Supplementary information

A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition

Georg Mohr, Sukrit Silas, Jennifer L. Stamos,
Kira S. Makarova, Laura M. Markham, Jun Yao, Patricia Lucas-Elío,
Antonio Sanchez-Amat, Andrew Z. Fire, Eugene V. Koonin, Alan M. Lambowitz

This file includes Supplementary Figures S1 to S7, Table S2 and legends to Data S1, Data S2, and Table S1

Mohr et al. Figure S1

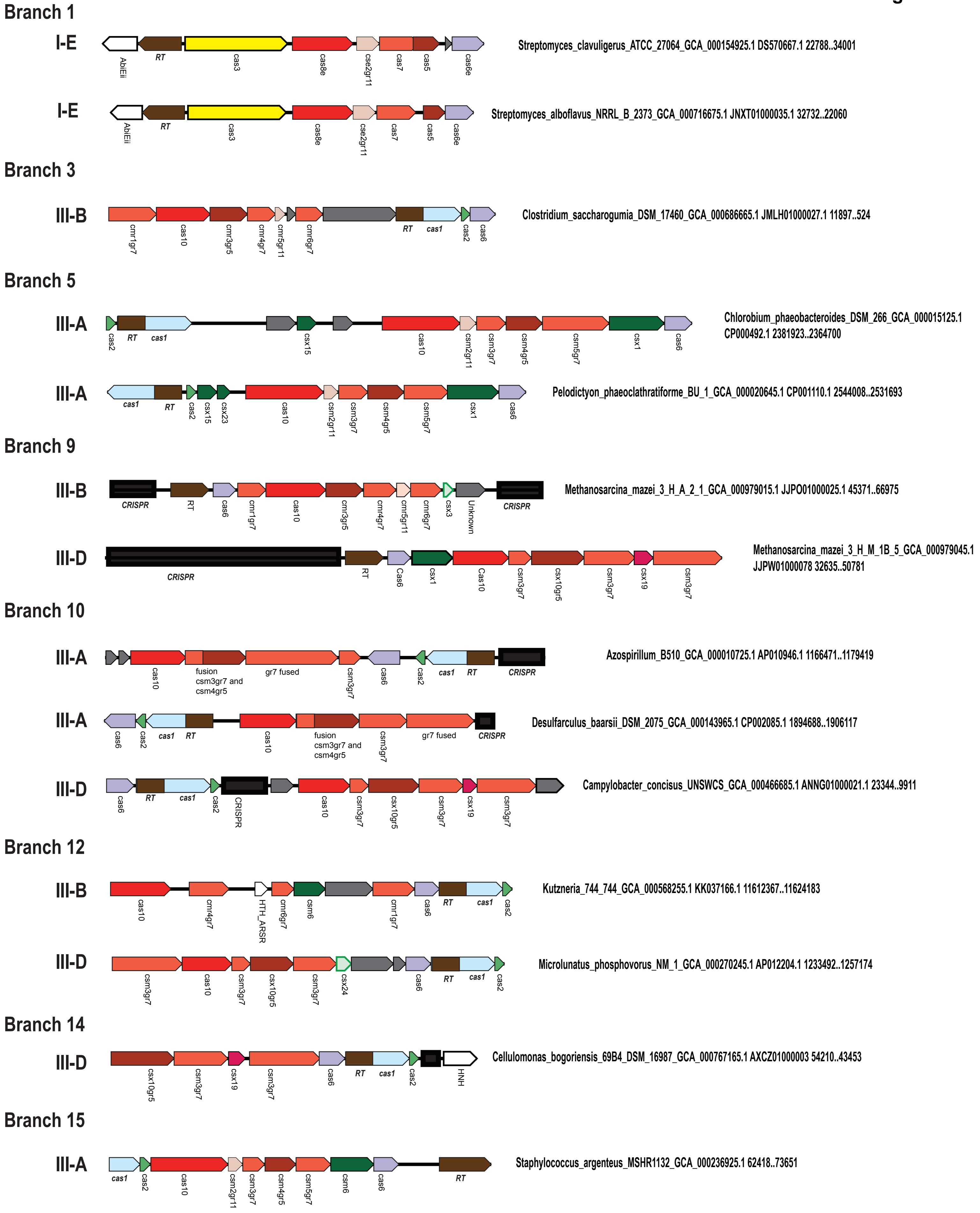

Branch 1
I-E
AbiEii
RT
cas3
cas8e
cse2gr11
cas7
cas5
cas6e
Streptomyces_clavuligerus_ATCC_27064_GCA_000154925.1 DS570667.1 22788..34001
I-E
AbiEii
RT
cas3
cas8e
cse2gr11
cas7
cas5
cas6e
Streptomyces_alboflavus_NRRL_B_2373_GCA_000716675.1 JNXT01000035.1 32732..22060
Branch 3
III-B
cmr1gr7
cas10
cmr3gr5
cmr4gr7
cmr5gr11
cmr6gr7
RT
cas1
cas2
cas6
Clostridium_saccharogumia_DSM_17460_GCA_000686665.1 JMLH01000027.1 11897..524
Branch 5
III-A
cas2
RT
cas1
csx15
cas10
csm2gr11
csm3gr7
csm4gr5
csm5gr7
csx1
cas6
Chlorobium_phaeobacteroides_DSM_266_GCA_000015125.1 CP000492.1 2381923..2364700
III-A
cas1
RT
cas2
csx15
csx23
cas10
csm2gr11
csm3gr7
csm4gr5
csm5gr7
csx1
cas6
Pelodictyon_phaeoclathratiforme_BU_1_GCA_000020645.1 CP001110.1 2544008..2531693
Branch 9
III-B
CRISPR
RT
cas6
cmr1gr7
cas10
cmr3gr5
cmr4gr7
cmr5gr11
cmr6gr7
csx3
Unknown
CRISPR
Methanosarcina_mazei_3_H_A_2_1_GCA_000979015.1 JJPO01000025.1 45371..66975
III-D
CRISPR
RT
Cas6
csx1
Cas10
csm3gr7
csx10gr5
csm3gr7
csx19
csm3gr7
Methanosarcina_mazei_3_H_M_1B_5_GCA_000979045.1 JJPW01000078 32635..50781
Branch 10
III-A
cas10
fusion csm3gr7 and csm4gr5
gr7 fused
csm3gr7
cas6
cas2
cas1
RT
CRISPR
Azospirillum_B510_GCA_000010725.1 AP010946.1 1166471..1179419
III-A
cas6
cas2
cas1
RT
cas10
fusion csm3gr7 and csm4gr5
csm3gr7
gr7 fused
CRISPR
Desulfarculus_baarsii_DSM_2075_GCA_000143965.1 CP002085.1 1894688..1906117
III-D
cas6
RT
cas1
cas2
CRISPR
cas10
csm3gr7
csx10gr5
csm3gr7
csx19
csm3gr7
Campylobacter_concisus_UNSWCS_GCA_000466685.1 ANNG01000021.1 23344..9911
Branch 12
III-B
cas10
cmr4gr7
HTH_ARSR
cmr6gr7
csm6
cmr1gr7
cas6
RT
cas1
cas2
Kutzneria_744_744_GCA_000568255.1 KK037166.1 11612367..11624183
III-D
csm3gr7
cas10
csm3gr7
csx10gr5
csm3gr7
csx24
cas6
RT
cas1
cas2
Microlunatus_phosphovorus_NM_1_GCA_000270245.1 AP012204.1 1233492..1257174
Branch 14
III-D
csx10gr5
csm3gr7
csx19
csm3gr7
cas6
RT
cas1
cas2
HNH
Cellulomonas_bogoriensis_69B4_DSM_16987_GCA_000767165.1 AXCZ01000003 54210..43453
Branch 15
III-A
cas1
cas2
cas10
csm2gr11
csm3gr7
csm4gr5
csm5gr7
csm6
cas6
RT
Staphylococcus_argenteus_MSHR1132_GCA_000236925.1 62418..73651

Figure S1. Gene order in CRISPR RT-containing CRISPR loci, Related to Figure 1.

The Figure shows additional representative examples of gene order in RT-containing CRISPR loci from branches 9 and 10 and from other branches for which gene order is not shown in Figure 1. Group II intron RTs, identified by conserved sequence motifs in the thumb/maturase domain and/or association with group II intron RNA secondary structure, RT fragments, and a branch 2 RT associated with Tn7 transposition machinery were excluded from the analysis. For each locus, species name, genome accession number, and the respective nucleotide coordinates are indicated. Genes are shown roughly to scale; CRISPR arrays are indicated as black rectangles. Homologous genes are color-coded, with the exception of numerous ancillary genes, which are all shown in light green and green outline, and unknown proteins are shown in gray. The gene names largely follow the established nomenclature (Makarova et al., 2015), but the RAMP proteins of groups 5 and 7 are denoted gr5 and gr7, respectively. The CRISPR-Cas system subtype is indicated for the loci encoding the respective effector genes.

Mohr et al. Figure S2

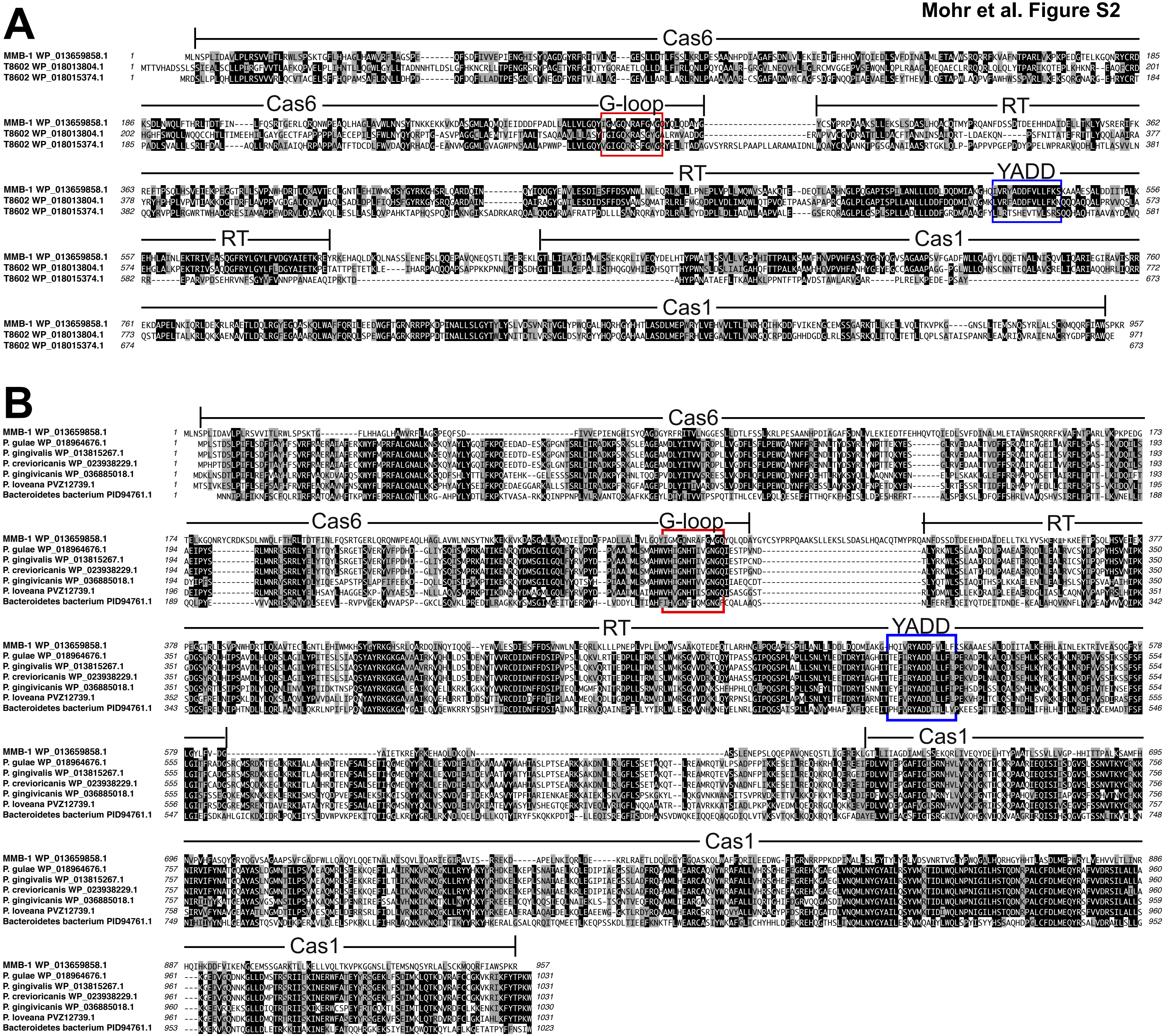


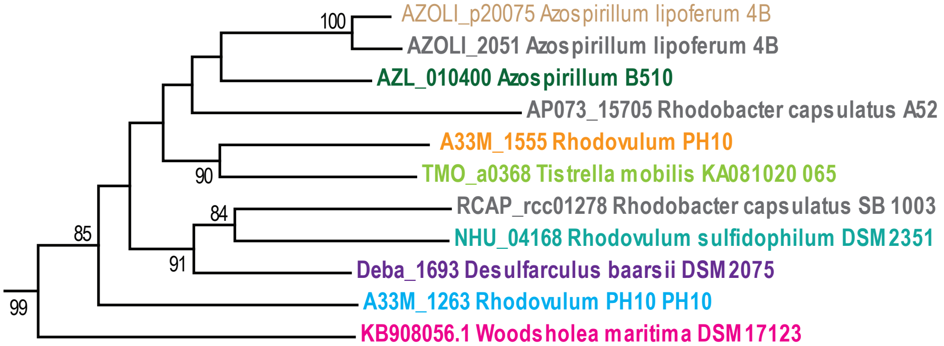
Figure S2. **Cas6-RT-Cas1 fusion proteins, Related to Figure 1.**

(A) Alignment of Cas6-RT-Cas1 fusion proteins from *M. mediterranea* (Top, WP_013659858.1) and *Tredinibacter tunerae* T8602 (WP_028885416.1, WP_028881221.1). The alignment was generated with CLUSTAL. Identical amino acids are shown in white with black background and similar amino acids are shown with grey background. Protein domains are demarcated with black lines. The conserved G-loop is indicated with a red box, and the Y/FADD motif 5 of the RT is indicated with a blue box. (B) Alignment of Cas6-RT-Cas1 fusion proteins from *M. mediterranea* (Top, WP_013659858.1) and *Porphyronomas sp*. (*P. gulae* WP_018964676.1, *P. gingivalis* WP_013815267.1, *P. crevioricanis* WP_023938229.1, *P. gingivicanis* WP_036885018.1, *P. loveana* PVZ12739.1, and Bacteroidetes bacterium PID94761.1) The alignment was generated with JNET and refined manually.

Mohr et al. Figure S3

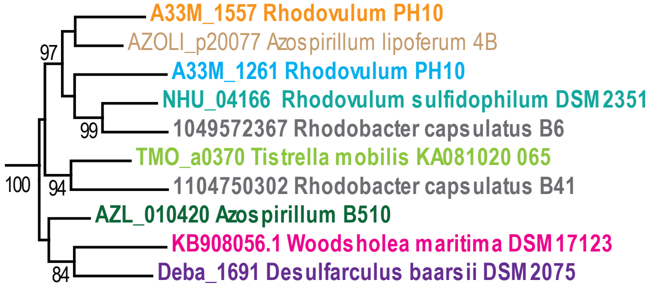
Cas6

Branch 10 subtree


100
AZOLI_p20075 Azospirillum lipoferum 4B
AZOLI_2051 Azospirillum lipoferum 4B
AZL_010400 Azospirillum B510
AP073_15705 Rhodobacter capsulatus A52
A33M_1555 Rhodovulum PH10
90
TMO_a0368 Tistrella mobilis KA081020 065
84
RCAP_rcc01278 Rhodobacter capsulatus SB 1003
NHU_04168 Rhodovulum sulfidophilum DSM2351
85
91
Deba_1693 Desulfarculus baarsii DSM2075
A33M_1263 Rhodovulum PH10 PH10
99
KB908056.1 Woodsholea maritima DSM17123

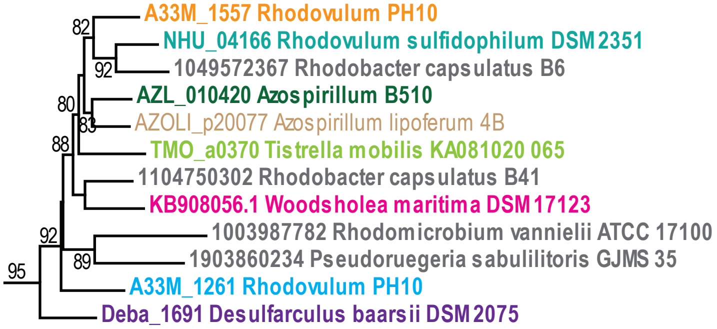
This work

Cas1


A33M_1557 Rhodovulum PH10
AZOLI_p20077 Azospirillum lipoferum 4B
97
A33M_1261 Rhodovulum PH10
NHU_04166 Rhodovulum sulfidophilum DSM2351
99
1049572367 Rhodobacter capsulatus B6
TMO_a0370 Tistrella mobilis KA081020 065
100
94
1104750302 Rhodobacter capsulatus B41
AZL_010420 Azospirillum B510
KB908056.1 Woodsholea maritima DSM17123
84
Deba_1691 Desulfarculus baarsii DSM2075

Silas et al, 2017

RT


A33M_1557 Rhodovulum PH10
82
NHU_04166 Rhodovulum sulfidophilum DSM2351
92
1049572367 Rhodobacter capsulatus B6
AZL_010420 Azospirillum B510
80
83
AZOLI_p20077 Azospirillum lipoferum 4B
88
TMO_a0370 Tistrella mobilis KA081020 065
1104750302 Rhodobacter capsulatus B41
KB908056.1 Woodsholea maritima DSM17123
92
1003987782 Rhodomicrobium vannielii ATCC 17100
89
1903860234 Pseudoruegeria sabulilitoris GJMS 35
95
A33M_1261 Rhodovulum PH10
Deba_1691 Desulfarculus baarsii DSM2075

Silas et al, 2017

Figure S3. **Parallel evolution of Cas6 and RT-Cas1 fusion proteins, Related to Figure 1.** Distinct, strongly supported subtrees extracted from the respective phylogenetic trees are shown for Cas6 (Branch 10) and the RT and Cas1 domains of the corresponding RT-Cas1 proteins (Silas et al., 2017b). The bootstrap support values for the internal branches are shown as percentages. Proteins encoded in the same locus are shown by the same color (gray branches denote the same species but different strains, due to random selection of representatives in the tree and incompleteness of some of the genomic sequences). The figure shows that, despite the absence of the physical fusion of Cas6 with RT-Cas1 and extensive shuffling of Cas6 genes within CRISPR-cas loci, the topologies of the three trees are largely congruent, indicating an early origin of the functional link between Cas6 and RT-Cas1.

Mohr et al. Figure S4

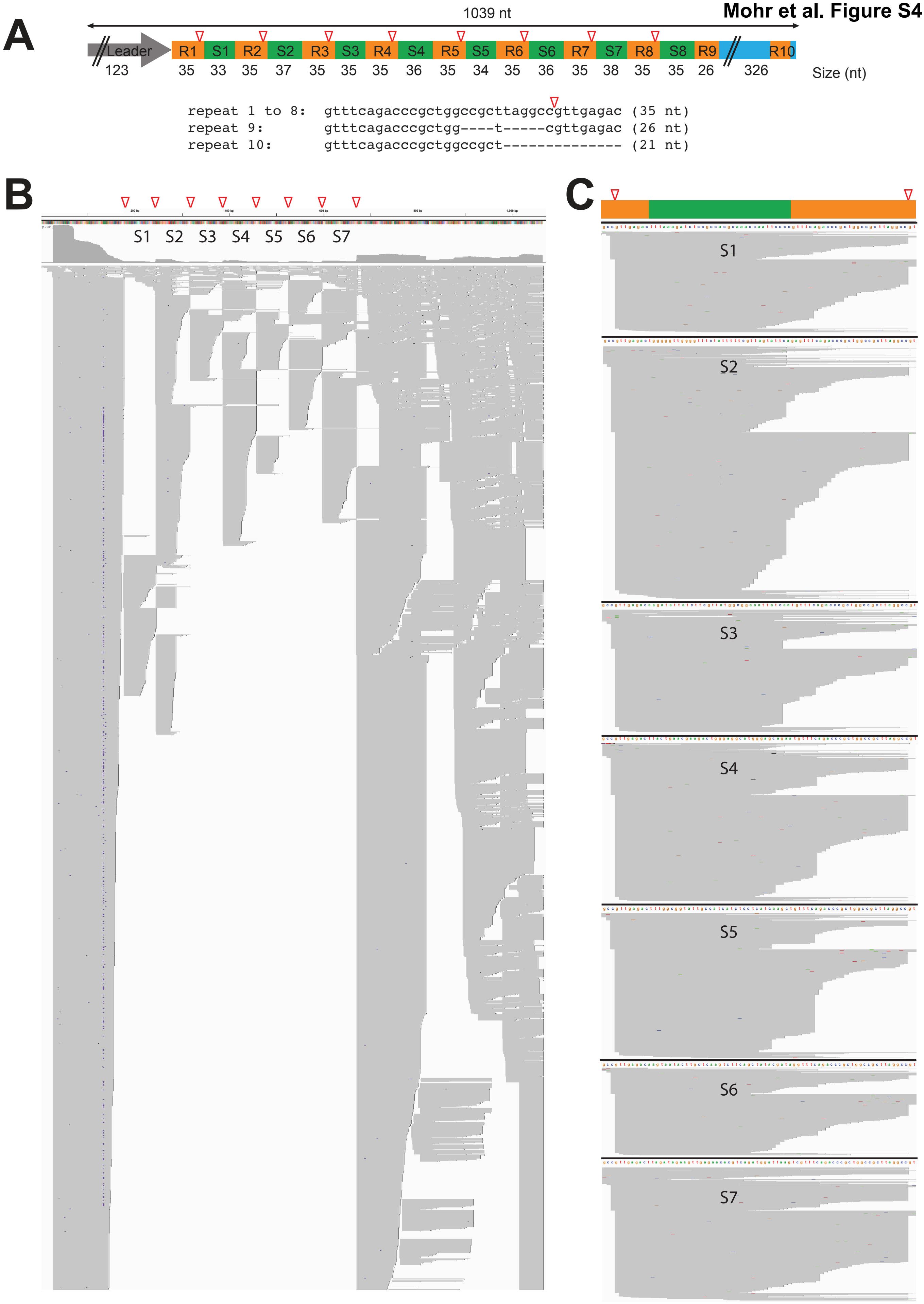


Figure S4. RNA-seq of MMB-1 CRISPR RNAs Processed by Cas6-RT-Cas1 *In Vitro*, Related to Figure 2.

(A) Schematic showing the full-length MMB-1 CRISPR03 array transcript used as a substrate for assaying Cas6 activity (redrawn from Figure 2). *In vitro* transcribed CRISPR03 RNA was incubated with WT Cas6-RT-Cas1 protein at 37°C for 1 hr. High-throughput sequencing libraries were prepared from purified reaction products using the TGIRT-seq protocol, as described (Nottingham et al., 2016). (B) RNA-seq reads aligned to the CRISPR03 sequence. Red triangles indicate matching 5’ and 3’ ends for the first 8 full CRISPR repeats. (C) Close up of reads mapped to individual crRNAs. The schematic at the top shows repeats in orange and spacers in green, red triangles indicate upstream and downstream cleavage sites.

Mohr et al. Figure S5

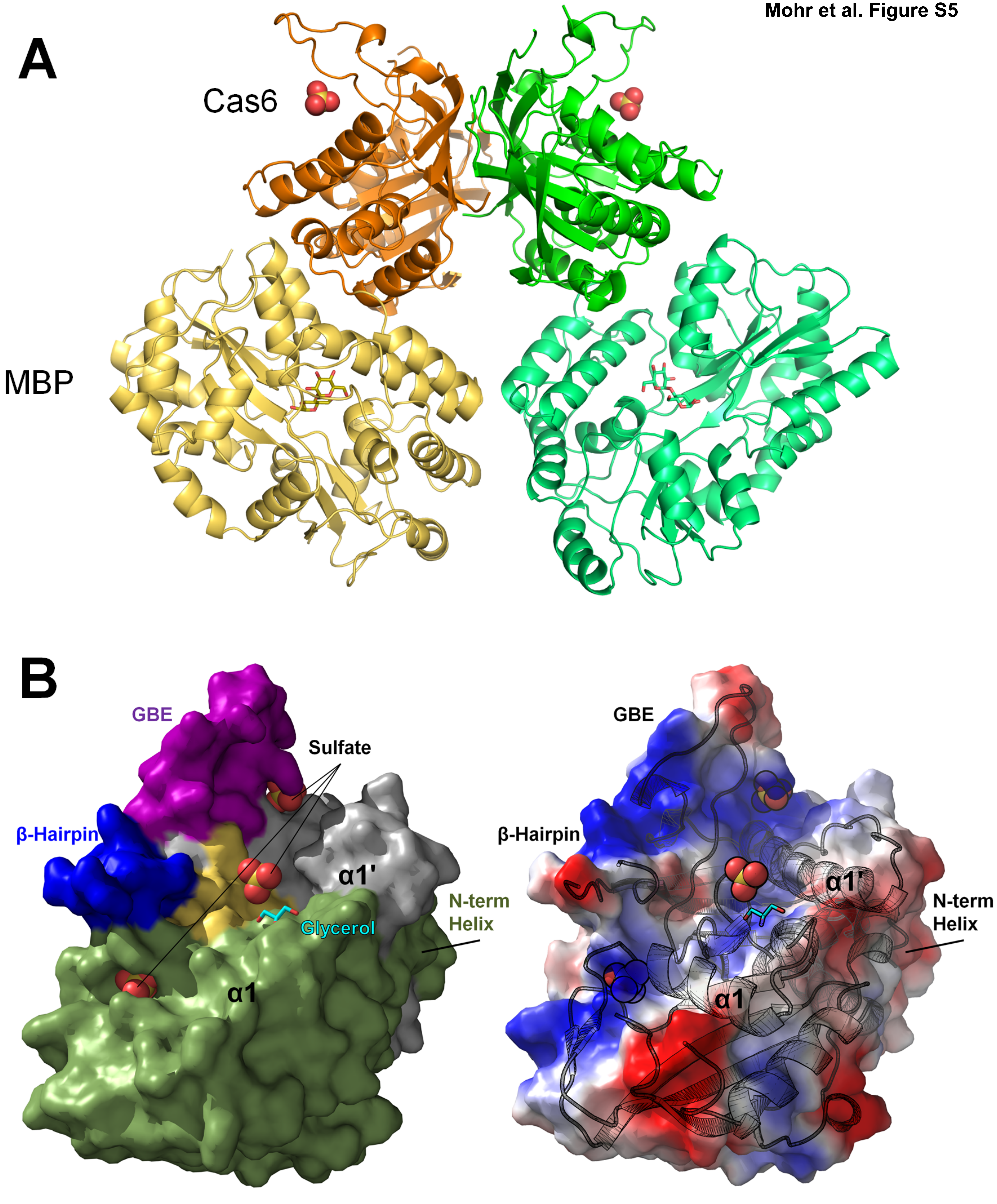


Figure S5. Crystallographic Dimer and Surface Features of the MMB-1 Cas6 crRNA Stem-Loop-Binding Face, Related to Figure 3.

(A) Contents of the asymmetric unit of the crystal structure, showing two MBP-MMB-1 Cas6 fusion proteins in orange and green, with the MBP portions in yellow and light green (maltose, stick form; sulfate ions, sphere representation). (B) Surface representation of the putative crRNA stem-loop binding region of MMB-1 Cas6, displaying a series of ordered sulfates (spheres) and a glycerol molecule (cyan stick) bound along a positively charged trench bordered on one side by the GBE and β-hairpin and on the other side by portions of the α1 and α1'-helices. *left,* colors as in Figure 3B; *right*, electrostatic surface potential, colors as in Figure 3C.

Mohr et al. Figure S6

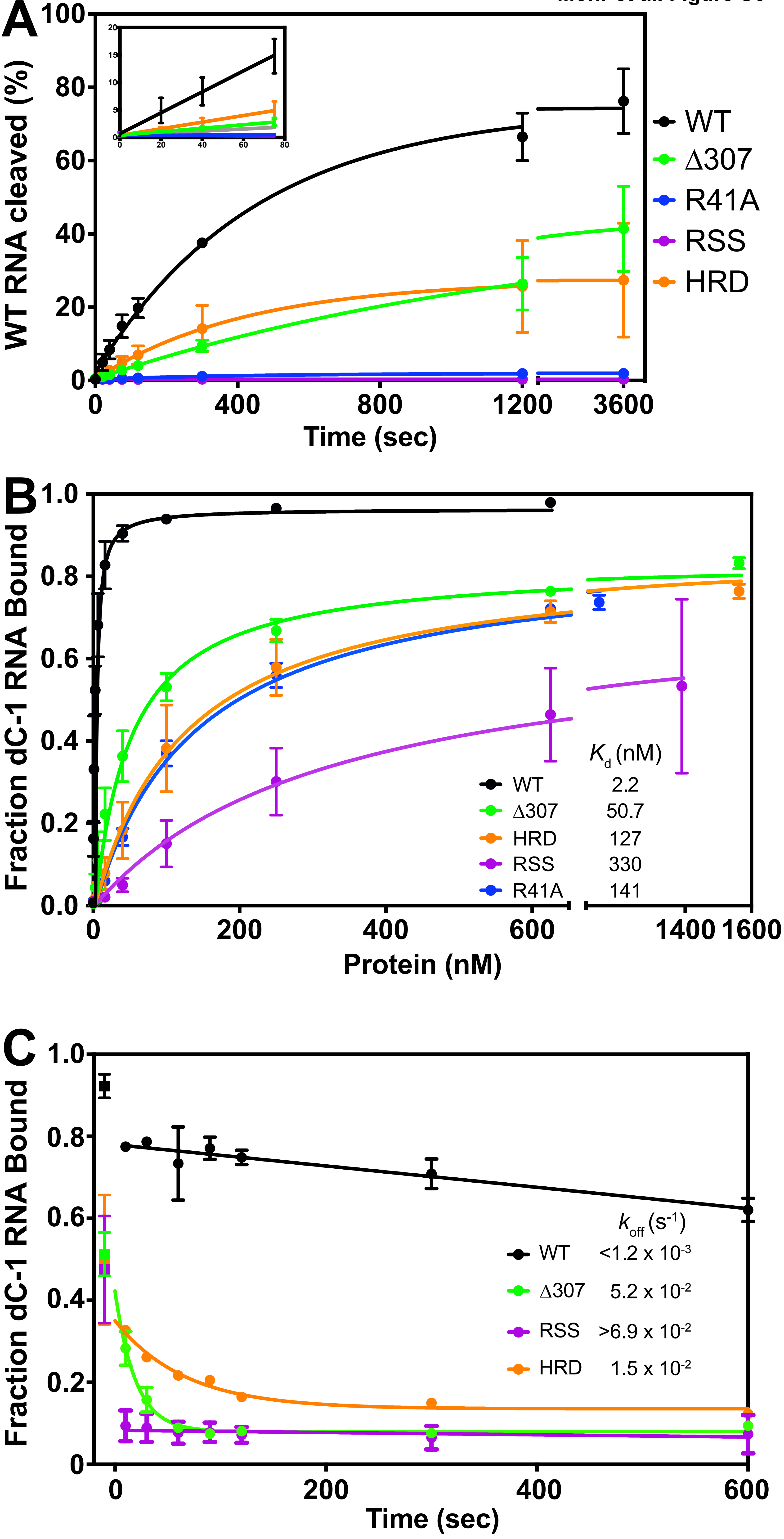


Figure S6. Assays of Cas6 Activity Under Excess RNA Conditions and Binding of Pre-crRNA Repeat by WT and Mutant Cas6-RT-Cas1 Proteins, Related to Figure 4.

(A) Pre-crRNA cleavage assay under excess RNA conditions. A 5' ^{32}P-labeled 35-nt RNA oligonucleotide (1 μM) consisting of the WT CRISPR repeat was incubated with WT and mutant Cas6-RT-Cas1 proteins (40 nM) for times up to 1 hr (3,600 sec), and the products were analyzed on a denaturing 12% polyacrylamide gel. The data were fit to a single exponential model. The inset shows the initial time points with a pseudo-linear fit. (B) Equilibrium binding assay. The non-cleavable 5' ^{32}P-labeled dC-1 RNA oligonucleotide (0.5 nM) was incubated with Cas6-RT-Cas1 mutant proteins at various concentrations (0 to 1,525 nM) for 5 min at 37°C, and protein-bound RNA was measured by nitrocellulose filter binding, as described in STAR Methods. Data were fit to a one-site specific binding model and the K_d values were obtained from the fit, as described in STAR Methods. (C) Complex dissociation (k_{off}) assay. 5' ^{32}P-labeled dC-1 RNA oligonucleotide (2 nM) was incubated with the indicated WT or mutant Cas6-RT-Cas1 protein (1 μM) for 5 min on ice followed by 5 min at 37°C. An initial sample was removed and used to ascertain the amount of protein-bound RNA (squares on the left). The remainder was immediately diluted with an equal volume of buffer containing excess (5 μM) dC-1 oligonucleotide. Samples were taken successively from each reaction for up to 10 min and quantified by nitrocellulose filter binding, as described in STAR Methods. Data for WT Cas6-RT-Cas1 and the RSS mutant were fit to a line, and data for the Δ307-957 and HRD mutants were fit to a single exponential model.

Data shown represent three technical repeats.

Mohr et al. Figure S7

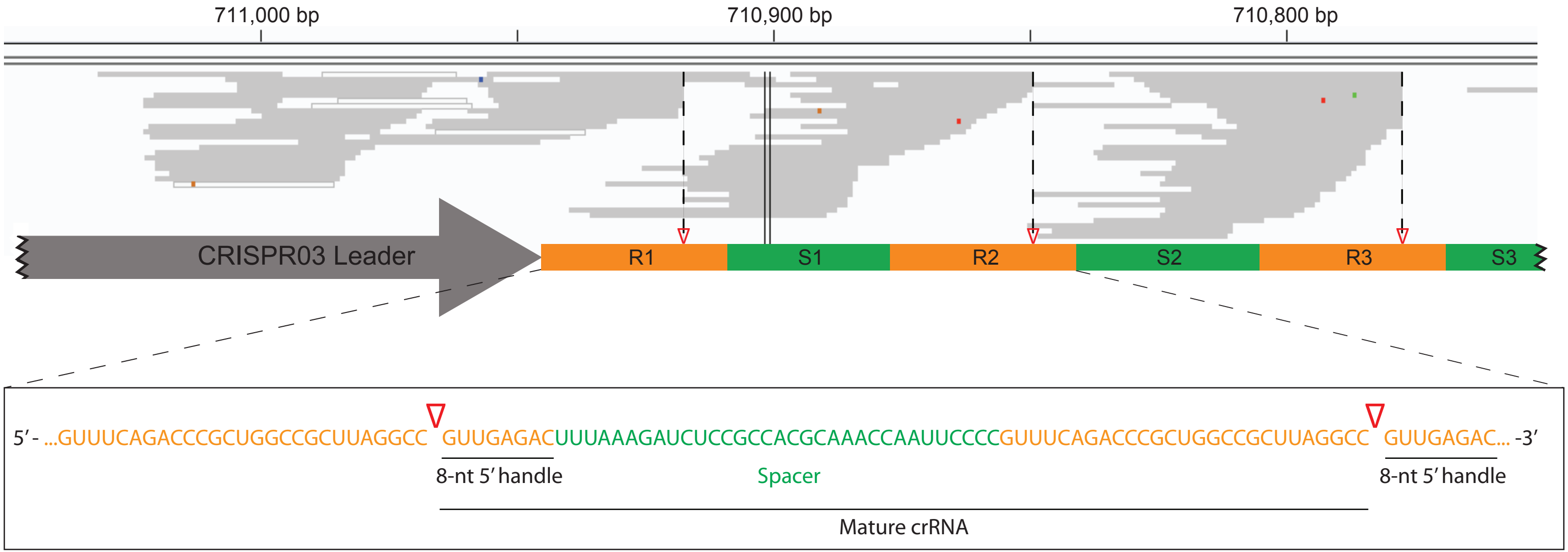

711,000 bp
710,900 bp
710,800 bp
CRISPR03 Leader
R1
S1
R2
S2
R3
S3
5′ - ...GUUUCAGACCCGCUGGCCGCUUAGGCC GUUGAGACUUUAAAGAUCUCCGCCACGCAAACCAAUUCCCCGUUUCAGACCCGCUGGCCGCUUAGGCC GUUGAGAC... -3′
8-nt 5′ handle
Spacer
8-nt 5′ handle
Mature crRNA

Figure S7. RNA-seq of MMB-1 CRISPR RNAs Processed by Cas6-RT-Cas1 *In Vivo*, Related to Figure 5.

The expected pre-crRNA processing site 8-nt upstream of the 3’ end of the CRISPR03 direct repeat sequence was observed in total RNA sequencing libraries of WT MMB-1. RNA-seq reads were mapped to the genomic copy of the CRISPR03 array with CRISPR features indicated in the schematic below. Dotted lines with red triangles indicate Cas6 cleavage sites. The bottom shows the sequence of the first native repeat-spacer-repeat unit of the pre-crRNA (orange), with the experimentally determined cleavage sites marked by red triangles. Colored short horizontal bars in the reads indicate sequence variation from the genomic sequence. The two parallel vertical lines in the center indicate the size of a single nucleotide.

Forward primer for spacer amplification: CGACGCTCTTCCGATCTNNNNNCTGAAATGATTGGAAAAAATAAGG	Andrew Fire	Silas et al. 2016
Reverse primer for spacer amplification: ACTGACGCTAGTGCATCACGTGGCGGAGATCTTTAA	Andrew Fire	Silas et al. 2016
Illumina forward barcoding primer: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCACTGACGCTAGTGCATCA where N8 is index	Andrew Fire	Silas et al. 2016
Illumina reverse barcoding primer: AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT where N8 is index	Andrew Fire	Silas et al. 2016
Pre-adenylated 3’ adapter oligo: /5rApp/NNNNNNAGATCGGAAGAGCACACGTCT/3ddC/	Andrew Fire	Silas et al. 2018
Reverse transcription primer: /5Phos/AGATCGGAAGAGCGTCGTGT/iSp18/CACTCA/iSp18/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Andrew Fire	Silas et al. 2018
Universal PCR primer: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Andrew Fire	Silas et al. 2018
Indexing primers: CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCG where N6 is index	Andrew Fire	Silas et al. 2018
MMB_repeat, rGrUrUrUrCrArGrArCrCrCrGrCrUrGrGrCrCrGrCrUrUrArGrGrCrCrGrUrUrGrArGrArC	This paper	N/A
MMB_repeat_DNA, GTTTCAGACCCGCTGGCCGCTTAGGCCGTTGAGAC	This paper	N/A
MMB_repeat_dC-1, rGrUrUrUrCrArGrArCrCrCrGrCrUrGrGrCrCrGrCrUrUrArGrGrCdCrGrUrUrGrArGrArC	This paper	N/A
MMB_repeat_StemFlip, rGrUrUrUrCrArGrArCrCrCrGrC rUrCrCrGrGrGrCrUrUrArCrCrGrGrGrUrUrGrArGrArC	This paper	N/A
MMB_full_UGU>CUG, rGrUrUrUrCrArGrArCrCrCrGrC rGrGrGrCrCrGrCrUrUrArGrGrCrCrUrCrUrGrArGrArC	This paper	N/A
MMB_repeat_27nt_cleaved, rGrUrUrUrCrArGrArCrCrCrGrCrUrGrGrCrCrGrCrUrUrArGrGrCrC	This paper	N/A
MMB_repeat_8nt_cleaved, rGrUrUrGrArGrArC	This paper	N/A
Spacer, rArGrCrGrUrGrCrGrUrUrCrCrArGrArCrArUrUrCrArGrCrCrCrUrCrUrArGrUrArGrA	This paper	N/A
MMB1crisp5b, CACTCGACCGGAATTATCGACGAA	Integrated DNA Technologies	Silas et al. 2016
MMB1crisp5b+T75’, ATGAATTCGTAATACGACTCACTATAGGGCACTCGACCGGAATTATCGACGAA	Integrated DNA Technologies	This paper
MMB1crisp3full, CTAGCTCTCGAGAGGCCTTCGTCA	Integrated DNA Technologies	This paper
MMB1crisp3, TCTGAAACTCTGAATACTAACGAAAAATAG	Integrated DNA Technologies	Silas et al. 2016

Table S2. Oligonucleotides Used in This Study, Related to KEYRESOURCES Table.

Table S1. Cas6 Sequence Information, Related to Figure 1.

The table includes information on species name, nucleotide and protein accession numbers, Cas6 protein lengths, Cas6 position and branch in the tree, subtype of the associated CRISPR-Cas system and presence of RT domain in the same locus.

Data S1. Cas6 Phylogenetic Tree in the Newick Format, Related to Figure 1.

This tree is schematically shown in Figure 1. The Newick format is compatible with any tree viewing software.

Data S2. Cas6 Genomic Loci Annotation, Related to Figure 1.

The table includes detailed information on *cas6* loci. This information is provided for *cas6* representatives including all those that were used for phylogenetic analysis (see Figure 1, Table S1 and Data S1).