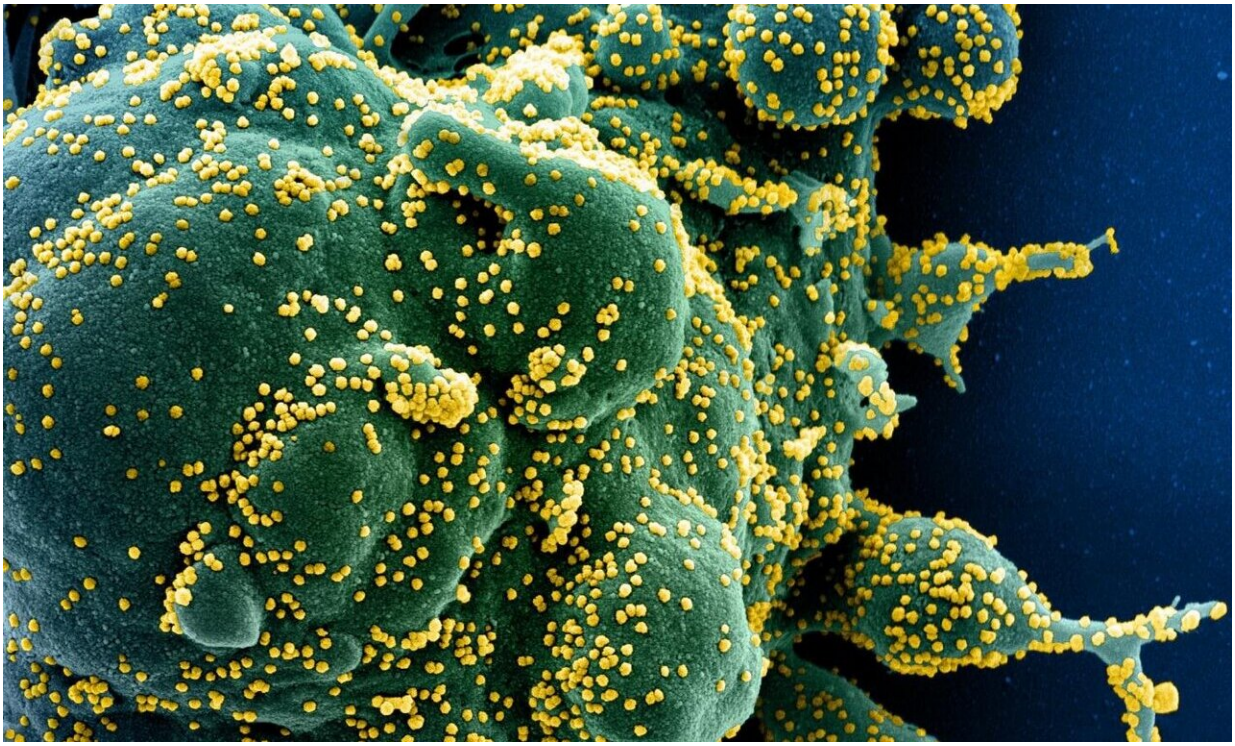
Data analysis identifies the 'mother' of all SARS-CoV-2 genomes

November 5 2020



Colorized scanning electron micrograph of an apoptotic cell (green) heavily infected with SARS-COV-2 virus particles (yellow), isolated from a patient sample. Image captured at the NIAID Integrated Research Facility (IRF) in Fort Detrick, Maryland. Credit: NIH/NIAID

In the field of molecular epidemiology, the worldwide scientific community has been sleuthing to solve the riddle of the early history of

SARS-CoV-2.

Since the first SARS-CoV-2 virus infection was detected in December 2019, tens of thousands of its genomes have been sequenced worldwide, revealing that the coronavirus is mutating, albeit slowly, at a rate of 25 mutations per genome per year.

But despite major efforts, no one to date has identified the first case of human transmission, or "patient zero" in the COVID-19 pandemic. Finding such a case is necessary to better understand how the virus may have jumped from its animal host first to infect humans as well as the history of how the SARS-CoV-2 viral genome has mutated over time and spread globally.

"The SARS-CoV-2 virus is carrying an RNA genome that has already infected more than 35 million people across the world," said Sudhir Kumar, director of the Institute for Genomics and Evolutionary Medicine, Temple University. "We need to find this common ancestor, which we call the progenitor genome."

This progenitor genome is the mother of all SARS-CoV-2 coronaviruses infecting people today.

In the absence of patient zero, Kumar and his Temple University research team now may have found the next best thing to aid the worldwide molecular epidemiology detective work. "We set out to reconstruct the genome of the progenitor by using a big dataset of coronavirus genomes obtained from infected individuals," said Sayaka Miura, a senior author of the study.

They found the "mother" of all SARS-CoV-2 genomes and its early offspring strains have subsequently mutated and spread to dominate the world pandemic. "We have now reconstructed the progenitor genome

and mapped where and when the earliest mutations happened," said Kumar, the corresponding author of a preprint study, which can be found on the bioRxiv server.

In doing so, their work has provided new insights into the early mutational history of SARS-CoV-2. For example, their study reports that a mutation of the SARS-CoV-2 spike protein (D416G), often implicated in increased infectivity and spread, occurred after many other mutations, weeks after COVID-19 started. "It is nearly always found alongside many other protein mutations, so its role in increased infectivity remains difficult to establish," said Sergei Pond, a senior co-author of the study.

Besides their findings on SARS-CoV-2's early history, Kumar's group has developed mutational fingerprints to quickly recognize strains and sub-strains infecting an individual or colonizing a global region.

Order to a pandemic

To identify the progenitor genome, they used a mutational order analysis technique, which relies on a clonal analysis of mutant strains and the frequency in which pairs of mutations appear together in the SARS-CoV-2 genomes.

First, Kumar's team sifted through data on almost 30,000 complete genomes of the SARS-CoV-2, the virus that causes COVID-19. Altogether, they analyzed 29,681 SARS-CoV-2 genomes, each containing at least 28,000 bases of sequence data. These genomes were sampled between 24 December 2019 and 07 July 2020, representing 97 countries and regions worldwide.

Many previous attempts in analyzing such large datasets were not successful because of "the focus on building an evolutionary tree of SARS-CoV-2," says Kumar. "This coronavirus evolves too slow, the

number of genomes to analyze is too large, and the data quality of genomes is highly variable. I immediately saw parallels between the properties of these genetic data from coronavirus with the genetic data from the clonal spread of another nefarious disease, cancer."

Kumar's group has developed and investigated many techniques for analyzing genetic data from tumors in cancer patients. They adapted and innovated those techniques and built a trail of mutations that automatically traces back to the progenitor. "Basically, the genome before the first mutation was that of the progenitor," said Kumar. "The mutation tracking approach is beautiful and predicts a phylogeny of "major strains" of SARS-CoV-2. It is a great example of how big data coupled with biologically-informed data mining reveals important patterns."

Progenitor genome

Kumar's team uncovered a predicted sequence of the progenitor (mother) genome of all SARS-CoV-2 genomes (proCoV2). In the proCoV2 genome, they identified 170 non-synonymous (mutations that cause an amino acid change in a protein) and 958 synonymous substitutions compared with the genome of a closely-related coronavirus, RaTG13, found in a Rhinolophus affinis bat. While the intermediary animal from bats to humans is still unknown, this amounted to a 96.12% sequence similarity between proCoV2 and RaTG13 sequences.

Next, they identified 49 single nucleotide variants (SNVs) that occurred with a greater than 1% variant frequency from their dataset. These were further examined to look at their mutational patterns and global spread.

"The tree of mutations predicts a tree of strains," said Kumar. "You can also do the tree of strains first, and predict the order of mutations. However, this way is greatly affected by the quality of sequences. When

the mutation rate is low, it becomes hard to distinguish between error due to low quality and a real mutation. The approach we took is much more robust against sequencing errors because analysis of pairs of positions across genomes is more informative."

An earlier timeline emerges

When comparing the inferred proCoV2 sequence with genomes in their collection revealed no full matches at the nucleotide level, Kumar's research team knew the original timing of the start of the pandemic was off.

"This progenitor genome had a sequence different from what some folks are calling the reference sequence, which is what was observed first in China and deposited into the GISAID SARS-CoV-2 database," said Kumar.

The closest match was to genomes sampled 12 days after the earliest sampled virus that became available on 24 December 2019. Multiple matches were found in all sampled continents and detected as late as April 2020 in Europe. Overall, 120 genomes Kumar's group analyzed all contained only synonymous differences from proCoV2. That is, all their proteins were identical to the corresponding proCoV2 proteins in the amino acid sequence. A majority (80 genomes) of these protein-level matches were from coronaviruses sampled in China and other Asian countries.

These spatiotemporal patterns suggested that proCoV2 already possessed the full repertoire of protein sequences needed to infect, spread and persist in the global human population.

They found the proCoV2 virus and its initial descendants arose in China, based on the earliest mutations of proCoV2 and their locations.

Furthermore, they also demonstrated that a population of strains with as many as six mutational differences from proCoV2 existed at the time of the first detection of COVID-19 cases in China. With estimates of SARS-CoV-2 mutating 25 times per year, this meant that the virus must already have been infecting people several weeks before the December 2019 cases.

Mutational signatures

Because there was strong evidence of many mutations before the ones found in the reference genome, Kumar's group had to come up with a new nomenclature of mutational signatures to classify SARS-CoV-2 and account for these by introducing a series of Greek letter symbols to represent each one.

For example, they found that the emergence of μ and α SARS-CoV-2 genome variants came before the first reports of COVID-19. This strongly implies the existence of some sequence diversity in the ancestral SARS-CoV-2 populations. All 17 of the genomes sampled from China in December 2019, including the designated SARS-CoV-2 reference genome, carry all three μ and three α variants. Interestingly, the six genomes containing μ variants but not α variants were sampled in China and the United States in January 2020. Therefore, the earliest sampled genomes (including the designated reference) were not the progenitor strains.

It also predicts the progenitor genome had offspring that were spreading worldwide during the earliest phases of COVID-19. It was ready to infect right from the start.

"The progenitor had all the ability it needed to spread," said Sergei Pond. "There is little evidence of selection on lineages between bats and humans, although there is strong selection on coronaviruses in bats."

Hitchhiking mutations

Furthermore, they found confounding evidence that there was always another mutation that accompanied the D416G spike protein mutation.

"Many people are interested in mutations in the spike protein because of its functional properties," said Kumar. "But what we are observing is that in addition to the spike protein, there were several additional changes within the genome that are always found along with the changes in the spike protein (D416G). We call these a beta group of mutations, and the spike mutation is one of them. Whatever we think the spike mutation is doing, it is best not to forget that other mutations may also be involved. Alternatively, these mutations may be simply hitchhiking together, we yet cannot tell."

"What is also interesting is that the genome containing the spike protein mutation underwent many other mutations. And what we call epsilon mutations (there are 3 of these) occurred on the background of the spike mutation, and they change arginine residues in a very important protein, the nucleocapsid (N) protein. The epsilon mutations are widespread in Europe, and they are always found with the spike protein mutation. So, epsilon mutations started a dominant offshoot in both Europe and Asia."

A global spread

Altogether, they have identified seven major evolutionary lineages that arose after the pandemic began, some of which arose in Europe and North America after the genesis of the ancestral lineages in China.

"Asian strains founded the whole pandemic," said Kumar. "But over time, it is the sub-strain containing the epsilon mutation, that may have occurred outside of China (first observed in the middle east and

Europe), is infecting Asia much more."

Their mutational-based analyses also established that North American coronaviruses harbor very different genome signatures than those prevalent in Europe and Asia.

"This is a dynamic process," said Kumar. "Clearly, there are very different pictures of spread that are painted by the emergence of new mutations, the three epsilons, gamma, and delta, which we found to occur after the spike protein change. We need to find out if any functional properties of these mutations have sped up the pandemic."

Next steps

Moving forward, they will continue to refine their results as new data becomes available.

"There are more than 100,000 SARS-CoV-2 genomes that have been sequenced now," said Pond. Kumar says that "the power of this approach is that the more data you have, the more easily you can tell the precise frequency of individual mutations and mutation pairs. These variants that are produced, the single nucleotide variants, or SNVs, their frequency, and history can be told very well with more data. Therefore, our analyses infer a credible root for the SARS-CoV-2 phylogeny."

Their results are being automatically updated online as new genomes are reported (which now exceeds 50,000 samples and can be found at http://igem.temple.edu/COVID-19).

"These findings and our intuitive mutational fingerprints of SARS-CoV-2 strains have overcome daunting challenges to develop a retrospective on how, when and why COVID-19 has emerged and spread, which is a prerequisite to creating remedies to overcome this

pandemic through the efforts of science, technology, public policy and medicine," said Kumar.

More information: Sudhir Kumar et al. An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic, (2020). DOI: 10.1101/2020.09.24.311845

Provided by Temple University