Published as: *Science*. 2010 September 10; 329(5997): 1355–1358.

Sequence- and structure-specific RNA processing by a CRISPR endonuclease

Rachel E. Haurwitz[2,‡], **Martin Jinek**[2,‡], **Blake Wiedenheft**[1,2], **Kaihong Zhou**[1,2], and **Jennifer A. Doudna**[1,2,3,4,*]

[1]Howard Hughes Medical Institute, University of California, Berkeley, CA 94720

[2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

[3]Department of Chemistry, University of California, Berkeley, CA 94720

[4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Abstract

Many prokaryotes contain genomic clustered regularly interspaced short palindromic repeats (CRISPRs) that confer resistance to invasive genetic elements. Central to this immune system is the production of CRISPR-derived RNAs (crRNAs) following transcription of the CRISPR locus. Here we identify the endoribonuclease (Csy4) responsible for pre-crRNA processing in *Pseudomonas aeruginosa*. A 1.8 Å crystal structure of Csy4 in complex with its cognate RNA reveals an unexpected recognition mechanism whereby Csy4 makes sequence-specific interactions in the major groove of the CRISPR repeat stem-loop. Together with electrostatic contacts to the phosphate backbone, these enable Csy4 to selectively bind and cleave pre-crRNAs. The active site of Csy4 comprises two invariant residues, a serine and a histidine. The RNA recognition mechanism identified here explains sequence- and structure-specific processing by a large family of CRISPR-specific endoribonucleases.

In bacteria and archaea, a specific class of short RNAs participates in an adaptive immune pathway that targets viruses and plasmids (1–5). Fragments of foreign DNA are integrated into clustered regularly interspaced short palindromic repeat (CRISPR) loci (6), from which they are transcribed as parts of long RNAs that include a repetitive sequence element characteristic of the host organism. These precursor CRISPR RNAs (pre-crRNAs) are post-transcriptionally processed into 60-nucleotide crRNAs that serve as homing oligonucleotides to prevent propagation of invading viruses or plasmids harboring cognate sequences (1, 7). Each of the eight known CRISPR subtypes (8–12) also includes a set of open reading frames encoding CRISPR-associated (Cas) proteins, which can be difficult to identify or classify due to extreme primary sequence divergence (3, 5, 11). *Pseudomonas aeruginosa* UCBPP-PA14 (hereafter Pa14), a Gram-negative opportunistic pathogen harboring a CRISPR/Cas system of the *Yersinia* subtype, contains six Cas genes flanked by two CRISPR elements (Fig. 1A). While Cas1 is found universally among CRISPR-containing organisms, and Cas3 is evident in most subtypes, Csy1-4 genes are unique to the *Yersinia* subtype. Both CRISPR elements comprise a characteristic arrangement of 28-nucleotide near-identical repeats interspersed with unique ~32-nucleotide spacers, some of which match sequences found in bacteriophage or plasmids (13). Pre-crRNA transcripts are post-transcriptionally cleaved to yield crRNAs that each contain one unique sequence flanked by sequences derived from the repeat element (1–2, 14–17).

*To whom correspondence should be addressed: doudna@berkeley.edu, Phone: (510) 643-0225; Fax: (510) 643-0080.
‡These authors contributed equally to this work.

To identify the protein(s) responsible for producing crRNAs from pre-crRNAs in the *Yersinia* CRISPR subtype, we recombinantly expressed each of the six Cas proteins from Pa14 and tested them for endoribonuclease activity using an *in vitro* transcribed pre-crRNA (data not shown). Based on the observation of sequence-specific pre-crRNA processing when pre-crRNA was incubated with Csy4, we concluded that Csy4 is the endoribonuclease responsible for crRNA biogenesis (Fig. 1B). CRISPR transcript cleavage is a rapid, metal ion-independent reaction, as observed for crRNA processing within two other CRISPR/Cas subtypes (1–2). Csy4 cleaves Pa14 pre-crRNA within the repeat element at the base of a predicted stem-loop structure, generating 60-nucleotide crRNAs consisting of a 32-nucleotide unique (phage-derived) sequence flanked at the 5' and 3' ends by eight and 20 nucleotides of the repeat sequence, respectively (Fig. 1A). Csy4 does not cleave pre-crRNA from *Streptococcus thermophilus*, which has a repeat stem-loop of a distinct sequence from Pa14 (Fig. 1B).

For Csy4 to be effective, we hypothesized that its RNA recognition mechanism must be highly specific in order to target only CRISPR-derived transcripts and not other cellular RNAs. To test this, Csy4 was expressed in *Escherichia coli* either alone or together with a synthetic Pa14 CRISPR RNA consisting of eight repeat sequences (derived from the CRISPR locus proximal to the Cas1 ORF) and seven identical spacer sequences. In spite of its highly basic composition (pI=10.2), Csy4 did not associate with endogenous cellular nucleic acids. However, when co-expressed with a Pa14 CRISPR transcript, the protein co-purified with a protected ~19-nucleotide crRNA fragment (Fig. 1C). These observations underscored the specificity of Csy4 recognition, leading us to explore the protein/RNA interactions required for Csy4 substrate recognition and cleavage. Csy4 binding and activity assays were performed *in vitro* using RNA oligonucleotides corresponding to different regions of the 28-nucleotide Pa14 CRISPR repeat sequence (data not shown). Using this approach, we identified a 16-nucleotide minimal RNA fragment, consisting of the repeat-derived stem-loop and one downstream nucleotide, that is sufficient for Csy4-catalyzed cleavage. Cleavage assays utilizing this minimal RNA as a substrate showed that Csy4 activity requires the presence of a 2'-hydroxyl group on the ribose immediately upstream of the cleavage site. A 2'-deoxyribonucleotide substitution at this position completely abrogates cleavage, but does not disrupt Csy4 binding (fig. S1, A and B).

To obtain structural insights into pre-crRNA recognition and cleavage, we co-crystallized Csy4 in complex with the non-cleavable 16-nucleotide minimal RNA substrate described above (fig. S2). Three unique crystal forms of the complex were obtained; one contained wild-type Csy4 and two contained a catalytically active point mutant (S22C) of Csy4 (fig. S3). The crystal structures of the Csy4-RNA complex in the three crystal forms were solved to a resolution of 2.3 Å, 2.6 Å and 1.8 Å, respectively (table S1). In all three crystal forms, the RNA binds to Csy4 in an almost identical manner in which the protein makes extensive interactions with the ssRNA-dsRNA junction at the base of the crRNA stem as well as with the major groove of the RNA hairpin (Fig. 2A). The RNA hairpin is clamped into a highly basic groove between the main body of the protein and an arginine-rich helix (α3, residues 108–120) that inserts into the RNA major groove (Fig. 2B).

In the complex, Csy4 adopts a two-domain architecture consisting of an N-terminal ferredoxin-like domain (residues 1–94) and a C-terminal domain (residues 95–187) that mediates most of the interactions with the RNA (fig. S4A). At the sequence level, Csy4 shares less than 10% identity with the two other known endoribonucleases involved in crRNA biogenesis, CasE from *Thermus thermophilus* (18) and Cas6 from *Pyrococcus furiosus* (2). The crystal structures of CasE and Cas6 in their nucleic acid-free states showed that these proteins possess a duplicated ferredoxin fold. The N-terminal ferredoxin fold is preserved in Csy4; structural superpositions using the DALI server (19) indicate that Csy4 in

its RNA-binding conformation superposes with CasE and Cas6 with root-mean-square deviation (rmsd) of 3.8 Å (over the N-terminal 111 Cα atoms) and 3.9 Å (over 104 Cα atoms), respectively. Although the C-terminal domain of Csy4 (residues 95–187) shares the same secondary structure connectivity as a ferredoxin-like fold, its conformation is markedly different due to the position of the bound RNA (fig. S4B).

The crRNA substrate forms a stem-loop structure, as predicted for this subclass of crRNA repeats (20). Nucleotides 6–10 and 16–20 base pair to produce a regular A-form helical stem. The GUAUA pentaloop contains a sheared G11-A15 base pair and an extruded nucleotide U14. This closely resembles the structures adopted by GNR(N)A pentaloops found in the yeast U6 small nuclear RNA intramolecular stem-loop (21) and in bacteriophage lambda boxB RNA (22). In the Csy4-RNA complex, the RNA stem-loop straddles the β-hairpin formed by strands β6-β7 of Csy4, with the C6-G20 base-pair directly stacking onto the aromatic side chain of Phe 155 (Fig. 2C). Thus, in the context of the full-length crRNA repeat, the role of Phe 155 is to recognize the ssRNA-dsRNA junctions in the pre-crRNA substrate. This anchors the RNA stem and orients it at the proper angle to permit sequence-specific interactions in the major groove.

Arg 102 and Gln 104, located in a linker segment connecting the body of Csy4 to the arginine-rich helix, make sequence-specific hydrogen bonding contacts in the major groove of the RNA stem to nucleotides G20 and A19, respectively (Fig. 2B). The Csy4-crRNA interaction is further stabilized by the insertion of the arginine-rich helix into the major groove of the RNA hairpin in the proximity of the extruded nucleotide U14 (Fig. 2C). The side chains of Arg 114, Arg 115, Arg 118, Arg 119 and His 120 contact the phosphate groups of nucleotides 7–12. Additionally, the sidechain of Arg 115 hydrogen-bonds to the base of G6 as the only sequence-specific interaction between the arginine-rich helix and the RNA hairpin. Interestingly, the binding of the Arg-rich helix to the major groove of the crRNA hairpin is reminiscent of the recognition mechanism employed by certain highly basic viral peptides for dsRNA binding, as observed in the N-peptide/boxB RNA interaction in lambdoid phages (23) (fig. S5, A and B) and in lentiviral Rev-RRE and Tat-TAR complexes (24–25).

Csy4 recognizes the hairpin element of the CRISPR repeat sequence and cleaves immediately downstream of it. In the Csy4-RNA complex structure, where RNA cleavage is abrogated by the introduction of a 2'-deoxy modification in nucleotide G20, ordered electron density is only evident for the scissile phosphate between G20 and C21. The ribose and cytosine moieties of C21 are not resolved and presumably disordered. The scissile phosphate binds in a pocket located between the β6-β7 hairpin turn on one side and helix α1 on the other (Fig. 3A), hydrogen-bonding to the side chain of His 29 and the backbone amide of Gln 149. Ser 148 is adjacent to the 2' ribose carbon atom of nucleotide G20 (4.6 Å) and may make a hydrogen-bonding interaction with the 2'-hydroxyl group of G20 in a *bona fide* pre-crRNA substrate.

Both His 29 and Ser 148 are invariant among Csy4 homologs (fig. S6). In addition, a strongly conserved tyrosine (Tyr 176) is also positioned near the scissile phosphate. To confirm the functional roles of these residues in pre-crRNA processing, we introduced point mutations at each of these residues and tested the mutant proteins for cleavage activity *in vitro* (Fig. 3B). Mutations of His 29 or Ser 148 (to alanine and cysteine, respectively) completely abolished cleavage activity without disrupting RNA binding (Fig. 3B, data not shown). By contrast, mutation of Tyr 176 to phenylalanine did not disrupt activity, indicating that Tyr 176 may play a role in orienting His 29 but does not directly participate in catalysis. Additional mutants were generated to probe the requirements for individual sequence- and structure-specific interactions in pre-crRNA processing. Alanine substitution

of Arg 102 abolished pre-crRNA processing *in vitro*, whereas mutation of Gln 104 to alanine did not significantly disrupt activity (Fig. 3B). This suggests that Arg 102, which recognizes the terminal C6–G20 base pair, is key to properly orienting the RNA substrate but that Gln 104 is dispensable for crRNA processing. An alanine mutation at Phe 155 severely impaired crRNA biogenesis, suggesting that this residue also plays an important role in substrate recognition.

These mutational data suggest that the interaction between Csy4 and the closing base pair of the RNA stem is crucial for pre-crRNA processing, but that the sequence-specific interaction between Csy4 and the penultimate base pair in the stem is less important. To probe the Csy4-RNA interactions required for cleavage, we incubated Csy4 with a panel of short RNA oligonucleotides containing a variety of mutations in the CRISPR repeat stem-loop sequence (fig. S7). These data confirm that Csy4 requires a C6–G20 base pair closing the RNA stem and that Csy4 can accommodate different nucleotides at the penultimate RNA base pair.

Csy4-catalyzed pre-crRNA cleavage requires the presence of the 2'-hydroxyl group in the nucleotide immediately upstream of the cleavage site. This suggests that the catalytic mechanism of Csy4 proceeds through a 2'-3' cyclic intermediate or yields a 2'-3' cyclic product. In this context, the observation of a strictly conserved serine residue (Ser 148) adjacent to the 2' position is unprecedented and points to Ser 148 playing a role in activating or positioning the 2'-hydroxyl for a nucleophilic attack on the scissile phosphate. How the serine sidechain may function in this context remains to be determined. Histidine, on the other hand, is commonly found in metal ion-independent active sites, such as in RNase A (26). Although mutation of His 29 to alanine leads to loss of activity in Csy4, mutation to lysine partially restores activity (fig. S8), suggesting that His 29 acts as a proton donor in stabilizing the 5' oxygen leaving group.

Phylogenetic analysis of CRISPR loci suggests that CRISPR repeat sequences and structures have co-evolved with the Cas genes (20). The similarity of Csy4 at the fold level to the CRISPR-processing endonucleases CasE and Cas6 suggests that collectively they are likely to have descended from a single ancestral endoribonuclease enzyme that has diverged throughout evolution. The structure described here reveals how Csy4 and related endonucleases from the same subfamily utilize an exquisite recognition mechanism to discriminate crRNA substrates from other cellular RNAs and illustrates the importance of co-evolution in shaping molecular recognition mechanisms in the CRISPR pathway.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References and Notes

1. Brouns SJ, et al. Science. 2008 Aug 15.321:960. [PubMed: 18703739]
2. Carte J, Wang R, Li H, Terns RM, Terns MP. Genes and Development. 2008; 22:3489. [PubMed: 19141480]
3. Haft DH, Selengut J, Mongodin EF, Nelson KE. PLoS Comput Biol. 2005 Nov.1:e60. [PubMed: 16292354]
4. Hale CR, et al. Cell. 2009 Nov 25.139:945. [PubMed: 19945378]
5. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. Biology Direct. 2006; 1:1. [PubMed: 16542032]
6. Barrangou R, et al. Science. 2007 Mar 23.315:1709. [PubMed: 17379808]
7. Marraffini LA, Sontheimer EJ. Science. 2008; 322:1843. [PubMed: 19095942]
8. Horvath P, Barrangou R. Science. 2010 Jan 8.327:167. [PubMed: 20056882]

9. Jansen R, van Embden JDA, Gaastra W, Schouls LM. Molecular Microbiology. 2002 Mar.43:1565. [PubMed: 11952905]
10. Sorek R, Kunin V, Hugenholtz P. Nat Rev Microbiol. 2008 Mar.6:181. [PubMed: 18157154]
11. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. Trends Biochem Sci. 2009 Jul 29.
12. Marraffini LA, Sontheimer EJ. Nat Rev Genet. 2010 Feb 2.11:181. [PubMed: 20125085]
13. Grissa I, Vergnaud G, Pourcel C. BMC Bioinformatics. 2007; 8:172. [PubMed: 17521438]
14. Tang TH, et al. Proceedings of the National Academy of Sciences of the United States of America. 2002 MAY 28.99(7536)
15. Lillestol RK, Redder P, Garrett RA, Brugger K. Archaea. 2006; 2:59. [PubMed: 16877322]
16. Lillestol RK, et al. Mol Microbiol. 2009 Apr.72:259. [PubMed: 19239620]
17. Tang TH, et al. Molecular Microbiology. 2005 JAN.55:469. [PubMed: 15659164]
18. Ebihara A, et al. Protein Sci. 2006 Jun.15:1494. [PubMed: 16672237]
19. Holm L, Sander C. J Mol Biol. 1993 Sep 5.233:123. [PubMed: 8377180]
20. Kunin V, Sorek R, Hugenholtz P. Genome Biol. 2007 Apr 18.8:R61. [PubMed: 17442114]
21. Huppler A, Nikstad LJ, Allmann AM, Brow DA, Butcher SE. Nat Struct Biol. 2002 Jun.9:431. [PubMed: 11992125]
22. Legault P, Li J, Mogridge J, Kay LE, Greenblatt J. Cell. 1998 Apr 17.93:289. [PubMed: 9568720]
23. Cai Z, et al. Nature Structural Biology. 1998 Mar.5:203.
24. Ye X, Gorin A, Ellington AD, Patel DJ. Nat Struct Biol. 1996 Dec.3:1026. [PubMed: 8946856]
25. Anand K, Schulte A, Vogel-Bachmayr K, Scheffzek K, Geyer M. Nat Struct Mol Biol. 2008 Dec. 15:1287. [PubMed: 19029897]
26. Raines RT. Chem Rev. 1998 May 7.98:1045. [PubMed: 11848924]
27. We thank W. Westphal for help with purification of Csy4 constructs; J. Doudna-Cate and members of the Doudna lab for critical reading of the manuscript; and C. Ralston and J. Holton (Beamlines 8.2.2 and 8.3.1, Advanced Light Source, Lawrence Berkeley National Laboratory) for assistance with X-ray data collection. R.E.H. is supported by the US National Institutes of Health training grant 5 T32 GM08295. M.J. is supported by a Human Frontier Science Program Fellowship. B.W. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. This work was supported in part by a grant from the NSF. J.A.D. is a Howard Hughes Medical Institute Investigator.

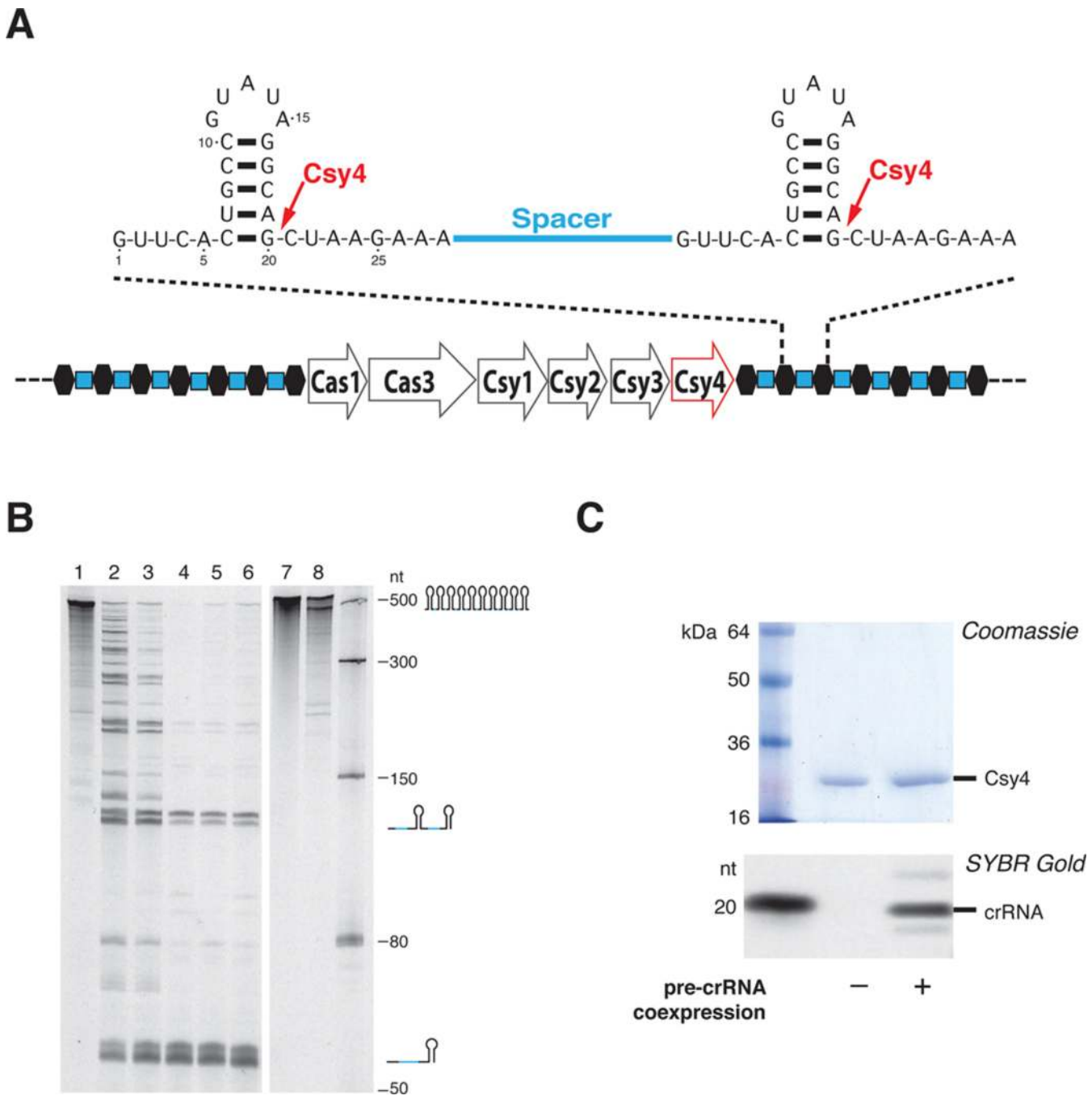

Fig. 1. Csy4 specifically cleaves only its cognate pre-crRNA substrate

(**A**) Schematic of the CRISPR/Cas locus in Pa14. Two CRISPR loci flank the six Cas genes. Enlarged is a schematic showing the predicted stem-loop in the 28-nucleotide direct repeat (black lettering) separated by a 32-nucleotide unique spacer sequence (blue). Red arrows denote the cleavage site. (**B**) Csy4 (lanes 2–6, 8) was incubated with *in vitro* transcribed Pa14 pre-crRNA (lanes 1–6) for 30 seconds (lane 2), one minute (lane 3) and five minutes (lanes 4–6) in buffer containing no exogenous metal ions (lanes 2–4) and in buffer supplemented with 2.5 mM $MgCl_2$ (lane 5) or 2.5 mM EDTA (lane 6). Csy4 was also incubated with *in vitro* transcribed pre-crRNA from *Streptococcus thermophilus* (lanes 7–8) for 5 minutes (lane 8). Products were acid phenolchloroform extracted, separated on 15%

denaturing PAGE and visualized with SYBR Gold staining. (C) Csy4 was heterologously expressed in *E. coli* in the presence (+) or absence (−) of a plasmid expressing a Pa14 CRISPR transcript. Csy4 was affinity purified; co-purifying RNA was extracted and analyzed by denaturing PAGE and staining with SYBR Gold. The ~19-nucleotide RNA corresponds to a protected fragment of the CRISPR repeat.

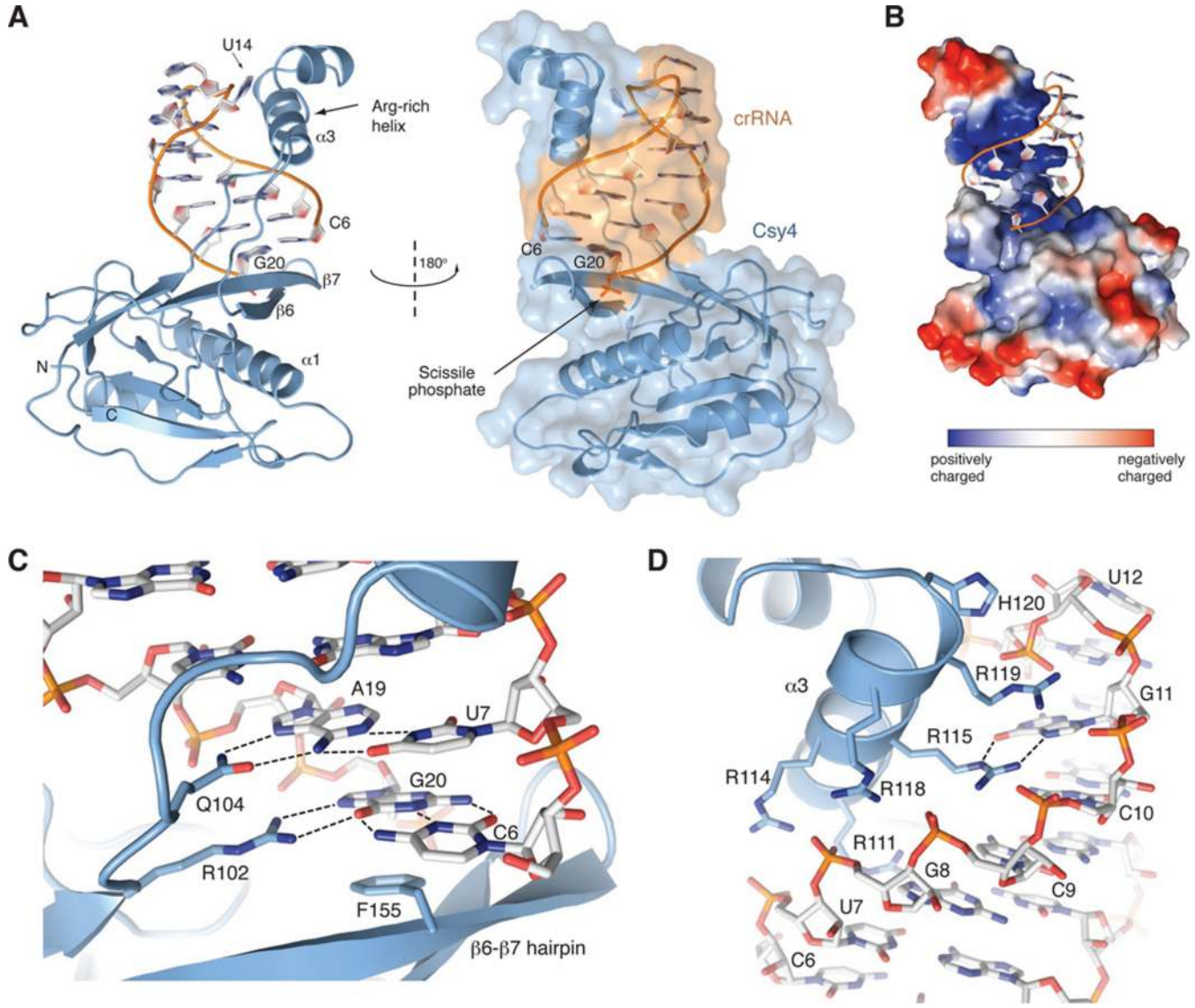

Fig. 2. The crystal structure of Csy4 bound to RNA substrate

(**A**) Front and back views of the complex. Csy4 is colored in blue and the RNA backbone is colored in orange. (**B**) Csy4 is shown as a surface representation colored according to electrostatic potential. The RNA is shown in ribbon representation with the phosphate backbone colored in orange. The Csy4-RNA complex is shown in the same orientation as in the right panel of (A). (**C**) Magnified view of the interactions between Csy4 and the major groove of the RNA hairpin. Hydrogen bonding is depicted with dashed lines. (**D**) Expanded view of the interactions between the arginine-rich helix α3 (blue) and the RNA phosphate backbone (shown in stick format, orange).

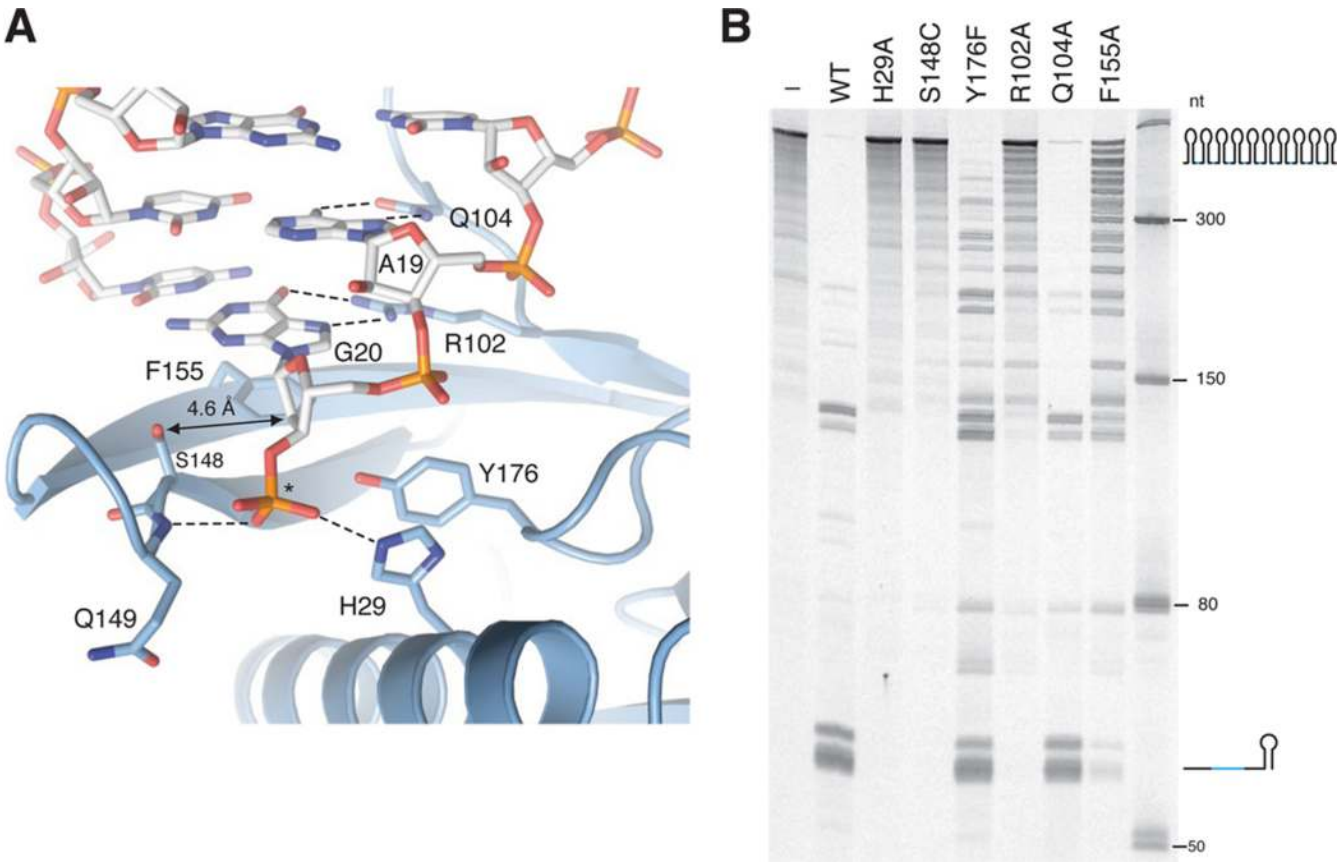

Fig. 3. Functional analysis of catalytic residues in Csy4

(**A**) Detailed view of the catalytic center. Only the phosphate group of nucleotide C21 (the scissile phosphate, indicated with an asterisk) is visible in electron density maps. Strictly conserved residues found in the proximity of the scissile phosphate are shown in stick format. The distance between the hydroxyl group of Ser 148 and the 2' ribose carbon of G20 is indicated with an arrow. (**B**) Cleavage activity of Csy4. Wild-type (WT) Csy4 and a series of single point mutants were incubated with *in vitro* transcribed pre-crRNA for 5 minutes at 25°C. Products were acid phenol-chloroform extracted, resolved by denaturing PAGE and visualized by staining with SYBR Gold.