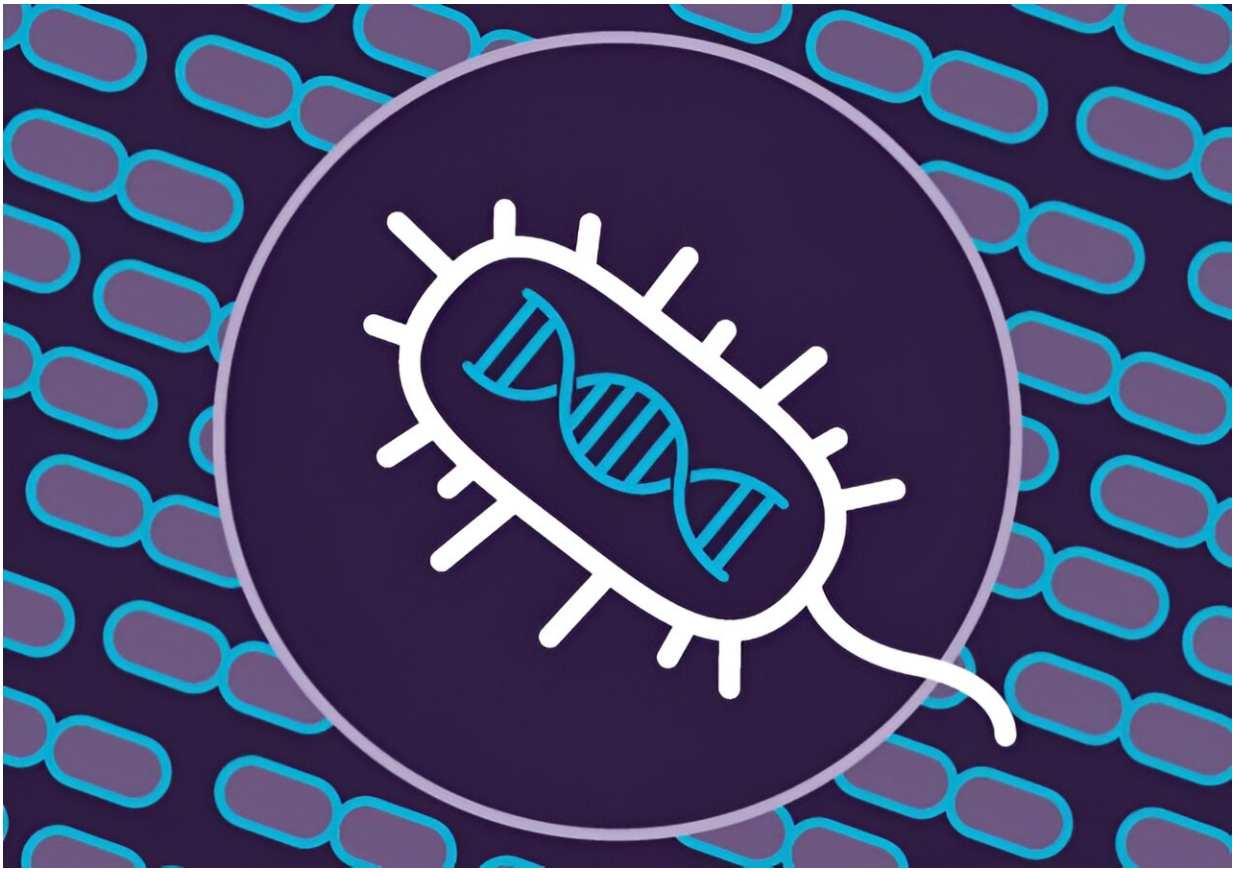
Search algorithm reveals nearly 200 new kinds of CRISPR systems

November 24 2023, by Allessandra DiCorato



Credit: Jessica Hernandez, Broad Communications

Microbial sequence databases contain a wealth of information about enzymes and other molecules that could be adapted for biotechnology.

But these databases have grown so large in recent years that they've become difficult to search efficiently for enzymes of interest.

Now, scientists at the Broad Institute of MIT and Harvard, the McGovern Institute for Brain Research at MIT, and the National Center for Biotechnology Information (NCBI) at the National Institutes of Health have developed a new search algorithm that has identified 188 kinds of new rare CRISPR systems in bacterial genomes, encompassing thousands of individual systems. The work appears in *Science*.

The algorithm, which comes from the lab of CRISPR pioneer Feng Zhang, uses big-data clustering approaches to rapidly search massive amounts of genomic data. The team used their algorithm, called Fast Locality-Sensitive Hashing-based clustering (FLSHclust) to mine three major public databases that contain data from a wide range of unusual bacteria, including ones found in coal mines, breweries, Antarctic lakes, and dog saliva.

The scientists found a surprising number and diversity of CRISPR systems, including ones that could make edits to DNA in human cells, others that can target RNA, and many with a variety of other functions.

The new systems could potentially be harnessed to edit mammalian cells with fewer off-target effects than current Cas9 systems. They could also one day be used as diagnostics or serve as molecular records of activity inside cells.

The researchers say their search highlights an unprecedented level of diversity and flexibility of CRISPR and that there are likely many more rare systems yet to be discovered as databases continue to grow.

"Biodiversity is such a treasure trove, and as we continue to sequence more genomes and metagenomic samples, there is a growing need for

better tools, like FLSHclust, to search that sequence space to find the molecular gems," said Zhang, a co-senior author on the study and a core institute member at the Broad.

Zhang is also an investigator at the McGovern Institute for Brain Research at MIT, the James and Patricia Poitras Professor of Neuroscience at MIT with joint appointments in the departments of Brain and Cognitive Sciences and Biological Engineering, and an investigator at the Howard Hughes Medical Institute. Eugene Koonin, a distinguished investigator at the NCBI, is co-senior author on the study as well.

Searching for CRISPR

CRISPR, which stands for Clustered Regularly Interspaced Short Palindromic Repeats, is a bacterial defense system that has been engineered into many tools for genome editing and diagnostics.

To mine databases of protein and nucleic acid sequences for novel CRISPR systems, the researchers developed an algorithm based on an approach borrowed from the big data community. This technique, called locality-sensitive hashing, clusters together objects that are similar but not exactly identical.

Using this approach allowed the team to probe billions of protein and DNA sequences—from the NCBI, its Whole Genome Shotgun database, and the Joint Genome Institute—in weeks, whereas previous methods that look for identical objects would have taken months. They designed their algorithm to look for genes associated with CRISPR.

"This new algorithm allows us to parse through data in a time frame that's short enough that we can actually recover results and make biological hypotheses," said Soumya Kannan, who is a co-first author on

the study. Kannan was a graduate student in Zhang's lab when the study began and is currently a postdoctoral researcher and Junior Fellow at Harvard University. Han Altae-Tran, a graduate student in Zhang's lab during the study and currently a postdoctoral researcher at the University of Washington, was the study's other co-first author.

"This is a testament to what you can do when you improve on the methods for exploration and use as much data as possible," said Altae-Tran. "It's really exciting to be able to improve the scale at which we search."

New systems

In their analysis, Altae-Tran, Kannan, and their colleagues noticed that the thousands of CRISPR systems they found fell into a few existing and many new categories. They studied several of the new systems in greater detail in the lab.

They found several new variants of known Type I CRISPR systems, which use a guide RNA that is 32 base pairs long rather than the 20-nucleotide guide of Cas9. Because of their longer guide RNAs, these Type I systems could potentially be used to develop more precise gene-editing technology that is less prone to off-target editing.

Zhang's team showed that two of these systems could make short edits in the DNA of human cells. And because these Type I systems are similar in size to CRISPR-Cas9, they could likely be delivered to cells in animals or humans using the same gene-delivery technologies being used today for CRISPR.

One of the Type I systems also showed "collateral activity"—broad degradation of nucleic acids after the CRISPR protein binds its target. Scientists have used similar systems to make infectious disease

diagnostics such as SHERLOCK, a tool capable of rapidly sensing a single molecule of DNA or RNA. Zhang's team thinks the new systems could be adapted for diagnostic technologies as well.

The researchers also uncovered new mechanisms of action for some Type IV CRISPR systems, and a Type VII system that precisely targets RNA, which could potentially be used in RNA editing. Other systems could potentially be used as recording tools—a molecular document of when a gene was expressed—or as sensors of specific activity in a living cell.

Mining data

The scientists say their algorithm could aid in the search for other biochemical systems. "This search algorithm could be used by anyone who wants to work with these large databases for studying how proteins evolve or discovering new genes," Altae-Tran said.

The researchers add that their findings illustrate not only how diverse CRISPR systems are, but also that most are rare and only found in unusual bacteria.

"Some of these microbial systems were exclusively found in water from coal mines," Kannan said. "If someone hadn't been interested in that, we may never have seen those systems. Broadening our sampling diversity is really important to continue expanding the diversity of what we can discover."

More information: Han Altae-Tran et al, Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering, *Science* (2023). DOI: 10.1126/science.adi1910

Provided by Broad Institute of MIT and Harvard