TOOL TO EASE SARS-COV-2 GENOME MUTATION ANALYSIS

Relevant for: Science & Technology | Topic: Achievements of Indians in science & technology

Convergence: Our analysis revealed 13 hotspot residues across the SARS-CoV-2 genome that occur at least in 40,000 or more samples, says Amit Dutt (right) | Photo Credit: Bhasker D

An automated computational tool – Infectious Pathogen Detector (IPD) – developed earlier by researchers at the Mumbai-based ACTREC, Tata Memorial Centre, to identify the presence of 1,060 different pathogens in any genome sequence sample and perform mutation and phylogenetic analysis has become even more useful with the addition of a module for SARS-CoV-2 virus.

The IPD tool has been already designed to perform analysis of diverse genomic datasets, which came handy while analysing diverse data sets of SARS-CoV-2 genome that have been uploaded to the GISAID database from across the globe. The diversity of SARS-CoV-2 genome sequence data in the GISAID database arises because of different sequencing platforms being located across the world. Different sequencing platforms being used generate either high-density but shorter read-length or low-density but higher read-length.

"To parse this plural kind of dataset requires distinct downstream pipelines which make the analysis complicated and difficult to compare against each other," says Dr. Amit Dutt from the Tata Memorial Centre and the lead author of a paper published in the journal *Briefings in Bioinformatics.* "But we have automated the entire process thereby allowing users to analyse in a stringent and statistically disciplined manner the SARS-CoV-2 genome data without being restricted by the platform used to generate the data."

Explaining the uniqueness of the IPD tool, Dr. Dutt says that it can automatically determine the abundance of SARS-CoV-2 genome sequences, carry out mutation analysis with respect to the Wuhan sequence and finally, based on the mutations seen in each sample, assign it to the respective phylogenetic clade. Assigning a sample to a phylogenetic clade is based on the complete profile of mutations seen in the sample.

"Researchers can either upload sequence data to the IPD server which then automatically analyses the data for mutations and then assign the sample to the respective phylogenetic clade or download the tool before using it for bulk analysis," says Dr. Dutt. Using the tool, the researchers analysed over 2,00,000 SARS-CoV-2 genome sequences available in the GISAID database. Only those with high-quality sequence data were included for analysis as the tool automatically rejects those with inferior quality. In over 2,00,000 sequences analysed, they found 2.58 million mutations in all with 6.6 nonsynonymous mutations (that do not alter the amino acid sequence) and five synonymous mutations (that alter the amino acid sequence) per sample. The results are posted on *bioRxiv* preprint server. Preprints are yet to be peer-reviewed.

"Our analysis revealed 13 hotspot residues across the SARS-CoV-2 genome that occur at least in 40,000 or more samples. This includes the D614G, one of the first mutations described in the spike protein," says Dr. Dutt. "Interestingly, none of the more recent spike glycoprotein mutations — N439K, S477Y, E484K, and N501Y — were found to be significantly abundant in the current variants in Britain, Brazil and South Africa."

The 13 hotspot mutations are occurring at a high frequency as seen in their presence in at least

40,000 samples. "So there is some kind of repetitive convergent evolution taking place. The 13 hotspot mutations which have been selected for are occurring independently," he cautions. "Besides hotspot mutations, we also see mutations in specific sub-clades. So there is adaptive and convergent evolution."

They found that the mutation rate of both nonsynonymous and synonymous mutations in 3,361 Indian COVID-19 sequence samples was comparable with the global rate. They also found 4,422 unique mutations that have not been reported outside India. "The hotspot mutations were seen in the Indian samples as well, including the D614G spike protein mutation. However, no significant occurrence of N439K, E484K, or N501Y mutations were found, except in two samples that harboured the S477Y spike protein mutation," he says.

According to Sanket Desai, the first author of the journal paper and the preprint, mutations are taking place randomly and selection will happen over time. It is just a matter of time before mutations that give the virus better fitness emerge. Viruses with such mutations will have either more transmissibility, as seen in the Britain variant or immune escape as seen in the South African variant.

Chances of the rise of dangerous mutations that render the virus greater fitness are high due to persistence of the pandemic in some countries. With just over 5,100 sequences from India, of which only 4,041 are complete and high-coverage, there is no way of knowing if new variants first reported in Britain, Brazil and South Africa are already present in India and whether new mutations so far unreported elsewhere that render better fitness to the virus have already emerged here.

Despite the COVID-19 task force mandating 5% of positive samples to be sequenced from all the States and Union Territories, the Indian SARS-CoV-2 Genomics Consortium (INSACOG) is far from reaching the target percentage.

With the SARS-CoV-2 genome being just about 30 kb in size, it is possible to pool up to 1,000 samples into one and carry out the sequencing at high coverage of 1,000x in one go and still be far less than 15 Gb sequencing capacity of platforms routinely used in Indian labs. High throughput will also help cut down the sequencing cost per sample and help have the data after analysis in about 10 days.

*Our Digital Subscription plans do not currently include the e-paper, crossword and print.

Dear reader,

We have been keeping you up-to-date with information on the developments in India and the world that have a bearing on our health and wellbeing, our lives and livelihoods, during these difficult times. To enable wide dissemination of news that is in public interest, we have increased the number of articles that can be read free, and extended free trial periods. However, we have a request for those who can afford to subscribe: please do. As we fight disinformation and misinformation, and keep apace with the happenings, we need to commit greater resources to news gathering operations. We promise to deliver quality journalism that stays away from vested interest and political propaganda.

Dear subscriber,

Thank you!

Your support for our journalism is invaluable. It's a support for truth and fairness in journalism. It has helped us keep apace with events and happenings.

The Hindu has always stood for journalism that is in the public interest. At this difficult time, it becomes even more important that we have access to information that has a bearing on our health and well-being, our lives, and livelihoods. As a subscriber, you are not only a beneficiary of our work but also its enabler.

We also reiterate here the promise that our team of reporters, copy editors, fact-checkers, designers, and photographers will deliver quality journalism that stays away from vested interest and political propaganda.

Suresh Nambath

Please enter a valid email address.

Here are some of the most interesting research papers to have appeared in top science journals last week.

You can support quality journalism by turning off ad blocker or purchase a subscription for unlimited access to The Hindu.

Sign up for a 30 day free trial.