Managing Mountains of Genomic Data

Tools invented at Fred Hutch are helping researchers analyze the enormous global database of genomic sequences of the virus that causes COVID-19 to better understand how pathogens evolve and spread

April 24, 2025 By Fred Hutch News Service and John Higgins

From the earliest days of the pandemic that shocked the world in 2020, researchers at Fred Hutch Cancer Center tracked the rapid evolution and spread of the virus that causes COVID-19 by studying how its genomic sequence — the order of its genetic building blocks — changed over time.

Computational biologist Trevor Bedford, PhD, and evolutionary biologist Jesse Bloom, PhD, became go-to media sources, providing expert information that influenced consequential policy decisions about what to shut down and when.

Researchers have now amassed an enormous database of genomic sequences of SARS-CoV-2 and its many variants, sampled from more than 16 million patients around the world over the last five years.

But the sheer volume of genomic sequences — a dataset that is orders of magnitude bigger than what's available for any other pathogen — has overwhelmed the capacity of common analytical methods to make sense of it in a timely and practical manner.

Two recently published papers from postdoctoral researchers in the Bedford and Bloom labs showcase new tools invented at Fred Hutch that provide researchers traction to make that mountain of data more manageable.

The Bedford study, recently published in the journal Nature, uses a new statistical technique to track the spread of a virus through a population.

The analysis of more than 100,000 genomic sequences of SARS-CoV-2 collected in Washington state between March 2021 and December 2022 captures fine-grain details about transmission between regions and age groups, including important information about the role young children played in transmission.

Meanwhile the Bloom study, recently published in the journal Cell, shows that the same method developed in his lab to analyze the effects of thousands of SARS-CoV-2 mutations in a single, safe

experiment can also be applied to understand the evolutionary capacity of viruses for which only a small number of sequences are available.

This study uses the Bloom method to safely measure the effects of mutations to a key protein from Nipah virus, a rarer but scarier and deadlier pathogen that some researchers worry could pose a risk of triggering a new outbreak or pandemic.

A new way to see the forest when the trees grow too thick

Bedford's early career focused on using genomic sequences to construct family trees of viruses and visualize their evolution as easily as someone traces the branches of their ancestry from parents to grandparents to great-grandparents.

He built phylogenetic trees to better understand influenza, Ebola, MERS and other viruses and co-founded an open-source website in 2015 called Nextstrain that posts the phylogenetic trees of global pathogens.

In February 2020, Bedford used phylogenetic trees to analyze genomic sequences of a newly emerged virus, SARS-CoV-2, that already was overwhelming hospitals in Wuhan, China.

His analysis revealed a probable transmission chain in Washington state that began in mid-January with the first diagnosed U.S. patient, a man who returned from Wuhan. The genomic sequence of a second Seattle-area patient in late February had several new mutations indicating that the virus had been spreading quickly and largely undetected in Washington.

Bedford sounded the alarm, prompting a rapid shutdown of the region that likely saved thousands of lives in the state.

Nextstrain became an essential tool for visualizing and tracing the evolution of SARS-CoV-2, but as the number of SARS-CoV-2 genomic sequences has swelled, phylogenetic trees have become increasingly unwieldy.

When those family trees grow beyond a few hundred or a few thousand sequences, it gets harder computationally to infer what's going on. There's just too many branches and twigs and offshoots to untangle.

"It just becomes computationally very costly to reconstruct that tree," said Cécile Tran-Kiem, PhD, lead author of the Nature study and a postdoctoral researcher in the Bedford Lab in the Vaccine and Infectious Disease Division at Fred Hutch.

She figured out a different way to track the spread of the virus that doesn't require building trees showing new branches that represent new mutations.

Her method instead looks for pairs of identical genomic sequences to figure out how different

subgroups of the population, such as those defined by geography or age, contribute to transmission.

The method takes advantage of a mismatch between the rate of transmission and the rate of mutation in SARS-CoV-2.

“We expect the virus to mutate every 11 to 12 days, but transmission maybe happens every six days, more or less,” Tran-Kiem said. “This means if we’re looking at viruses that haven’t mutated yet, we’re looking at people who are pretty close in a transmission chain.”

The more often pairs of identical genomic sequences straddle two groups — with one half of the pair in one group and the other half in the other group — the greater the probability of higher transmission between those groups.

The method finds statistical patterns and connections that might otherwise be missed because the virus is moving too fast in the population for the tree method based on new mutations to keep pace.

Testing the method against real-world data from Washington state

The idea made computational sense, but Tran-Kiem wanted to test it against real-world data to see if her method reached the same conclusions about the spread of the virus as conventional methods that epidemiologists use to track the spread of disease.

“Classically, it’s been shown that patterns of transmission of respiratory pathogens correlate well with mobility data — the virus tends to go where you move,” she said.

Tran-Kiem applied the method to 114,298 SARS-CoV-2 genomes collected in Washington state. The patterns of paired sequences were consistent with expectations from mobility data collected from smartphone records.

She also was able to track transmission between age groups, which also was consistent with expectations based on social contact surveys.

But her approach could potentially overcome the limitations of tracking smartphones (not everyone has one) or conducting surveys (memories of previous contacts may be inaccurate or incomplete). It also saves money by reducing the need for phone and survey data, which is costly to collect.

As predicted, her method showed that adjacent counties were more likely to be linked by identical pairs of sequences than regions that are further apart, except for adjacent counties separated by the Cascade Mountain Range. Also as expected, transmission generally flowed from Western to Eastern Washington.

Some data didn’t match expectations, however.

For example, Tran-Kiem and her colleagues found two counties that shared many more pairs of identical sequences than they should have based on their geography.

They realized that what they were seeing in the genomic data was transmission between male prisons. They discussed their results with epidemiologists and physicians working with the Washington State Department of Corrections.

They said that certain policies and procedures regarding the transfer of prisoners and staff between prisons could explain patterns that did not show up in conventional mobility data. The state has only two women’s prisons, which are in adjacent counties and didn’t stand out.

By analyzing the timing of sequence collection, she also could use the method to better understand transmission between age groups and provide some context to the highly debated role played by young children during the pandemic.

Young kids are notorious germ-factories who spread colds to classmates and bring them home from school, so initially it was not unreasonable to think they might also play a big role in transmitting COVID-19.

But while it is possible for young children to transmit the virus to adults, that’s not usually how the virus spread in Washington state.

During the alpha and delta variant waves, she found that children ages 0 to 9 could have been a source of infection for the elderly, but not younger adults.

But that pattern disappeared during the omicron wave and overall, she found no indication that young school age children were a major source of transmission in Washington state, even after schools reopened.

“We know for sure they did infect adults, we’re just saying that overall, when there was transmission between children and adults, it didn’t tend to be the kids who infected the adults. It rather tended to be in the other direction.”

Information like that — which is more precise than data about where someone’s phone has been or who they remember being around before they got sick — could be useful in the future when policymakers must make decisions about whether to close schools and for how long.

A versatile tool for data sets great and small

The abundance of genetic information about COVID-19 has helped scientists monitor the emergence of new variants and identify thousands of new mutations that may or may not improve the virus’ survival — vital information that could help us make more effective vaccines and annual

boosters.

But finding out which of those mutations matters takes considerably more time using the usual experimental methods, which test one mutation at a time.

Over the last few years, Bloom's team in the Basic Sciences Division have figured out a new way to speed that process that allows them to safely run thousands of experiments simultaneously.

They call the method pseudovirus-based deep mutational scanning, which involves genetically engineering a pseudovirus that can infect a cell like the real thing, but only once. The pseudovirus cannot reproduce and spread like an actual virus, making it much safer for experimentation.

To make the pseudovirus, they strip a live virus commonly used in research and gene therapy down to its backbone, removing its ability to replicate.

Think of it like studying the pistons of a Porsche by installing them in a modified Honda Civic that can't escape the garage because it's up on blocks with no tires.

They also add some features to the pseudovirus to streamline its use for deep mutational scanning experiments that can simultaneously measure the effects of thousands of mutations on the virus' ability to infect cells and escape virus-killing antibodies.

In a study published last summer in the journal Nature, Bloom and his colleagues describe how they used this method to identify recent mutations to the infamous "spike" protein in SARS-CoV-2 in that have helped the virus dodge our vaccines and outrun our growing natural immunity.

The analysis partially explains why some variants have had more success in humans than others, providing clues about the next steps the virus will probably take to evade our defenses.

The latest paper from the Bloom Lab, however, shows how the method can also be used for viruses on the other end of the data spectrum, which occur in humans so rarely that the genomic sequences available to researchers are numbered in the dozens rather than millions.

The lead author, Brendan Larsen, PhD, a postdoctoral researcher in the Bloom Lab, studied the Nipah virus, which occasionally jumps from fruit bats to humans and causes outbreaks nearly every year in Southeast Asia.

"It's one of the deadliest viruses that we know about," Larsen said. "Luckily, it doesn't spill over too much from bats to humans."

Nipah causes mild to severe disease, including swelling of the brain and death with a mortality rate as high as 70%.

The 2011 movie Contagion, which squeamish folks were advised not to watch during the COVID-19 pandemic, featured a fictional cross of Nipah and influenza that soon spreads around the world,

terrorizing humanity.

So far, outbreaks of the real Nipah virus have not spread widely in humans, but it is considered an emerging pathogen with the potential to trigger a pandemic.

Nipah belongs to a family called paramyxoviruses that Larsen studied when he was a PhD student at the University of Arizona, where he caught and swabbed hundreds of bats for his research.

“I never found Nipah-like viruses in any of these samples, but it began my interest in paramyxoviruses in general,” he said.

Paramyxoviruses have an unusual way of infecting host cells, using a protein on its surface to grab and bind a host cell, and a separate protein to meld with and infect the host cell.

“Most viruses do the grabbing and the melding with a single protein,” Larsen said. “All paramyxoviruses, including Nipah, require these two separate proteins to coordinate together, so there’s this added level of complexity.”

Larsen saw an opportunity to study the grabby protein using the same pseudovirus-based deep mutational scanning (DMS) method the lab has used to manage the huge SARS-CoV-2 database.

“Brendan had this idea to do this for Nipah virus, which is much less studied than COVID or influenza viruses,” Bloom said. “It’s definitely a very general tool we can apply to a lot of different viral proteins.”

Avoiding biohazards and information hazards

Because there is no vaccine or licensed treatment specific for Nipah, the live virus is typically studied only in a Biosafety Level 4 laboratory, the highest level of security possible where scientists work in full-body, air-supplied positive pressure suits.

The Bloom Lab’s method, however, enables the safe study of key features of the virus without handling the real thing.

Rather than experiment with human cells, Larsen used Chinese hamster ovary cells, a common lab cell culture line. He engineered the cells to express the receptor for Nipah virus found in flying foxes, a species of bat that Nipah already infects.

The Bloom Lab takes that precaution to reduce “information hazard” — the risk that their discovery of significant mutations could give ideas to bad actors about to make Nipah even worse for humans.

“We don’t really understand how the virus is mutating and evolving in the wild because there’s fewer than 100 whole genomes total that we have sequenced for all of Nipah,” Larsen said. “It’s really hard to say what the virus is going to do next when we have such limited information. That’s

where DMS comes in and really gives us a much better overview of the protein."

He focused on the protein that grabs and binds to the host cell, but found mutations that likely influence its coordination with the other key protein that melds with the host cell.

"We know they interact to infect cells, but we have no idea how," Larsen said. "That's the holy grail of paramyxovirus research. We think we've found some mutations or sites that are probably involved in that interaction."

He also identified mutations that could enhance its already strong binding ability.

No formal antibody therapies exist for Nipah, but some have been tried as compassionate, last-resort measures.

Larsen tested the effects of six antibodies with the potential to kill Nipah to see if the virus has already devised mutations that would enable it to escape those antibodies, or if Nipah could tolerate such mutations if they evolved.

Doctors can use this data to see if a patient's particular Nipah sequence has mutations that will render some antibodies less effective than others.

"We've learned that resistance to these sorts of antibodies can be a big deal, so understanding the possible mechanisms of resistance ahead of time has a lot of value," Bloom said.

This work was supported by grants from the National Institutes of Health, the Centers for Disease Control and Prevention, Gates Ventures, the private office of Bill Gates (Seattle Flu Study award), Howard Hughes Medical Institute, Fast Grants (part of Emergent Ventures at the Mercatus Center, George Mason University), Washington State Department of Health, shared resources of the Fred Hutch/University of Washington/Seattle Children's Cancer Consortium, Fred Hutch Scientific Computing, Burroughs Welcome Fund, and the University of Washington Arnold and Mabel Beckman Cryo-EM Center.

John Higgins, a staff writer at Fred Hutch Cancer Center, was an education reporter at The Seattle Times and the Akron Beacon Journal. He was a Knight Science Journalism Fellow at MIT, where he studied the emerging science of teaching. Reach him at jhiggin2@fredhutch.org or @jhigginswriter.bsky.social.

This article was originally published April 22, 2025, by Fred Hutch News Service. It is republished with permission.

© 2026 Smart + Strong All Rights Reserved.

http://beta.docker.cancerhealth.com/article/managing-mountains-genomic-data