
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 84 (2020) 104330

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

# Emerging genetic diversity among clinical isolates of SARS-CoV-2: Lessons for today

Javaid Ahmad Sheikh[a,1], Jasdeep Singh[b,1], Hina Singh[b], Salma Jamal[b], Mohd. Khubaib[b], Sunil Kohli[c], Ulrich Dobrindt[d,*], Syed Asad Rahman[e,*], Nasreen Zafar Ehtesham[f], Seyed Ehtesham Hasnain[b,g,**]

[a] Department of Biotechnology, Jamia Hamdard, New Delhi, India
[b] JH-Institute of Molecular Medicine, Jamia Hamdard, New Delhi, India
[c] Department of Medicine, Hamdard Institute of Medical Sciences & Research, Jamia Hamdard, New Delhi, India
[d] Institute of Hygiene, Mendelstrasse 7, 48149 Münster, Germany
[e] BioInception Pvt. Ltd, Swift House Ground Floor, 18 Hoffmanns Way, Chelmsford, Essex CM1 1GU, United Kingdom
[f] ICMR-National Institute of Pathology, Safdarjung Hospital Campus, New Delhi, India
[g] Dr Reddy's Institute of Life Sciences, University of Hyderabad Campus, Hyderabad, India

ARTICLE INFO

*Keywords:*
COVID-19
SARS-CoV-2
Coronavirus
Phylogenetics
Molecular divergence

ABSTRACT

Considering the current pandemic of COVID-19, it is imperative to gauge the role of molecular divergence in SARS-CoV-2 with time, due to clinical and epidemiological concerns. Our analyses involving molecular phylogenetics is a step toward understanding the transmission clusters that can be correlated to pathophysiology of the disease to gain insight into virulence mechanism. As the infections are increasing rapidly, more divergence is expected followed possibly by viral adaptation. We could identify mutational hotspots which appear to be major drivers of diversity among strains, with RBD of spike protein emerging as the key region involved in interaction with ACE2 and consequently a major determinant of infection outcome. We believe that such molecular analyses correlated with clinical characteristics and host predisposition need to be evaluated at the earliest to understand viral adaptability, disease prognosis, and transmission dynamics.

Humanity is at the verge of a serious crisis due to COVID-19 (coronavirus disease-2019) pandemic caused by a novel coronavirus SARS-CoV-2 (severe acute respiratory syndrome-coronavirus-2). The disease was reported earlier from Wuhan, China in December 2019 and in a matter of weeks, witnessed a meteoric spread to almost every part of the world (Zhu et al., 2020). New emerging epicentres like Italy, Iran, Spain and USA have amplified the crisis. The looming disaster of its spread to over populated countries with fragile health infrastructure necessitates even more urgently to implement population-centred approaches to understand this outbreak and fight the pandemic. Next generation sequencing (NGS) has helped the community to explore the evolutionary diversity and mutational propensity of the virus. Efforts of medical and scientific fraternity have been unprecedented with genetic data being available at exhilarating speed along with rapid epidemiological and clinical data sharing - all aiming to translate these into developing effective interventions (Baden and Rubin, 2020).

We analysed the genome sequence of available clinical isolates (n = 257) to comprehend the phylogenetic diversity of the isolates (Shu and McCauley, 2017). Genome sequences of SARS-COV-2 were retrieved from GISAID (https://www.gisaid.org) as on 11th Mar 2020 followed by manual curation to remove redundancy (100% identical sequences) using mmseq2 software (Mirdita et al., 2019). Multiple sequence alignments (MSA) of all coronavirus whole genomes were performed using MUSCLE and MAFFT software (Madeira et al., 2019; Rozewicki et al., 2019). Phylogenetic plots were constructed using iTOL (https://itol.embl.de) (Letunic and Bork, 2019) software based on the neighbour-joining method. Coding ORF's in SARS-CoV-2 genomes were predicted by Glimmer (Delcher et al., 1999) and homology search was performed using mmseq2. Variable region analysis among SARS-CoV-2 ORF's was done using in-house protocols. Mutations were computed

* Corresponding authors.
** Corresponding author at: JH-Institute of Molecular Medicine, Jamia Hamdard, New Delhi, India.
*E-mail addresses:* ulrich.dobrindt@ukmeunster.de (U. Dobrindt), asad.rahman@bioinceptionlabs.com (S.A. Rahman), seyedhasnain@gmail.com, vc@jamiahamdard.ac.in (S.E. Hasnain).
[1] Javaid Ahmad Sheikh and Jasdeep Singh contributed equally.

https://doi.org/10.1016/j.meegid.2020.104330
Received 29 March 2020; Received in revised form 21 April 2020; Accepted 22 April 2020
Available online 24 April 2020

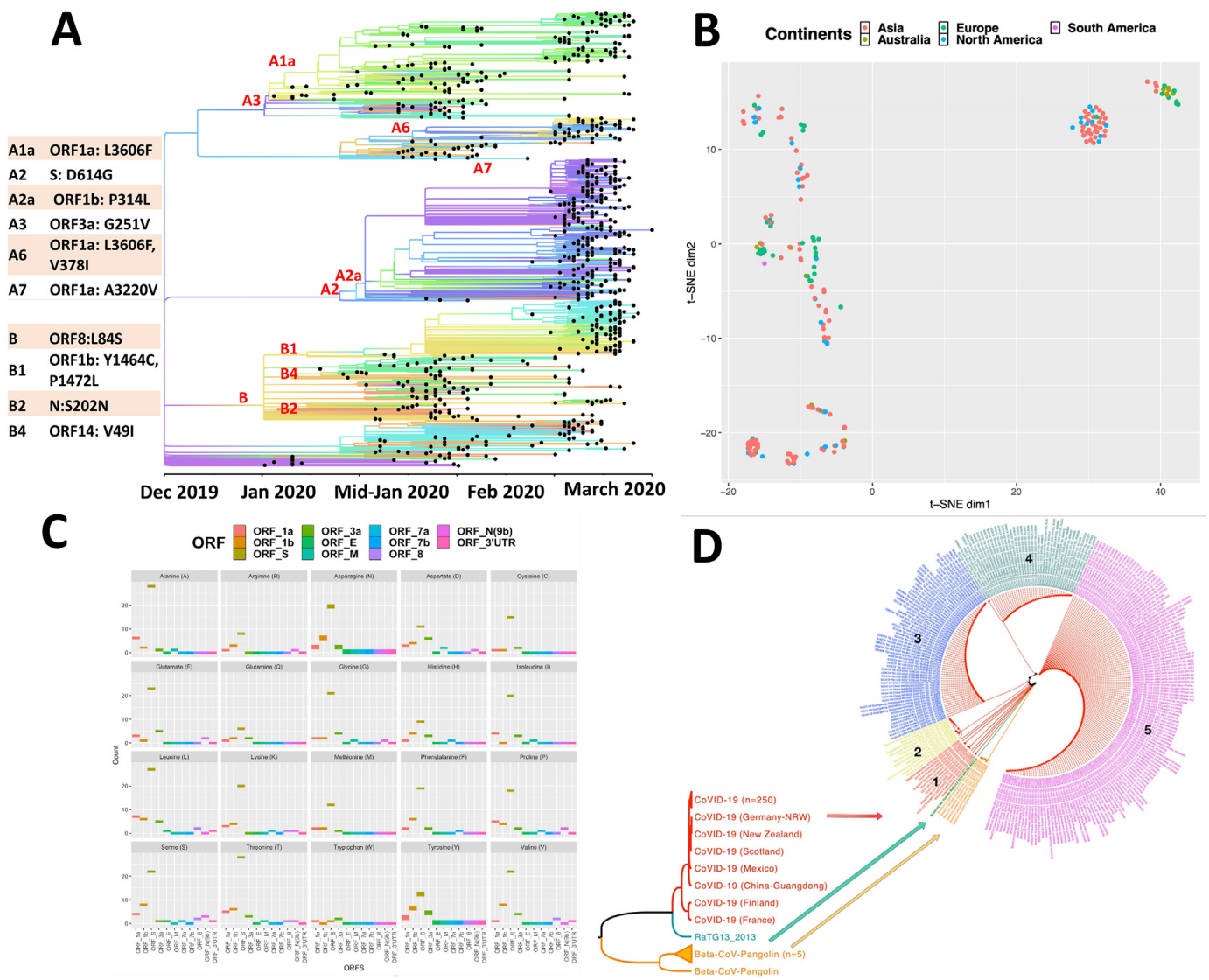


**Fig. 1. Visual depiction of emerging genomic diversity among clinical isolates of SARS-CoV-2** (A) SARS-CoV-2 separating into different clades since its identification in December 2019 till March 2020 using Wuhan-Hu-1/2019 (NC_045512) as reference strain (courtesy: nextstrain.org). Amino acid mutations in individual ORF's are also shown that lead to clade separation. (B) t-distributed stochastic neighbour embedding (tSNE), a non-linear dimensionality reduction Machine learning (ML) technique plot for COVID-19 strains (n = 257, based on the global ORF similarity between genomes) highlights two dominant clusters; Europe and North America + Asia). Diffuse geographical clustering was observed for other strains across continents - an indication of the effects of human migration in the modern population. (C) Mutation analysis of all ORF's from analysed COVID-19 strains (n = 257). Plots for individual amino acids in ORF's show that every type of amino acid has undergone mutations, predominantly in ORF's 1ab, S, 3a and E. (D) Diversity in clinical isolates indicate that SARS-CoV-2 has divulged into at least five different clades, evolutionary distinct from Bat CoV RaTG13 and Beta-CoV-pangolins.

with Wuhan-Hu-1/2019 isolate as a reference strain using SeaView (Gouy et al., 2010). MSA of the ORF's were performed by MAFFT and MUSCLE as above. R software was used to generate the plots (Chan, 2018). tSNE clustering was done using Rtsne employing R (https://cran.r-project.org/) package (Krijthe, 2015).

The mutations arising during human to human spread could provide insights into transmission dynamics and, compounded with clinical and epidemiological data can predict disease prognosis (Wertheim, 2020). Moreover, the mutational propensity of RNA viruses can impede the development of interventions which is an emerging challenge before the scientific community. Our analyses employing advanced machine learning approaches revealed a great deal of genetic diversity emerging among clinical isolates (Fig. 1A) along the time line. This genome clustering highlights evolution of SARS-CoV-2 into distinct clades that were driven through specific mutations in its ORF's as it spread around the globe from Dec 2019 to March 2020 (Fig. 1A). One fascinating aspect that emerged is the observed diversity of viral strains in every continent (Fig. 1B). However, there was no geographical clustering of the isolates and every continent seems to have multiple introductions of different viral strains. Earlier pandemics had geographical signatures in their sequences that abetted the tracing of new infections (Miller et al., 2009). Those markers are obscure yet in the current pandemic possibly due to rampant globalisation, a major bottleneck in tracing the lineages. The intriguing difference in the significantly more than average transmission and mortality in the Lombardy region of Italy, compared to other European countries or that of African continent or even China, could not be correlated to any specific molecular divergence pattern and very likely a reflection of the advanced age/co-morbidity. Interestingly, we observed 5' terminal of the genome to be more variable and prone to mutations, as compared to 3' terminal. It appears that ORF1ab, spike, ORF3a and E are key drivers of diversity among strains (Fig. 1C) with receptor binding domain (RBD) of spike emerging as mutational

hotspot (Andersen et al., 2020; Jamal et al., 2020; Tai et al., 2020; Wan et al., 2020). Interaction of RBD with ACE2 (angiotensin-converting enzyme 2) could determine the outcome of infection indicating the clinical importance of these mutations (Shang et al., 2020; Wang et al., 2020; Yan et al., 2020). Our phylogenetic analyses reveal at least five different clades circulating as of date and more divergence is expected with time (Fig. 1D). We believe that these emerging mutations, once corroborated with viral pathogenesis and clinical characteristics along with epidemiological correlates would be valuable in predicting disease progression and also tracing pathogen mobility and re-emergence.

## Declaration of Competing Interest

The Authors declare no conflict of interest.

## Acknowledgement

JAS acknowledges Start-up grant by UGC, India. JS, SJ and HS are financially supportéd under Young Scientist and Women Scientist schemes of Dept. of Health Research, Government of India. MK is a Silver Jubilee Post-Doctoral Fellow at Jamia Hamdard. SEH and NZE would like to acknowledge funding support from Centre of Excellence Grant BT/PR12817/COE/34/23/2015 and DBT North-East Grants BT/PR23099/NER/95/632/2017 and BT/PR23155/NER/95/634/2017 by Department of Biotechnology, Ministry of Science and Technology, Government of India. SEH is a JC Bose National Fellow, Department of Science and Technology, Government of India and Robert Koch Fellow, Robert Koch Institute, Germany.

## References

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26, 450–452. https://doi.org/10.1038/s41591-020-0820-9.

Baden, L.R., Rubin, E.J., 2020. Covid-19 - The search for effective therapy. N. Engl. J. Med. https://doi.org/10.1056/NEJMe2005477. NEJMe2005477.

Chan, B.K.C., 2018. Data analysis using R programming. Adv. Exp. Med. Biol. 1082, 47–122. https://doi.org/10.1007/978-3-319-93791-5_2.

Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L., 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27, 4636–4641. https://doi.org/10.1093/nar/27.23.4636.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224. https://doi.org/10.1093/molbev/msp259.

Jamal, S., Singh, J., Sheikh, J.A., Singh, H., Khubaib, M., Kohli, S., Dobrindt, U., Rahman, S.A., Ehtesham, N.Z., Hasnain, S.E., 2020. Molecular Analyses of Over Hundred Sixty Clinical Isolates of SARS-CoV-2: Insights on Likely Origin, Evolution and Spread, and Possible Intervention. Preprints 2020030320. https://doi.org/10.20944/preprints202003.0320.v1.

Krijthe, J.H., 2015. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. https://github.com/jkrijthe/Rtsne.

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256–W259. https://doi.org/10.1093/nar/gkz239.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., Lopez, R., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47, W636–W641. https://doi.org/10.1093/nar/gkz268.

Miller, M.A., Viboud, C., Balinska, M., Simonsen, L., 2009. The signature features of influenza pandemics—implications for policy. N. Engl. J. Med. 360, 2595–2598. https://doi.org/10.1056/NEJMp0903906.

Mirdita, M., Steinegger, M., Soding, J., 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35, 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057.

Rozewicki, J., Li, S., Amada, K.M., Standley, D.M., Katoh, K., 2019. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 47, W5–W10. https://doi.org/10.1093/nar/gkz342.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., Li, F., 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature. https://doi.org/10.1038/s41586-020-2179-y.

Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. Eur. Surveill. 22 (13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S., Zhou, Y., Du, L., 2020. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell. Mol. Immunol. https://doi.org/10.1038/s41423-020-0400-4.

Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94 (7). https://doi.org/10.1128/JVI.00127-20. (e00127-20).

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y., Wang, Q., Zhou, H., Yan, J., Qi, J., 2020. Structural and functional basis of SARS-CoV-2 Entry by using human ACE2. Cell. https://doi.org/10.1016/j.cell.2020.03.045. S0092-8674(20)30338-X.

Wertheim, J.O., 2020. A glimpse into the origins of genetic diversity in SARS-CoV-2. Clin. Infect. Dis. https://doi.org/10.1093/cid/ciaa213. ciaa213.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q., 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444–1448. https://doi.org/10.1126/science.abb2762.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., China Novel Coronavirus Investigating and Research Team, 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382, 727–733. https://doi.org/10.1056/NEJMoa2001017.