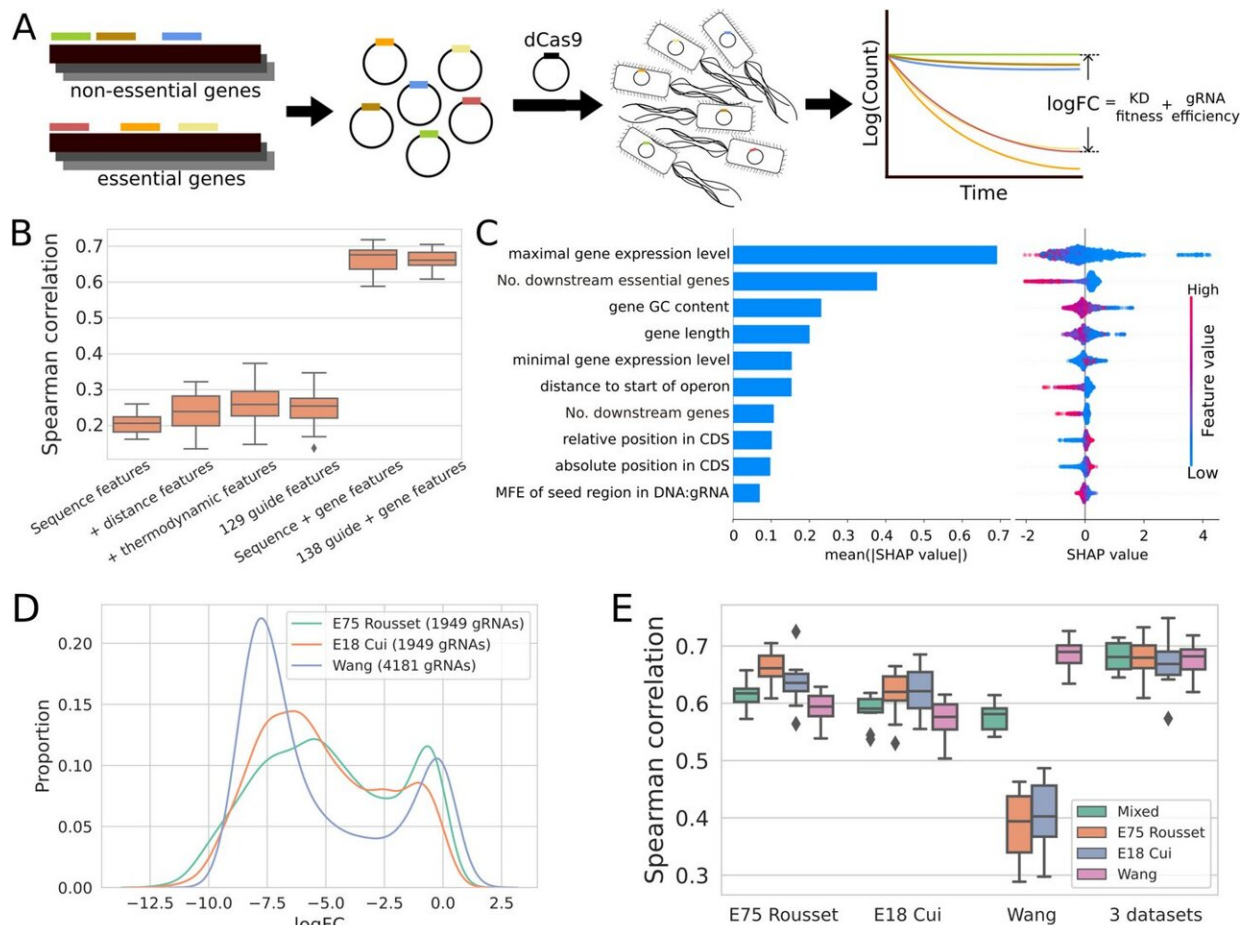# Bioinformatics: Researchers develop a new machine learning approach

January 11 2024, by Andreas Fischer

Automated machine learning and data fusion predicts depletion in CRISPRi essentiality screens. Credit: *Genome Biology* (2024). DOI: 10.1186/s13059-023-03153-y

To combat viruses, bacteria and other pathogens, synthetic biology offers new technological approaches whose performance is being validated in experiments. Researchers from the Würzburg Helmholtz Institute for RNA-based Infection Research and the Helmholtz AI Cooperative applied data integration and artificial intelligence (AI) to develop a machine learning approach that can predict the efficacy of CRISPR technologies more accurately than before.

The findings were published in the journal *Genome Biology*.

The genome or DNA of an organism incorporates the blueprint for proteins and orchestrates the production of new cells. Aiming to combat pathogens, cure genetic diseases or achieve other positive effects, molecular biological CRISPR technologies are being used to specifically alter or silence genes and inhibit protein production.

One of these molecular biological tools is CRISPRi (from "CRISPR interference"). CRISPRi blocks genes and gene expression without modifying the DNA sequence. As with the CRISPR-Cas system also known as "gene scissors," this tool involves a ribonucleic acid (RNA), which serves as a guide RNA to direct a nuclease (Cas). In contrast to gene scissors, however, the CRISPRi nuclease only binds to the DNA without cutting it. This binding results in the corresponding gene not being transcribed and thus remaining silent.

Until now, it has been challenging to predict the performance of this method for a specific gene. Researchers from the Würzburg Helmholtz Institute for RNA-based Infection Research (HIRI) in cooperation with the University of Würzburg and the Helmholtz Artificial Intelligence Cooperation Unit (Helmholtz AI) have now developed a machine learning approach using data integration and artificial intelligence (AI) to improve such predictions in the future.

The approach

CRISPRi screens are a highly sensitive tool that can be used to investigate the effects of reduced gene expression. In their study, the scientists used data from multiple genome-wide CRISPRi essentiality screens to train a machine learning approach. Their goal: To better predict the efficacy of the engineered guide RNAs deployed in the CRISPRi system.

"Unfortunately, genome-wide screens only provide indirect information about guide efficiency. Hence, we have applied a new machine learning method that disentangles the efficacy of the guide RNA from the impact of the silenced gene," explains Lars Barquist.

The computational biologist initiated the study and heads a bioinformatics research group at the Würzburg Helmholtz Institute, a site of the Braunschweig Helmholtz Center for Infection Research in cooperation with the Julius-Maximilians-Universität Würzburg.

Supported by additional AI tools ("Explainable AI"), the team established comprehensible design rules for future CRISPRi experiments. The study authors validated their approach by conducting an independent screen targeting essential bacterial genes, showing that their predictions were more accurate than previous methods.

"The results have shown that our model outperforms existing methods and provides more reliable predictions of CRISPRi performance when targeting specific genes," says Yanying Yu, Ph.D. student in Lars Barquist's research group and first author of the study.

The scientists were particularly surprised to find that the guide RNA itself is not the primary factor in determining CRISPRi depletion in essentiality screens. "Certain gene-specific characteristics related to gene

expression appear to have a greater impact than previously assumed," explains Yu.

The study also reveals that integrating data from multiple data sets significantly improves the predictive accuracy and enables a more reliable assessment of the efficiency of guide RNAs. "Expanding our training data by pulling together multiple experiments is essential to create better prediction models. Prior to our study, lack of data was a major limiting factor for prediction accuracy," summarizes junior professor Barquist.

The approach now published will be very helpful in planning more effective CRISPRi experiments in the future and serve both biotechnology and basic research. "Our study provides a blueprint for developing more precise tools to manipulate bacterial gene expression and ultimately help to better understand and combat pathogens," says Barquist.

More information: Yanying Yu et al, Improved prediction of bacterial CRISPRi guide efficiency from depletion screens through mixed-effect machine learning and data integration, *Genome Biology* (2024). DOI: 10.1186/s13059-023-03153-y