Indian J Med Res 152, July & August 2020, pp 12-15
DOI: 10.4103/ijmr.IJMR_3667_20

Quick Response Code:

Perspective

Genetic characterization of SARS-CoV-2 & implications for epidemiology, diagnostics & vaccines in India

Within the last 10 years, the world has witnessed two major pandemics, that of the 2009 influenza A/pH1N1 and the currently ongoing SARS-CoV-2 pandemic. The SARS-CoV-2, causative agent of the coronavirus disease 2019 (COVID-19) pandemic, like the severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1) outbreak of 2002-2004 and the Middle East respiratory syndrome (MERS) CoV outbreak of 2012, is a bat-derived betacoronavirus[1]. In any new host, viruses have the potential to transmit, evolve rapidly and also possess quasispecies diversity, which is the main driving force for the long-term survival of viruses in nature[2,3].

Coronaviruses, unlike most other RNA viruses, in general, have a moderate mutation rate due to the proofreading activity of the exoribonuclease, which increases the fidelity of RNA synthesis and replication, subsequently generating less genomic diversity[4]. Most variants involve random non-functional changes that seldom become fixed[5] and are majorly useful for tracing transmission chains. Despite this, accumulation of mutations in the relatively large genome (~29.8 kb) of the SARS-CoV-2 can have implications during the rapid development and evaluation of diagnostics and in formulating strategies for control such as vaccines, antivirals and antibody therapies[6].

Viral genome sequencing and genetic characterization have thus emerged as an essential tool for the epidemiological investigation of the COVID-19 virus outbreak. The integration of the genomic data with such investigations can allow in-depth analysis of transmission dynamics and the evolution of viral aetiology.

Whole genome sequencing - Global perspective

Real-time sequencing of viral genomes can help to understand the transmission history of pandemics and provide insights into how the pathogen is evolving, the mutation rates, *etc*. These data have provided useful epidemiological insights into the history of the pandemic, for example, multiple introductions into different geographical areas.

Comprehensive information of the SARS-CoV-2 strains circulating in different parts of the world and in particular regions/countries has been reported since the first full genome of the Wuhan strain was submitted to the global database on January 5, 2020[7]. The repository has seen unprecedented activity and as of August 26, 2020, more than 89,000 genome sequences of SARS-CoV-2 from across the world have been shared on the publicly available platform Global Initiative on Sharing All Influenza Data (GISAID) (*https://www.gisaid.org/*). In GenBank, >13,000 complete genome sequences of the virus have been submitted (*https://www.ncbi.nlm.nih.gov/genbank/*). By contrast, even after a decade of the emergence of influenza A swine-origin pandemic H1N1, the number of sequences available for the A/H1N1 pdm2009 is lower at around 60,000 in GISAID and 51,000 in GenBank. Thus the SARS-CoV-2 virus could be among the most genetically characterized viruses in the world.

The first sequence of the earliest Wuhan strain played a major role in determining the ancestry of this virus. Several reports have demonstrated that the genome is closest to SARS-like CoVs from horseshoe bats and the receptor-binding domain of its spike protein is closest to that of pangolin viruses[8,9]. Although the direct ancestral viruses have not been identified, these observations reflect the likely bat-origin of the virus with possible recombination[9].

As of August 19, 2020, there were >20,000 non-synonymous substitutions identified based on the genomes deposited in the GISAID database (CoV-GLUE resource, *http://cov-glue.cvr.gla.ac.uk/#/home*).

At present, except for the mutation D614G in the spike protein of the virus that is reported to have a role in enhanced transmissibility[10,11], there is no evidence that these point mutations have any significance in the functional context of within-host infections or transmission rates. However, the rapid diversification of the strains has enabled delineation into clades and sub-clades. Differing nomenclatures based on diverse approaches were proposed by the different platforms such as GISAID and NextStrain, while a dynamic nomenclature was proposed by Rambaut *et al*[12]. At present, the absence of a universally accepted nomenclature is creating misperceptions in interpretations of the virus phylogeny among reported strains.

The evolution of the virus using varied tools, including phylogeny-based molecular clocks[13] and network analysis[14], agrees on a common ancestral time towards the end of 2019 as well as more or less concur on the viral evolutionary substitution rate[5]. Several studies have demonstrated rates of mutation similar to SARS-CoV-1[15,16], with some amount of variation[17]. Hence, continued studies to estimate the rates of evolution with larger datasets would help provide further insight into the evolutionary dynamics of the SARS-CoV-2.

Whole genome sequencing efforts in India

Efforts in India for a 1000 SARS-CoV-2 genome sequencing project was led by the Council for Scientific and Industrial Research (CSIR). The number of full genomes submitted from India to the GISAID was >2400 as on August 26, 2020. While uploading the sequence data into the database, efforts are being made to include meta-data such as epidemiological and clinical data by coordinating with the National Centre for Disease Control at New Delhi. However, a lot of precautions would be needed to correlate genome and clinical meta-data, specifically linking mutations with the disease outcomes, in the absence of experimental validations.

Studies undertaken at the ICMR-National Institute of Virology (ICMR-NIV), Pune, helped characterize the first two genomes of SARS-CoV-2 from the earliest cases in India[18], followed by those from the cases that were reported from a group of Italian tourists and their contacts in north India[19]. Epidemiological correlations with the molecular data are vital for tracking the transmission of the virus, identifying inter/intra-State movements, and to understand the mechanisms of the spread of the virus[20]. In addition, monitoring the hotspots of evolution and evidence for selective pressures would also be vital for tracking evolution of the strains in terms of adaptation to varying environments. As of now, the sequence data from the different States of India are non-uniform, with some States such as Gujarat, Delhi, Odisha, Telangana, Maharashtra, Karnataka, West Bengal, Madhya Pradesh and Tamil Nadu being better represented than the others. Thus, focus would now be needed to sequence the strains from the unrepresented and lesser represented States to help explore the establishment of the clades, transmissions within the country and the evidence of indigenous evolution.

Diagnostic perspective

The relatively rapid sequencing of the full genome of SARS-CoV-2 early during the pandemic facilitated the development of specific laboratory protocols for the detection of COVID-19. The protocol of the first RT-qPCR test that was based on the envelope (*E*), RNA-dependent RNA polymerase *(RdRp)* and nucleocapsid (*N*) genes of SARS-CoV-2 was published promptly by the end of January 2020[21].

A few studies have indicated that some of the currently available nucleic acid detection assays can result in false positives[22] as the SARS-CoV-2 is closely related to other coronaviruses[23]. At present, multiple RT-qPCR tests are available with multiplex or singleplex composition, which detect the *ORF1b*, *E*, *N*, *RdRp* or *S* (spike glycoprotein) gene segments with varied sensitivity, specificity and run time. In accordance with the WHO recommendation[21], the ICMR-NIV, Pune, adopted the gold standard RT-qPCR tests, which enable the detection of three genes (*E*, *ORF1b* and *RdRP*) in a single reaction. This allows detection of viruses from the betacoronavirus group (*E* gene), as well as to identify the SARS-CoV-2 virus (*N* and *RdRP, ORF1b genes*). Such a design provides double confirmation in cases of infection and it also limits the risk of obtaining false-negative results in case of the detection of only a single target for SARS-CoV-2. Since the early phase of the pandemic, GISAID has been constantly monitoring high-quality genomes (defined as <1% ambiguous nucleotides and <0.05% unique non-synonymous mutations) for variations which could impact commonly used primer and probe sequences under the WHO protocol for COVID-19 diagnosis[24]. Up to one or two mutations in either the forward primer, probe or reverse primer regions were found functional. This was specifically noted for the *N* gene primers (*https://gisaid.org/hcov-19-analysis-update*). These criteria thus serve as a guide to the permitted variability of the targeted region beyond which sensitivity could be affected.

In India, even though most of the available diagnostics have focussed on RT-PCR, additional methods include using serological and full genome sequencing. An anti-SARS-CoV-2 IgG ELISA assay was designed indigenously for the detection of IgG antibodies against the SARS-CoV-2 virus in human serum/plasma using an indirect ELISA[25]. A serology-based point-of-care test for SARS-CoV-2 is under development. Likewise, an indigenous antigen capture ELISA would be standardized to test the COVID-19 antigen from infected patients. Diagnostics would also need to be adapted for testing of non-human hosts.

Vaccine perspectives

As a swift response to develop and manufacture anti-SARS-CoV-2 vaccines, the Indian Council of Medical Research (ICMR) has partnered with other institutes and three major companies, Serum Institute of India, Bharat Biotech and Zydus Cadila. The strategies being explored are the adenovirus vector-based vaccine, inactivated vaccine and plasmid DNA vaccine, respectively. Pre-clinical animal studies done at the ICMR-NIV, Pune, have been invaluable in vaccine development so far[26]. Infectious disease outbreaks heavily depend on choosing the best isolates for animal models that inform about the best vaccine candidates and treatments[6]. Thus, complete genome sequencing and comprehensive analysis of the phenotypic characteristics of any potential vaccine strain is of paramount importance. The monitoring of mutations in the virus's genetic make-up is equally important since these could potentially disrupt the efficacy of any vaccine by altering the antigenic structure of the virus. This would not be of consequence in the case of inactivated virus-based vaccines. On the other hand, a comparison of circulating strains of the different lineages shows an overall per cent nucleotide identity of 99.94 per cent (unpublished data). Thus it can be believed that the comparatively lower mutation rate of SARS-CoV-2 would assure the possibility of the development of efficacious and effective vaccines.

Conclusions and way forward

The pattern of emergence of mutations in a virus genome is a key for accurate diagnosis, genetic characterization and therapeutics, which, in turn, depict the potential course of the viral spread and the epidemic in real time. The genomic epidemiology that has revealed both the exchange across distant countries as well as within country currently has been beneficial for the mitigation and control of the SARS-CoV-2 outbreak. The rapid and open access deposition of virus genomes is also enabling precise investigations into patterns of human-to-human transmission. Further, the concept of reverse zoonotic disease transmission is another perspective that would need to be looked into in the time to come, which may, in turn, contribute to further transmission and possible outbreaks in humans in the future.

***Conflicts of Interest*:** None.

Priya Abraham[†,*], Sarah Cherian[1] & Varsha Potdar[2]
[1]Bioinformatics & Data management Group & [2]Human Influenza Group,
[†]ICMR-National Institute of Virology,
Pune 411 001, Maharashtra, India
[*]*For correspondence*:
director.niv@icmr.gov.in

Received August 28, 2020

References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; *579* : 270-3.
2. Schneider WL, Roossinck MJ. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *J Virol* 2001; *75* : 6566-71.
3. Domingo E, Baranowski E, Ruiz-Jarabo CM, Martín-Hernández AM, Sáiz JC, Escarmís C. Quasispecies structure and persistence of RNA viruses. *Emerg Infect Dis* 1998; *4* : 521-7.
4. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, *et al*. Discovery of an RNA virus 3'→5' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci U S A* 2006; *103* : 5108-13.
5. Grubaugh ND, Petrone ME, Holmes EC. We shouldn't worry when a virus mutates during disease outbreaks. *Nat Microbiol* 2020; *5* : 529-30.
6. Bauer DC, Tay AP, Wilson LOW, Reti D, Hosking C, McAuley AJ, *et al*. Supporting pandemic response

using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. *Transbound Emerg Dis* 2020; *67* : 1453-62.

7. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020; *579* : 265-9.

8. Boni MF, Lemey P, Jiang X, Lam TT, Perry BW, Castoe TA, *et al*. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 2020. doi: 10.1038/s41564-020-0771-4.

9. Lau SKP, Luk HKH, Wong ACP, Li KSM, Zhu L, He Z, *et al*. Possible bat origin of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 2020; *26* : 1542-7.

10. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, *et al*. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020; *182* : 812-27.

11. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, *et al*. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020. doi: 10.1101/2020.06.12.148726.

12. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, *et al*. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020. doi: 10.1038/s41564-020-0770-5.

13. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, *et al*. Transmission dynamics and evolutionary history of 2019-nCoV. *J Med Virol* 2020; *92* : 501-11.

14. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020; *117* : 9241-3.

15. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ* 2020; *98* : 495-504.

16. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, *et al*. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* 2004; *4* : 21.

17. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, *et al*. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 2020; *83* : 104351.

18. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, *et al*. Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res* 2020; *151* : 200-9.

19. Potdar V, Cherian SS, Deshpande GR, Ullas PT, Yadav PD, Choudhary ML, *et al*. Genomic analysis of SARS-CoV-2 strains among Indians returning from Italy, Iran & China, Italian tourists in India. *Indian J Med Res* 2020; *151* : 255-60.

20. Maitra A, Sarkar MC, Raheja H, Biswas NK, Chakraborti S, Singh AK, *et al*. Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. *J Biosci* 2020; *45* : 76.

21. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, *et al*. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* 2020; *25* : 2000045.

22. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, *et al*. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 2020; *9* : 221-36.

23. Wang M, Wu Q, Xu W, Qiao B, Wang J, Zheng H, *et al*. Clinical diagnosis of 8274 samples with 2019-novel coronavirus in Wuhan. *medRxiv* 2020. doi: 10.1101/2020.02.12.20022327.

24. Centers for Disease Control and Prevention. *CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel*. CDC-006-00019, Revision: 02. Atlanta: CDC; 2020.

25. Sapkal G, Shete-Aich A, Jain R, Yadav PD, Sarkale P, Lakra R, *et al*. Development of indigenous IgG ELISA for the detection of anti-SARS-CoV-2 IgG. *Indian J Med Res* 2020; *151* : 444-9.

26. Mohandas S, Jain R, Yadav PD, Shete-Aich A, Sarkale P, Kadam M, *et al*. Evaluation of the susceptibility of mice & hamsters to SARS-CoV-2 infection. *Indian J Med Res* 2020; *151* : 479-82.