

SARS-CoV-2 genome analysis of strains in Pakistan reveals GH, S and L clade strains at the start of the pandemic

Najia Karim Ghanchi¹, Kiran Iqbal Masood^{1*}, Asghar Nasir^{1*}, Waqasuddin Khan², Syed Hani Abidi³, Saba Shahid¹, Syed Faisal Mahmood⁴, Akbar Kanji¹, Safina Razzak¹, Zeeshan Ansar¹, Nazneen Islam¹, M. B. Dharejo⁵, Zahra Hasan^{**,1#} and Rumina Hasan^{**,1,6}

#Corresponding author, *contributed equally to this work

¹Department of Pathology and Laboratory Medicine, ²Department of Pediatrics and Child Health, ³Department of Biological and Biomedical Sciences, ⁴Department of Medicine, The Aga Khan University, Karachi, Pakistan. ⁵Department of Health, Government of Sindh, Pakistan, ⁶Infectious and Tropical Disease, London School of Hygiene and Tropical Medicine, London, UK

Running title: SARS-CoV-2 strains in Pakistan

#Corresponding author:

Zahra Hasan, PhD, FRCPath,

Professor, Department of Pathology and Laboratory Medicine,
The Aga Khan University, Karachi 74800,

Pakistan

Email: zahra.hasan@aku.edu

Tel: +92300 8247815

Abstract

Objectives: Pakistan has a high infectious disease burden with about 265,000 reported cases of COVID-19. We investigated the genomic diversity of SARS-CoV-2 strains and present the first data on viruses circulating in the country.

Methods: We performed whole-genome sequencing and data analysis of SARS-CoV-2 eleven strains isolated in March and May.

Results: Strains from travelers clustered with those from China, Saudi Arabia, India, USA and Australia. Five of eight SARS-CoV-2 strains were GH clade with Spike glycoprotein D614G, Ns3 gene Q57H, and RNA dependent RNA polymerase (RdRp) P4715L mutations. Two were S (ORF8 L84S and N S202N) and three were L clade and one was an I clade strain. One GH and one L strain each displayed Orf1ab L3606F indicating further evolutionary transitions.

Conclusions: This data reveals SARS-CoV-2 strains of L, G, S and I have been circulating in Pakistan from March, at the start of the pandemic. It indicates viral diversity regarding infection in this populous region. Continuing molecular genomic surveillance of SARS-CoV-2 in the context of disease severity will be important to understand virus transmission patterns and host related determinants of COVID-19 in Pakistan.

Background and Rationale

The global outbreak of the novel coronavirus 2019, SARS-CoV-2 (causative agent of COVID-19), has caused over 566,654 deaths as of 13 July 2020, with more than 7 million individuals infected globally. SARS-CoV-2 was first reported in Wuhan, China associated with acute respiratory infection (1). The SARS-CoV-2 RNA virus genome is about 28.9 kb in size and phylogenetic analysis shows it to belong to the *Sarbecovirus* subgenus of *Betacoronavirus* and the family *Coronaviridae* (2). Its sequence is closely related to the Bat CoV RaTG13 with a sequence similarity of 96.3% across the genome. The virus has a higher rate of infectivity than previous coronaviruses such as the MERS (Middle East Respiratory

Syndrome) virus and SARS (Severe Acute Respiratory Syndrome) virus outbreaks in 2014 and 2006 respectively (3).

SARS-CoV-2 is transmitted via droplets and small particles produced during coughing or sneezing and affects the upper respiratory tract including nose, throat, pharynx and lower respiratory tract (4). Most individuals present with fever, body ache, non-productive cough and shortness of breath, with a small proportion developing more severe disease including death. The case fatality rate (CFR) of SARS-CoV-2 infections has varied regionally. In China, the CFR ranged from 5.8 % in Wuhan to 0.7 % in the rest of the country. CFR at the peak of the pandemic in Mar was variable at 6.2% in Italy, 3.6% in Iran and 0.79% in South Korea (5). In later months, April to June, the CFR continued to vary globally. Reliable assessment of CFR may be limited by the degree of testing in the population however, the possibility of SARS-CoV-2 diversity may play a role in disease variability.

Since first identified as the Wuhan-1 strain in Hubei, China in January 2020, (6) SARS-CoV-2 strains have acquired mutations during its spread across regions that differentiate it into different global clades (7) (8). Initial SARS-CoV-2 strains were primarily L clade which subdivided to S and also sub-divided to V and G, as found in Asia, Oceania, Europe, South America, and North America; with clade I identified more recently. The GISAID database has made available greater than 63,000 SARS-CoV-2 global isolates (www.gisaid.org) to facilitate a comparison of strains.

Pakistan has a population of about 200 million with limited health resources. Karachi is its most populous city with about 20 million individuals. Up until mid-July approximately 262,000 COVID-19 cases have been diagnosed of which about 70,000 were from Karachi. Reduced rates of COVID-19 have been attributed to limited testing for SARS-CoV-2. Thus far, the case fatality ratio for SARS-CoV-2 in Pakistan has been 2% with some regional variations (9). Data from Pakistan regarding viral transmission has not been available. We sequenced SARS-CoV-2 isolates from Karachi and compared from Pakistan, global and Asian strains to understand transmission dynamics in the country.

Materials and Methods

Ethical approval

This study was approved by the Ethical Review Committee at the Aga Khan University (AKU), Karachi, Pakistan.

Diagnostic testing for SARS-CoV-2

Nasopharyngeal swab specimens were tested for SARS-CoV-2 by reverse transcription (RT) polymerase chain reaction (PCR) at the Section of Molecular Pathology, Clinical Laboratory, AKUH. Specimens received in March were tested using the WHO protocol for the 2019-nCoV RT-PCR assay (10). Specimens in May were tested using the Cobas[®] SARS-CoV-2 RT-PCR assay (Roche Diagnostics, USA). Respiratory samples archived at the Clinical Laboratory, Section of Molecular Pathology were used for virus extraction and testing. Laboratory data was (including age, gender) was utilized where available.

RNA extraction, Library preparation, NGS sequencing, and Consensus Sequence generation
RNA was extracted from eight respiratory samples positive for SARS-CoV-2 using the QIAmp Viral RNA Mini kit (Qiagen). Sequencing was performed as described previously (11). First-strand cDNA synthesis was performed with SuperScript III Reverse Transcriptase (SSIII), Thermo Fisher Scientific, USA. Briefly, 8 µl of viral RNA was mixed with 1 µl of 50 ng/µl of random hexamer and 1 µl of 10mM dNTPs. Then denatured at 65°C for 5 minutes. Then, 10X Buffer, 0.1 mM DTT, RNase Inhibitor (40 U/µl), 25mM MgCl₂ and 200 U SuperScript III Reverse Transcriptase was added. Second strand cDNA was synthesized with DNA polymerase I, Large Fragment, Klenow (Invitrogen, USA).

The library preparation was performed according to the Nextera XT DNA Library Preparation kit (Illumina) using 1 ng DNA. Bead-based normalization of the libraries was carried out as recommended by the manufacturer. Normalized libraries were equimolarly pooled and spiked with PhiX control prior to sequencing.

Sequencing was performed on the Illumina Miniseq platform using a 300 cycle Miniseq Reagent Kit v2 (Illumina). DRAGEN RNA Pathogen Detection App v3.5.7 on BaseSpace (12) was used to filter out human sequence reads by combining human (hg38) + SARS-CoV-2 virus reference (NC_045512.2) genome to generate consensus sequences (Supplementary Table 1). Quality metrics of genome data generated are provided in Supplementary Table 2.

Variant calling and Phylogenetic analysis

FASTQ files were aligned to the SARS-CoV-2 virus reference genome Wuhan-1 (NC_045512.2) by BWA (13). PICARD tools (<http://broadinstitute.github.io/picard/>) were used to remove redundant alignments and calculating alignment statistics. Variants were identified by Genome Analysis Toolkit (GATK) (14). The effect on protein-coding by the mutation is determined by an impact score (15).

For phylogenetic analysis, 7 SARS-CoV-2 sequences from our study, 3 Pakistani SARS-CoV-2 sequences (Supplementary Table 1) along with the 449 full-length SARS-CoV-2 reference sequences (Supplementary Table 3) from different pandemic countries obtained from the NCBI SARS-CoV-2 Resources (<https://www.ncbi.nlm.nih.gov/sars-cov-2/>) were subjected to Multiple Sequence Alignment (MSA) along using MAFFT online server (16). The MSA was subsequently used to generate a Maximum Likelihood (ML) phylogenetic tree using PhyML 3.0 (<http://www.atgc-montpellier.fr/phyml/>) with a GTR-based nucleotide substitution model and aLRT SH-Like branch support. The root of the tree and branch length variance was determined using the TreeRate tool (17) by applying a generalized midpoint rooting strategy. The tree was visualized and edited in Figtree software (<http://tree.bio.ed.ac.uk/software/figtree/>). The mean and individual pairwise distance between 7 SARS-CoV-2 sequences from our study and 3 previously deposited Pakistani SARS-CoV-2 sequences was calculated using MEGA 7 (18).

For genomic epidemiology of Asian- strains focused subgroup analysis of SAR-CoV-2 as of 16th July 2020, we downloaded 5,215 complete sequences of SARS-CoV-2 along with the required metadata from the GISAID considering the following parameters: 1) genome length > 29,000 bps, 2) further assigns labels of high-coverage <1% Ns – undefined bases, and 3) low-coverage >5% Ns. After the inclusion of our 7 strains to the fasta file, phylogenetic tree reconstruction was performed using NEXTSTRAIN's (<https://www.nextstrain.org/>) augur (<https://www.docs.nextstrain.org/projects/augur/en/stable/>) pipeline. Ancestral state reconstruction and branch length timing were performed with TreeTime (19). Finally, the collection of all annotated nodes and metadata was exported to the interactive phylodynamic visualizing tool Auspice's (<https://www.nextstrain.github.io/auspice/>) JSON format.

Results

We obtained seven full genome sequences of SARS-CoV-2 strains and one partial sequence (Supplementary Table 2). All eight isolates were from individuals with mild COVID-19. Five strains sequenced isolates were travelers from Iran and Turkey in March, at the start of the COVID-19 pandemic in Pakistan. Three strains isolated from cases in May were from a traveler to Iran, a case of local transmission and a religious pilgrim who attended regional event. We further compared three Pakistani isolates present in the NCBI database which were from a traveler to Iran, a case of local transmission and one in which travel history was not available. Clinical information was not available for these cases. Overall we had data for 10 strains; from seven male and three female cases. The age group of infected individuals based on available data was n=2, less than 18 y; n=4, 19-35 y; n=2, 36-50 y and n=2, > 50 y.

Phylogenetic analysis:

Initial phylogenetic analysis was done using ten SARS-CoV-2 strains from Pakistan (Supplementary Table 2) as compared with 449 global isolates from different pandemic countries obtained from the NCBI SARS-CoV-2 Resources (<https://www.ncbi.nlm.nih.gov/sars-cov-2/>) (Figure 1). Two strains (S2 and S3) clustered with those from Saudi Arabia, and India (Figure 1). S5 clustered with a strain from the USA. S19 clustered with a strain from Australia. S10 and S11 clustered with strains from Saudi Arabia. S21 was related to strains clustered from the USA and Australian. Additionally, Pakistani strain analysis revealed PAK/KHI1 and PAK/Gilgit1 sequences matched with Chinese isolate and PAK/Manga1 with an isolate from the USA.

The mean pairwise genetic distance between our sequences; sequences previously deposited from Pakistan, and sequences from India, Saudi Arabia, USA, and Australia was found to be 0.00, indicating phylogenetic relatedness between these sequences.

We further expanded the genomic epidemiological analysis by focusing on 5,215 Asian isolates available in the GISAID database. As depicted in Figure 2, in the absence of non-Asian sequences, sequences S2 and S3 clustered with the Japanese sequences (Supplementary Table 1). S5 clustered with a strain from Israel; S10 and S11 clustered with strains from Saudi Arabia. S19 matched a Pakistani isolate with the divergence rate of 0.0007953 and was a case of local transmission. S21 was most closely related to a strain from India. Additional Pakistani strain analysis revealed; PAK/KHI1 matched an Indonesian isolate, PAK/Gilgit1 with a Chinese isolate and PAK/Manga1 with an isolate from Singapore.

We further analyzed genome sequences of these ten full and one partial (S16) SARS-CoV-2 isolate. Of the eleven, five were G clade of the sub-clade GH; three were from March (S2, S3, and S5 from Iran and Turkey travelers) and two from May (S16 from an Iran traveler and S19 through a local transmission). Two isolates from travelers to Turkey (S10) and Iran (S11) in March belonged to the S clade. Three strains from May comprised two GH clade strains, one from an Iran traveler (S16) and the other from local transmission belonged to L clade, with 99% homology to the Wuhan-Hu-1 nCoV strain type. Of the three Pakistani isolates present in the NCBI database, PAK/KHI1 belonged to L clade, PAK/Gilgit-1 from a traveler to Iran belonged to I clade and PAK/Manga1 was an L clade strain transmitted locally.

Mutation Analysis:

Variant analysis of the genomes revealed 34 SNV comprising of 2 non-coding, 18 non-synonymous and 14 synonymous variants (Table 1). The Orf1ab gene region exhibited the highest number of variants in all isolates, with 19 different variants (Table 2). Two Orf1ab mutations associated with evolutionary changes at nucleotide positions 8782 (2839S) and 14408 (P4715L) were observed. Variation in Orf1ab-8782S and Orf8-L84S are associated with evolutionary changes dividing SARS-Co-V-2 into lineages L and S (20) of which, the L lineage is more prevalent. We found S clade strains (S10, S11) with ORF8 L84S and ORF1ab 2839S mutations. These two strains also had S202N in the nucleocapsid and 302T in Spike gene regions.

The Spike glycoprotein mutation D614G defines the virus clade 'G' which further splits into GH and GR clades (20). D614G with ORF3a-Q57H found in five isolates (S2, S3, S5, S16, S19) identified these as GH clade. In four of the GH isolates we found P4715L/P323L in the nsp12 protein. We found GH clade strains to have additional Orf1ab variants at positions T265I, E2089D and Q2702H. One GH (S19) and one L (S21) clade strain both had the Orf1ab variant L3606F, which is associated with divergence of L clade towards the S clade and also has been shown as an additional mutation found in G clade strains (7). The L clade (S21) isolate had an additional Orf1ab mutation in the nsp4 region producing the A4489V change.

Of the three SARS-CoV-2 isolates previously submitted in the NCBI database from Pakistan, one (PAK/Gilgit1) had ORF1ab-encoded non-structural protein 2 region mutation G1397A producing a non-synonymous change V378I mutation. This mutation has been used to identify strains belonging to clade IV or I lineage isolates (21, 22). Further, this isolate had four other variants in Orf1ab; two were synonymous and two were non-synonymous, R207C

and P2965L. The other two Pakistani isolates present in the NCBI database belonged to the L clade closely resembling the Wuhan-Hu-1 strain. One strain had three synonymous variants only and the other had genomic variants in the coding region.

Variants classified as having a moderate effect on the SARS-CoV-2 protein structure were found in the nucleocapsid N region (Table 1). Five mutations were found in the N gene region in isolates S10, S11, S19 and S21 from S, GH and L clades, respectively. One GH clade strain had P13T and K370N, one L clade strain had P13L, S180I and we observed S202N in two S clade strains. This indicates variability in the N gene as shown previously across global isolates (23). Structural proteins E and M genes showed most conserved regions across all the isolates with a silent variant at position 71Y in M proteins found in three GH clade strains. No variants were observed in ORF10, ORF6, and ORF7a region.

Discussion

This study provides insights into the SARS-CoV-2 strains circulating in Pakistan using a small data set with nine genomes from Karachi and two from Northern Pakistan. We demonstrate the presence of SARS-CoV-2 GH, S, I and L clade strains in March at the start of the epidemic. The presence of different clades is not surprising as initial strains were all associated with travelers who presented with symptoms. The import of cases from locations such as Iran, UK and Turkey in March while the pandemic was spreading globally provides an opportunity for variable strain types to be studied.

Considering S2 and S3 which were GH clade strains, global phylogenetic analysis paired these with strains from Saudi Arabia whilst Asian subsampling showed clustering with Japanese sequences, which might indicate that strains from these countries might be genetically similar and might be evolving at a similar rate.

S19 clustered with an Australian isolate in comparison with global strains. In the Asian strain-based Nextstrain, S19 matched a local Pakistani isolate which correlated with it being a case of local transmission. Phylogenetic analysis with global sequences showed S21 to be related with USA and Australian strains, while Asian-focused real-time strain evolution, as inferred by the time-resolved phylogenetic tree analysis, showed phylogenetic relationship with an Indian strain. This isolate was from a religious pilgrim who attended an event in Punjab province which was subsequently associated with a super spread of SARS-CoV-2 infections. The period when the sample was collected further supports virus evolution with

the nearest strain is at the Indo-Pak region. S10 and S11 (S clade) exhibit coincident time lineages with Saudi Arabian strains; these were from travelers from Iran and Turkey who returned from religious pilgrimage and have acquired Saudi Arabian strains similar to the hCoV-19 Saudi Arabian strains.

Clade GH has been associated with returning travelers from Iran to other countries (22). We found GH clade strains to be associated with travelers from Iran and Turkey concurring with previous reports (22).

S21, a L lineage isolates, found to be closest to the Hubei strain, was from a religious pilgrim who had attended an event in Punjab Province where it is believed that SARS-CoV-2 infections were spread through attendees who had traveled from China. This may explain similarity of the isolate with the Wuhan-Hu-1 SARS-CoV-2 reference strain. This L lineage isolates exhibited G11083T mutation which was first reported on January 2020 in China and is associated with super spreader events in the USA, Singapore, Japan, and Europe (7). The L clade further evolved into the V clade with the emergence of G26144T. L clade is highly variable and shown to harbor up to 12 SNVs in the early phase of the COVID-19 pandemic. Further, GH and L clade strains in our cohort were also identified from both overseas travel and local transmission events.

Of the two SARS-CoV-2 evolutionary lineages L and S, the L lineage continued to split initially equally into G and V versions, with G reaching 50% of isolates in March 2020 and then splitting further into GR and GH subclades (24). G clade strains are identified by the Spike protein D614G mutation and then sub-divided into the H clade with the ORF3a Q57H mutation (25). We found that S19, a GH clade, and S21, an L clade strain both had Orf1ab L3606F mutation. S19 in addition to having GH clade mutations D614G, Q57H, and P323L, had an additional 4 mutations. Of note, both S19 and S21 were strains from May, later in our study set and the presence Orf1ab L3606F mutation suggests additional evolutionary changes in these isolates.

The spike glycoprotein facilitates SARS-CoV-2 entry into host cells by binding to the ACE2 receptor (26). The D614G mutation has been associated with increased virulence and transmission of SARS-CoV2 isolates, demonstrating a higher viral load within infected individuals (27). It is speculated that mutation to Glycine at the 614 site may introduce structural instability into the spike protein (28). *In vitro* studies using a mutant Spike protein with D614G have shown that the virus may be more infective due to the correlated reduction

in viral shedding of the receptor binding S1 domain which is associated with the mutation, leading to increased S-protein incorporation into the virion (29). Thus far, G clade isolates have been found globally and are thought to comprise a high proportion of isolates in Europe and North America. It has been suggested that the S D614G mutation is associated with greater mortality observed in Belgium, Spain, Italy, France, Netherlands and Switzerland (30). However, it is difficult to draw conclusions from laboratory studies and the impact of the D614G variant on transmission between patients and across a population.

Variants in the non-coding 5'UTR and 3' UTR regions of the SARS-CoV-2 virus have been reported to affect viral replication and transcription. TAR DNA binding protein (TARDBP) is a predominantly nuclear RNA/DNA-binding protein that functions in RNA transcription, splicing, transport, and stability. The variant (241C > T) has been reported to result in the strong binding of TARDBP to the 5'UTR region of the SARS-CoV-2 virus (31). This has been implicated in facilitating the translation of viral proteins resulting in its effective propagation within the human host. Interestingly, variant (241C > T) of the 5'UTR region often coexists with spike glycoprotein variant (S protein, D614G) (32). This coexistence is also evident in our study as we found 4 GH clade and an I clade strain to have the 5' UTR variant +241 C>T.

Whilst the 2 S clade strains had the 3' UTR + 29742 G>A. Human microRNAs (miRNAs) are noncoding RNAs that bind to complementary sequences in the 3'-UTR of the target RNAs and regulate the stability of the RNA at a post-transcriptional level. In addition, they also modulate different stages of viral replication, either positively or negatively. The variant (29742G>A) in the 3'UTR region of the SARS-CoV-2 virus has been reported to suggested to affect the binding of the miR-1307 and potentially causing a weakened host immune response against the virus (33). Hence, non-coding variants in both 5' and 3' UTR regions may enhance the virulence of SARS-CoV-2 strains however it is yet to be proven clinically.

The most common missense variants in the Pakistan SARS-CoV-2 strains were Orf1ab P4715L, Q57H and D614G which have previously been mainly reported from Europe and United States (23). Mutations were observed Orf1ab in all isolates studied and comprised thirteen different variants, exhibiting the highest mutability as shown previously (23). Two isolates (S10, S11) had the S clade mutation ORF8 L84S and also the N gene S202N.

We found Orf1ab and N region to have the highest number of mutations as has been shown previously (34). The N protein mutation at Pro13 was present in both GH clade and L lineage isolates. This has mutation been reported as P13L previously in patients from the UK and Australia (35).

Overall, mutations in SARS-CoV-2 are an important mechanism to study variability and spread of the virus and should ideally be done in the context of the clinical disease caused by taking into account host factors as well. Our results are important as they highlight the diversity of SARS-CoV-2 strains at the start of the epidemic in Pakistan. The study was small due to resource constraints and the difficulty in access to NGS reagents in a global lock-down where shipments were stopped, it indicates that at the time of initial spread there were L, S, G and I clades circulating in the country. Initially, the epidemic was associated with travelers and local transmission of SARS-CoV-2 within a couple of weeks of the first known case. It would be of great importance to study the variability in the strains over the past months and also to see how they may differ between those associated with travelers and local transmission. We now hope to be a stage of flattening the COVID-19 curve with reducing rates of positive cases. It would be important to continue surveillance and conduct molecular epidemiological testing of the isolates to understand the disease in the context of strain diversity. This would also be important to understand local host immunity against the circulating SARS-CoV-2 and to investigate possible virulence characteristics associated with the changing viral genome and to relate these to disease severity in a country with a very high infectious disease burden.

Acknowledgements:

This study received funding support from Health Security Partners, USA and University Research Council, The Aga Khan University. We thank Drs. Roger Hewson and Barry Atkinson, Public Health England, UK for their assistance in establishing SARS-CoV-2 diagnostics. Thanks to the European Virus Archive - Global (EVAg), a European Union infrastructure project for making available control material for the study. Thanks for technical support to the Aga Khan University Hospital (AKUH) Clinical Laboratory sections of Molecular Pathology and Microbiology.

Author contributions: Study design and funding (ZH, RH), Research methodology and Data collection (ZH, NG, KI, AN, SS, AK, ZA, NI, MBD), Data analysis (WK, SHA, NG, AN, ZH), Writing (ZH, RH, NG, AN, SFM)

References

1. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-20.
2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565-74.
3. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? *Clin Microbiol Infect*. 2020;26(6):729-34.
4. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506.
5. Khafaie MA, Rahim F. Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2. *Osong Public Health Res Perspect*. 2020;11(2):74-80.
6. Huang WH, Teng LC, Yeh TK, Chen YJ, Lo WJ, Wu MJ, et al. 2019 novel coronavirus disease (COVID-19) in Taiwan: Reports of two cases from Wuhan, China. *J Microbiol Immunol Infect*. 2020.
7. Yang X, Dong N, Chan EW, Chen S. Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. *Emerg Microbes Infect*. 2020;9(1):1287-99.
8. Rambaut A, Holmes EC, Hill V, A OT, McCrone J, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv*. 2020.
9. Dil S, Dil N, Maken ZH. COVID-19 Trends and Forecast in the Eastern Mediterranean Region With a Particular Focus on Pakistan. *Cureus*. 2020;12(6):e8582.
10. 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(3).
11. Goya S, Valinotto LE, Tittarelli E, Rojo GL, Nabaes Jodar MS, Greninger AL, et al. An optimized methodology for whole genome sequencing of RNA respiratory viruses from nasopharyngeal aspirates. *PLoS One*. 2018;13(6):e0199714.
12. Illumina. DRAGEN 3.5.7 2020 [Available from: <https://support.illumina.com/content/dam/illumina-support/documents/downloads/software/dragen/dragen-v3.5.7-release-notes-100000097726-00.pdf>].
13. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
14. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11 0 1- 0 33.
15. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
16. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2019;20(4):1160-6.
17. Maljkovic Berry I, Athreya G, Kothari M, Daniels M, Bruno WJ, Korber B, et al. The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for

- optimizing the evolutionary rate in phylogenetic trees with longitudinal data. *Epidemics*. 2009;1(4):230-9.
18. Agbo EC, Duim B, Majiwa PA, Buscher P, Claassen E, te Pas MF. Multiplex-endonuclease genotyping approach (MEGA): a tool for the fine-scale detection of unlinked polymorphic DNA markers. *Chromosoma*. 2003;111(8):518-24.
 19. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37(5):1530-4.
 20. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the continuing evolution of SARS-CoV-2. *Nat Sci Rev*. 2020;7(6):1012-23.
 21. Gong YN, Tsao KC, Hsiao MJ, Huang CG, Huang PN, Huang PW, et al. SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerg Microbes Infect*. 2020;9(1):1457-66.
 22. Eden JS, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol*. 2020;6(1):veaa027.
 23. Koyama T, Daniel Platt D, Laxmi Paridaa L. Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization* 2020;98:495-504.
 24. Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*. 2020;7(6): 1012–23.
 25. GISAID Initiative 2020 [Available from: <https://www.gisaid.org/>].
 26. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*. 2020;181(2):271-80 e8.
 27. Daniloski Z, Guo X, Sanjana NE. The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types. *bioRxiv*. 2020.
 28. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Velesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*. 2020;181(2):281-92 e6.
 29. Zhang L, Jackson, CB., Mou, H., Ojha, A., Rangarajan, ES., Izzard, T., Farzan, M., Choe, H. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*. 2020.
 30. Easwarkhanth M, Al Madhoun A, Al-Mulla F. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *Int J Infect Dis*. 2020;96:459-60.
 31. Mukherjee M, Goswami S. Global cataloguing of variations in untranslated regions of viral genome and prediction of key host RNA binding protein-microRNA interactions modulating genome stability in SARS-CoV-2. *bioRxiv*. 2020.
 32. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*. 2020;112(5):3588-96.
 33. Chan AP, Choi YC, Schork NJ. Conserved Genomic Terminals of SARS-CoV-2 as Co-evolving Functional Elements and Potential Therapeutic Targets. *bioRxiv*. 2020.
 34. Laha S, Chakraborty J, Das S, Manna SK, Biswas S, Chatterjee R. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infect Genet Evol*. 2020;85:104445.
 35. Coppee F, Lechien JR, Declèves AE, Tafforeau L, Saussez S. Severe acute respiratory syndrome coronavirus 2: virus mutations in specific European populations. *New Microbes New Infect*. 2020;36:100696.
 36. 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.
 37. Mishra A, Pandey AK, Gupta P, Pradhan P, Dhamija S, Gomes J, et al. Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions [bioRxiv preprint

]. In press 2020.

38. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18(1):179.

39. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol*. 2020;92(6):667-74.

40. Sah R, Rodriguez-Morales AJ, Jha R, Chu DKW, Gu H, Peiris M, et al. Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Isolated in Nepal. *Microbiol Resour Announc*. 2020;9(11).

Figure Legends

Figure 1: Maximum-Likelihood phylogenetic tree of SARS-CoV-2 sequences from Karachi. The tree was constructed using Karachi sequences (Supplementary Table 2) along with the 449 full-length SARS-CoV-2 reference sequences (Supplementary Table 3). Karachi sequences are shown in green colour. The root of the tree was determined using TreeRate tool by applying generalized midpoint rooting strategy. The tree was visualized and edited in Figtree software.

Figure 2: Time-resolved phylogenetic distribution of genomic epidemiology of SAR-CoV-2 focused on Asian subsampling (Screenshot of the current Nextstrain display in SVG format are protected by CC-BY license). Tree option layout is selected as radial, branch length is set as time interval while branch labels as amino acid substitution (**A**) Asian subsampling, and (**B**) Asian subsampling with highlighted Pakistani SARS-CoV-2 isolates (total 16, 7 are from this study inclusive).

Table 1. Description of single nucleotide variants (SNV) found in SARS-CoV-2 isolates from Karachi

Sample IDs	No .	Clade	Position (+)	Type	Gene	Gene region	Amino acid Change	Nucleotide change	Impact	References
S19, S2, S3, S5, PAK/Gil git1	5	GH, I	241	up_v, nc	5'UTR	5'UTR	-	C > T	Modifier distance=25	(36)
PAK/Gil git1	1	I	884	ms	Orf1ab	nsp2	p.R207C	c.619Cgt>Tgt	moderate	(25)
S3	1	GH	934	syn	Orf1ab	nsp2	p.223D	c.669gaC>gaT	Low	This study
PAK/Gil git1	1	I	1348	syn	Orf1ab	nsp2	p.361P	c.1083ccC>c cT	Low	(25)
PAK/Gil git1	1	I	1397	ms	Orf1ab	nsp2	p.V378I	c.1132Gta>A ta	Moderate	(23)
PAK/KH II	1	L	1912	syn	Orf1ab	nsp2	p.549S	c.1647tcC>tc T	Low	(25)
S5	1	GH	1059	ms	Orf1ab	nsp2	p.265T >I	c.794aCc>aT c	Moderate	(34)

S19	1	GH	2416	syn	Orflab	nsp2	p.717Y	c.2151taC>ta T	Low	(37)
S19, S2,S3,S5	4	GH	3037	syn	Orflab	nsp3	p.924F	c.2772ttC>tt T	Low	(36)
S21	1	L	6312	ms	Orflab	nsp3	p.2016 T>K	c.6047aCa>a Aa	Moderate	(25)
S19	1	GH	6532	ms	Orflab	nsp3	p.2089 E>D	c.6267gaG>g aT	Moderate	(25)
S19	1	GH	8371	ms	Orflab	nsp3	p.2702 Q>H	c.8106caG>c aT	Moderate	(25)
S10,S11	2	S	8782	syn	Orflab	nsp4	p.2839S	c.8517agC>a gT	Low	(38)
PAK/Gil git1	1	I	9159	ms	Oflab	nsp4	p.P2965 L	c.8894cCt>cT t	Moderate	(25)
PAK/KH II/2020	1	L	10582	syn	Orflab	3C-like proteinase	p.3439 D	10317gaC>ga T	Low	17
S19,S21	2	GH, L	11083	ms	Orflab	nsp6	p.3606	c.10818ttG>tt	Moderate	(39)

							L>F	T		
S21	1	L	13730	ms	Orf1ab	RdRp	p.4489 A>V	c.13466gCt> gTt	Moderate	(25)
S19,S2, S3,S5	4	GH	14408	ms	Orf1ab	RdRp	p.4715P >L	c.14144cCt>c Tt	Moderate	(38)
S19	1	GH	17187	syn	Orf1ab	helicase	p.5641 L	c.16923ctA>c tG	Low	This study
S2, S3, S16	3	GH	18877	syn	Orf1ab	3'-to-5' exonuclease	p.6205 L	c.18613Cta> Tta	Low	This study
S10,S11	2	S	22468	syn	S	S	p.302T	c.906acG>ac T	Low	This study
S19	1	GH	22477	syn	S	S	p.305S	c.915tcC>tcT	Low	This study
S2,S3,S5, S16, S19	5	GH	23403	ms	S	S	p.614D >G	c.1841gAt>g Gt	Moderate	(29, 30)
S21	1	L	23929	syn	S	S	p.789Y	c.2367taC>ta T	Low	This study
S19, S2, S3, S5,	5	GH	25563	ms	Orf3a	Orf3a	p.57Q> H	c.171caG>ca T	Moderate	(35, 37)

S16										
PAK/KH II/2020	1	L	26022	syn	Orf3a	Orf3a	p.210D	c.630gaC>ga T	Low	17
S2, S3, S16	3	GH	26735	syn	M	M	p.71Y	c.213taC>taT	Low	This study
S10,S11	2	S	28144	ms	Orf8	Orf8	p.84L> S	c.251tTa>tCa	Moderate	(40)
S19	1	GH	28310	ms	N	N	p.13P> T	c.37cCc>Acc	Moderate	(25)
S21	1	L	28311	ms	N	N	p.13P> L	c.38cCc>cTc	Moderate	(35)
S21	1	L	28812	ms	N	N	p.180S >I	c.539aGt>aTt	Moderate	(25)
S10,S11	2	S	28878	ms	N	N	p.202S >N	c.605aGt>aAt	Moderate	(20)
S19	1	GH	29383	ms	N	N	p.370K >N	c.1110aaG>a aT	Moderate	(37)
S10,S11	2	S	29742	dn_v, nc	3'UTR	3'UTR	-	G > A	Modifier distance=68	(37)

Nc, non-coding; up_v, upstream variant; syn, synonymous variant; ms, missense variant; dn_v, downstream variant; variants depicted for 8 SARS-CoV-2 isolates from Karachi based on full NGS data and S16 for which partial (27%) sequence data was available; in addition to strains PAK/KHI1/2020 and PAK/Gilgit1/2020. No SNVs were found in PAK/Manga1/2020; ORF, open reading frame.

Table 2: Number of variants detected in the gene regions of SARS-CoV-2 virus

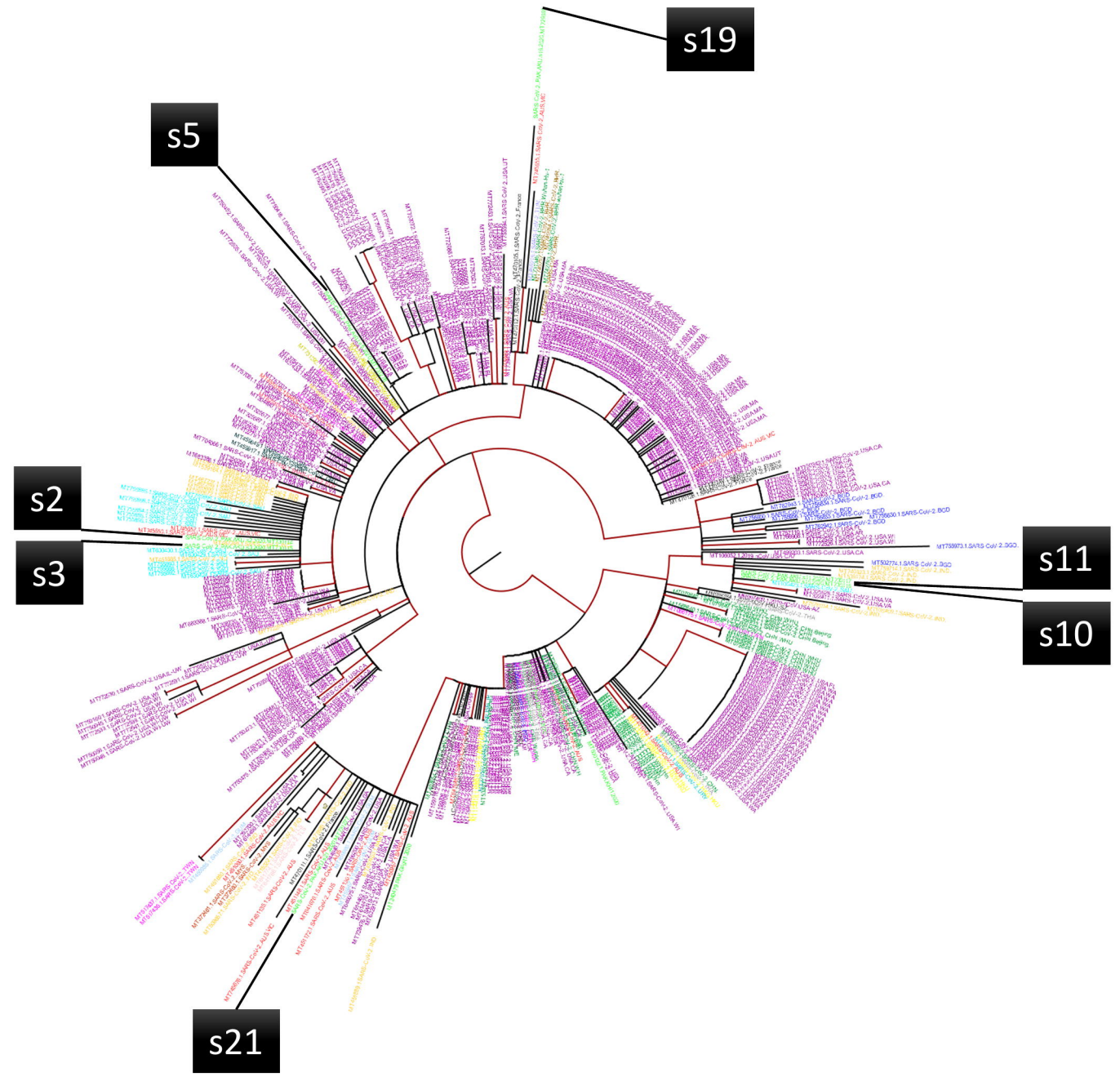
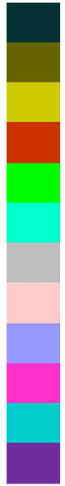
Gene	Gene region	Number of samples	Number of Variants	Synonymous variants	Missense variants
5' UTR	untranslated region	5	1	NA	NA
Orf1ab	Nsp2	5	7	4	3
	Nsp3	5	4	1	3
	Nsp4	3	2	1	1
	3C-like proteinase	1	1	1	0
	Nsp6	2	1	0	1
	RdRp	5	2	0	2
	Helicase	1	1	1	0
	EndoRNase	3	1	1	0
S	surface glycoprotein	8	4	3	1
Orf3a	ORF3a protein	6	2	1	1
M	membrane glycoprotein	3	1	1	0
Orf8	ORF8 protein	2	1	0	1
N	nucleocapsid phosphoprotein	4	5	0	5
3' UTR	untranslated region	2	1	NA	NA

UTR: untranslated region; ORF: open reading frame; Nsp: non-structural protein; RdRp: RNA-dependent RNA polymerase; S: spike glycoprotein; M: membrane glycoprotein; N: nucleocapsid phosphoprotein; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2

- Australia
- Bangladesh
- Bahrain
- China
- Egypt
- France
- Greece
- Guam
- Hong Kong
- India
- Israel



- Italy
- Jamaica
- Morocco
- Malaysia
- Pakistan
- Saudi Arabia
- Thailand
- East Timor
- Tunisia
- Taiwan
- Uruguay
- USA



Phylogeny

Country ^

