint at https://doi.org/10.1101/2024.05.24.595836 (2024).

475. Luo, D. & Saltzman, W. M. Synthetic DNA delivery systems. *Nat. Biotechnol.* **18**, 33–37 (2000).

476. 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).

477. Lampson, B. C., Inouye, M. & Inouye, S. Retrons, msDNA, and the bacterial genome. *Cytogenet. Genome Res.* **110**, 491–499 (2005).

478. Simon, A. J., Ellington, A. D. & Finkelstein, I. J. Retrons and their applications in genome engineering. *Nucleic Acids Res.* **47**, 11007–11019 (2019).

479. Chappell, S. A., Edelman, G. M. & Mauro, V. P. Ribosomal tethering and clustering as mechanisms for translation initiation. *Proc. Natl. Acad. Sci.* **103**, 18077–18082 (2006).

480. Wannier, T. M. *et al.* Improved bacterial recombineering by parallelized protein discovery. *Proc. Natl. Acad. Sci.* **117**, 13689–13698 (2020).

481. Aronshtam, A. Dominant negative mutator mutations in the mutL gene of Escherichia coli. *Nucleic Acids Res.* **24**, 2498–2504 (1996).

482. Nyerges, Á. *et al.* A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species. *Proc. Natl. Acad. Sci.* **113**, 2502–2507 (2016).

483. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).

484. Zhang, Y. *et al.* A gRNA-tRNA array for CRISPR-Cas9 based rapid multiplexed genome editing in Saccharomyces cerevisiae. *Nat. Commun.* **10**, 1053 (2019).

485. Liu, J.-J. *et al.* CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218–223 (2019).

486. Knapp, D. J. H. F. *et al.* Decoupling tRNA promoter and processing activities enables specific Pol-II Cas9 guide RNA expression. *Nat. Commun.* **10**, 1490 (2019).

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

488. Rogers, J. K. *et al.* Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Res.* **43**, 7648–7660 (2015).

489. 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).

490. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

491. Baker Brachmann, C. *et al.* Designer deletion strains derived fromSaccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).

492. Tian, S. & Das, R. Primerize-2D: automated primer design for RNA multidimensional chemical mapping. *Bioinformatics* **33**, 1405–1406 (2017).

493. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).

494. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **19**, 67–80 (2018).

495. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science* **342**, 357–360 (2013).

496. Nyerges, Á. *et al.* Directed evolution of multiple genomic loci allows the prediction of antibiotic resistance. *Proc. Natl. Acad. Sci.* **115**, (2018).

497. Barbieri, E. M., Muir, P., Akhuetie-Oni, B. O., Yellman, C. M. & Isaacs, F. J. Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. *Cell* **171**, 1453-1467.e13 (2017).

498. Isaacs, F. J. *et al.* Precise Manipulation of Chromosomes in Vivo Enables Genome-Wide Codon Replacement. *Science* **333**, 348–353 (2011).

499. Tong, Y. *et al.* Highly efficient DSB-free base editing for streptomycetes with CRISPR-BEST. *Proc. Natl. Acad. Sci.* **116**, 20366–20375 (2019).

500. Tong, Y., Jørgensen, T. S., Whitford, C. M., Weber, T. & Lee, S. Y. A versatile genetic engineering toolkit for E. coli based on CRISPR-prime editing. *Nat. Commun.* **12**, 5206 (2021).

501. Volke, D. C., Martino, R. A., Kozaeva, E., Smania, A. M. & Nikel, P. I. Modular (de)construction of complex bacterial phenotypes by CRISPR/nCas9-assisted, multiplex cytidine base-editing. *Nat. Commun.* **13**, 3026 (2022).

502. Yuan, Q. & Gao, X. Multiplex base- and prime-editing with drive-and-process CRISPR arrays. *Nat. Commun.* **13**, 2771 (2022).

503. Aulicino, F. *et al.* Highly efficient CRISPR-mediated large DNA docking and multiplexed prime editing using a single baculovirus. *Nucleic Acids Res.* **50**, 7783–7799 (2022).

504. Li, H. *et al.* Multiplex precision gene editing by a surrogate prime editor in rice. *Mol. Plant* **15**, 1077–1080 (2022).

505. Simon, A. J., Morrow, B. R. & Ellington, A. D. Retroelement-Based Genome Editing and Evolution. *ACS Synth. Biol.* **7**, 2600–2611 (2018).

506. Jiang, W. *et al.* High-efficiency retron-mediated single-stranded DNA production in plants. *Synth. Biol.* **7**, ysac025 (2022).

507. 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).

508. 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).

509. Alper, H., Jin, Y.-S., Moxley, J. F. & Stephanopoulos, G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metab. Eng.* **7**, 155–164 (2005).

510. Kang, M. J. *et al.* Identification of genes affecting lycopene accumulation in *Escherichia coli* using a shot-gun method. *Biotechnol. Bioeng.* **91**, 636–642 (2005).

511. Jin, Y.-S. & Stephanopoulos, G. Multi-dimensional gene target search for improving lycopene biosynthesis in Escherichia coli. *Metab. Eng.* **9**, 337–347 (2007).

512. Chen, H., Bjerknes, M., Kumar, R. & Jay, E. Determination of the optimal aligned spacing between the Shine – Dalgarno sequence and the translation initiation codon of Escherichia coli m RNAs. *Nucleic Acids Res.* **22**, 4953–4957 (1994).

513. Engler, C., Kandzia, R. & Marillonnet, S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE* **3**, e3647 (2008).

514. Cunningham, F. X., Sun, Z., Chamovitz, D., Hirschberg, J. & Gantt, E. Molecular structure and enzymatic function of lycopene cyclase from the cyanobacterium Synechococcus sp strain PCC7942. *Plant Cell* **6**, 1107–1121 (1994).

515. Ferreira, R., Skrekas, C., Nielsen, J. & David, F. Multiplexed CRISPR/Cas9 Genome Editing and Gene Regulation Using Csy4 in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **7**, 10–15 (2018).

516. DiCarlo, J. E. *et al.* Yeast Oligo-Mediated Genome Engineering (YOGE). *ACS Synth. Biol.* **2**, 741–749 (2013).

517. Roy, K. R. *et al.* Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* **36**, 512–520 (2018).

518. Guo, X. *et al.* High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR–Cas9 in yeast. *Nat. Biotechnol.* **36**, 540–546 (2018).

519. Liang, Z., Metzner, E. & Isaacs, F. J. Advanced eMAGE for highly efficient combinatorial editing of a stable genome. Preprint at https://doi.org/10.1101/2020.08.30.256743 (2020).

520. Hitoshi, N., Ken-ichi, Y. & Jun-ichi, M. Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene* **108**, 193–199 (1991).

521. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14**, 321–327 (2011).

522. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci.* **108**, 10098–10103 (2011).

523. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).

524. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39**, 9275–9282 (2011).

525. Barrangou, R. & Marraffini, L. A. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Mol. Cell* **54**, 234–244 (2014).

526. Smith, L. M. *et al.* CRISPR-Cas immunity is repressed by the LysR-type transcriptional regulator PigU. *Nucleic Acids Res.* **52**, 755–768 (2024).

527. Smith, L. M. *et al.* The Rcs stress response inversely controls surface and CRISPR–Cas adaptive immunity to discriminate plasmids and phages. *Nat. Microbiol.* **6**, 162–172 (2021).

528. Westra, E. R., Buckling, A. & Fineran, P. C. CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* **12**, 317–326 (2014).

529. Perez-Rodriguez, R. *et al.* Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli: Envelope stress triggers CRISPR silencing. *Mol. Microbiol.* **79**, 584–599 (2011).

530. Muzellec, B., Teleńczuk, M., Cabeli, V. & Andreux, M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Preprint at https://doi.org/10.1101/2022.12.14.520412 (2022).

531. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

532. Zhang, J., Mahdi, A. A., Briggs, G. S. & Lloyd, R. G. Promoting and Avoiding Recombination: Contrasting Activities of the *Escherichia coli* RuvABC Holliday Junction Resolvase and RecG DNA Translocase. *Genetics* **185**, 23–37 (2010).

533. Almendros, C., Mojica, F. J. M., Díez-Villaseñor, C., Guzmán, N. M. & García-Martínez, J. CRISPR-Cas Functional Module Exchange in Escherichia coli. *mBio* **5**, e00767-13 (2014).

534. Burroughs, A. M. & Aravind, L. New biochemistry in the Rhodanese-phosphatase superfamily: emerging roles in diverse metabolic processes, nucleic acid modifications, and biological conflicts. *NAR Genomics Bioinforma.* **5**, lqad029 (2023).

535. Holm, L. Dali server: structural unification of protein families. *Nucleic Acids Res.* **50**, W210–W215 (2022).

536. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

537. Sousa, A., Bourgard, C., Wahl, L. M. & Gordo, I. Rates of transposition in *Escherichia coli*. *Biol. Lett.* **9**, 20130838 (2013).

538. Alexander, D. L. *et al.* Random Mutagenesis by Error-Prone Pol Plasmid Replication in Escherichia coli. in *Directed Evolution Library Creation* (eds. Gillam, E.

M. J., Copp, J. N. & Ackerley, D.) vol. 1179 31–44 (Springer New York, New York, NY, 2014).

539. Reeh, S., Pedersen, S. & Friesen, J. D. Biosynthetic regulation of individual proteins in relA + and relA strains of Escherichia coli during amino acid starvation. *Mol. Gen. Genet. MGG* **149**, 279–289 (1976).

540. Hansen, A.-M. *et al.* Structural Basis for the Function of Stringent Starvation Protein A as a Transcription Factor. *J. Biol. Chem.* **280**, 17380–17391 (2005).

541. Hansen, A. *et al.* SspA is required for acid resistance in stationary phase by downregulation of H-NS in *Escherichia coli*. *Mol. Microbiol.* **56**, 719–734 (2005).

542. Ishihama, A. & Saitoh, T. Subunits of RNA polymerase in function and structure. *J. Mol. Biol.* **129**, 517–530 (1979).

543. Wang, F. *et al.* Structural basis for transcription inhibition by E. coli SspA. *Nucleic Acids Res.* **48**, 9931–9942 (2020).

544. Travis, B. A. *et al.* Structural Basis for Virulence Activation of Francisella tularensis. *Mol. Cell* **81**, 139-152.e10 (2021).

545. Lou, J. *et al.* The stringent starvation protein SspA modulates peptidoglycan synthesis by regulating the expression of peptidoglycan synthases. *Mol. Microbiol.* **118**, 716–730 (2022).

546. Levchenko, I., Seidel, M., Sauer, R. T. & Baker, T. A. A Specificity-Enhancing Factor for the ClpXP Degradation Machine. *Science* **289**, 2354–2356 (2000).

547. Hansen, A., Lehnherr, H., Wang, X., Mobley, V. & Jin, D. J. *Escherichia coli* SspA is a transcription activator for bacteriophage P1 late genes. *Mol. Microbiol.* **48**, 1621–1631 (2003).

548. Drahos, D. J. & Hendrix, R. W. Effect of bacteriophage lambda infection on synthesis of groE protein and other Escherichia coli proteins. *J. Bacteriol.* **149**, 1050–1063 (1982).

549. Hansen, A.-M. & Jin, D. SspA up-regulates gene expression of the LEE pathogenicity island by decreasing H-NS levels in enterohemorrhagic Escherichia coli. *BMC Microbiol.* **12**, 231 (2012).

550. Westra, E. R. *et al.* H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Mol. Microbiol.* **77**, 1380–1393 (2010).

551. Pougach, K. *et al.* Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol.* **77**, 1367–1379 (2010).

552. Pul, Ü. *et al.* Identification and characterization of *E. coli* CRISPR- *cas* promoters and their silencing by H-NS. *Mol. Microbiol.* **75**, 1495–1512 (2010).

553. Masters, M. *et al.* The pcnB gene of Escherichia coli, which is required for ColE1 copy number maintenance, is dispensable. *J. Bacteriol.* **175**, 4405–4413 (1993).

554. Santin, Y. G. *et al.* In vivo TssA proximity labelling during type VI secretion biogenesis reveals TagA as a protein that stops and holds the sheath. *Nat. Microbiol.* **3**, 1304–1313 (2018).

555. St-Pierre, F. *et al.* One-Step Cloning and Chromosomal Integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).

556. Hossain, A. A., McGinn, J., Meeske, A. J., Modell, J. W. & Marraffini, L. A. Viral recombination systems limit CRISPR-Cas targeting through the generation of escape mutations. *Cell Host Microbe* **29**, 1482-1495.e12 (2021).

557. Su, M.-T., Venkatesh, T. V. & Bodmer, R. Large- and Small-Scale Preparation of Bacteriophage λ Lysate and DNA. *BioTechniques* **25**, 44–46 (1998).

558. Salgado, H. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* **34**, D394–D397 (2006).

559. Pettersen, E. F. *et al.* UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

560. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

561. Granger, B. E. & Perez, F. Jupyter: Thinking and Storytelling With Code and Data. *Comput. Sci. Eng.* **23**, 7–14 (2021).

562. Joshi, N. & Fass, J. sickle - A windowed adaptive trimming tool for FASTQ files using quality.

563. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

564. Shimoyama, Y. pyCirclize: Circular visualization in Python.

565. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (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:
Santiago Caetano Lopez
36C484B7600D4CD...
Author Signature

7/24/2024
Date