Supplementary Materials for

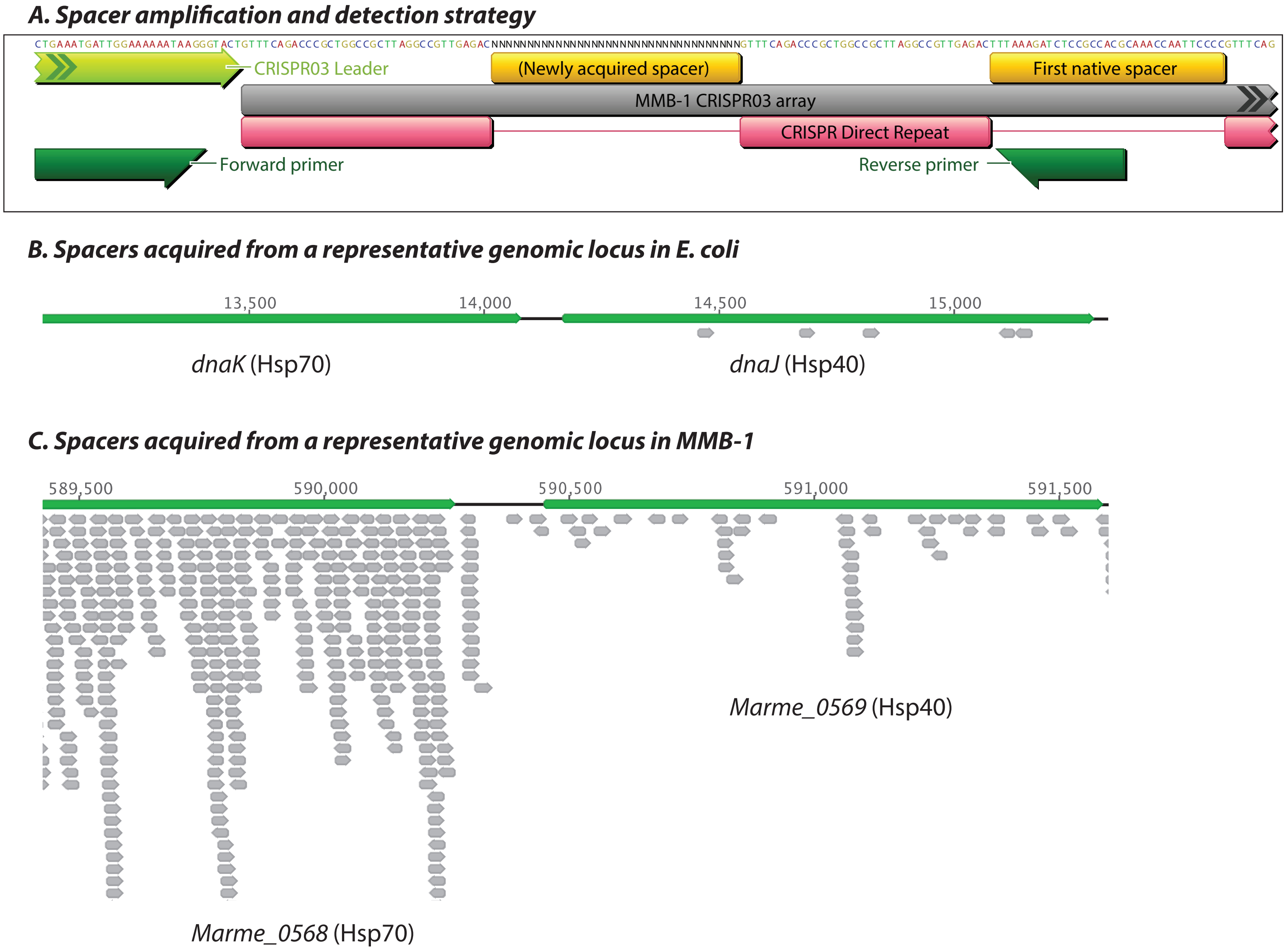
Direct CRISPR spacer acquisition from RNA by a natural reverse-transcriptase-Cas1 fusion protein

Sukrit Silas, Georg Mohr, David J. Sidote, Laura M. Markham, Antonio Sanchez-Amat, Devaki Bhaya, Alan M. Lambowitz, Andrew Z. Fire

correspondence to: afire@stanford.edu, lambowitz@austin.utexas.edu

This PDF file includes:

Figs. S1 to S10
Tables S1 and S2



A. Spacer amplification and detection strategy

CTGAAATGATTGGAAAAAATAAGGGTACTGTTTCAGACCCGCTGGCCGCTTAGGCCGTTGAGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTTTCAGACCCGCTGGCCGCTTAGGCCGTTGAGACTTTAAAGATCTCCGCCACGCAAACCAATTCCCCGTTTCAG
CRISPR03 Leader
(Newly acquired spacer)
First native spacer
MMB-1 CRISPR03 array
CRISPR Direct Repeat
Forward primer
Reverse primer

B. Spacers acquired from a representative genomic locus in E. coli

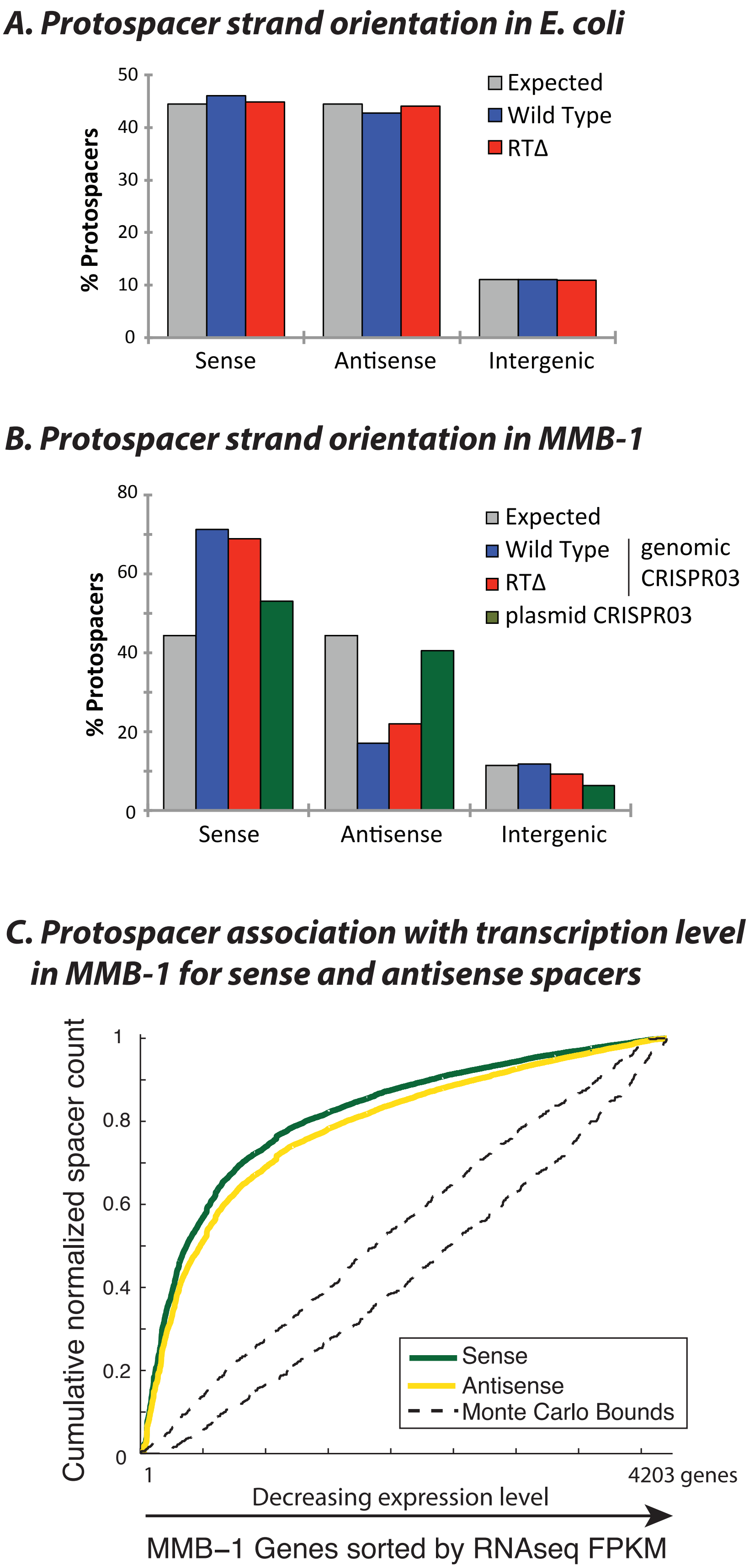
C. Spacers acquired from a representative genomic locus in MMB-1

D. Total number of genome and plasmid derived spacer acquisition events in all experiments with wild-type RT-Cas1 in this study

Experiment	Genomic	Plasmid
ectopic *E. coli* assay	2836	475
over-expression in MMB-1	84538	3096

Fig. S1. Acquisition of new spacers by wild-type RT-Cas1 in *E. coli* and *M. mediterranea* MMB-1.

(A) Schematic showing the leader-proximal region of an expanded CRISPR03 array amplified by PCR in our spacer-detection assay. The leader sequence was identified by directional RNA sequencing of MMB-1 to determine the polarity of the CRISPR arrays. RNAseq data also confirmed that mature crRNAs with 8-nt 5'-repeat-derived handles (*17*) were being generated. The native spacers in both CRISPR arrays in this system were 34-36 bp long and did not match any other sequence in GenBank. **(B)** and **(C)** Alignments of a subset of newly acquired spacers to a conserved gene pair shows the diversity of acquired spacers in both systems. **(B)** Alignments of a subset of newly acquired spacers from ectopic *E. coli* assays to the *dnaK* and *dnaJ* genes. **(C)** Alignments of a subset of newly acquired spacers from MMB-1 overexpression assays to Marme_0568 and Marme_0569 (*dnaK* and *dnaJ* homologs respectively). Marme_0568 is ~5 fold more highly expressed than Marme_0569 (RNAseq data from this study) and is sampled ~20 times more frequently by the RT-Cas1 spacer acquisition machinery in MMB-1. **(D)** Total counts of newly acquired genomic and plasmid protospacers detected in all experiments with wild-type spacer acquisition components in *E. coli* and MMB-1.



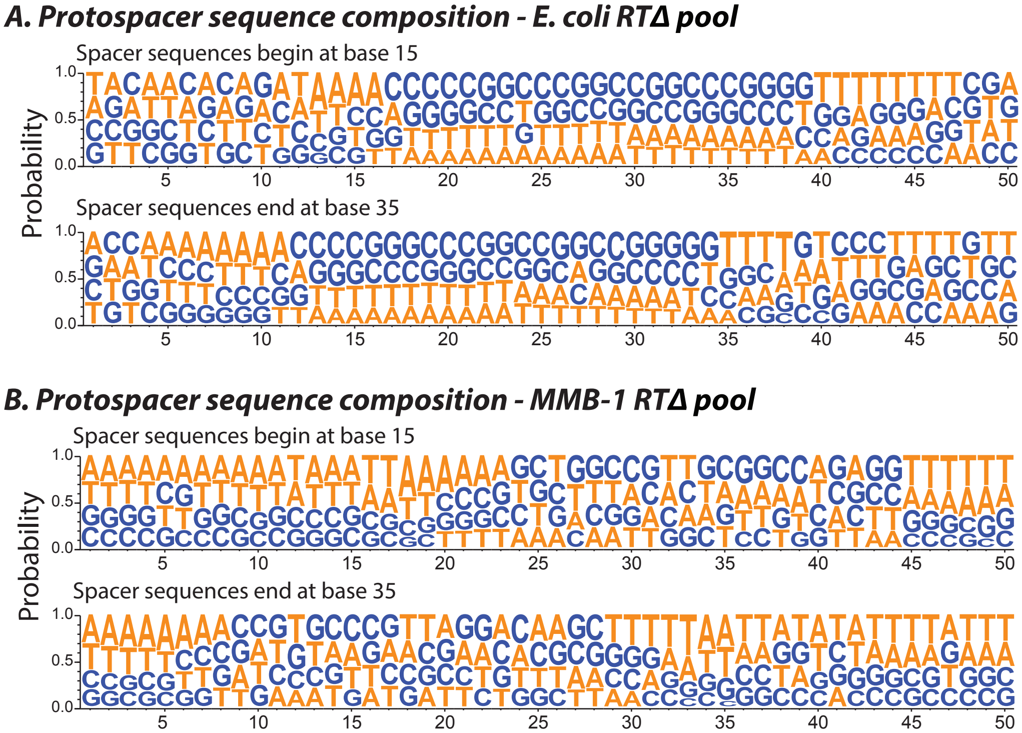
A. Protospacer strand orientation in E. coli

B. Protospacer strand orientation in MMB-1

C. Protospacer association with transcription level in MMB-1 for sense and antisense spacers

Fig. S2. RT-independent sense-strand bias in spacer acquisition by RT-Cas1 in MMB-1 but not *E. coli.*

(A) Percentage of spacers from *E. coli* ectopic assay (data from Fig. 2D) acquired from coding and template strands of *E. coli* genes, and from intergenic regions (note that all regions not annotated as genes are considered intergenic for this analysis; a fraction of these are transcribed, e.g., intergenic sequences within operons). **(B)** Percentage of spacers isolated from the endogenous copy of MMB-1 CRISPR03 (data from Fig. 3C) acquired from sense and antisense strands of MMB-1 genes, and from intergenic regions. The bias for the sense strand persists in the RTΔ-Cas1 acquired spacer pool. The larger dataset of spacers isolated from the plasmid-supplied copy of CRISPR03 (data from Fig. S7C) exhibits a less pronounced bias for the coding strand; these data were collected using a modified spacer detection protocol for transconjugants with plasmid copies of CRISPR03 (see *Materials and Methods*). **(C)** Cumulative distribution of spacers among MMB-1 genes sorted by RNAseq FPKM (RNAseq data from Fig. 3E), with most highly expressed genes listed first (note that these expression profiles were obtained from different MMB-1 transconjugants than Fig. 3E). Wild-type RT-Cas1-acquired spacers isolated from plasmid copies of CRISPR03 (data from Fig. S7C) were split into two pools: 43,766 spacers mapping to the sense strand of MMB-1 genes, and 32,573 spacers mapping to the antisense strand. Monte Carlo bounds were calculated as in Figs. 2F, 3E.



A. Protospacer sequence composition - E. coli RTΔ pool

Spacer sequences begin at base 15

Spacer sequences end at base 35

B. Protospacer sequence composition - MMB-1 RTΔ pool

Spacer sequences begin at base 15

Spacer sequences end at base 35

Fig. S3. Protospacer sequence composition for RTΔ constructs.

Nucleotide probabilities at each position along the protospacers acquired by the RTΔ version of RT-Cas1 in **(A)** *E. coli*, and **(B)** MMB-1, including 15 bp of flanking sequence on each side. Due to varying protospacer lengths, two panels are shown with spacer 5’ and 3’ ends anchored at positions 15 and 35, respectively.

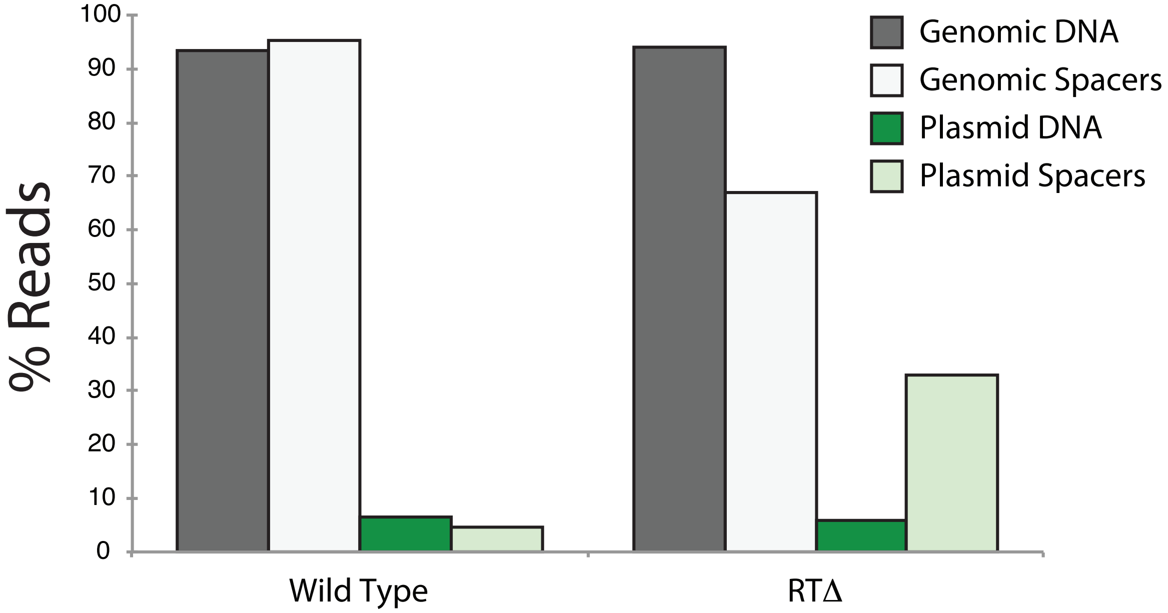


Fig. S4. Proportion of genome and plasmid derived spacers in MMB-1.

A total of 497 spacers mapping to the MMB-1 genome, and 24 to the pKT230 expression vector were recovered in experiments with MMB-1 strains where wild-type RT-Cas1 associated genes were overexpressed. We sequenced DNA from one such transconjugant using Nextera technology (Illumina, Inc.) to measure the plasmid copy number and observed no enrichment for plasmid-derived spacers. Upon deletion of the RT domain of RT-Cas1, Nextera profiling of total DNA revealed that the plasmid copy number had remained unchanged, but the proportion of plasmid-derived spacers had increased 6-fold from 4.6% to 33% (369 spacers mapping to the MMB-1 genome and 181 to the pKT230 expression vector). In contrast, spacer acquisition by the native *E. coli* Cas1/Cas2 complex is 100-1000x biased towards plasmid DNA (*52*).

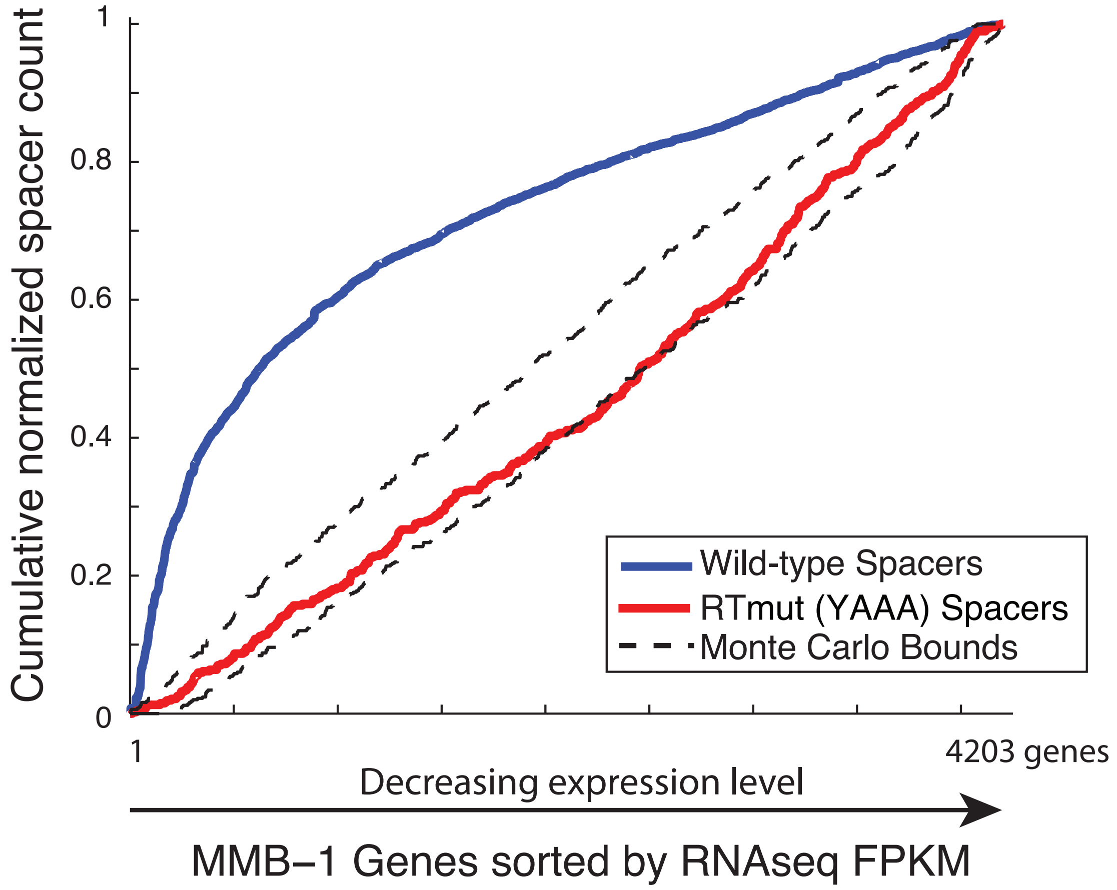
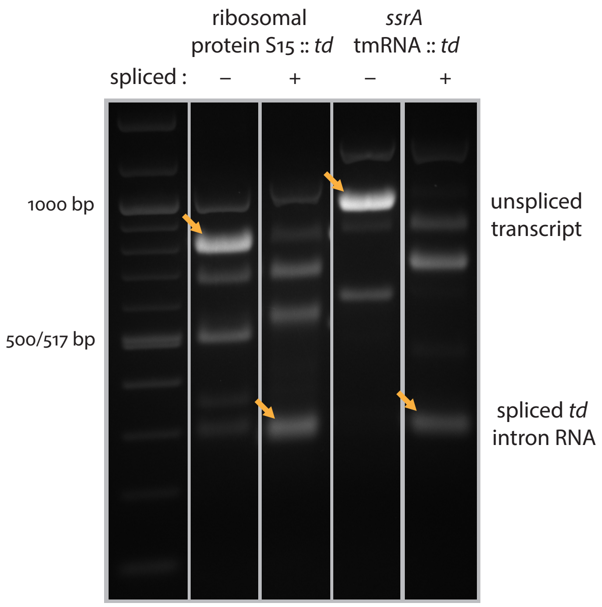


Fig. S5. Protospacer association with transcription level for RT active site mutant. Cumulative distribution of spacers among MMB-1 genes sorted by RNAseq FPKM (RNAseq data from Fig. 3E), with most highly expressed genes listed first (note that these expression profiles were obtained from different MMB-1 transconjugants and growth conditions than in Fig. 3E, in particular a lower incubation temperature: 23°C). 3,631 wild-type RT-Cas1, and 472 RT active site mutant (YAAA)-acquired spacers isolated from plasmid copies of CRISPR03 mapping to MMB-1 genes are included. Monte Carlo bounds were calculated as in Figs. 2F, 3E.

A. td intron splicing in vitro



B. Estimation of td intron splicing efficiency in vivo

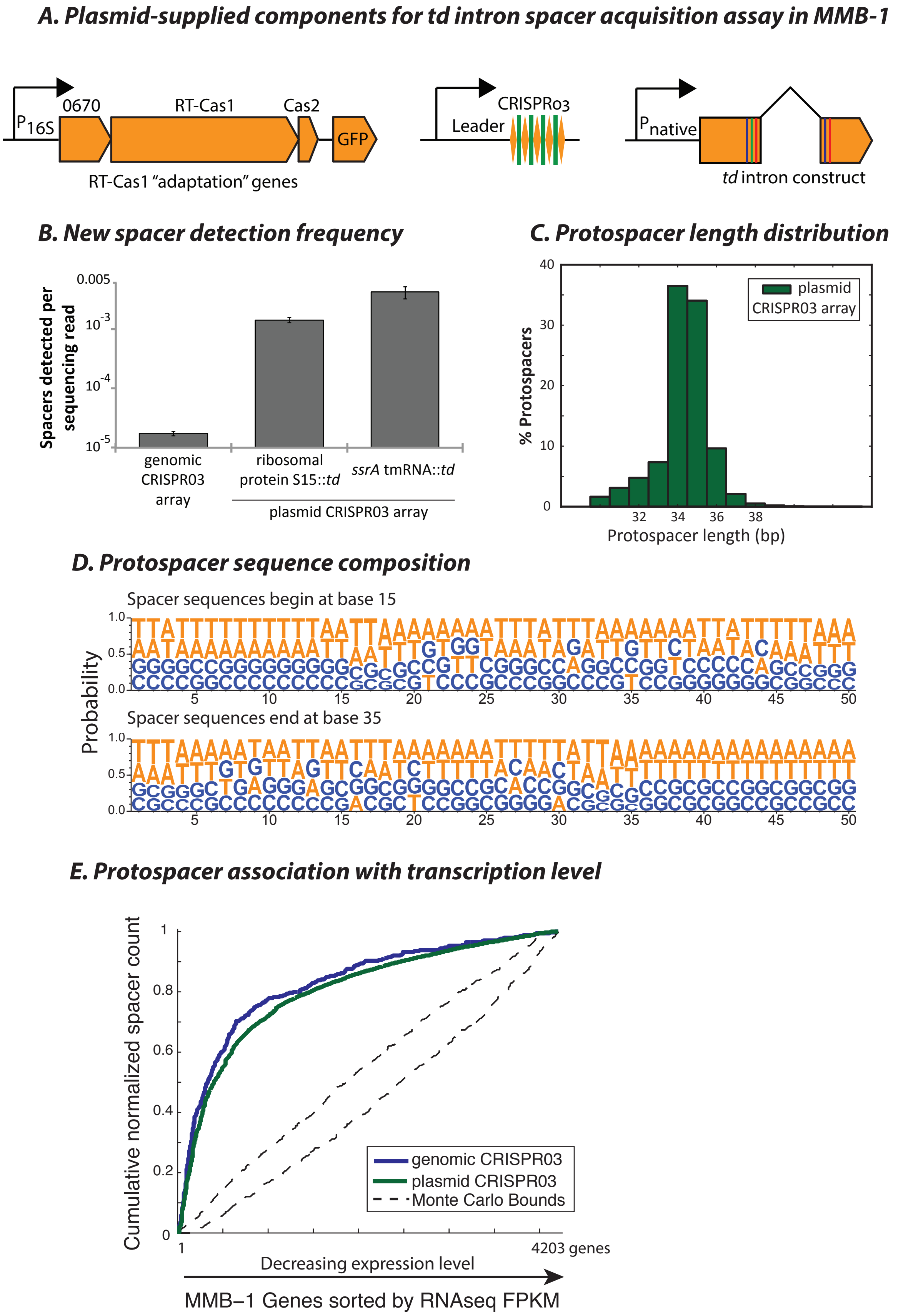
Construct	Spliced	Genomic	% Spliced
ribosomal protein S15 :: *td* - 1	5228	11945	30.44
ribosomal protein S15 :: *td* - 2	3152	6667	32.10
ssrA tmRNA :: *td* - 1	867	4445	16.32
ssrA tmRNA :: *td* - 2	2264	9939	18.55

C. Estimation of spliced sequence present in DNA form

Construct	Spliced	Genomic
ribosomal protein S15 :: *td* - 1	0	6 x 10^6
ribosomal protein S15 :: *td* - 2	0	5.9 x 10^6
ssrA tmRNA :: *td* - 1	0	4.3 x 10^6
ssrA tmRNA :: *td* - 2	0	5.4 x 10^6

Fig. S6. Verification of *td* intron splicing.

(A) Electrophoresis of spliced and unspliced *in vitro* transcripts from *td* intron containing copies of the MMB-1 ribosomal protein S15 and *ssrA* tmRNA genes shows efficient splicing activity. All lanes have been cropped and placed together from the same gel. **(B)** Numbers of reads of spliced and unspliced transcripts in MMB-1 clones obtained from two independent conjugations (denoted 1 and 2) per construct, as determined by RT-PCR and high-throughput sequencing. **(C)** Numbers of reads from targeted DNA sequencing analyses of the same bacterial cultures used in (B) to empirically determine whether *td* exon-exon junctions are present in DNA form outside of the CRISPR locus (see *Materials and Methods*).



A. Plasmid-supplied components for td intron spacer acquisition assay in MMB-1

B. New spacer detection frequency

C. Protospacer length distribution

D. Protospacer sequence composition

E. Protospacer association with transcription level

Fig. S7. RT-Cas1 mediated spacer acquisition into plasmid copies of CRISPR03 in MMB-1. **(A)** Gene arrangement of MMB-1 expression constructs. To demonstrate spacer acquisition from RNA, a self-splicing *td* intron was inserted within plasmid copies of two genes that were frequently sampled by the spacer acquisition machinery – the gene encoding ribosomal protein S15, and the *ssrA* gene encoding tmRNA. The unstructured "mRNA like domain" of the tmRNA was chosen as it was highly over-represented in our initial spacer pools. Bases that were mutated to provide flanking exon sequences favorable for *td* intron splicing are depicted as colored bars within the exons of the intron-containing construct. **(B)** Spacer detection frequency from plasmid-encoded CRISPR03 arrays using a modified spacer detection protocol (see *Materials and Methods*), as compared with spacer acquisition into the endogenous CRISPR03 array (data for the latter redrawn from Fig. 3B). Bars indicate values of two biological replicates for each *td* intron-containing construct. **(C)** Histogram showing normalized counts of MMB-1 protospacers isolated from plasmid copies of CRISPR03, distributed by mappable length. Pooled data from several experiments are presented. **(D)** Nucleotide probabilities at each position along the wild-type RT-Cas1-acquired protospacers in (C) including 15 bp of flanking sequence on each side. Due to varying protospacer lengths, two panels are shown with spacer 5' and 3' ends anchored at positions 15 and 35, respectively. **(E)** Cumulative distribution of spacers in (C) among MMB-1 genes sorted by RNAseq FPKM (RNAseq data from Fig. 3E) with most highly expressed genes listed first (note that these expression profiles were obtained from different MMB-1 transconjugants than in Fig. 3E). 77,050 wild-type RT-Cas1-acquired spacers isolated from plasmid copies of CRISPR03 mapping to MMB-1 genes are included and are distributed similarly to the 455 wild-type RT-Cas1 acquired spacers isolated from the endogenous CRISPR03 array (data for the latter redrawn from Fig. 3E). Monte Carlo bounds were calculated as in Figs. 2F, 3E.

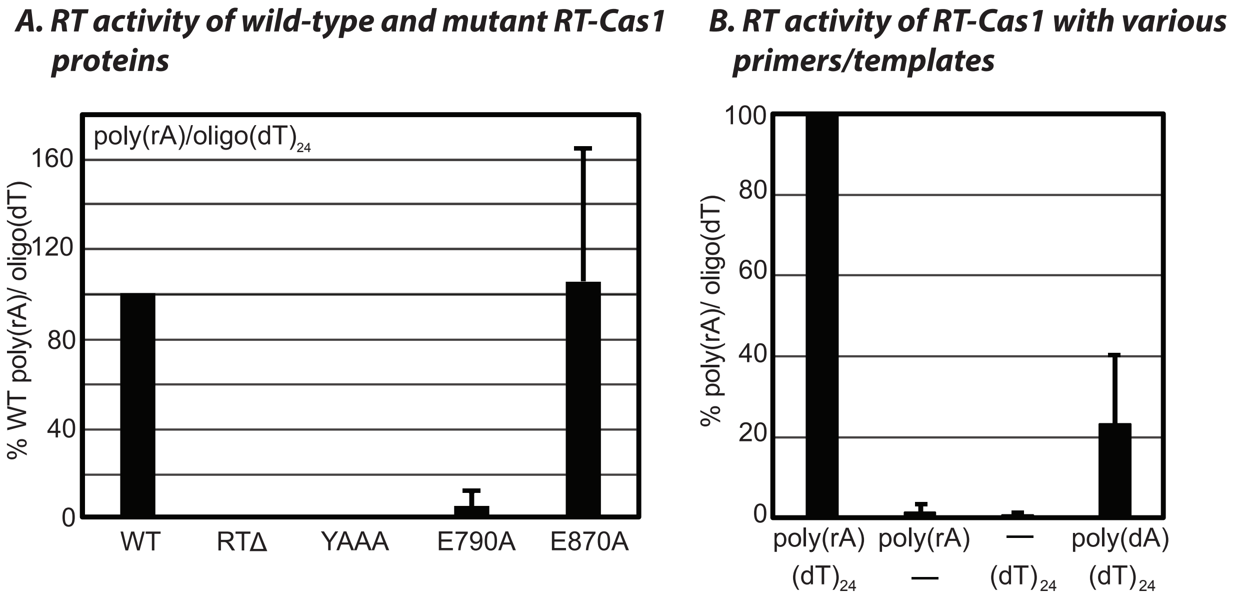


Fig. S8. MMB-1 RT-Cas1 is an active reverse transcriptase *in vitro*.

(A) Wild-type (WT) and mutant RT-Cas1 proteins (1-2 µM final concentration) were assayed for RT activity by polymerization of radiolabeled dTTP in 30-min time courses using the artificial template-primer substrate poly(rA)/oligo(dT)$_{24}$. The bar graphs show RT activity measured as moles of ^{32}P-dTTP polymerized per minute per mole protein, based on the initial rate of ^{32}P-dTTP incorporation and normalized to RT activity of WT RT-Cas1 assayed in parallel. Two independent protein preparations were assayed in duplicate. Wild-type RT-Cas1 protein has RT activity that is abolished by deletion of the RT domain (RTΔ) or mutations at the RT active site (YADD → YAAA at aa pos. 532-533). Note that the two Cas1 active site mutants, E790A and E870A, behave differently in RT assays: E870A has high RT activity comparable to that of the wild-type protein, but E790A has very little activity, suggesting interaction between the RT and Cas1 domains. **(B)** RT assays of WT RT-Cas1 with different template-primer substrates show that the putative RT activity requires both the poly(rA) template and oligo(dT) DNA primer, excluding terminal transferase activity, and that the wild-type protein also has some DNA-dependent DNA polymerase activity when assayed with poly(dA)/oligo(dT)$_{24}$. Error bars in (A) and (B) indicate standard deviations for at least 3 replicates in each case.

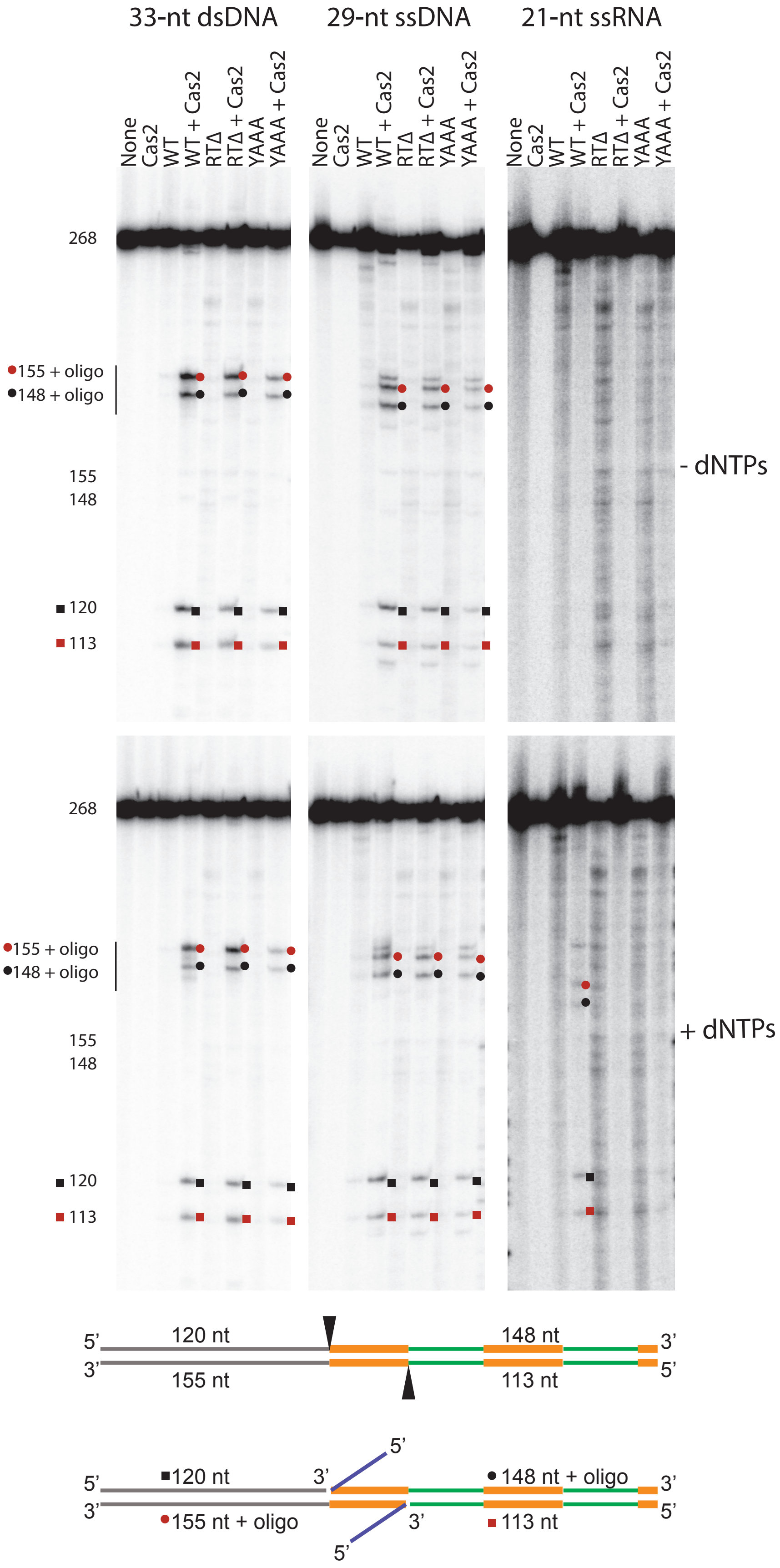

33-nt dsDNA
29-nt ssDNA
21-nt ssRNA
None
Cas2
WT
WT + Cas2
RTΔ
RTΔ + Cas2
YAAA
YAAA + Cas2
268
155 + oligo
148 + oligo
155
148
120
113
- dNTPs
+ dNTPs
5'
3'
120 nt
155 nt
148 nt
113 nt
120 nt
155 nt + oligo
148 nt + oligo
113 nt

Fig. S9. CRISPR DNA cleavage and oligonucleotide ligation *in vitro*.

Wild-type (WT) and mutant RT-Cas1 proteins with or without Cas2 were incubated with the internally labeled 268 bp CRISPR DNA and 33-nt dsDNA (left), 29-nt ssDNA (middle), or 21-nt RNA (right) oligonucleotides in the absence (top panels) or presence (bottom panels) of dNTPs. RT-Cas1 has non-specific nuclease activity indicated by degradation products of the labeled CRISPR DNA in the absence of Cas2. The cleavage of CRISPR DNA and ligation of DNA oligonucleotides requires both Cas1 and Cas2. The RT mutations (RTΔ and YAAA) inhibit ligation of RNA but not DNA oligonucleotides, and dNTPs are required for ligation of RNA but not DNA oligonucleotides (also see Fig. 5). Red and black dots and squares indicate the expected cleavage/ligation products as indicated in the schematic below. A larger band of unknown composition is seen above the 155-nt + oligo product in some lanes. The numbers to the left indicate the sizes of the CRISPR DNA cleavage and ligation products determined from a DNA sequencing ladder run in parallel lanes of the same gel. The schematic at the bottom shows the structure and size of the CRISPR DNA substrate and the cleavage-ligation products, with cleavage sites indicated by arrowheads. The products resulting from ligation of the DNA or RNA oligonucleotide (blue) to 5’ ends of the downstream fragments of both strands are indicated by red and black circles, and the corresponding upstream fragments are indicated by red and black squares.

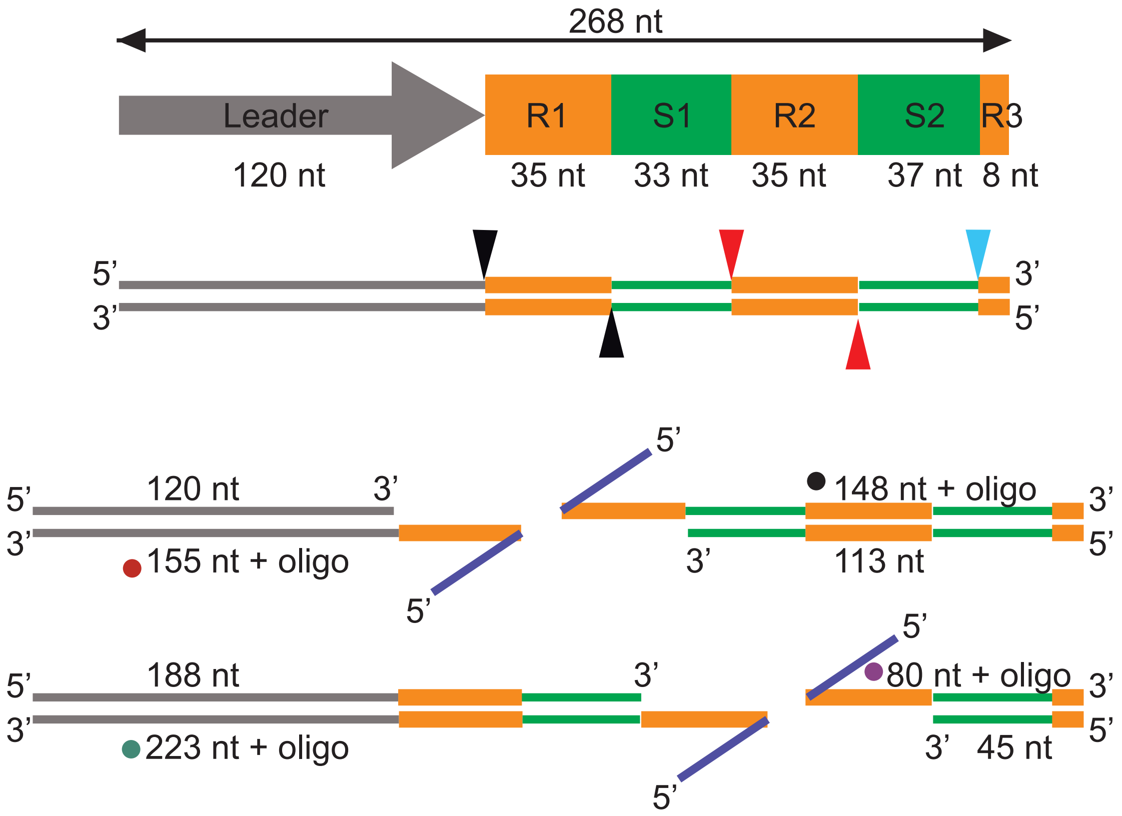


Fig. S10. Schematic showing the products resulting from RT-Cas1 catalyzed cleavage-ligation reactions with the CRISPR DNA substrate.

Cleavage and ligation at the 5’ ends of the first repeat junctions (black) produces 5’ fragments of 120 and 113 nt, and 3’ fragments of 148 and 155 nt plus the ligated oligonucleotides (black and red dots). The same reaction at the 5’ ends of the second repeat produces 5’ fragments of 45 and 188 nt, and 3’ fragments of 80 and 223 nt plus the ligated oligonucleotide (purple and turquoise dots). Labeled products of the expected size for cleavage and ligation at the second repeat junctions can be seen as weak bands in Fig. 5C, lane 4, Fig. 5E, lanes 6, 7, 9, and 10, and Fig. 5F, lanes 6, 8 and 9. Oligonucleotides of various sequences and sizes (ssDNA 19-59 nt; RNA 21-50 nt) can function as substrates for the cleavage/ligation reaction.

Table S1. Exclusive association of RT-Cas1 genes with Type III CRISPR systems in diverse bacteria.

This analysis is based on all the RT-Cas1 fusion protein records that were available prior to 2014 and may not be exhaustive. *(See attached file)*

Table S2. High-throughput sequencing data presented in this study.

(See attached file)