Reconstructing transcriptional histories by CRISPR acquisition of retron-based genetic barcodes

Authors: Santi Bhattarai-Kline[1], Elana Lockshin[2], Max G. Schubert[3,4], Jeff Nivala[5], George Church[3,4], Seth L. Shipman[1,6]*

Affiliations:

[1]Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, USA

[2]Department of Neurobiology, Duke University Medical Center, Durham, NC, USA

[3]Department of Genetics, Harvard Medical School, Boston, MA, USA

[4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA

[5]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

[6]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

*Correspondence to: seth.shipman@gladstone.ucsf.edu

ABSTRACT

Biological processes depend on the differential expression of genes over time, but methods to make true physical recordings of these processes are limited. Here we report a strategy for making time-ordered recordings of transcriptional events into living genomes. We do this via engineered RNA barcodes, based on prokaryotic retrons, which are reverse-transcribed into DNA and integrated into the genome using the CRISPR-Cas system. This approach enables the targeted recording of time-ordered transcriptional events in cells. The unidirectional integration of barcodes by CRISPR integrases enables reconstruction of transcriptional event timing based on a physical record via simple, logical rules rather than relying on pre-trained classifiers or post-hoc inferential methods.

INTRODUCTION

DNA is the universal storage medium for cellular life. In recent years, an emerging field of biotechnology has begun repurposing DNA to store data that has no cellular function. The same qualities of DNA that are beneficial in a biological context – high density, ease of copying, and durability – also enable flexible storage of text, images, and sound[1–3]. Extending this general concept, researchers have developed data storage systems contained within living organisms that allow the recording of biological signals into DNA, such as endogenous transcription and environmental stimuli. One particular avenue of interest for such systems is in the longitudinal recording of biological processes within cells[4–6].

These recordings address a fundamental limitation in standard methods to interrogate complex biological processes that require the destruction of cells and, thus, can only provide measurements at single points in time (e.g. RNA-Seq). Because biological processes are not perfectly synchronized at the cellular level, any individual cell collected in the middle of a biological process could be either ahead or behind in the progression of events relative to any other cell collected at that same time. This cellular heterochronicity makes it impossible to definitively reconstruct time-dependent processes from the destructive measurement of parallel samples. Indeed, cell-to-cell heterochronicity has actually been exploited in computational methods to infer position in a biological process among cells within a single sample (e.g. single-cell RNA-Seq pseudotime)[7]. However, these methods of inference make assumptions about the relationship between cells that are not explicitly known, and often require user-imposed constraints or the incorporation of prior biological knowledge[8].

An approach known as molecular recording provides an alternative to statistical inference. Molecular recorders are biological devices that continuously record cellular processes, storing a physical record of the data permanently in cellular DNA, so that it may be retrieved at the very end of an experiment or process. Approaches to build molecular recorders have relied on different methods of modifying DNA, including site-specific recombinases and CRISPR-Cas

nucleases[4,9,10]. Another approach to molecular recording, which we have worked to develop, leverages CRISPR-Cas integrases[11].

CRISPR-Cas systems function as adaptive immune systems in bacteria and archaea. During the first phase of the immune response to infection by phage or mobile genetic elements, called adaptation, the CRISPR proteins Cas1 and Cas2 integrate a piece of foreign DNA into a genomic CRISPR array. The CRISPR array consists of a leader sequence followed by unique spacer sequences derived from foreign DNA, which are all separated by identical sequences called repeats. The sequence information stored in the spacers serves as an immunological memory of previous infection. This machinery, comprised of the CRISPR array, Cas1, and Cas2, is a ready-made storage device. When the Cas1-Cas2 complex integrates a spacer into the CRISPR array, it is added next to the leader sequence and the previous spacers are shifted away from the leader[12,13]. Thus, spacers which are further away from the leader sequence were acquired further in the past, and those closer to the leader acquired more recently. This unidirectionality of CRISPR arrays has been previously used to encode information into CRISPR arrays through the delivery of chemically synthesized oligos[2,11] or by modulating the copy number of a reporter plasmid in response to a biological stimulus[3,5]. However, the ability to record the timing of multiple different biological signals into the CRISPR array of a single cell has not yet been demonstrated.

Here, we demonstrate successful recording of temporal relationships by adding a new molecular component to the system, a retroelement called a retron. Recently determined to function in bacteria as a defense system against phage infection[14], a typical retron consists of a single operon that controls the expression of: (1) a small, highly structured noncoding RNA (retron ncRNA), (2) a retron reverse transcriptase (retron RT) that specifically recognizes and reverse transcribes part of its cognate ncRNA, and (3) one or more effector proteins which are implicated in downstream functions[14–16]. The compact size, specificity, and flexibility of retrons to produce customizable DNA *in vivo* make them an attractive tool for biotechnology. Retrons have been

used in applications such as genome editing in several host systems[17–19] and early analog molecular recorders[20]. By combining the functions of retrons and CRISPR-Cas integrases, we have built a system to make temporal recordings of transcriptional events.

To record transcriptional events, we engineered retrons to produce a set of compact, specific molecular tags, which can be placed under the control of multiple promoters of interest inside a single cell. When a tagged promoter is active, the tag sequence is transcribed into RNA, recognized by the retron reverse transcriptase, and reverse transcribed to generate a DNA 'receipt' of transcription. That DNA 'receipt' can then bound by Cas1-Cas2 and integrated into the cell's CRISPR array, creating a permanent record of transcription. If another tagged promoter subsequently becomes active, a different DNA 'receipt' can be generated and integrated into the CRISPR array following the first spacer. By producing a linear record of these 'receipts' in the genome, we have built a biological device that records the temporal history of specific gene expression events within single cells (Fig 1a).

RESULTS

Retron reverse-transcribed DNA can be acquired by CRISPR integrases

The first challenge to building a temporal recorder of gene expression was to generate specific DNA barcodes following a transcriptional event, which can be permanently stored in a cell's genome through integration by Cas1-Cas2. For integration, Cas1-Cas2 require DNA of at least 35 bases from end-to-end, with a 23 base complementary core region, and a protospacer-adjacent motif (PAM)[21]. We designed variant ncRNA sequences of a native *E. coli* retron, Eco1[22,23] (Supp. Fig. 1a), for integration into the genome by the type I-E CRISPR system of *E. coli* BL21-AI cells[12] after they are reverse transcribed (Fig 1b).

We tested multiple variants for both reverse-transcription functionality and the ability of their RT-DNA to be acquired by the CRISPR adaptation machinery, and identified two that accomplish these aims. When overexpressed in *E. coli*, variants v32 and v35 (Supp. Fig. 1b,

Supp. Fig. 1c) produced robust levels of RT-DNA that could be easily visualized on a PAGE gel (Fig 1c), had perfect 3'-TTC PAM sequences, and were able to hybridize to create a 23-base core. Rather than a single copy of the retron RT-DNA hairpin forming the prespacer for acquisition, both v32 and v35 are designed such that two copies of the RT-DNA can form a duplex, which we hypothesized would be efficiently integrated into the CRISPR array (Fig. 1d, Supp. Fig. 1b, Supp. Fig. 1c). To measure the ability of variant retrons to be acquired, we overexpressed the variant ncRNA, Eco1 RT, and Cas1-Cas2 in BL21-AI cells which harbor a single CRISPR array in their genome. We then sequenced the CRISPR arrays of these cells to quantify integrations. In both cases, we found new spacers in these cells matched the sequence of the retron RT-DNA that was expressed (Fig 1e). Critically, arrays containing retron-derived spacers were only seen when cells also harbored a plasmid coding for Eco1 RT (Fig 1e), indicating that the retron-derived spacers were indeed a result of the production of RT-DNA, rather than being derived exclusively from plasmid DNA. Retron v35 was acquired at a higher rate than v32 (Fig 1e), and was selected for use in subsequent work.

We further modified v35 by extending the length of the non-hairpin duplex region referred to as the a1/a2 region (Fig 1f). We have previously shown that this modification to retrons both increases production of RT-DNA in bacteria and yeast and increases the efficiency of genome editing techniques which rely on retrons[24]. Consistent with our previous findings, extending the a1/a2 region of retron v35 resulted in an increase in the percentage of arrays which contained retron-derived spacers (Fig 1f). This suggests that, like RT-DNA-templated genome editing, the rate of acquisition of retron-derived spacers is dependent on the abundance of RT-DNA. To take advantage of this improved acquisition efficiency, we incorporated this modification into all future Eco1 constructs.

To better characterize the acquisition of spacers by Cas1-Cas2 over time, we expressed retron v35 and Cas1-Cas2 for 24 hours and sampled arrays at regular intervals throughout (Fig. 1g-i). This showed that the number of arrays that contain retron-derived spacers increased

114 regularly over time (Fig. 1g). As retron-derived spacers accumulated, they were accompanied by
115 spacers derived from the cell's genome and from plasmids, as previously described[12]. These non-
116 retron-derived spacers also increased in the population of arrays over time (Fig 1h). The
117 proportion of new retron-derived spacers remained relatively stable over time, making up between
118 1-10% of new spacer acquisitions (Fig 1i). Thus, the abundance of retron-derived spacers can be
119 used as a proxy for the duration of a transcriptional event. This result demonstrates a new
120 implementation of analog molecular recording, similar in function to those previously described[20],
121 but based on the marriage of retrons and CRISPR-Cas integrases.

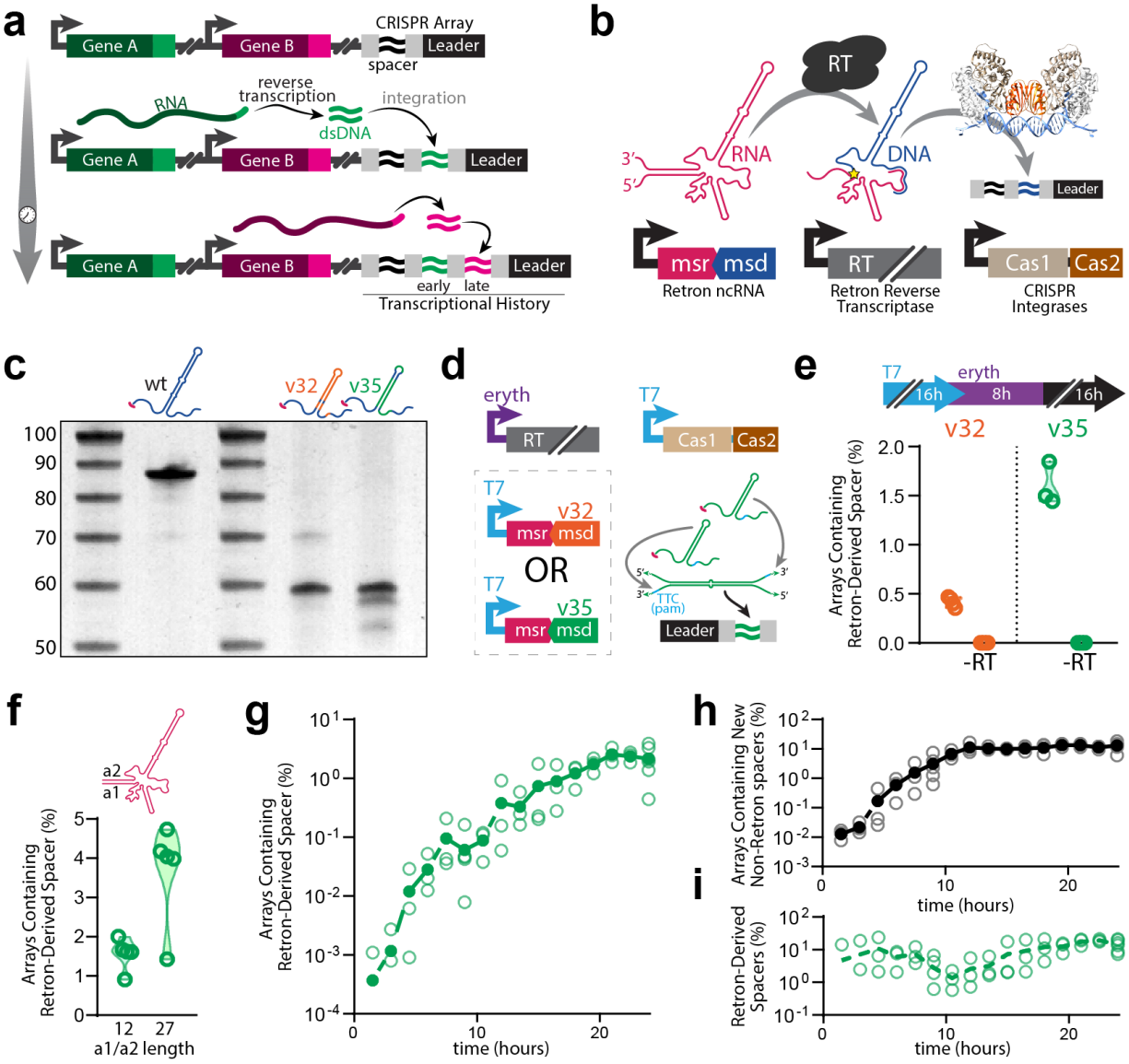


Figure 1. Retron reverse transcribed DNA can be acquired by CRISPR integrases. a. Schematic representation of retroelement-based transcriptional recording into CRISPR arrays. **b.** Schematic representation of biological components of the retron-based recorder. **c.** Urea-PAGE visualization of RT-DNA from retron Eco1 ncRNA variants. From left to right (excluding ladders): wild-type Eco1, Eco1 v32, Eco1 v35. **d.** Schematic of experimental promoters used to test retron-recorder parts and cartoon of hypothetical duplex RT-DNA prespacer structure. **e.** Quantification of arrays expanded with retron-derived spacers using Eco1 variants v32 (orange) and v35 (green). Open circles represent biological replicates. **f.** Quantification of arrays expanded with retron derived spacers with a wild-type (12 bp) and extended (27 bp) a1/a2 region. Open circles represent biological replicates. **g.** Time series of array expansions from retron-derived spacers. Open circles represent biological replicates, closed circles are the mean. **h.** Time series of array expansions from non-retron-derived spacers. Open circles represent biological replicates, closed circles are the mean. **i.** Proportion of total new spacers that are retron-derived. Open circles represent biological replicates, dashed line is the mean. All statistics in Supplementary Table 1.

Diversification of retron-derived transcript barcodes

A crucial advantage of retron-based molecular recording is the ability to follow multiple transcripts of interest within a single cell. This enables the recording of gene expression timing within genetically identical cells, rather than relying on a mixed population of cells each harboring different sensors. The specificity of retrons also enables more focused recordings compared to promiscuous RTs, which cannot be made to selectively reverse-transcribe individual transcripts[6,25]. Additionally, in contrast to recombinase-based molecular recording systems, the retron-based approach should enable a much larger set of sensors to coexist within a population of genetically identical cells. This is because the set of barcoded retrons is only limited by DNA sequence, rather than by the comparatively small number of well-characterized recombinases[4,26]. To construct a set of unique retron tags, we chose to use the loop in retron v35's RT-DNA hairpin as a six-base barcode (Fig 2a, Supp. Fig. 2). This barcoding strategy allows multiple otherwise identical ncRNAs to be reverse transcribed by the same RT, but remain easily distinguishable by sequence in CRISPR arrays. We synthesized a set of barcoded retrons, expressed them in cells along with Cas1-Cas2, and analyzed how efficiently they were acquired by sequencing CRISPR arrays (Fig 2b). We compared these barcoded variants to the original v35 retron, and included a dead-RT version of the v35 retron as a negative control. Overall, we observed small differences in the rate at which different barcoded retrons were acquired, but none of the retrons suffered a drastic reduction in acquisition efficiency (Fig 2b).

To test our ability to discriminate between the barcoded spacers derived from this set of retrons, we searched the sequence data from each sample expressing one barcoded retron for all of the other barcodes in the set. In our computational pipeline, we specify a tolerance of up to 3 bases of mismatches or indels (out of a 23 base search sequence) when determining the identity of a retron-derived spacer. This is to compensate for minor differences which may be found in mature spacers compared to their hypothetical sequence. As such, if our retron barcodes are faithfully preserved through all steps of the recording process (DNA coding sequence → RNA →

RT-DNA → CRISPR array), then we should be able to effectively distinguish between barcodes which differ by 4 bases or more. This proved to be true when we examined our original set of 9 barcodes for orthogonality *in-silico*. Barcodes which differed by less than 4 bases could not be differentiated and barcodes which differed by 4 bases or more could be distinguished from each other with perfect accuracy, forming a set of 6 mutually orthogonal barcodes (Fig 2c-d). This demonstrated that barcode sequences in retron-based transcriptional tags are faithfully preserved throughout the process of molecular recording, allowing for the facile construction of sets of mutually orthogonal tags.

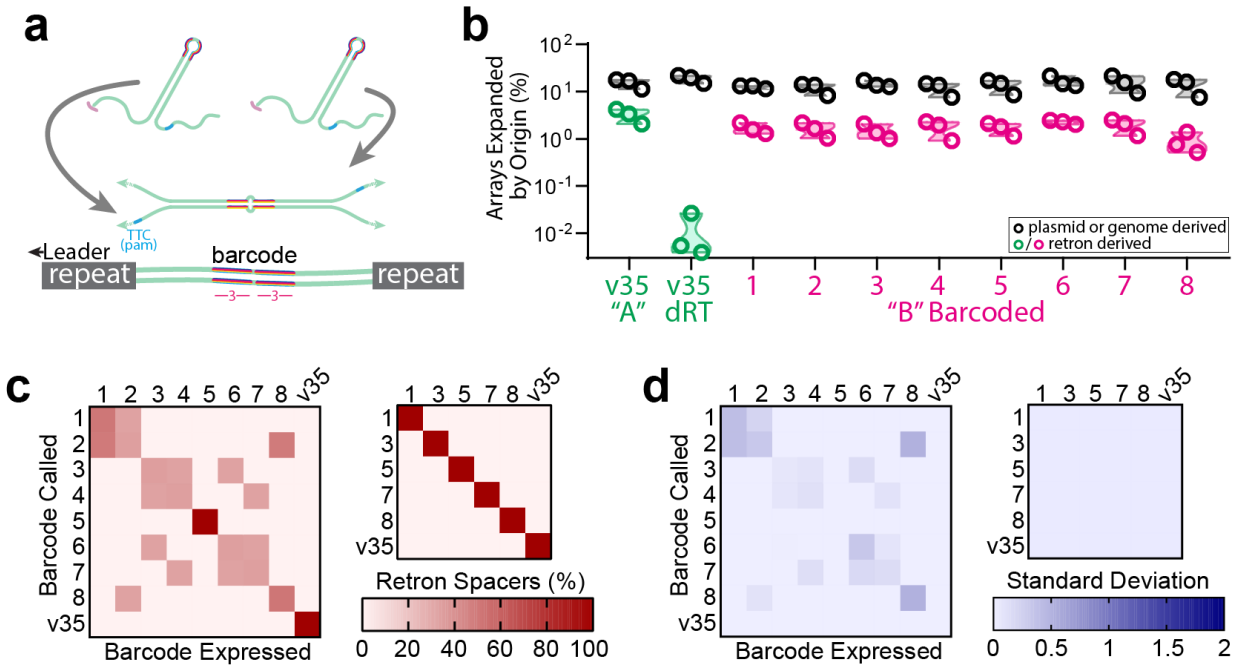


Figure 2. Diversification of retron-derived transcript barcodes. a. Hypothetical structure of duplexed RT-DNA prespacer with 6-base barcode and retron-derived spacer. **b.** Quantification of array expansions from barcoded variants of retron Eco1 v35, showing both retron-derived (green/pink) and non-retron derived (black) spacers for each variant. Open circles represent biological replicates. **c.** Left: Heatmap of *in silico* ability to distinguish between all barcoded Eco1 v35 variants. Right: Heatmap of *in silico* ability to distinguish between reduced set of barcoded Eco1 v35 variants. **d.** Heatmap of standard deviation between three separate trials of barcode discrimination test. Left: full set. Right: reduced set. All statistics in Supplementary Table 1.

Mechanism of retron-derived prespacer acquisition

While it has been demonstrated that Cas1-Cas2 requires two complementary strands of DNA to integrate a spacer into the CRISPR array[11], recent evidence suggests that Cas1-Cas2 are capable of binding ssDNA and may in fact bind the two strands of a prespacer separately, in a stepwise fashion[27]. To date, all experimentally characterized retrons have been shown to use the 2'-OH from a conserved guanosine to initiate reverse-transcription[15,23], leaving a biochemically unique 2'-5' RNA-DNA linkage. We hypothesized that this unique feature of RT-DNA prespacers might allow us to further interrogate the mechanism of prespacer loading and spacer acquisition by Cas1-Cas2.

Unlike many prespacers examined in prior work, which generally form perfect DNA duplexes, the duplex formed by our retron has three characteristic regions where the prespacer should contain mismatches. The first of these regions is a stretch of five bases which, after integration, is located closest to the leader sequence. We will refer to this region as the leader-proximal, or LP, region (Fig 3a). Next, there is a single base mismatch which falls near the middle of the final spacer. We will refer to this region as the middle, or M, region (Fig 3a). Finally, the last of the mismatched regions is found, in the mature spacer, in the five bases furthest from the leader sequence. We will refer to this as the leader-distal, or LD, region (Fig 3a). We found that in retron-derived spacers, the sequence of these mismatched regions either corresponded to one strand of the original hypothetical prespacer duplex or the other (Fig 3b). In this analysis, we will refer to the two strands of the hypothetical prespacer as the (+) and (-) strands. In Eco1-derived spacers, the sequence in the LP region overwhelmingly corresponded to the (-) strand. This (-) strand contains the PAM-proximal 3'-end, which determines directionality[11] and has been shown to be integrated second in the spacer integration process[27,28]. This pattern of preserving the PAM-derived 3'-end sequence in the LP region was also seen when cells were electroporated with a synthetic oligonucleotide version of the retron RT-DNA (Fig 3b).

At the opposite end of the spacer, however, retron-derived and oligo-derived spacers were not identical. In the LD region, oligo-derived spacers overwhelmingly mapped to the (+) strand, whereas the LD regions of retron RT-DNA-derived spacers predominantly mapped to the (-) strand (Fig 3b). Because the *in vivo*-produced RT-DNA contains a 2'-5' linkage and the oligo does not, we suspected that the 2'-5' linkage present in the Eco1 RT-DNA may interfere with the CRISPR adaptation process. To test this, we treated purified Eco1 RT-DNA with the eukaryotic debranching enzyme DBR1, which processes RNA lariats by cleaving 2'-5' bonds in RNA[29]. Treatment of Eco1 RT-DNA with DBR1 *in vitro* resulted in a characteristic downward shift in the size of Eco1 RT-DNA from the loss of a small number of ribonucleotides remaining at the branch point. DBR1 treatment also rendered Eco1 RT-DNA sensitive to the 5'-exonuclease recJ (Fig 3c). This indicates that DBR1 is able to remove the 2'-5' linkage and produce Eco1 RT-DNA with an unbranched 5'-end. When purified Eco1 RT-DNA was treated with DBR1 and electroporated back into cells expressing Cas1-Cas2, the LD sequences of retron-derived spacers closely resembled those of retron-derived spacers after oligo electroporation (Fig 3d), indicating that the presence of the 2'-5' linkage in Eco1 RT-DNA is responsible for its unique pattern of spacer sequences. One potential explanation for the spacer pattern observed in Fig 3b is that the integrases may use one molecule of RT-DNA as the (-) strand and one molecule of plasmid-derived ssDNA (the retron coding sequence) as the (+) strand when processing a prespacer for acquisition.

Beyond the apparent difference in prespacer processing due to the 2'-5' linkage, we were curious to see whether the efficiency of acquisition would increase if the 2'-5' linkage was removed. We approached this question by electroporating cells with three different prespacer types: purified RT-DNA, purified and debranched RT-DNA, and a synthetic oligo version of the RT-DNA. Debranched RT-DNA and oligos tended to be acquired more efficiently than the natively-branched RT-DNA, but this trend did not reach statistical significance (Fig 3e).

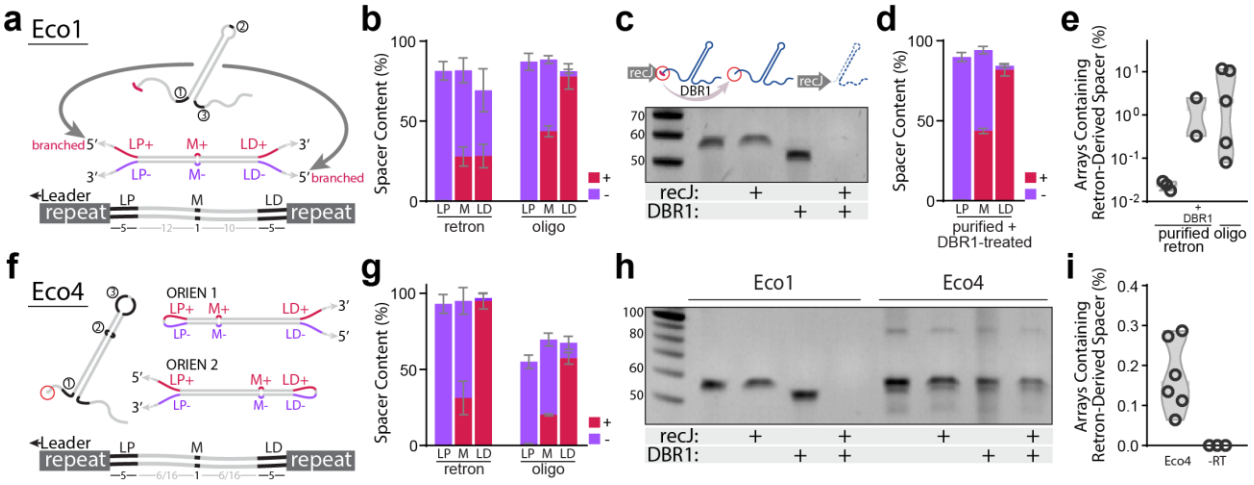
In some retrons, processing naturally occurs following reverse transcription to remove the 2'-5' linkage[30], so we tested such a retron to see whether this processing would change the pattern

or efficiency of retron-derived acquisitions. While the biosynthesis of the retron Eco4 RT-DNA still depends on priming from the 2'-hydroxyl of a conserved guanosine, its RT-DNA is cleaved 4 bases away from the 5' branch point by an ExoVII exonuclease complex-dependent mechanism[30,31], leaving a mature RT-DNA lacking a 2'-5' linkage (Supp. Fig. 3a)[30]. We expressed wildtype Eco4 ncRNA, Eco4 RT, and Cas1-Cas2 in cells and then sequenced their CRISPR arrays to measure acquisitions. Notably, unlike the variant Eco1 retron, acquisitions from Eco4 occurred in two different orientations (Fig. 3f, Supp. Fig. 3b-c). Although the wildtype Eco4 RT-DNA does not have any perfect PAM sites (3'-TTC), both orientations observed in Eco4-derived spacers had a near-perfect PAM (3'-GTC) which proved sufficient for integration. We could not determine whether these Eco4-derived spacers were derived from single hairpins or duplexes. We next analyzed the mismatched regions of the Eco4-derived spacers. As expected, almost all the LP regions mapped to the (-) strand, but unlike with variant Eco1, the LD region of Eco4-derived spacers almost entirely mapped to the (+) strand (Fig 3g). Oligo-derived Eco4 spacers produced similar patterns of acquisition (Fig 3g), indicating that retron Eco4, and likely other unbranched RT-DNAs, avoid the peculiarities caused by using a branched RT-DNA as a prespacer.

To confirm that Eco4 RT-DNA is debranched *in vivo*, we treated purified Eco4 RT-DNA with DBR1 and did not observe a size shift that would indicate removal of ribonucleotides (Fig. 3h). In addition, the RT-DNA was not recJ sensitive because there were fewer than 6 bases of single stranded DNA on the 5' end, which recJ needs for exonuclease activity.

The final test for Eco4 was to determine the overall efficiency of acquisition. We observed that retron-derived spacers from Eco4 were dependent on the presence of Eco4 RT, but their frequency was ultimately lower than Eco1-derived spacers (Fig 3i). Based on these baseline efficiencies, we have focused our efforts on engineering Eco1 for the purpose of molecular recording. However, these results demonstrate that other retrons can also be used for molecular

253 recording and, as is the case with Eco4, may possess unique qualities which affect their function
254 in these applications.



255 **Figure 3. Mechanism of retron-derived prespacer acquisition. a.** Hypothetical structure of duplexed
256 Eco1 v35 RT-DNA prespacer and retron-derived spacer, with mismatched regions highlighted. **b.**
257 Quantification of mismatch region sequences in spacers from cells expressing Eco1 v35 versus cells
258 electroporated with oligo mimic. Bars represent the mean of 4 and 5 biological replicates for the retron and
259 oligo-derived conditions, respectively (±SD). **c.** Urea-PAGE visualization of Eco1 RT-DNA. DBR1 treatment
260 resolves 2'-5' linkage. **d.** Quantification of mismatch region sequences in spacers from cells electroporated
261 with purified, debranched Eco1 v35 RT-DNA. Bars represent the mean of 3 biological replicates (±SD). **e.**
262 Quantification of array expansions from different prespacer substrates. Open circles are individual
263 biological replicates. **f.** Schematic of Eco4 RT-DNA, in both orientations, with mismatch sequences
264 highlighted. **g.** Quantification of mismatch region sequences in cells expressing Eco4 versus cells
265 electroporated with oligo mimic. Bars represent the mean of 3 biological replicates (±SD). **h.** Urea-PAGE
266 visualization of Eco4 RT-DNA. DBR1 does not cause size shift of Eco4 RT-DNA. **i.** Quantification of array
267 expansions from retron Eco4. Open circles represent individual biological replicates. All statistics in
268 Supplementary Table 1.

269 **Temporal recordings of gene expression**

270 Having built and characterized the requisite tools, we set out to make a temporal recording
271 of gene expression using retron-based tags. First, we first constructed a signal plasmid and a
272 recording plasmid. The signal plasmid, pSBK.134, harbored two copies of the Eco1 v35 ncRNA
273 with different barcodes in the loop, which we will refer to as "A" and "B", under different inducible

promoters. “A” was under the control of the anhydrotetracycline-inducible promoter, pTet*, and ncRNA “B” was under the control of the choline chloride-inducible promoter, pBetI (Fig 4a)[32]. The recording plasmid contained the coding sequence for retron Eco1 RT, expressed from the constitutive promoter J23115, and the coding sequences for Cas1 and Cas2, both under the control of a T7/lac promoter. The recorded responses to induction of pTet* and pBetI were well matched, with 24 hours of induction of each promoter yielding similar numbers of “A” and “B” derived spacers (Fig 4b).

To record a time-ordered biological event, we transformed *E. coli* BL21-AI cells with both the signal and recording plasmids, and grew them under two different experimental conditions for a total of 48 hours. In the first temporal recording condition, cells were grown for 24 hours with inducers driving the expression of Eco1 RT, Cas1-2, and ncRNA “A”. The cells were then grown for another 24 hours while expressing ncRNA “B”, along with the Eco1 RT and Cas1-Cas2 (Fig 4c). In the second condition, the order of expression of ncRNA “A” and “B” was reversed (ncRNA “B” was expressed for the first day and ncRNA “A” for the second) (Fig 4d). Samples were taken at 24 and 48 hours. Examination of the expanded arrays revealed a significant increase in the percentage of cells that received a retron-derived spacer in the 24 hours where its chemical inducer was present, compared to the 24 hours where it was absent. This held true for both ncRNAs “A” and “B” under both the “A”-before-“B” and “B”-before-“A” expression schemes (Fig 4c,d). The number of non-retron-derived spacers also increased consistently over 48 hours (Fig 4e).

To further test the generalizability of the system, we made a recording of a different set of promoters driving the same retron ncRNAs. For this second arrangement, the recording plasmid remained the same, but in the signal plasmid, pSBK.136, ncRNA “A” was placed under the control of the sodium salicylate-inducible promoter, pSal, and “B” under the control of pTet* (Fig. 4f). In this setting, 24 hours of induction of each promoter resulted in a much higher rate of acquisitions from retron “A” driven by pSal than acquisitions of retron “B” from pTet* (Fig 4g). Notably, in one

biological replicate the recording system appeared to break, resulting in nearly non-existent acquisitions; this sample was excluded from further analysis following its identification as an outlier by Grubbs' test (Fig. 4g). Next, we tested two experimental conditions: "A"-before-"B" and "B"-before-"A" (Fig. 4h, 4i). Despite the mismatched promoter strengths, when arrays were examined from 24- and 48-hour timepoints, more arrays were expanded with retron-derived spacers in the presence of their respective inducers than in their absence (Fig. 4h-i). In addition, the numbers of non-retron-derived spacers again increased over 48 hours (Fig 4j).

This analysis of spacer acquisitions from the signal plasmid was enabled by a timepoint sampling in the middle of the overall transcriptional sequence. However, the aim of this work is to reconstruct the timing of transcriptional events using only data acquired at an endpoint. Therefore, we defined logical rules that should govern the ordering of spacers in the CRISPR arrays, and allow us to reconstruct the order of transcription of separate ncRNAs. Because spacers are acquired unidirectionally, with newer spacers closer to the leader sequence, we postulated that if transcript "A" is expressed before transcript "B", arrays of the form "A" → "B" → Leader should be more numerous than "B" → "A" → Leader. Accordingly, if "B" is expressed before "A", then the opposite should be true: the number of "B" → "A" → Leader arrays should be greater than the number of "A" → "B" → Leader arrays.

Another feature of using CRISPR arrays for recording is that Cas1-2 also acquire spacers derived from the plasmid and genome[12]. These untargeted acquisitions can also be used to interpret temporal information[3,5]. If we assume that these untargeted spacers (denoted "N") are acquired at a constant rate throughout the experiment, we can define a set of rules that govern the order of "N" → "A" → Leader versus "A" → "N" → Leader arrays and of "N" → "B" → Leader versus "B" → "N" → Leader arrays. For the "A"-before-"B" case, since "A" is expressed in the first half of an experiment, arrays of the form "A" → "N" → Leader should be more numerous than "N" → "A" → Leader. And since "B" is expressed in the second half of the experiment, "N" → "B" → Leader arrays should be more numerous than "B" → "N" → Leader arrays. Likewise, in the "B"-

before-"A" condition, "N" → "A" → Leader arrays should be more numerous than "A" → "N" → Leader arrays and "B" → "N" → Leader arrays should be more numerous than "N" → "B" → Leader arrays. Restating these as mathematical statements, we can take the difference between possible array types (e.g. "A" → "B" → Leader minus "B" → "A" → Leader) as the numerator and the sum of the two possibilities (e.g. "A" → "B" → Leader plus "B" → "A" → Leader) as the denominator (Fig. 4k) to yield a number between -1 and 1 for each ordering rule (A/B, A/N, and B/N). By the convention of our ordering rules, positive values would indicate that "A" was present before "B", and a negative output would indicate that "B" was present before "A". The magnitude of the output ($0 \leq |x| \leq 1$) is a measure of how strongly the rule is satisfied in a given direction.

To test these predictions, we sequenced the CRISPR arrays of all of our samples at the 48-hour endpoint. Between 6 biological replicates of samples with signal plasmid pSBK.134, the samples in which "A" was expressed before "B" yielded positive values when subjected to analysis by our ordering rules, correctly identifying the order of expression. Likewise, for samples where "B" was expressed before "A", the rules yielded negative values, again correctly identifying the order (Fig 4l). We also calculated a composite score by taking a weighted average of all three rules. The score consists of the average between the A/B rule and the sum of the A/N rule and B/N rule. We devised this formulation based on what the ordering rules represent in an ideal system. By definition, the A/B rule represents the degree of order between A and B. Likewise, the A/N and B/N rules represent the degree of order between N and A or B, respectively. In an ideal system, where the rate of acquisition of N is assumed to constant, the sum of the A/N score and B/N score can be used as a proxy for the order of A with respect to B. Thus, if we assume that the rate of acquisition of N is constant, we can average the sum of the A/N and B/N scores with the A/B score to generate a composite score which integrates all three rules and is representative of the degree of temporal order between A and B. When applied to our *in vivo* recording data, this method accurately determined that each experiment yielded directional acquisition of spacers and corrected recalled the order of events for both directions. Critically, this demonstrates our ability

to accurately reconstruct the order of two transcriptional events in an endpoint biological sample, using only logical rules derived from first principles. When each replicate was examined separately, though all rules were not uniformly satisfied, the order of expression could be consistently determined (Supp. Fig. 4a-b).

When this analysis was applied to samples with the signal plasmid pSBK.136 (which had mismatched “A” and “B” promoter strengths), we were still able to accurately reconstruct the order of events from endpoint data (Fig. 4m, Supp. Fig. 4c-d), demonstrating that the temporal analysis of gene expression is robust. Ultimately, this paradigm enables the reconstruction of temporal histories within genetically-identical populations of cells, based on a physical molecular record.

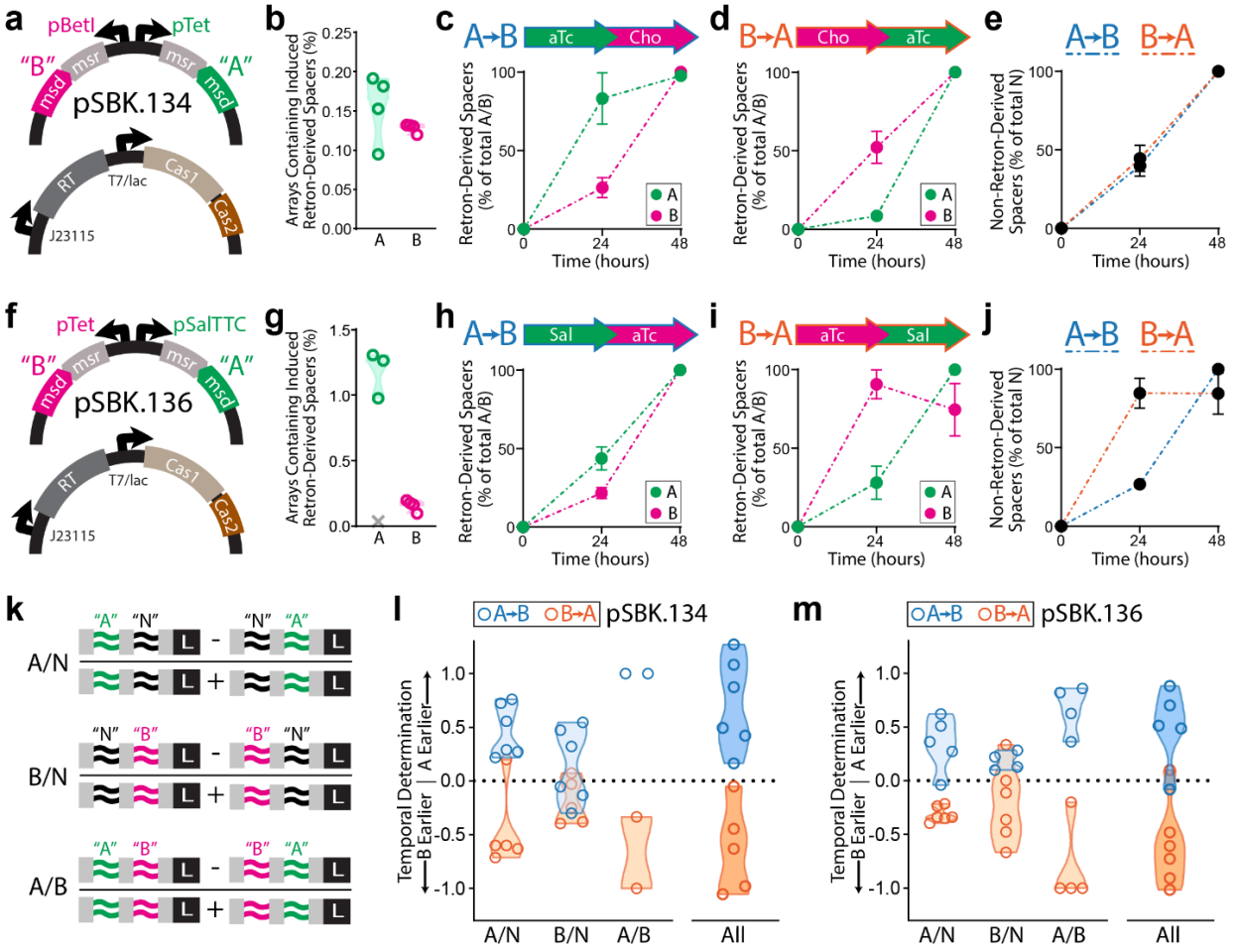


Figure 4. Temporal recordings of gene expression. a. Schematic of the signal plasmid pSBK.134 used to express ncRNAs "A" and "B", and the recording plasmid used to express Eco1 RT and Cas1 and 2. **B.** Accumulation of retron-derived spacers from pSBK.134 after 24 hours of induction from their respective promoters. Open circles represent individual biological replicates. **C.** Accumulation of retron-derived spacers when ncRNAs were induced in the order "A" then "B" from pSBK.134. Filled circles represent the mean of four biological replicates (±SEM). **D.** Accumulation of retron-derived spacers when ncRNAs were induced in the order "B" then "A" from pSBK.134. Filled circles represent the mean of four biological replicates (±SEM). **E.** Accumulation of non-retron-derived spacers in cells harboring pSBK.134, in both induction conditions. Filled circles represent the mean of four biological replicates (±SEM). **F.** Schematic of the signal plasmid pSBK.136 used to express ncRNAs "A" and "B", and the recording plasmid used to express Eco1 RT and Cas1 and 2. **g.** Accumulation of retron-derived spacers from pSBK.136 after 24 hours of induction from their respective promoters. Outlier sample determined by Grubbs' test is denoted as a grey "X". Open circles represent individual biological replicates. **h.** Accumulation of retron-derived spacers when ncRNAs were induced in the order "A" then "B" from pSBK.136. Filled circles represent the mean of three biological replicates (±SEM). **i.** Accumulation of retron-derived spacers when ncRNAs were induced in the order "B" then "A" from pSBK.136. Filled circles represent the mean of four biological replicates (±SEM). **j.** Accumulation of non-retron-derived spacers in cells harboring pSBK.136, in both induction conditions. Filled circles represent the mean of four biological replicates (±SEM). **k.** Graphical representation of the ordering rules used to determine order of expression. Array types are illustrated. **l.** Ordering rule analysis of recording experiments with signal plasmid pSBK.134. Open circles are individual biological replicates. **m.** Ordering rule analysis of recording experiments with signal plasmid pSBK.136. Open circles are individual biological replicates. All statistics in Supplementary Table 1.

DISCUSSION

Here, we have described a technique for the recording and reconstruction of transcriptional history in a population of cells. We achieved this by engineering an RNA molecular tag, which is specifically reverse-transcribed to produce a DNA 'receipt' of transcription that is permanently saved in a CRISPR array. We demonstrated the flexibility and potential for continued development of these tools by making the recording retron more efficient with modifications to the structure of the retron ncRNA, and developed a toolkit of barcoded retrons for future application to more complex systems. Beyond this, we investigated the ability of the CRISPR adaptation system to utilize RT-DNA as a prespacer, and discovered that the retron 2'-5' linkage causes a marked difference in the type of spacers acquired. Finally, we used this system to record and reconstruct time-ordered biological events in populations of cells.

We believe that this framework of selective tagging and recording of biological signals in an RNA → DNA → CRISPR direction is a powerful, modular, and extensible method of making temporal recordings in cells. Using only *a priori* ordering rules, we can detect and interpret time-ordered biological signals from a single endpoint sample. We also see this work as a first step toward the use of retrons, and other programmable retroelements, for creative applications beyond their current uses in recombineering and HDR-based genome-editing.

METHODS

Bacterial Strains and Growth Conditions

This work uses the following *E. coli* strains: NEB 5-alpha (NEB C2987), BL21-AI (ThermoFisher C607003), bMS.346, and bSLS.114. bMS.346 was generated from *E. coli* MG1655 by inactivating *exoI* and *recJ* genes with early stop codons as in previous work[33]. Additionally, the *araB*::T7RNAP-*tetA* locus was transferred from BL21-AI by P1 phage transduction[34]. bSLS.114 (which has been used previously[24]) was generated from BL21-AI by deleting the retron Eco1 locus

by lambda Red recombinase mediated insertion of an FRT-flanked chloramphenicol resistance cassette. This cassette was amplified from pKD3[35] with homology arms added to the retron Eco1 locus. This amplicon was electroporated into BL21-AI cells expressing lambda Red genes from pKD46[35], and clones were isolated by selection on chloramphenicol (10 µg/mL) 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. Experimental cultures were grown with shaking in LB broth at 37ºC with appropriate inducers and antibiotics. Inducers and antibiotics were used at the following working concentrations: 2 mg/mL L-arabinose (GoldBio A-300), 1 mM IPTG (GoldBio I2481C), 400 µM erythromycin, 100 ng/mL anhydrotetracycline, 100 µM choline chloride, 1 mM sodium salicylate, 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; used at 10 µg/mL for selection during recombineering). Additional strain information can be found in Supplemental Table 2.

Plasmid Construction

All cloning steps were performed in *E. coli* NEB 5-alpha. pWUR 1+2, containing Cas1 and Cas2 under the expression of a T7lac promoter, was a generous gift from Udi Qimron[12]. Eco1 wildtype ncRNA and Eco1 RT, along with Cas1+2, were cloned into pRSF-DUET (Sigma 71341) to generate pSLS.405. Eco1 variant ncRNA sequences v32 and v35 were cloned into pRSF-DUET along with Cas1+2 to generate pSLS.407 and pSLS.408, respectively. Extended a1/a2 v35 ncRNA expression plasmid pSLS.416 was generated from pSLS.408 by site-directed mutagenesis. Retron Eco1 RT and retron Eco4 RT were cloned into pJKR-O-mphR to generate pSLS.402 and pSLS.400, respectively. pJKR-O-mphR was generated previously[36] (Addgene plasmid # 62570). Barcoded, extended a1/a2 v35 ncRNA expression plasmids pSBK.009-016 were generated from pSLS.416 by site-directed mutagenesis. Wildtype retron Eco4 ncRNA was cloned into pRSF-DUET along with Cas1+2 to generate SLS.419. pSBK.134 and pSBK.136 were

generated in three steps. First, barcoded, extended a1/a2 v35 ncRNA sequences were cloned into the 'Marionette' plasmids pAJM.717, pAJM.718, and pAJM.771. pAJM.717, pAJM.718, and pAJM.771 were gifts from Christopher Voigt[32] (pAJM.717 - Addgene plasmid # 108517 // pAJM.718 - Addgene plasmid # 108519 // pAMJ.771 - Addgene plasmid # 108534). Then, in two steps, two ncRNA expression cassettes (for barcoded ncRNAs "A" and "B") from the Marionette plasmids were cloned into pSol-TSF (Lucigen F843213-1) facing in opposite directions. pSBK.079 was generated by cloning the resistance marker AmpR in place of the KanR marker into the plasmid pSLS.425, which was synthesized by Twist biosciences. Additional plasmid information can be found in Supplemental Table 3.

RT-DNA Purification and PAGE Visualization

Retron RT-DNA was expressed in *E. coli* bMS.346 and purified in two steps. First, DNA was extracted from cells using a plasmid midiprep kit (Qiagen 12943). This purified DNA was then treated for 30 minutes at 37C with RNAse A/T1 mix (ThermoFisher EN0551) and, if required, DBR1 (OriGene TP300024) and/or $RecJ_f$ (NEB M0264). This sample was then used as the input for the Zymo Research ssDNA/RNA Clean & Concentrate kit (Zymo D7011). Samples eluted from the ssDNA kit were resolved using TBE-urea PAGE (ThermoFisher EC6885BOX). Gels were stained with SYBR Gold for imaging (ThermoFisher S11494) and imaged on a Bio-Rad Gel Doc imager.

Retron Acquisition Experiments

Cells were transformed sequentially: first with the RT expression plasmid (pSLS.400 or pSLS.402), and second with the ncRNA and Cas1+2 expression plasmid (eg. pSLS.416). For the -RT condition, cells were only transformed with an ncRNA and Cas1+2 expression plasmid (e.g. pSLS.416). For testing acquisition of retron-derived spacers in figures 1e-f, cells with RT, ncRNA, and Cas1+2 expression plasmids were grown overnight (16 hours) in 3 mL LB with antibiotics and

459 inducers IPTG and arabinose, from individual clones on plates. In the morning, 240 uL of overnight
460 culture was diluted into 3 mL fresh media with antibiotics, IPTG, and arabinose and grown for 2
461 hours. After 2 hours, 320 uL of culture was diluted into 3 mL fresh media with antibiotics and
462 erythromycin (no erythromycin was used in the -RT condition) and grown for 8 hours. After 8
463 hours, culture was diluted 1:1000 into 3 mL LB with antibiotics and without inducers and grown
464 overnight (16 hours). In the morning, 25 uL of culture was mixed with 25 uL of water, heated to
465 95C for 5 minutes to lyse cells, cooled, and frozen at -20C for later analysis. For data presented
466 in Figures 2b-d and 3i, cells were grown overnight (16 hours) in 3 mL LB with antibiotics and
467 inducers IPTG and arabinose, from individual clones on plates. In the morning, 240 uL of overnight
468 culture was diluted into 3 mL fresh media with antibiotics, IPTG, and arabinose and grown for 2
469 hours. After 2 hours, 320 uL of culture was diluted into 3 mL fresh media with antibiotics and
470 erythromycin and grown for 2 (rather than 8) hours. At this point, 25 uL of culture was mixed with
471 25 uL of water, heated to 95C for 5 minutes to lyse cells, cooled, and frozen at -20C for later
472 analysis.
473 For the 24-hour time course experiment, the experiment was broken into two halves: the first 9
474 hours, and the final 15 hours. For the entirety of the time course, cells were grown in media with
475 antibiotics and inducers (arabinose, IPTG, and erythromycin). For the first 9-hour samples,
476 cultures were grown starting from single colonies added to 0.5 mL of media. These cultures were
477 sampled every 1.5 hours until hour 9, with 1 mL of media added at hour 3 and 1.5 mL of media
478 added at hour 6. For the final 15-hour samples, 3 mL of media was inoculated with single colonies
479 from plates and grown for 9 hours. Starting at hour 9, samples were taken every 1.5 hours until
480 hour 24. At hour 16.5, 200 uL of culture was diluted into 1.5 mL of fresh media and the experiment
481 continued in the new tube. At hour 21, 1 mL media was added to the culture.

Oligo Prespacer Feeding

For spacer acquisition experiments using exogeneous DNA prespacers (purified RT-DNA or synthetic oligos), cells containing pWUR1+2 were grown overnight from individual colonies on plates. In the morning, 100 uL of overnight culture was diluted into 3 mL LB with antibiotics, IPTG, and arabinose. Cells were grown with inducers for 2 hours. For each electroporation, 1 mL of culture was pelleted and resuspended in water. Cells were washed a second time by pelleting and resuspension, then pelleted one final time and resuspended in 50 uL of prespacer DNA solution at a concentration of 6.25 uM of single-stranded RT-DNA. All wash steps were done using ice cold water, all centrifugation steps were done in a centrifuge chilled to 4C, and samples kept on ice until electroporation was complete. The cell-DNA mixture was transferred to a 1 mm gap cuvette (Bio-Rad 1652089) and electroporated using a Bio-Rad gene pulser set to 1.8 kV and 25 uF with pulse controller at 200 Ohms. After electroporation, cells were recovered in 3 mL of LB without antibiotics for 2 hours. Then, 25 uL of culture was mixed with 25 uL of water, heated to 95C for 5 minutes to lyse cells, cooled, and frozen at -20C for later analysis.

Temporal Recordings

Cells were transformed sequentially, first with pSBK.134 or pSBK.136 and then with pSBK.079. For recording, single colonies were picked from plates and grown overnight in 3 mL of LB with antibiotics and without inducers. In the morning, 150 uL of culture was diluted into 3 mL of LB with antibiotics and appropriate inducers (Fig. 4) and grown for 8 hours. After 8 hours, 60 uL of culture was diluted into 3 mL of LB with appropriate inducers and grown overnight (16 hours). In the morning, 150 uL of culture was diluted into 3 mL of LB with appropriate inducers (for second day of expression) and grown for 8 hours. Samples were collected at this 24-hour timepoint. 25 uL of culture was mixed with 25 uL of water, heated to 95C for 5 minutes to lyse cells, cooled, and frozen at -20C for later analysis. After 8 hours, 60 uL of culture was diluted into 3 mL of LB with appropriate inducers and grown overnight (16 hours). In the morning, 25 uL of culture was mixed

with 25 uL of water, heated to 95C for 5 minutes to lyse cells, cooled, and frozen at -20C for later analysis.

Analysis of Spacer Acquisition

Analysis of spacer acquisition was conducted by sequencing a library of all CRISPR arrays in an experimental population using an Illumina MiSeq instrument. Libraries were created by amplifying a region of the genomic CRISPR array using PCR, then indexed using custom indexing oligos. Up to 192 conditions were run per flow cell. A list of oligo prespacers and primers can be found in Supplemental Table 4.

Processing and Analysis of MiSeq Data

Sequences were analyzed using custom Python software, which will be available on GitHub upon peer-reviewed publication. In brief, newly acquired spacer sequences were extracted from array sequences based on their position between identifiable repeats and compared to preexisting spacers in the array. In this preliminary analysis, metrics were collected including number of expansions in arrays (unexpanded, single, double, and triple expanded) and proportion of each present in the library. Sequenced arrays were sorted into subcategories based on these characteristics (e.g. doubly expanded with first three repeats identifiable) for further analysis. Next, to determine number of retron-derived spacers and the order of spacers in multiply expanded arrays, two different analyses were used: one strict and one lenient. In the strict analysis (used in figures 1, 2, and 3) a retron-derived spacer is defined to be a spacer which contains the 23-base core region of the hypothetical prespacer structure from a given retron (with three mismatches or indels allowed). In the lenient analysis (used in figures 4 and 5) a retron-derived spacer is defined to be a spacer which contains an 11-base region of the hypothetical prespacer consisting of the 7-base barcode region and 2 bases on either side (with one mismatch

or indel allowed). The order of spacers in multiply expanded arrays is then reported (e.g. Leader-NNA) and these data are used to complete the ordering rule analysis.

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 will be available in the NCBI SRA upon peer-reviewed publication.

Code Availability

Custom code to process or analyze data from this study will be made available on github upon peer-reviewed publication.

ACKNOWLEDGEMENTS

Work was supported by funding from the Simons Foundation Autism Research Initiative (SFARI) Bridge to Independence Award Program, the Pew Biomedical Scholars Program, the NIH/NIGMS (1DP2GM140917-01), and the UCSF Program for Breakthrough Biomedical Research. S.L.S. acknowledges additional funding support from the L.K. Whittier Foundation. We thank Kathryn Claiborn for editorial assistance as well as Sierra Lear and Santiago Lopez for comments on the manuscript.

AUTHOR CONTRIBUTIONS

S.L.S. conceived the study with contributions from J.N. and G.M.C. S.B.K. and S.L.S. designed and carried out experiments, with contributions from E.L. and M.G.S. S.B.K. and S.L.S. analyzed data. S.B.K. wrote the manuscript with input from all co-authors.

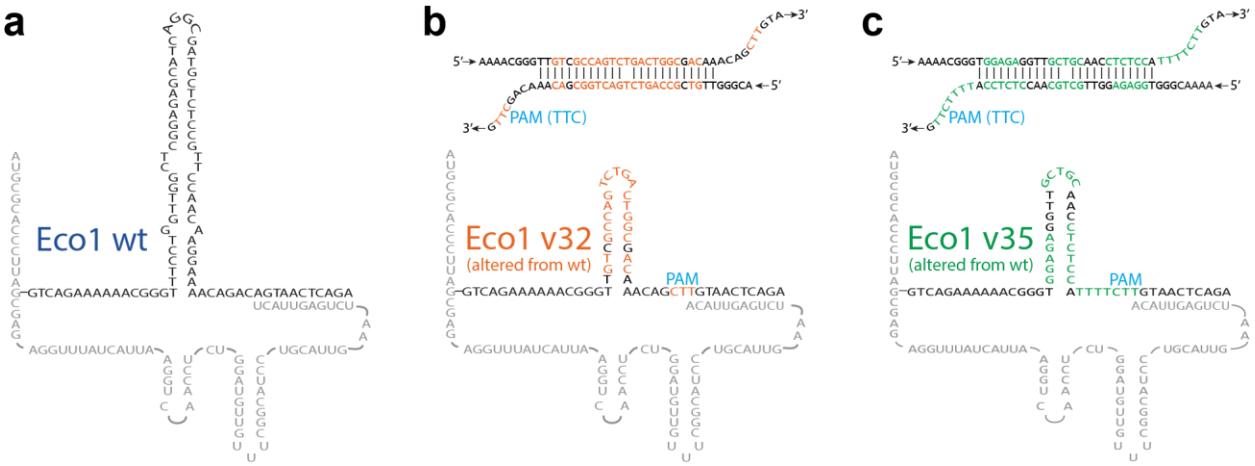
COMPETING INTERESTS

S.L.S., G.M.C, M.G.S., and J.N. are named inventors on a patent application related to the technologies described in this work.

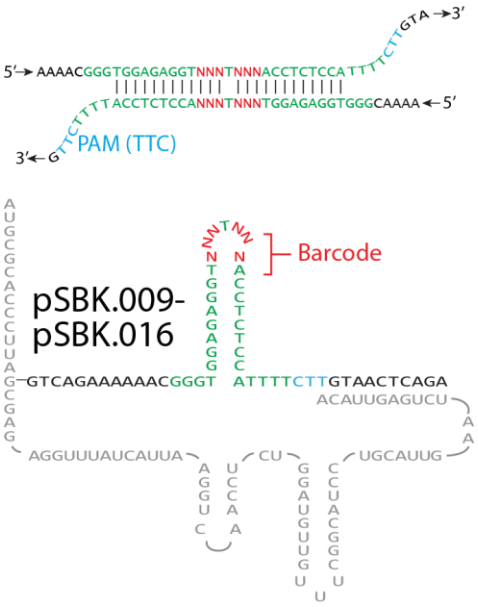
CORRESPONDING AUTHOR

Correspondence to Seth L. Shipman.

Supplementary Figures

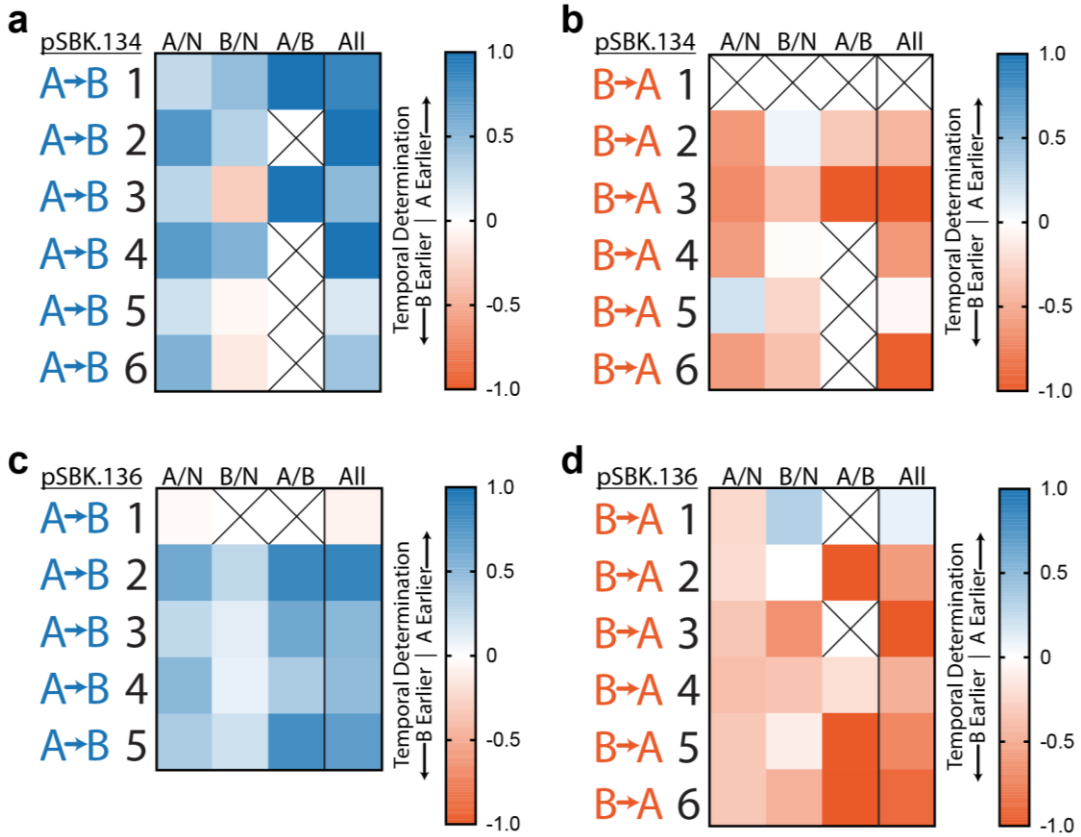


Supplementary Figure 1. Accompaniment to Figure 1. a. Hypothetical Eco1 wild-type ncRNA-linked RT-DNA structure. **b.** Hypothetical Eco1 v32 ncRNA-linked RT-DNA structure and hypothetical duplexed RT-DNA prespacer structure. Nucleotides that are altered from wild-type Eco1 are shown in orange. **c.** Hypothetical Eco1 v35 ncRNA-linked RT-DNA structure and hypothetical duplexed RT-DNA prespacer structure. Nucleotides that are altered from wild-type Eco1 are shown in green.



Supplementary Figure 2. Accompaniment to Figure 2. a. Hypothetical barcoded Eco1 v35 ncRNA-linked RT-DNA structure and hypothetical duplexed RT-DNA prespacer structure. Bases used to barcode retrons are shown in red.

Supplementary Figure 3. Accompaniment to Figure 3. a. Hypothetical wild-type Eco4 ncRNA-linked RT-DNA structure. ExoVII-dependent RT-DNA cleavage site is shown as a red slash. **b.** Eco4-derived spacer sequences and orientations. Bases are colored to match Figure 3f. **c.** Proportion of Eco4-derived spacers in each orientation. Open circles are individual biological replicates.



Supplementary Figure 4. Accompaniment to Figure 4. a. Ordering rules for pSBK.134 "A"-before-"B" replicates. The scores for each rule, and the composite score, are shown for each individual replicate. X-containing boxes indicate that no informative arrays, for that particular rule, were present in that replicate. **b.** As in panel (a), ordering rules for pSBK.134 "B"-before-"A" replicates. **c.** As in panel (a), ordering rules for pSBK.136 "A"-before-"B" replicates. **d.** As in panel (a), ordering rules for pSBK.136 "B"-before-"A" replicates.

References:

1. Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628–1628 (2012).

2. 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).

3. Yim, S. S., McBee, R. M., Song, A. M., Huang, Y., Sheth, R. U. & Wang, H. H. Robust direct digital-to-biological data storage in living cells. *Nat Chem Biol* **17**, 246–253 (2021).

4. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, (2016).

5. 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).

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

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

8. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. & Dudoit, S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

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

10. Park, J. *et al.* Recording of elapsed time and temporal information about biological events using Cas9. *Cell* **184**, 1047-1063.e23 (2021).

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

12. 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).

13. Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W. & Doudna, J. A. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528–534 (2014).

14. Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y. & Sorek, R. Bacterial Retrons Function In Anti-Phage Defense. *Cell* **183**, 1551-1561.e12 (2020).

15. Lampson, B. C., Inouye, M. & Inouye, S. Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res* **110**, 491–499 (2005).

16. Bobonis, J. *et al. Bacterial retrons encode tripartite toxin/antitoxin systems*. http://biorxiv.org/lookup/doi/10.1101/2020.06.22.160168 (2020) doi:10.1101/2020.06.22.160168.

17. Simon, A. J., Morrow, B. R. & Ellington, A. D. Retroelement-Based Genome Editing and Evolution. *ACS Synth. Biol.* **7**, 2600–2611 (2018).

18. Schubert, M. G., Goodman, D. B., Wannier, T. M., Kaur, D., Farzadfard, F., Lu, T. K., Shipman, S. L. & Church, G. M. High-throughput functional variant screens via in vivo production of single-stranded DNA. *Proc Natl Acad Sci USA* **118**, e2018181118 (2021).

19. Sharon, E., Chen, S.-A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K. & Fraser, H. B. Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).

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

21. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. & Wang, Y. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* **163**, 840–853 (2015).

22. Lampson, B. C., Sun, J., Hsu, M. Y., Vallejo-Ramirez, J., Inouye, S. & Inouye, M. Reverse transcriptase in a clinical strain of Escherichia coli: production of branched RNA-linked msDNA. *Science* **243**, 1033–1038 (1989).

23. Simon, A. J., Ellington, A. D. & Finkelstein, I. J. Retrons and their applications in genome engineering. *Nucleic Acids Res* **47**, 11007–11019 (2019).

24. Lopez, S. C., Crawford, K. D., Bhattarai-Kline, S. & Shipman, S. L. *Improved architectures for flexible DNA production using retrons across kingdoms of life*. http://biorxiv.org/lookup/doi/10.1101/2021.03.26.437017 (2021) doi:10.1101/2021.03.26.437017.

25. Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A. M. & Fire, A. Z. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016).

26. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *PNAS* **109**, 8884–8889 (2012).

27. Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S. J. J. & Joo, C. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141–145 (2020).

28. Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. *J. Biol. Chem.* **295**, 3403–3414 (2020).

29. Chapman, K. B. & Boeke, J. D. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**, 483–492 (1991).

30. Lim, D. Structure and biosynthesis of unbranched multicopy single-stranded DNA by reverse transcriptase in a clinical Eschehchia coli isolate. *Molecular Microbiology* **6**, 3531–3542 (1992).

31. Jung, H., Liang, J., Jung, Y. & Lim, D. Characterization of cell death in Escherichia coli mediated by XseA, a large subunit of exonuclease VII. *J Microbiol.* **53**, 820–828 (2015).

32. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. *Nat Chem Biol* **15**, 196–204 (2019).

33. Mosberg, J. A., Gregg, C. J., Lajoie, M. J., Wang, H. H. & Church, G. M. Improving Lambda Red Genome Engineering in Escherichia coli via Rational Removal of Endogenous Nucleases. *PLOS ONE* **7**, e44638 (2012).

34. Moore, S. D. Assembling New Escherichia coli Strains by Transduction Using Phage P1. in *Strain Engineering: Methods and Protocols* (ed. Williams, J. A.) 155–169 (Humana Press, 2011). doi:10.1007/978-1-61779-197-0_10.

35. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *PNAS* **97**, 6640–6645 (2000).

36. Rogers, J. K., Guzman, C. D., Taylor, N. D., Raman, S., Anderson, K. & Church, G. M. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research* **43**, 7648–7660 (2015).