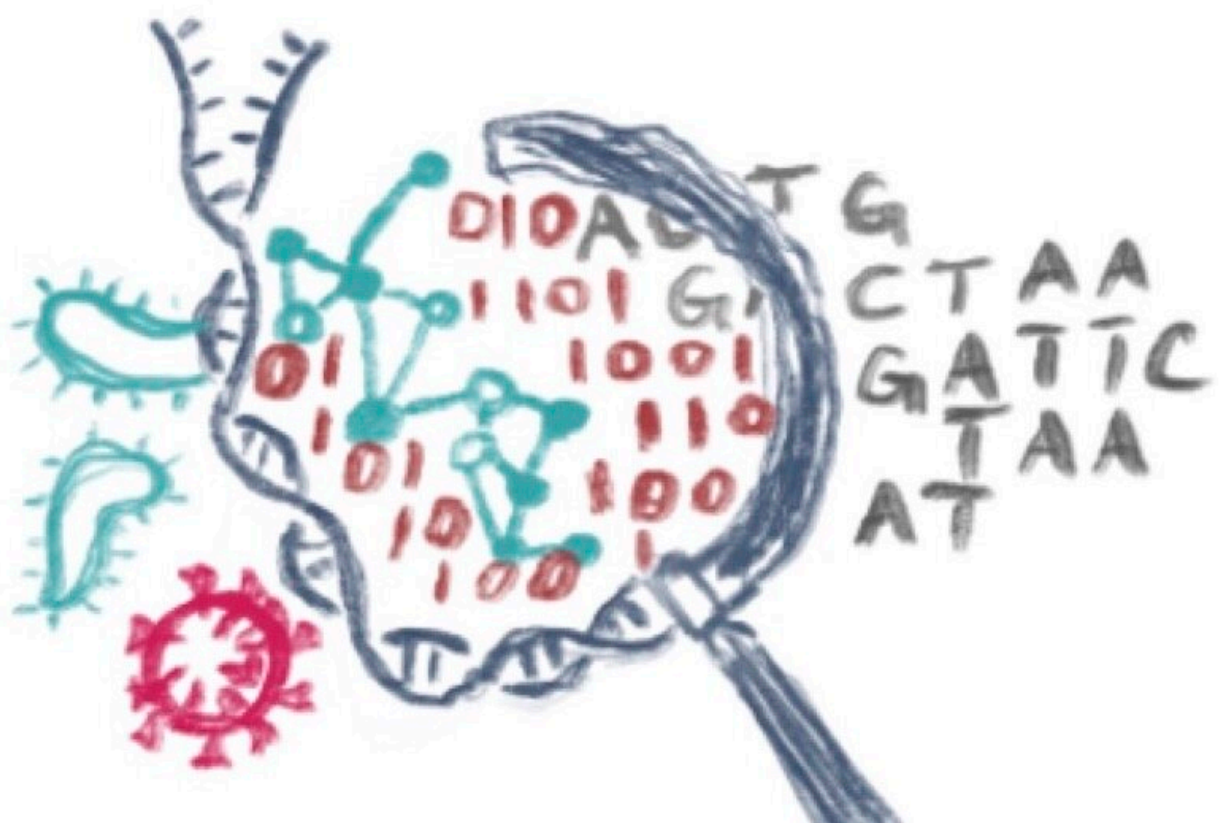# Open-source program IDs synthetic, naturally occurring gene sequences

June 21 2022, by Mike Williams

Rice University computer scientists and their collaborators have developed SeqScreen, a program to screen short DNA sequences, whether synthetic or natural, to determine their toxicity. Credit: Treangen Lab/Rice University

It's a given that certain bacteria and viruses can cause illness and disease,

but the real culprits are the sequences of concern that lie within the genomes of these microbes.

Calling them out is about to get easier.

Years of work by Rice University computer scientists and their colleagues have led to an improved platform for DNA screening and pathogenic sequence characterization, whether naturally occurring or synthetic, before they have the chance to impact public health.

Computer scientist Todd Treangen of Rice's George R. Brown School of Engineering and genomic specialist Krista Ternus of Signature Science LLC led the study that produced SeqScreen, a program to accurately characterize short DNA sequences, often called oligonucleotides.

Treangen said SeqScreen is intended to improve the detection and tracking of a wide range of pathogenic sequences.

"SeqScreen is the first open-source software toolkit that is available for synthetic DNA screening," Treangen said. "Our program improves upon the previous state of the art for companies, individuals and government agencies for their DNA screening practices."

The study, which began as high-risk, high-payoff research project, appears in the journal *Genome Biology*.

SeqScreen takes advantage of work by partners at Austin, Texas-based company Signature Science to curate a database of thousands of gene sequences representing 32 types of virulence functions. "This curated database took years of biocuration and review to develop, and is at the core of the training data of SeqScreen's machine learning algorithm," Treangen said.

The company collaborated with Treangen last year to find SARS-CoV-2 mutations that may have made the Omicron variant more resistant to antibodies, including those from vaccinations. "SeqScreen came first, and some of its ideas carried over to the COVID project," he said. "But SeqScreen is much broader in scope."

"We focus on identifying functions of sequences of concern—which we call FunSoCs—whereas previous screening approaches were more concerned with looking at 'are you this bacterium?' or 'are you this virus?'" Treangen said. "SeqScreen doesn't focus on the names of which bacteria or viruses are in your sample. Rather, we want to know if there are sequences in that sample that could be harmful, such as toxins that can destroy human cells."

Focusing on functions of concern is important, he said, because bacteria readily exchange DNA via horizontal gene transfer.

"We have highlighted examples in the publication of bacteria whose genomes are essentially identical, except one has a sequence of concern, such as a toxin, that the other does not," Treangen said. "SeqScreen really hones in on the presence or absence of functions that represent virulence factors."

He said SeqScreen will also aid in the detection of novel or emerging pathogens from the environment.

More information: Advait Balaji et al, SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning, *Genome Biology* (2022). DOI: 10.1186/s13059-022-02695-x

Provided by Rice University