Ultrafast platform enables real-time tracking of COVID-19 strains for contact tracing

By Nick Gallagher
May 10, 2021

An international team has created a digital system for logging and categorizing new COVID-19 genome sequence data that, at the time of writing, was over 3,000 times faster than its closest competitor, allowing researchers to identify and target potentially dangerous viral strains more efficiently as they emerge in real time.

The open-access platform, known as Ultrafast Sample placement on Existing tRees (UShER), creates a central tree of COVID-19 genomic data with collapsible branches for different strains of interest, allowing users to easily add new RNA samples to the existing database without needing to recreate the entire tree. The tool was detailed in a paper published Monday in *Nature Genetics*.

"This can give you a very early indication of something that might be potentially concerning about certain strains," Yatish Turakhia, a postdoctoral researcher at the Genomics Institute at the University of California, Santa Cruz, and lead author on the paper, told *Fastinform*. "You

are basically able to draw relationships between all of these infections occurring worldwide."

The COVID-19 pandemic was one of the first to benefit from large-scale genomic tracking. There is more data about the virus's genome than any other in history, the researchers said. But as publicly available samples ballooned into the hundreds of thousands last year, scientists found that analyzing those samples or adding new ones to the mix was becoming increasingly difficult, requiring more time and computing power.

Some previous databases required users to build a new genomic tree from scratch by downloading all of the old data and integrating new samples into it. "This works when you have a few thousand sequences, but we figured out really quickly that if you wanted to maintain a tree of all the [COVID-19] sequences, this would never work. It would take you weeks," Russell Corbett-Detig, an assistant professor of biomolecular engineering at the University of California, Santa Cruz, and a co-corresponding author on the paper, told *Fastinform*. "By the time you had your new tree, it was already out of date."

After a preprocessing stage, the system can place a single RNA sample into an appropriate category in roughly half a second. That's far ahead of other platforms, which can take anywhere from half an hour to a few days to process and place a sample, according to the researchers. The platform is also just as accurate as other systems, correctly

placing samples into their corresponding nodes — or the sub-branches corresponding to different viral variants — 97.2% of the time.

Genomic tracking is critical for identifying virus variants across geographical regions — what researchers refer to as "genomic surveillance" — facilitating real-time monitoring by epidemiologists as they seek to understand how different strains are moving through the population. Certain variants may present unique challenges, such as a speedier transmission rate, a more severe set of symptoms or a higher resistance to vaccines. These samples may also provide valuable information to contact tracers, who can identify an uncommon strain in a particular city and warn nearby individuals ahead of time that they may have been exposed so that they can quarantine and prepare for possible infection.

In order to categorize a new sample, the platform automatically assesses which branch of the tree most closely aligns with its unique set of mutations. The highly efficient process allows people who may not have access to cutting-edge computing equipment to still contribute their findings to the database. The tree already includes "the vast majority of sequences that have been released," Corbett-Detig said. The researchers also partnered with the Centers for Disease Control and Prevention to develop a module that teaches scientists how to access the UShER server and upload their sequences to the platform.

In their paper, the authors noted that as of September 2020, more than 2,000 groups had gathered nearly 98,000 genome samples to help track the spread of the virus. Those numbers have continued growing rapidly over the past year, far beyond the figures presented in the paper. "We had bets on how much data would exist, and I think all of our bets were off," Turakhia said. "We all underestimated how much data we were going to be accumulating over the course of this pandemic."

"Working with SARS-CoV-2, everybody wants the answer yesterday," Corbett-Detig added. "The landscape changes very quickly."

There are now over 1 million samples from 172 countries and territories entered into the global initiative on sharing avian influenza data, or GISAID, a repository first created to track influenza variants that has since expanded to be the world's most popular platform for accessing coronavirus data.

Throughout the development and testing of UShER, the research team has been motivated by the global, collaborative effort to address the pandemic. "For the first time, we've been able to trace the evolutionary history of any species with such fine granularity and such great detail, right from the onset — the first sequence in Wuhan to this point. It's never happened before," Turakhia said.

The study "Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic" published May 10 in Nature Genetics,

was authored by Yatish Turakhia, Bryan Thornlow, Angie S. Hinrichs and David Haussler, University of California, Santa Cruz; Nicola De Maio, European Bioinformatics Institute; Landen Gozashti, University of California, Santa Cruz and Harvard University; Robert Lanfear, Australian National University; and Russell Corbett-Detig, University of California, Santa Cruz and National Research University Higher School of Economics.