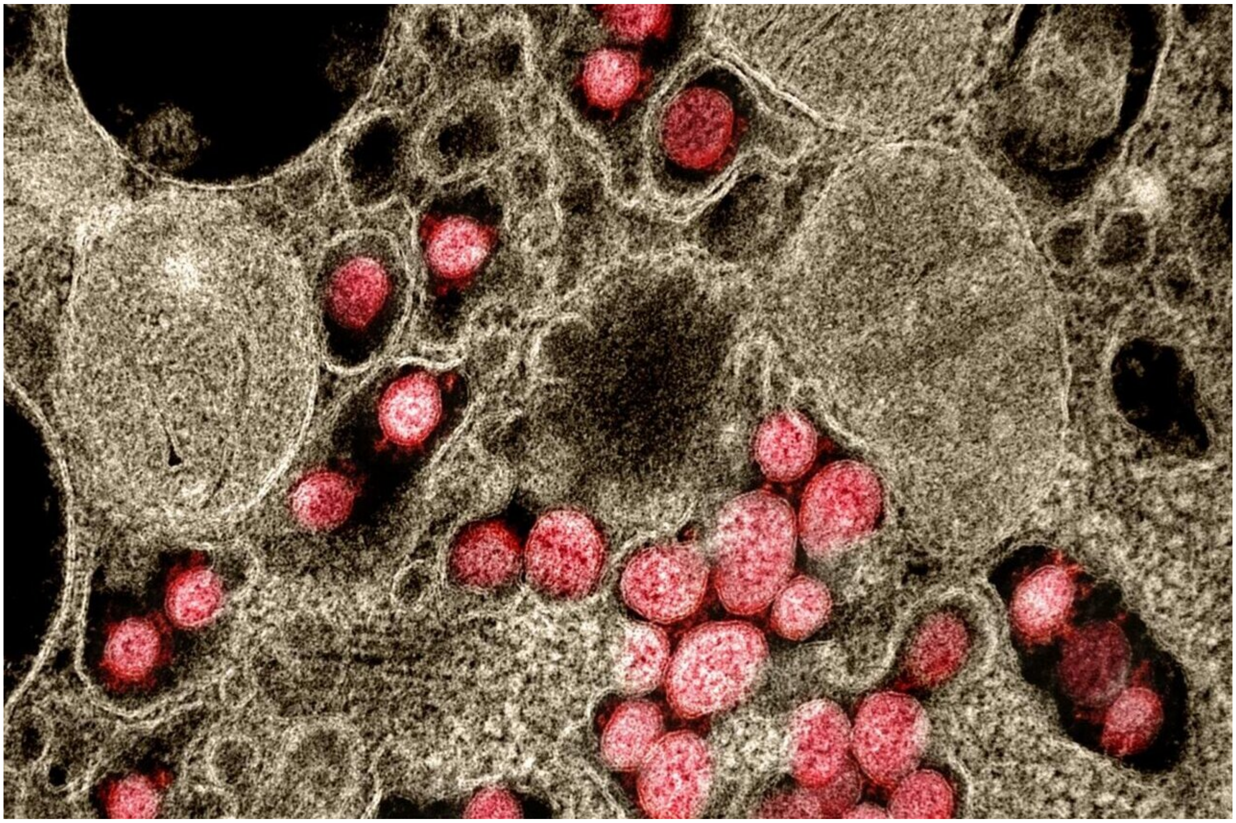
PHYS.ORG

# A comprehensive map of the SARS-CoV-2 genome

May 11 2021



Transmission electron micrograph of SARS-CoV-2 virus particles isolated from a patient. Credit: NIAID

In early 2020, a few months after the Covid-19 pandemic began, scientists were able to sequence the full genome of the virus that causes the infection, SARS-CoV-2. While many of its genes were already known at that point, the full complement of protein-coding genes was unresolved.

Now, after performing an extensive comparative genomics study, MIT researchers have generated what they describe as the most accurate and complete gene annotation of the SARS-CoV-2 genome. In their study, which appears today in *Nature Communications*, they confirmed several protein-coding genes and found that a few others that had been suggested as genes do not code for any proteins.

"We were able to use this powerful comparative genomics approach for evolutionary signatures to discover the true functional protein-coding content of this enormously important genome," says Manolis Kellis, who is the senior author of the study and a professor of computer science in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) as well as a member of the Broad Institute of MIT and Harvard.

The research team also analyzed nearly 2,000 mutations that have arisen in different SARS-CoV-2 isolates since it began infecting humans, allowing them to rate how important those mutations may be in changing the virus' ability to evade the immune system or become more infectious.

## Comparative genomics

The SARS-CoV-2 genome consists of nearly 30,000 RNA bases. Scientists have identified several regions known to encode protein-coding genes, based on their similarity to protein-coding genes found in related viruses. A few other regions were suspected to encode proteins, but they had not been definitively classified as protein-coding genes.

To nail down which parts of the SARS-CoV-2 genome actually contain genes, the researchers performed a type of study known as comparative genomics, in which they compare the genomes of similar viruses. The SARS-CoV-2 virus belongs to a subgenus of viruses called Sarbecovirus, most of which infect bats. The researchers performed their analysis on SARS-CoV-2, SARS-CoV (which caused the 2003 SARS outbreak), and 42 strains of bat sarbecoviruses.

Kellis has previously developed computational techniques for doing this type of analysis, which his team has also used to compare the human genome with genomes of other mammals. The techniques are based on analyzing whether certain DNA or RNA bases are conserved between species, and comparing their patterns of evolution over time.

Using these techniques, the researchers confirmed six protein-coding genes in the SARS-CoV-2 genome in addition to the five that are well established in all coronaviruses. They also determined that the region that encodes a gene called ORF3a also encodes an additional gene, which they name ORF3c. The gene has RNA bases that overlap with ORF3a but occur in a different reading frame. This gene-within-a-gene is rare in large genomes, but common in many viruses, whose genomes are under selective pressure to stay compact. The role for this new gene, as well as several other SARS-CoV-2 genes, is not known yet.

The researchers also showed that five other regions that had been proposed as possible genes do not encode functional proteins, and they also ruled out the possibility that there are any more conserved protein-coding genes yet to be discovered.

"We analyzed the entire genome and are very confident that there are no other conserved protein-coding genes," says Irwin Jungreis, lead author of the study and a CSAIL research scientist. "Experimental studies are needed to figure out the functions of the uncharacterized genes, and by determining which ones are real, we allow other researchers to focus their attention on those genes rather than spend their time on something that doesn't even get translated into protein."

The researchers also recognized that many previous papers used not only incorrect gene sets, but sometimes also conflicting gene names. To remedy the situation, they brought together the SARS-CoV-2 community and presented a set of recommendations for naming SARS-CoV-2 genes, in a separate paper published a few weeks ago in *Virology*.

## Fast evolution

In the new study, the researchers also analyzed more than 1,800 mutations that have arisen in SARS-CoV-2 since it was first identified. For each gene, they compared how rapidly that particular gene has evolved in the past with how much it has evolved since the current pandemic began.

They found that in most cases, genes that evolved rapidly for long periods of time before the current pandemic have continued to do so, and those that tended to evolve slowly have maintained that trend. However, the researchers also identified exceptions to these patterns, which may shed light on how the virus has evolved as it has adapted to its new human host, Kellis says.

In one example, the researchers identified a region of the nucleocapsid protein, which surrounds the viral genetic material, that had many more mutations than expected from its historical evolution patterns. This protein region is also classified as a target of human B cells. Therefore, mutations in that region may help the virus evade the human immune system, Kellis says.

"The most accelerated region in the entire genome of SARS-CoV-2 is sitting smack in the middle of this nucleocapsid protein," he says. "We speculate that those variants that don't mutate that region get recognized by the human immune system and eliminated, whereas those variants that randomly accumulate mutations in that region are in fact better able to evade the human immune system and remain in circulation."

The researchers also analyzed mutations that have arisen in variants of concern, such as the B.1.1.7 strain from England, the P.1 strain from Brazil, and the B.1.351 strain from South Africa. Many of the mutations that make those variants more dangerous are found in the spike protein, and help the virus spread faster and avoid the immune system. However, each of those variants carries other mutations as well.

"Each of those variants has more than 20 other mutations, and it's important to know which of those are likely to be doing something and which aren't," Jungreis says. "So, we used our comparative genomics evidence to get a first-pass guess at which of these are likely to be important based on which ones were in conserved positions."

This data could help other scientists focus their attention on the mutations that appear most likely to have significant effects on the virus' infectivity, the researchers say. They have made the annotated gene set and their mutation classifications available in the University of California at Santa Cruz Genome Browser for other researchers who wish to use it.

"We can now go and actually study the evolutionary context of these variants and understand how the current pandemic fits in that larger history," Kellis says. "For strains that have many mutations, we can see which of these mutations are likely to be host-specific adaptations, and which mutations are perhaps nothing to write home about."

**More information:** Irwin Jungreis et al, SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus

genomes, *Nature Communications* (2021). DOI: 10.1038/s41467-021-22905-7

Provided by Massachusetts Institute of Technology

Citation: A comprehensive map of the SARS-CoV-2 genome (2021, May 11) retrieved 11 May 2024 from https://phys.org/news/2021-05-comprehensive-sars-cov-genome.html