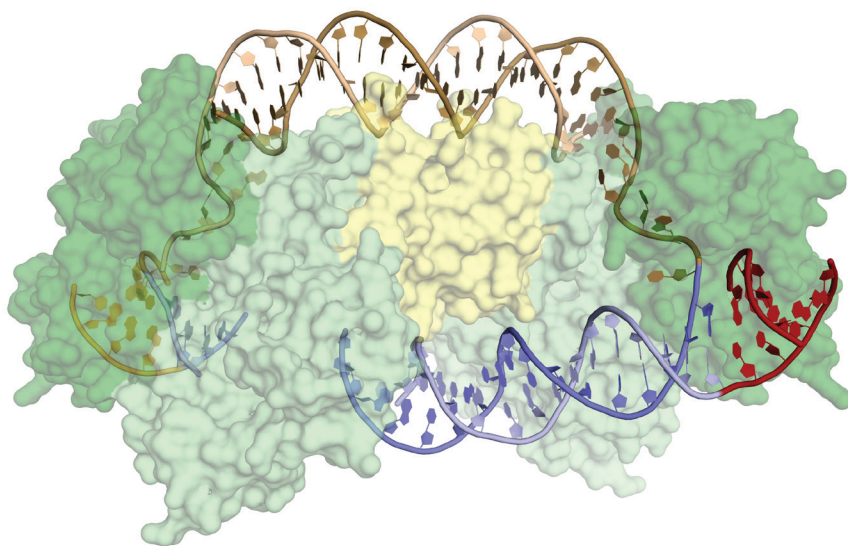
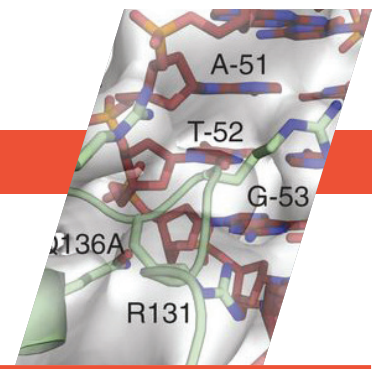


The CRISPR Target-Recognition Mechanism



Surface representation of the Cas1-Cas2 complex, consisting of four Cas1 proteins (light and dark green) and two Cas2 proteins (yellow). Donor DNA (brown) is being integrated into the target DNA (blue), at a precise location in the CRISPR array, following a short leader sequence (red).

CRISPR-associated (Cas) proteins have revolutionized gene editing by vastly simplifying the insertion of short snippets of new (“donor”) DNA into very specific locations of target DNA. Now, researchers have discovered how Cas proteins are able to recognize their target locations with such great specificity. X-ray crystallography was used to solve the structures of Cas1 and Cas2—responsible for DNA-snippet capture and integration—as they were bound to synthesized DNA strands designed to mimic different stages of the process. The resulting structures not only show how the system works in its native context (as part of a bacterial immune system), they also inform the development of the CRISPR-Cas system as a general-purpose molecular recording device—a tool for encoding information in genomes.

Cas1 appears to have evolved from a more “promiscuous” (less selective) type of

enzyme that catalyzes the movement of DNA sequences from one position to another (a transposase). At some point, Cas1 acquired an unusual degree of specificity for a particular location in the bacterial genome: the CRISPR array. This specificity is critical to the bacteria, both for acquiring immunity as well as for avoiding genome damage caused by the insertion of viral fragments at the wrong location. The researchers wanted to know how the Cas1-Cas2 proteins recognize the target sequence, in order to compare them with previously studied transposases and integrases (enzymes that catalyze the integration of donor DNA into target DNA) and to find out whether the proteins can be altered to recognize new sequences for custom applications.

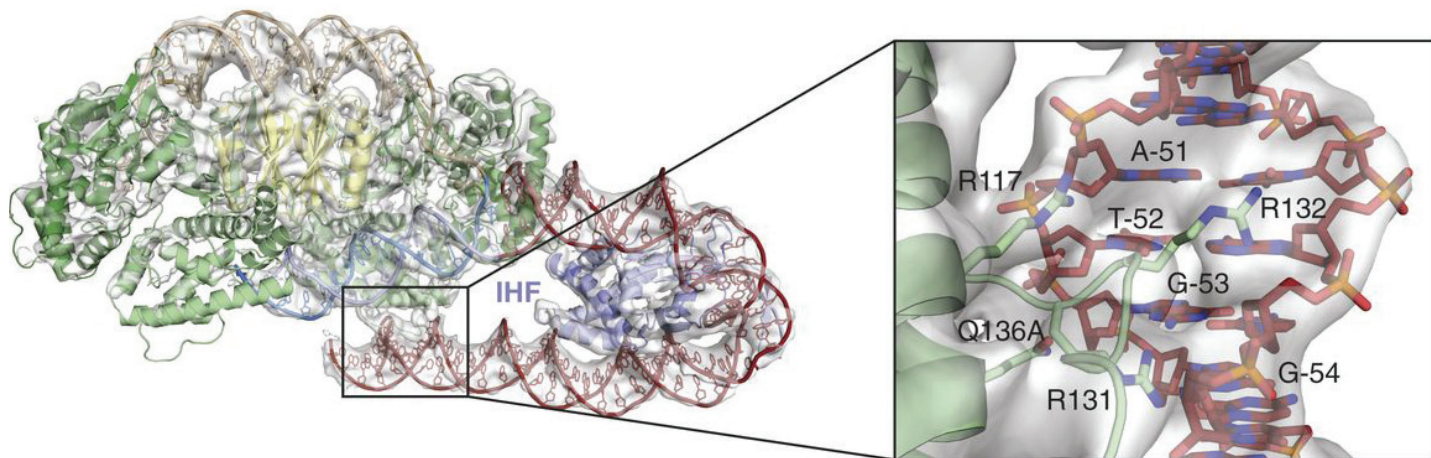
Previous structures of Cas1 and Cas2, both alone and bound to donor DNA, had been solved using data from the ALS. These structures were informative, but in

For the Record

Bacterial DNA is characterized by regions of Clustered Regularly Interspaced Short Palindromic Repeats—CRISPR arrays—where fragments of DNA from viruses are stored for future reference in fighting off infection. The CRISPR-associated (Cas) proteins Cas1 and Cas2 catalyze the storage process and are integral to the CRISPR immune system. But the Cas1-Cas2 complex also has potential as a genetic recording device. Notably, the complex was recently used by a group at Harvard [*Nature* 547, 345 (2017)] to encode a short movie (Muybridge’s *Horse in Motion*) into the genomes of a population of *E. coli*.

The technique could conceivably be adapted to record information about cellular states that might be useful in diagnosing diseases. Additionally, from the research perspective, Cas1-Cas2 can be used to uniquely barcode individual cells in a diverse population, allowing for tracking cell lineages in studies of development, cancer, and immunology. To use the proteins outside of their native context, however, we need to be able to predict potential target sites and possibly engineer Cas1-Cas2 to recognize new sequences. By understanding how Cas1-Cas2 recognizes targets, we can approach these challenges in a more rational way.

the absence of target DNA, they revealed little about how target specificity is achieved. Previous work had also revealed that an accessory protein, IHF (integration host factor), binds adjacent to the recognition site and is critical for the activity of Cas1-Cas2 in vivo.



Upstream sequence recognition by Cas1. The IHF accessory protein (blue) induces a 180° turn in the target DNA, directing it back toward the Cas1-Cas2 complex. Inset: Closeup of the upstream recognition sequence that interacts with one of the Cas1 proteins.

In this work, the researchers crystallized Cas1-Cas2 in complex with pre-formed DNA strands that mimicked reaction intermediates and products. X-ray crystallography studies were performed at ALS Beamline 8.3.1 and at Stanford Synchrotron Radiation Lightsource (SSRL). The structures showed substantial distortions in the target DNA, but there were surprisingly few sequence-specific contacts with the Cas1-Cas2 complex, and the resulting flexibility of the DNA produced disorder in the crystals.

Attempts to model the DNA across the disordered sections led to the realization that the DNA had to be even more distorted. Cryo-electron microscopy experiments, coupled with the crystallography data, confirmed that IHF introduces an additional sharp bend in the DNA, bringing an upstream recognition sequence into contact with Cas1 to increase both the specificity and efficiency of integration.

The lack of direct sequence recognition might reflect the evolutionary origins of Cas1 as a transposase. The bending of target DNA in transposases and integrases serves to eject DNA from active sites after integration, whereas in the CRISPR-Cas system, the feature provides the sequence specificity needed to begin integration.

The role played by IHF also represents an unexpected variation. Furthermore, in transposases, IHF helps recognize foreign DNA, whereas here it helps recognize target DNA, reflecting the shift in Cas1 use from facilitating infection to conferring immunity.

Bacterial transposases are robust tools for DNA tagging, insertion, and deletion, but they are promiscuous in their target selection and require sequence-specific

interactions with donor DNA that limit their use in some systems. The architecture of the CRISPR integration complex described here suggests that subtle adjustment of the distance between Cas1 active sites could reprogram the system to recognize different target sites. Changes in its architecture could thereby be exploited for genome tagging applications and may also explain the natural divergence of CRISPR arrays in bacteria.

Publication about this research: A.V. Wright, J.-J. Liu, G.J. Knott, K.W. Doxzen, E. Nogales, and J.A. Doudna, "Structures of the CRISPR genome integration complex," *Science* **357**, 1113 (2017). doi:10.1126/science.aao0679

Research conducted by: A.V. Wright, G.J. Knott, and K.W. Doxzen (UC Berkeley); J.-J. Liu (UC Berkeley and Berkeley Lab); and E. Nogales and J.A. Doudna (UC Berkeley, Berkeley Lab, and Howard Hughes Medical Institute).

Research funding: National Science Foundation, National Institutes of Health, and Howard Hughes Medical Institute. Operation of the ALS is supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences Program.

Published by the
**ADVANCED LIGHT SOURCE
COMMUNICATIONS GROUP**

