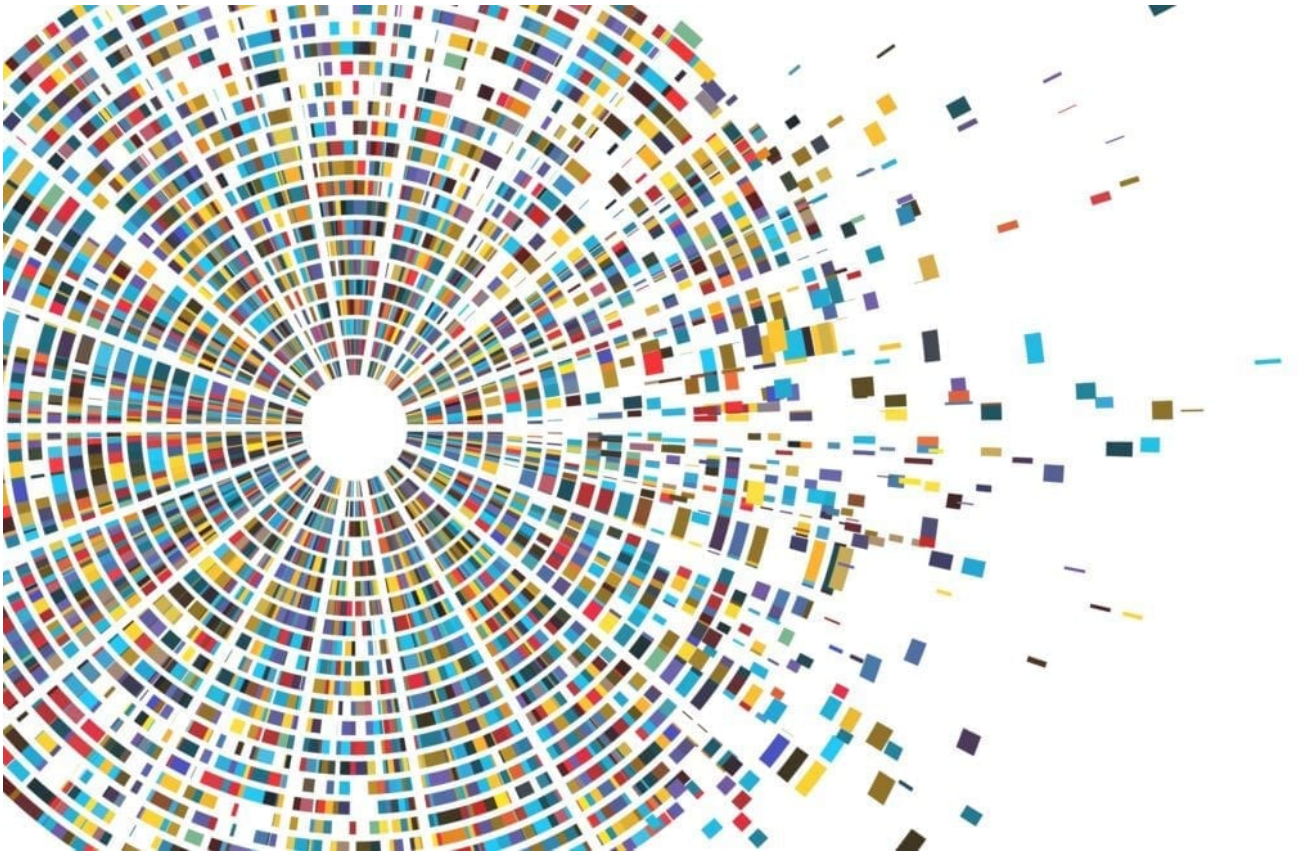
Supercomputers, Genome Sequencing, and the Race to Decode Coronavirus

16 March 2020



Update June 23, 2020: *Researchers around the world continue to hunt for an effective coronavirus vaccine, sifting through immense genomics datasets and gathering information from an ever-growing pool of COVID-19 patients. To support those efforts, Lenovo and Intel partnered to offer supercomputing resources to BGI, including the use of Lenovo's Genomics Optimization And Scalability Tool (GOAST)—a custom-built architecture optimized to accelerate genomics analytics. The ongoing collaboration remains a valuable part of the battle against the pandemic, complementing other supercomputer-assisted research.*Billions of coronavirus particles can swarm through just one drop of respiratory fluid. Each of those particles—many with subtle variations—contains some 30,000 DNA bases. That staggering biological density is buried in the viral genome, which defines every aspect of the virus, from virulence to transmissibility. Researchers in China and across the globe already sequenced a handful of COVID-19 genomes, arming clinicians and scientists with the essential foundation to begin fighting back.

But mapping the genomes among circulating coronaviruses is just the beginning. Exponentially more challenging, and equally important to understanding this pandemic, is sequencing the genomes of the people infected. Like the virus, questions swarm: Why is this virus more contagious than others? What mechanisms lead to pneumonia and

hospitalization in some and a mild cough in others? How will individual patients respond to different treatments or vaccines? The answers are written somewhere in the interaction between an individual's genome and the viral strains infecting them. Once mapped, that DNA-driven interplay points the way to diagnostics, vaccines, and immunotherapies.

Researchers at BGI Genomics—who also developed the first diagnostic test kits for the disease—and others around the world are engaging in the painstaking process of population-scale genomics for COVID-19. To engineer an effective vaccine or other protective measures, scientists need massive datasets to pinpoint potentially protective genetic differences. Working at that scale generates terabytes to petabytes of data—too much to process without large-scale analyses supported by high-performance computing (HPC).

The speed of genome sequencing has risen in stride with the rapid acceleration of computational power. A process that initially spanned more than a decade and cost billions for a single genome can now be executed in a matter of hours on clusters of supercomputers running fully optimized hardware architecture. While researchers stress that the road to a vaccine is likely very long, an unprecedented array of tools may accelerate timelines.

"BGI Genomics has sequenced hundreds of clinical samples to analyze and distinguish the complications of the infection," said Xiangqian Jin, CIO of BGI Genomics. "Having access to the latest high-performance computing and genomics analytics technologies are important factors in improving analysis efficiency."



BGI Genomics researcher working on the T7 sequencer

To support BGI's efforts and empower researchers leading the battle against coronavirus, Intel and Lenovo partnered to donate a dedicated supercomputer cluster as well as the software and hardware expertise to maximize its use.

"We are humbled to contribute to the critical efforts of genomics researchers and healthcare providers on the front line of the fight against the novel coronavirus," said Mileidy Giraldo, Ph.D., Global Lead for Genomics R&D at Lenovo. Dr. Giraldo spent years as a bioinformatics scientist at NIH contributing to vaccine design for infectious diseases and now helps bridge the gap between scientists and engineers developing hardware and software for life sciences. "We are donating equipment and expertise, but the real breakthrough, the real contribution will come from what the researchers at BGI will accomplish and what the rest of the biomedical community will in turn learn and develop based on BGI's findings."Imagine if the genome of every human being on earth could be represented by their own thousand-page book. This strange text would use only four letters: AGCT. The limited alphabet—signifying the base pairings in DNA—provides the instructions for every single feature that makes you who you are: hair color, height, and even susceptibility to a disease like COVID-19. Most of those instructions are spelled identically person to person, but all-important variations are hidden across a handful of pages.

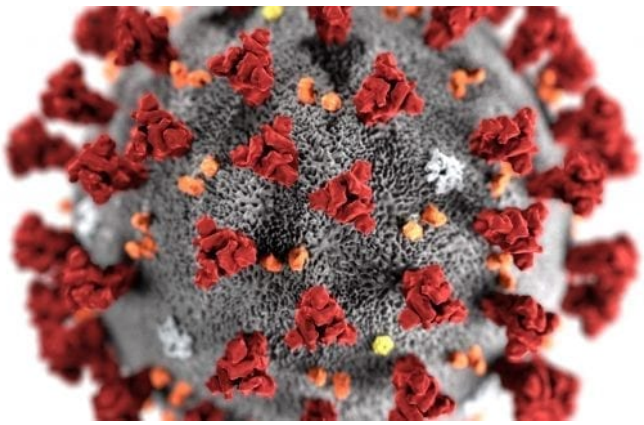


To understand how a feature like susceptibility to infection will manifest in a given person, researchers must identify the precise pages (i.e. the genes) giving the relevant instructions. This can only be done by comparing those pages across as many patients as possible—pinpointing useful commonalities—and then leveraging data linking those variations to susceptibility or resistance to infection.

This tremendously difficult task of genomics translation and interpretation sits at the core of scientists' battle against the novel coronavirus. Decoding the complex interplay between the relevant human gene products and the spike-crowned virus can reveal ways to inhibit or completely shut down the process. Scientists will also hunt for common pages in the coronavirus's own book—regions of the genome where the virus cannot tolerate mutations or variations. Those regions point to exploitable weaknesses—a kind of Achilles' heel that could open the virus to an effective vaccine or treatment.There's a kernel of truth to the trope

in outbreak movies where scientists desperately hunt for one person immune to the disease. Natural immunities can, in fact, provide key insights to engineer an effective vaccine.

“What’s missing in those movies is a realistic time scale,” Dr. Giraldo said. “Movies don’t show all the genomics work, spanning months to years, needed to compare large datasets within and across patients to create a candidate vaccine. Then, the movies ignore the time involved in clinical trials, testing many tweaks of a vaccine until we find one with high efficacy and low risk.”



A rendering of the coronavirus revealing the ring of crown-like spikes that inspired the virus name (corona means “crown” in Latin)

Consider two patient reactions to the virus: one develops life-threating pneumonia and the other gets only a passing cough. What underlying differences explain this disparity? A compromised immune system? A genetic predisposition? Past exposure to a different illness? Age? Sex? Nutrition? The dominance of one strain of the virus? Answering those questions for just two patients is already a challenge—applied to thousands, the complexity soars.

But only a massive amount of data can even begin to sift through the countless variations in both genes and environmental influences. The more clinical and genomic data scientists have, the better they can isolate the key commonality across patients.

Population-scale genomics provides a path composed of billions upon billions of data points. With the novel coronavirus, scientists hope to compare DNA across tens of thousands of diagnosed patients. This is an overwhelming computational challenge, one requiring the processing power and data storage capacity available only inside an HPC environment.Remember that genome book? The one that took an entire decade to read the first time? Genomics researchers worldwide generally analyze an entire genome in about 150 hours — a fantastic leap, certainly, but still unequal to the speed demanded by the COVID-19 pandemic. Even isolating and sequencing the bits that code for protein and propagate viruses—a handful of pages called exomes—usually takes at least 4 hours.

Now, BGI researchers can access HPC clusters optimized to assemble and analyze hundreds of whole genomes and thousands of exomes.

“With this donation, our hope is to extend the existing resources BGI researchers already have at their disposal so that the biomedical community can do more and get there faster,” Dr. Giraldo said. “I can’t think of a better example of using technology to tackle humanity’s greatest challenges than one where a multidisciplinary team of scientists, clinicians, and engineers have come together to pool their combined brain power to fight back the coronavirus global pandemic.”

> I can’t think of a better example of using technology to tackle humanity’s greatest challenges than one where a multidisciplinary team of scientists, clinicians, and engineers have come together to pool their combined brain power to fight back the coronavirus global pandemic.

Building on a powerful solution first developed by Intel, Lenovo developed an optimized hardware and system architecture to radically reduce those genome processing times. Lenovo’s solution for population-level genomics—the Genomics Optimization and Scalability Tool (GOAST)—leverages the Broad Institute’s open-source Genome Analysis Toolkit (GATK) software on an optimized hardware recipe. Identifying the right optimizations and hardware building blocks to accelerate genomics required testing hundreds of HPC configurations.

“The year-long process focused entirely on the real needs of scientists,” Giraldo said. “Researchers’ time is better spent by focusing on the science; not on the underlying hardware. So we performed a systematic permutation test of all the hardware building blocks available to us to find the right hardware recipe that reduced execution time. Our tests used the same software used by researchers in the lab to make this tool immediately deployable.”

The results? A whole human genome sequenced in five and a half hours, and exomes in just four minutes—up to a 40-fold speed-up. Supported by a dedicated supercomputing cluster, BGI researchers will soon be hard at work using GOAST to study COVID-19 on the long road to a vaccine.

In the short term, predicting virulence based on a patient’s dominant strains may also help hospitals more effectively triage patients—knowing who is at greatest risk as soon as they reach a clinic and what therapies may be effective. In the long term, even beyond a vaccine, the COVID-19 genome contains hints of its source. Knowing its genomic history and point of origin can help predict and prevent future outbreaks.

All in all, a staggeringly dense and high-stakes puzzle to solve.

“The equipment and technology will speed up the rapid identification of COVID-19 infected people and the study of virus genome characteristics, providing strong support for accurate diagnosis, treatment and epidemic prevention of COVID-19,” Jin said.