Machine learning helps determine success of advanced genome editing
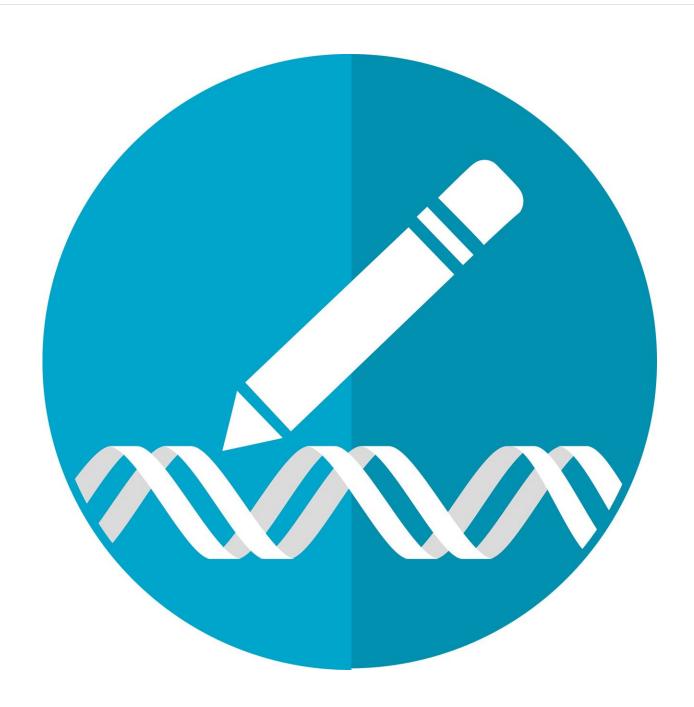
February 16 2023

A new tool to predict the chances of successfully inserting a gene-edited sequence of DNA into the genome of a cell, using a technique known as prime editing, has been developed by researchers at the Wellcome Sanger Institute. An evolution of CRISPR-Cas9 gene editing technology, prime editing has huge potential to treat genetic disease in humans, from cancer to cystic fibrosis. But thus far, the factors determining the success of edits are not well understood.

The study, published today (February 16) in *Nature Biotechnology*, assessed thousands of different DNA sequences introduced into the genome using prime editors. These data were then used to train a machine learning algorithm to help researchers design the best fix for a given genetic flaw, which promises to speed up efforts to bring prime editing into the clinic.

Developed in 2012, CRISPR-Cas9 was the first easily programmable gene editing technology. These "molecular scissors" enabled researchers to cut DNA at any position in the genome in order to remove, add or alter sections of the DNA sequence. The technology has been used to study which genes are important for various conditions, from cancer to rare diseases, and to develop treatments that fix or turn off harmful mutations or genes.

Base editors were an innovation expanding on CRISPR-Cas9 and were called "molecular pencils" for their ability to substitute single bases of DNA. The latest gene editing tools, created in 2019, are called prime editors. Their ability to perform search and replace operations directly on the genome with a high degree of precision has led to them being dubbed "molecular word processors."

The ultimate aim of these technologies is to correct harmful mutations in people's genes. More than 16,000 small deletion variants—where a small number of DNA bases have been removed from the genome—have been causally linked to disease. This includes cystic fibrosis, where 70% of cases are caused by the deletion of just three DNA bases. In 2022, base edited T-cells were successfully used to treat a patient's leukemia, where chemotherapy and bone marrow transplant had failed.

In this new study, researchers at the Wellcome Sanger Institute designed 3,604 DNA sequences of between one and 69 DNA bases in length. These sequences were inserted into three different human cell lines, using different prime editor delivery systems in various DNA repair contexts. After a week, the cells were genome sequenced to see if the edits had been successful or not.

The insertion efficiency, or success rate, of each sequence was assessed to determine common factors in the success of each edit. The length of sequence was found to be a key factor, as was the type of DNA repair mechanism involved.

Jonas Koeppel from the Wellcome Sanger Institute and first author of the study said, "The variables involved in successful prime edits of the genome are many, but we're beginning to discover what factors improve the chances of success. Length of sequence is one of these factors, but it's not as simple as the longer the sequence the more difficult it is to insert. We also found that one type of DNA repair prevented the insertion of short sequences, whereas another type of repair prevented the insertion of long sequences."

To help make sense of these data, the researchers turned to machine learning to detect patterns that determine insertion success, such as length and the type of DNA repair involved. Once trained on the existing data, the algorithm was tested on new data and was found to accurately

predict insertion success.

Juliane Weller from the Wellcome Sanger Institute and a first author of the study said, "Put simply, several different combinations of three DNA letters can encode for the same amino acid in a protein. That's why there are hundreds of ways to edit a gene to achieve the same outcome at the protein level. By feeding these potential gene edits into a machine learning algorithm, we have created a model to rank them on how likely they are to work. We hope this will remove much of the trial and error involved in prime editing and speed up progress considerably."

The next steps for the team will be to make models for all known human genetic diseases to better understand if and how they can be fixed using prime editing. This will involve other research groups at the Sanger Institute and its collaborators.

Dr. Leopold Parts from the Wellcome Sanger Institute and senior author of the study said, "The potential of prime editing to improve human health is vast, but first we need to understand the easiest, most efficient and safest ways to make these edits. It's all about understanding the rules of the game, which the data and tool resulting from this study will help us to do."

More information: Leopold Parts, Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants, *Nature Biotechnology* (2023). DOI: 10.1038/s41587-023-01678-y. www.nature.com/articles/s41587-023-01678-y

Provided by Wellcome Trust Sanger Institute