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 94 (2021) 105003

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Genetic and evolutionary analysis of SARS-CoV-2 circulating in the region surrounding Islamabad, Pakistan

Sana Tamim [a,*], Nidia S. Trovao [b], Peter Thielen [c], Tom Mehoke [c], Brian Merritt [c], Aamer Ikram [a], Muhammad Salman [a], Muhammad Masroor Alam [d], Massab Umair [a], Nazish Badar [a], Adnan Khurshid [d], Nayab Mehmood [d]

[a] Department of Virology/Immunology, National Institute of Health, Park Road, Chak Shahzad, Islamabad 45500, Pakistan
[b] Fogarty International Centre, National Institute of Health, Bethesda, MD, USA
[c] Johns Hopkins University Applied Physics Laboratory, USA
[d] WHO Regional Reference Laboratory, Polio eradication Initiative, National Institute of Health, Islamabad, Pakistan

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2
Genomics
Pakistan
Viral evolution
Bayesian analysis
Phylodynamics

ABSTRACT

Genomic epidemiology of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) has provided global epidemiological insight into the COVID-19 pandemic since it began. Sequencing of the virus has been performed at scale, with many countries depositing data into open access repositories to enable in-depth global phylogenetic analysis. To contribute to these efforts, we established an Oxford Nanopore Technologies (ONT) sequencing capability at the National Institutes of Health (NIH), Pakistan. This study highlights multiple SARS-CoV-2 lineages co-circulating during the peak of a second COVID-19 wave in Pakistan (Nov 2020-Feb 2021), with virus origins traced to the United States of America and Saudi Arabia. Ten SARS-CoV-2 positive samples were used for ONT library preparation. Sequence and phylogenetic analysis determined that the patients were infected with lineage B.1.1.250, originally identified in the United Kingdom and Bangladesh during March and April of 2020, and in circulation until the time of this study in Europe, USA and Australia. Lineage B.1.261 was originally identified in Saudi Arabia with widespread local dissemination in Pakistan. One sample clustered with the parental B.1 lineage and the other with lineage B.6 originally from Singapore. In the future, monitoring the evolutionary dynamics of circulating lineages in Pakistan will enable improved tracing of the viral spread, changing trends of their expansion trajectories, persistence, changes in their demographic dynamics, and provide guidance for better implementation of control measures.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in Wuhan, China, in December 2019 and causes mild to severe respiratory symptoms. Severe disease is more frequent in patients with comorbidities or old age, and may require hospitalization (Huang et al., 2020; Wu et al., 2020) or lead to death. On March 11th 2020 the World Health Organization (WHO) declared coronavirus disease 2019 (COVID-19) a pandemic (World Health Organization, 2020). The pandemic represents the third major human outbreak from viruses belonging to the coronaviridae family, betacoronavirus genus, and to date has caused more than 121 million confirmed cases and 2.7 million deaths worldwide (World Health Organization, 2020). The initial seeding of COVID-19 cases occurred in countries with travelers from China, and later cryptic transmission events resulted in local spread of the virus (Bedford et al., 2020; Nabeshima et al., 2021). Pakistan's first case was reported on February 28th 2020 and was imported by a traveler from Iran. Since then, sporadic cases were observed with travel history from Iran and the Middle East. The first wave of viral transmission in Pakistan peaked between May and September 2020.

Building the health sector's capacity for diagnosis and early detection for rapid outbreak response was a main priority during the first wave of the COVID-19 pandemic in Pakistan. As of July 17th 2020 Pakistan had a total testing capacity of 71,780 tests per day, with 133 testing laboratories nationwide [4] (www.dw.com/en/pakistan-coronavirus-testing/a-54221822). The containment response plan shifted from complete lockdown to smart lockdown in 20 major cities across Pakistan as coronavirus hotspots and contact tracing culminated into

* Corresponding author at: Department of Virology/Immunology, National Institute of Health, Chak Shahzad, Park Road, Islamabad, Pakistan.
E-mail address: s.tamim@nih.org.pk (S. Tamim).

https://doi.org/10.1016/j.meegid.2021.105003
Received 26 March 2021; Received in revised form 8 July 2021; Accepted 11 July 2021
Available online 14 July 2021

Table 1
Demographic characteristics of COVID-19 cases analyzed in the study.

Sample Name	Date of collection	Location	Gender	Age	Patient status	Duration of symptoms	Other infected family members
hCOV-19/pakistan/NIH-421800/2020	04-12-2020	Rawalakot, Kashmir	F	27	recovered	12 days	yes (5 members)
hCoV-19/Pakistan/NIH-417328/2020	02-12-2020	Talagang	M	60	deceased	2 days	no
hCoV-19/Pakistan/NIH-453246/2020	18-12-2020		F	45	recovered		
hCoV-19/Pakistan/NIH-446374/2020	22-12-2020	Rawalpindi	F	63	recovered	14 days	yes (3 members)

Travel history within Pakistan	Ct value	Quantity of amplified product (ng/ul)	Quantity of indexed product (ng/ul)	Row reads	Coverage	Depth	Nextstrain clade	No of Mutations	Pangolin lineage assigned	Genbank Accession number
yes	13	74.8	4.18	78,279	25,049	940×	20D	17	B.1.1.250	MW535197
yes	14	102	3.98	66,368	26,458	787×	20E	19	B.1.261	MW534548
	19	18.5	3.12	6965	16,673	82×	20D	11	B.6	MW542138
no	17	22.4	1.54	11,059	18,169	111×	20C	10	B.1	MW542139

only temporary solutions as an epidemiological strategy to track COVID 19 cases. In November 2020, the number of SARS-CoV-2 cases started surging again, and as of February 4th Japan, 2021, more than 550,000 SARS-CoV-2 cases had been confirmed in Pakistan, with an average of 1000 to 5000 new cases per day and a death rate of 2659 deaths/million [4] (www.ourworldindata.org/coronavirus-data-explorer). Sindh province has been the most affected province, accounting for 42.83% (n = 149,782) of all reported cases, while densely populated Punjab had 27.23% (n = 149,782). Khyber Pakhtunkhwa reported 11.56% (n = 63,615) cases and Balochistan had 3.38% (n = 18,612). The case count in Azad Jammu and Kashmir was 1.5% (n = 8631) while Gilgit Baltistan reported 0.8% (n = 4884) cases. The Islamabad Capital Territory had 7.29% (n = 40,111) COVID-19 patients during the same time period.

SARS-CoV-2 has a slower evolutionary rate than other RNA viruses (0.8×10^{-3} substitutions/site/year) (Rambaut et al., 2020a), and more than 800 Pango lineages (as defined by lineage assignment tool Pangolin) have been reported to circulate in different geographical areas worldwide (https://cov-lineages.org). The recent emergence of variants of concern identified in the United Kingdom (B.1.1.7) (Rambaut et al., 2020a, 2020b), South Africa (B.1.3.51) (Tegally et al., 2020) and Brazil (P.1) (Japan. 2021), with suspected increased transmissibility rate (50–70%) (Rambaut A 2020) has alarmed national and international public health groups and emphasized the significance of genomic surveillance in addition to laboratory investigation.

In this study, we established the ARTIC network protocol for full genome sequencing of SARS-CoV-2 isolates in Pakistan using the Oxford Nanopore Technologies MinION platform. The sequences obtained from clinical specimens were analyzed for their genetic variation, evolutionary history and spatio-temporal dynamics.

2. Materials and methods

2.1. Sample collection

Nasopharyngeal swab samples were collected from symptomatic patients from the major tertiary care hospitals in Islamabad as part of the laboratory-based COVID-19 surveillance during the month of December 2020 (Table 1). Samples were received in Department of Virology, NIH on the date of collection for diagnostic purposes. The use of human specimens was approved by the Institute's Research Committee, which waived written consent requirements for viral genome sequencing on the condition that the clinical information of patient will remain anonymous and the urgency of the pandemic crisis situation.

2.2. Library preparation for MinION sequencing technology

Viral RNA was isolated from nasopharyngeal swabs through a TANBead Nucleic extractor (SLA-16/32, SLA-E132 Series) to conduct lysis, washing, and elution steps. RNA extracts of samples positive for SARS-CoV-2 (ct value <20) were reverse transcribed with SSIV VILO cDNA master mix and used as primary input for overlapping tiled PCR reactions (400–600 nt reads) spanning the viral genome using New England Biolabs Q5 High-Fidelity 2× Master Mix. (M0492L) (**primers provided in** Supplementary Table T1). Amplicon pools were generated using the ARTIC Network amplicon sequencing protocol v2, with the v3 primer pools (Quick, 2020). Following purification, PCR product pools were quantified using a Promega Quantus fluorometer. NEBNext Ultra II End-Repair/dA-tailing (New England Biolabs) was performed prior to input into the Oxford Nanopore native indexing reagent set (EXP-NBD114). A no-template control (NTC) was processed in each step. Equal quantities of indexed products were combined for equal representation in the final library mixture. 210 ng of DNA library was loaded onto a primed MinION Spot On flowcell (R9.4). The sequencing run was set to sequence for 72 h in the MinKNOW software platform (version-20.10.3), with high accuracy basecalling enabled. Consensus genomes were generated using a modified ARTIC network pipeline v1.0.0 within

the BaseStack software platform (https://github.com/jhuapl-bio/Basestack). Variant polishing was evaluated for Nanopolish v0.13.2, Medaka v0.11.5, and samtools 1.9, with read length filtering and reference alignment against the Wuhan-Hu-1 genome (GenBank accession number MN908947.3). Resulting consensus sequences were analyzed through PANGOLIN v2.1.7 (https://github.com/cov-lineages/pangolin/) and Nextstrain (http://nextstrain.org/ncov) (Hadfield et al., 2018). The consensus sequences of genomes from Pakistan were deposited in Genbank with following accession numbers MW535197, MW534548, MW542138, MW542139.

We analyzed the evolutionary and spatio-temporal dynamics of two samples (NIH-421800/2020 and NIH-417328/2020), that had sufficient coverage across the genome (approx. 70%).

2.3. Phylogenetic classification through lineages

We used the phylogenetic assignment of named global outbreak lineages (PANGOLIN) [COG-UK (cog-uk.io)] to capture the genetic diversity patterns of sequences NIH-421800/2020 and NIH-417328/2020.

2.4. Selecting a genomic background dataset

For phylogenetic analyses, full-length viral genome sequences belonging to Pango lineages B.1.261 and B.1.1.250 were downloaded from GISAID [https://www.epicov.org/] on January 15 Japan, 2021. Accession numbers can be found in Supplementary Table T2. Multiple sequence alignment was performed using MAFFT v7.458 (Katoh et al., 2019) using parameters –reorder –anysymbol –nomemsave –adjustdirection –addfragments, and used Wuhan-Hu-1 (GenBank accession number: MN908947.3) sequence as a reference. Sequences with fewer than 75% unambiguous bases were excluded, as were duplicate sequences defined as having identical nucleotide composition and having been collected on the same date and in the same country. The resulting dataset was trimmed at the 5′ and 3′ ends resulting in a multi-sequence alignment with 29,782 nucleotides. This dataset was subjected to multiple iterations of phylogeny reconstruction with 1000 replicates of ultrafast bootstraps using IQ-TREE multicore software version v1.6.12 (Nguyen et al., 2015) with parameters -m GTR + G -bb 1000 -bnni -nt 50, and exclusion of outlier sequences whose genetic divergence and sampling date were incongruent using TempEst (Rambaut et al., 2016), resulting in a datasets with 34 and 107 sequences for the B.1.261 and B.1.1.250 datasets, respectively (Supplementary Figs. 1 and 2).

2.5. Phylodynamic reconstruction

Phylogenetic relationships were inferred for B.1.261 and B.1.1.250 datasets separately, with a Bayesian phylogenetic approach using Markov chain Monte Carlo (MCMC) available via the BEAST v1.10.5 package (Suchard et al., 2018) and the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA (http://biowulf.nih.gov). We used an uncorrelated relaxed molecular clock with branch rates drawn from a lognormal distribution to account for evolutionary rate variation, with an exponential growth coalescent model in which effective population size and rate of exponential growth are estimated. We assumed a general-time reversible (GTR) model of nucleotide substitution with gamma-distributed rate variation among sites. For sequences with only the year of viral collection available, the lack of tip date precision was accommodated by sampling uniformly across a one-year window from 1st January to 30th December 2020.

We estimated spatial diffusion dynamics among countries using a Bayesian discrete phylogeographic approach (Lemey et al., 2009). This approach conditions on the trait information recorded at the tips and models the transition history among those states as a continuous time Markov chain (CTMC) process, allowing the inference of unobserved states at the ancestral nodes in each tree of the posterior distribution. We used a non-reversible CTMC model (Edwards et al., 2011) and incorporated a Bayesian stochastic search variable selection to identify a sparse set of transition rates that adequately summarized the epidemiological connectivity (Lemey et al., 2009).

The MCMC chain was run separately at least five times for each of the datasets and for at least 200 million iterations with subsampling every 20,000 iterations, using the BEAGLE (Ayres et al., 2019) [https://academic.oup.com/sysbio/article/61/1/170/1680634] library to improve computational performance. All parameters reached convergence, as assessed visually using Tracer v.1.7.1 (Rambaut et al., 2018), with statistical uncertainty reflected in values of the 95% highest posterior density (HPD). At least 10% of the chain was removed as burn-in. Maximum clade credibility (MCC) trees were summarized using TreeAnnotator v1.10.5 and visualized in FigTree v1.4.4.

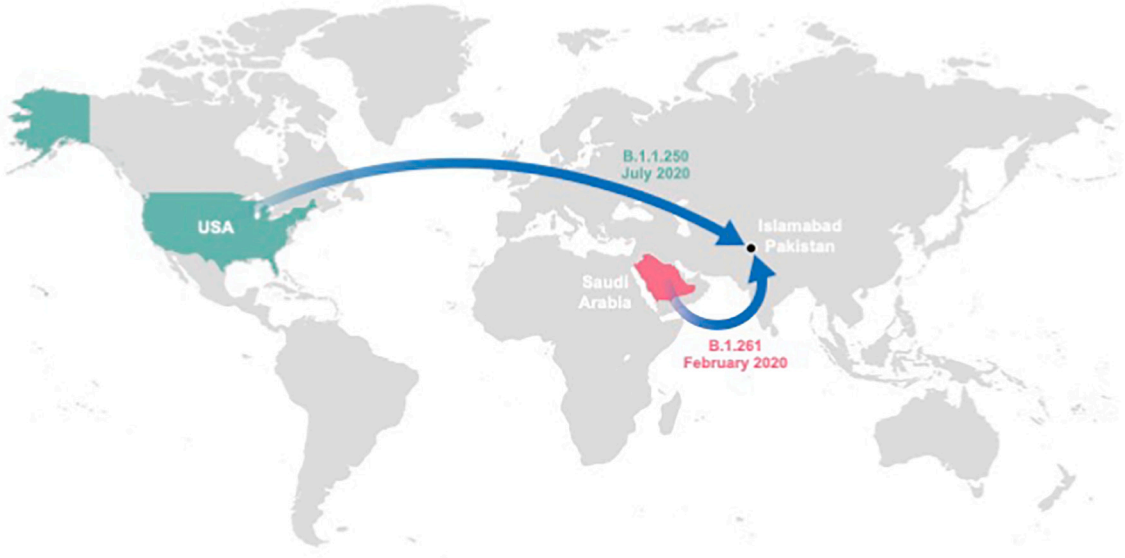

Fig. 1. Spatial dispersal of B.1.1.250 and B.1.261 lineages into the region surrounding Islamabad, Pakistan. The spatial dispersal of SARS-CoV-2 viruses from Saudi Arabia and the United States into the region around Islamabad, Pakistan was inferred from the respective MCC trees (Supplementary Figs. 2 and 3). Colored arrows project the MCC trees' branches that lead to the seeding events of the B.1.1.250 (green) and B.1.261 (magenta) lineages in Pakistan. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

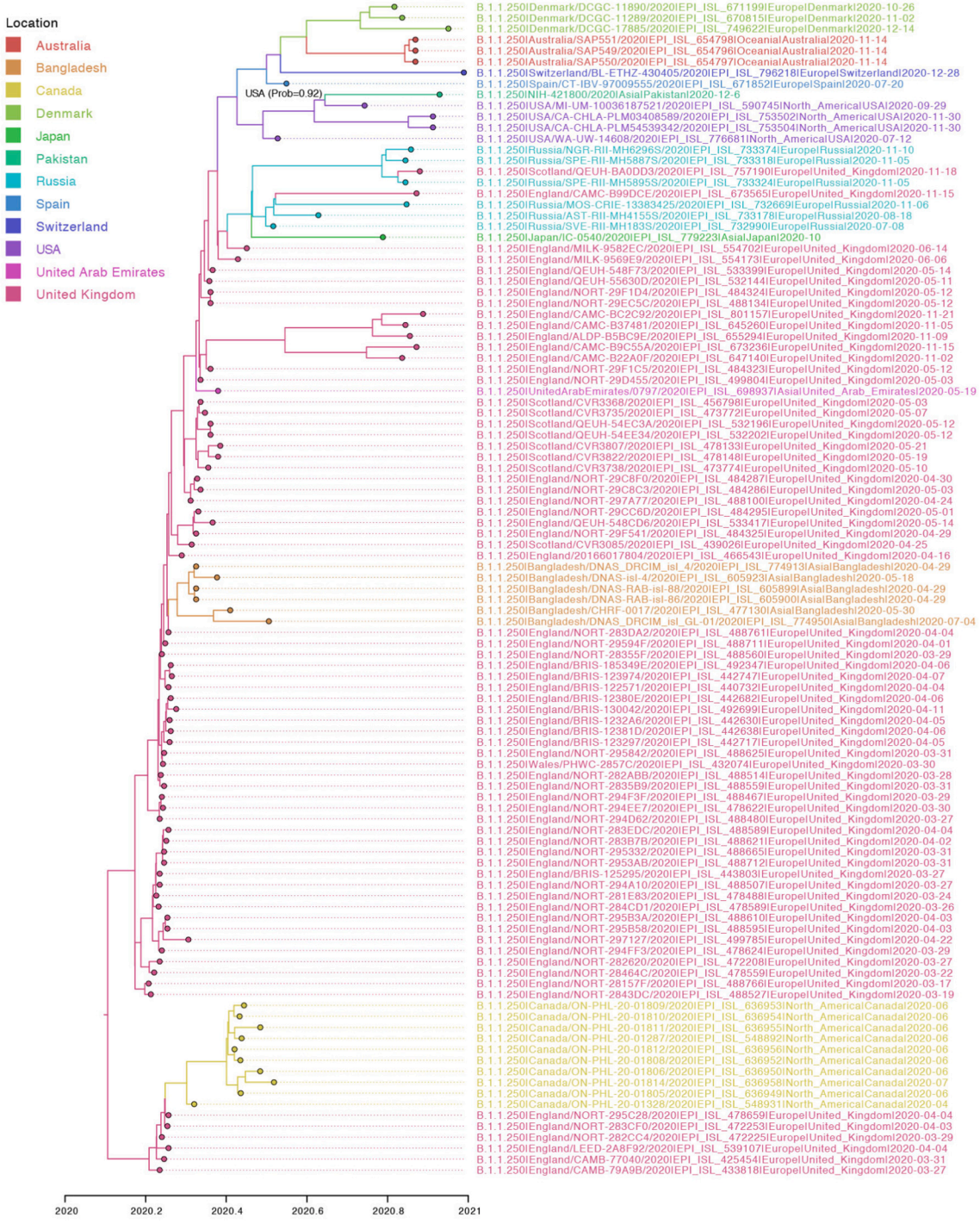

Fig. 2. Maximum clade credibility (MCC) tree inferred for the whole genomes of lineage B.1.1.250. The shade of the branches and tips indicates the inferred location state at the nodes. The ancestral node of NIH-421800/2020 is annotation with the inferred location and probability.

3. Results

In order to develop a MinION sequencing protocol at our location, we selected ten SARS-CoV-2 samples with diagnostic real-time PCR cycle threshold (Ct) values less than 20 for targeted sequencing. Following preparation with the ARTIC network tiled amplicon approach, the final library of 15 µl at 14 ng/µl was loaded onto a MinION flowcell set to run for 72 h. After 36 h, the sequencing run was terminated due to reduced sequence data generation. Analysis through the ARTIC sequencing pipeline showed that, among the 10 samples, four had coverage of 25,049, 26,458, 16,673 and 18,169 with sequence depth of 940×, 787×, 82× and 111×, respectively (Table 1). Six of the ten samples did not produce sufficient sequencing output to generate consensus genomes.

3.1. Mutational Analysis

Sample NIH-421800/2020 belonged to clade 20D and had 17 mutations, four synonymous mutations (ORF1a: C1912T, C3037T, ORF1b: C1356T, S: C24904T) and nine missense mutations, three in ORF1a (T1246I, G3278S, I3587T), two in the S gene (D614G, T1117I) and one each in ORF1b (F1296L), ORF3a (G174C), ORF8 (T26I) and the N (G204R) gene. Sample NIH-417328 represented clade 20A and had eight

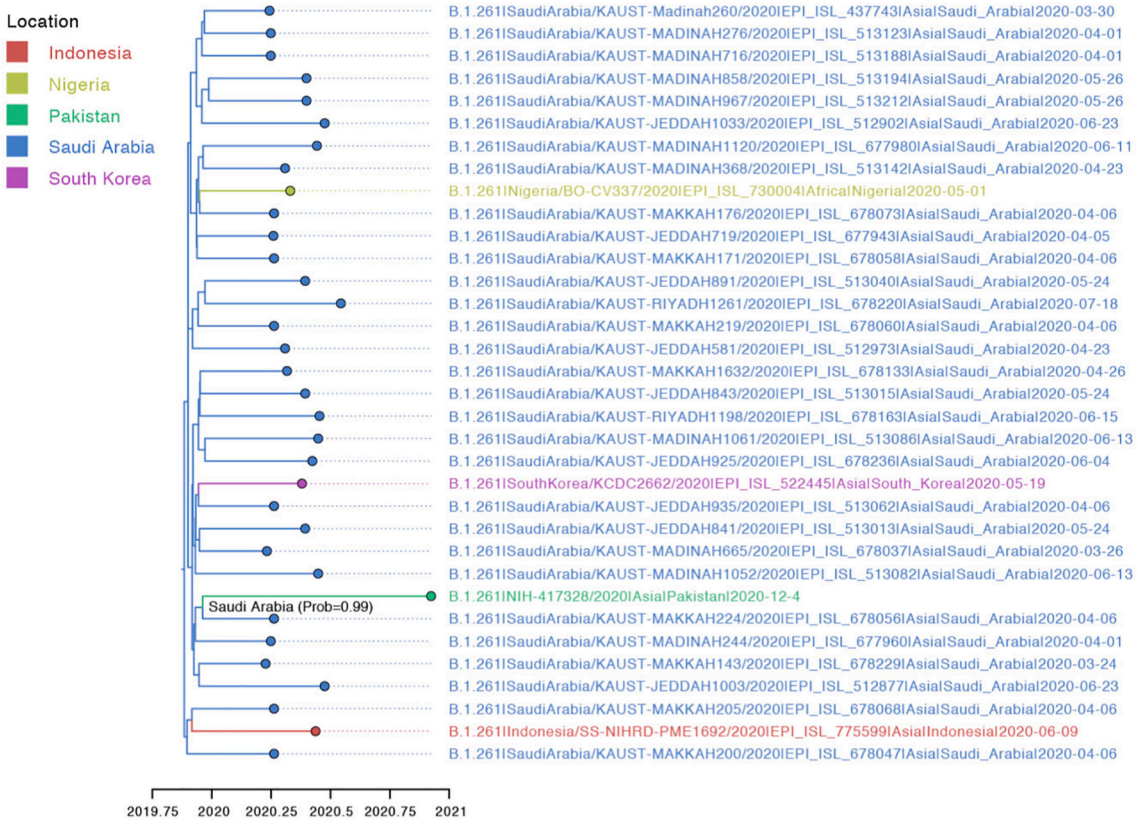

Fig. 3. Maximum clade credibility (MCC) tree inferred for the whole genomes of lineage B.1.261. The shade of the branches and tips indicates the inferred location state at the nodes. The ancestral node of NIH-417328/2020 is annotated with the inferred location and probability.

synonymous mutations in ORF1ab and M genes and nine non-synonymous mutations spread across its genome in ORF1a: S2926F, ORF1b: M2667I, S: F157L, L1024R; ORF3a: Q57H; E: S67F, ORF8: V62L; N: Q163K, R209I.

Samples NIH-453246 and NIH-446374 belonged to clades 20D and 20C respectively and both showed three synonymous mutations. Among the non-synonymous mutations, sample NIH-453246 had eight mutations, of which 5 were in ORF1a (A565V, T1246I, P1692S, P2144L, L3667F) and one each in the S (T95I), M (H125Y) and N (G204R) genes. NIH 446374 had seven non-synonymous mutations in ORF1a (T265I, T3646A), ORF3a (Q57H, T151I), ORF8 (T80I) and the N gene (P13S, A220V).

3.2. Lineage determination

PANGOLIN analysis identified sample NIH-421800/2020 as lineage B.1.1.250, which had previously been observed in isolates from Bangladesh, the United States, and the United Kingdom during March–September 2020 (Fig. 2). NIH-417328/2020 was identified as part of lineage B.1.261 which also circulated in Saudi Arabia and South Korea between March and June 2020 (Fig. 3). Both lineages were assigned with high support (probability = 1).

3.3. Evolutionary and spatio-temporal patterns of SARS-CoV-2 in Pakistan

Both lineages were found to be evolving at similar rates [B.1.1.250: 9.83×10^{-4} (95% Highest Posterior Probability (HPD): 5.98×10^{-4}–1.52×10^{-3}) substitutions/site/year; B.1.261: 3.05×10^{-4} (95% HPD 7.00×10^{-5}–6.70×10^{-4}) substitutions/site/year), estimates that are consistent with previous studies (Motayo et al., 2021). The genetic diversity of NIH-421800/2020 traces back to a common ancestor that existed on July 29th, 2020 in United States (location probability = 0.92) (Figs. 1 & 2). Additionally, NIH-417328/2020 was inferred to have a common ancestor around February 1st, 2020 in Saudi Arabia (location probability = 0.99) (Figs. 1 & 3 and Supplementary Table T3).

4. Discussion

Through implementation of Oxford Nanopore sequencing, we identified SARS-CoV-2 lineages in circulation in the Pakistani population. There is abundant genomic data available from Europe, United States and South East Asia, whereas very few sequences from Pakistan have been reported. NIH has established this sequencing technique through collaboration with the United States NIH Fogarty International Center (NIH/FIC) and the Johns Hopkins University Applied Physics Laboratory (JHU/APL). Through this effort, we successfully established the ARTIC network sequencing protocol locally.

The two patients whose samples were used for phylogenetic analysis had no travel history outside of Pakistan and had only local travel to Islamabad, since it has the largest tertiary care hospital. Both cases contracted COVID-19 in Islamabad. NIH-421800/2020, collected from a patient who was part of a family cluster, had the D614G mutation that may have accelerated the spread of B.1.1.250, one of the most contagious variants during the first wave (Leung et al., 2020). The patient infected with NIH-417328 had co-morbidities (Table 1) and succumbed to the infection despite being a non-D614G mutation variant.

We speculate that these samples were from two separate local sporadic transmission events since we reconstructed distinct phylogenetic origins and the viruses harbored unique mutations in their S and ORF1ab genes (i.e singleton mutations). The D614G mutation of S gene identified in NIH-421800/2020 was first detected in February in Europe and was the predominant lineage of the first wave of COVID-19 globally and replaced Wuhan-Wu-1 (Korber et al., 2020). It has also been

reported circulating in Karachi (Shakeel et al., 2021), during the first wave of the pandemic in the country, however with different genetic makeup (i.e. B.1.36, B.1.160, B.1.255 Lineages).

Individual contact tracing is useful only during phases of the outbreak when transmission chains can be easily traced (Leo et al., 2003; Faye et al., 2015; Kim et al., 2017). However, during widespread pandemics, when patient numbers overwhelm local contact tracing capacity, genomic surveillance may be a practical strategy to trace the viral spread, as has been proven effective in previous viral outbreaks (Bahl et al., 2011; Baillie et al., 2012; Dudas et al., 2017; Grubaugh et al., 2017). Viral lineage dynamics can fluctuate at the regional-level during epidemics, and genomic surveillance can provide insights into local genetic diversity as well as identifying previously undetected lineages and potential phenotypic determinants of transmission and pathogenicity. Transmission events or seeding of particular lineages in a population can be more easily traced through patients with travel history. The phylogeographic analysis estimated the origins of the most recent common ancestor for NIH-421800/2020 and NIH-417328/2020 in Saudi Arabia and the United States, respectively, months earlier than their collection dates. This could suggest some degree of undetected cryptic transmission in Pakistan, since transmission probably did not occur directly between two locations, and most likely involved unsampled intermediary locations before being detected by genomic surveillance. We could not determine the entry points of these lineages in the country or account for their previous circulation in Pakistan due to lack of within country SARS-CoV-2 genetic data. Additional evolutionary relationships could be established with the inclusion of complete viral sequences from an increased number of samples, which would help determine the diversity and regional distribution with respect to host population.

The actual genetic diversity of SARS-CoV-2 in Pakistani population could be addressed with large sample numbers across the country, and this study is limited by the overall number of sequences described. A large sample size will support the identification of the predominant lineages in circulation and avoid sampling bias. There is a demand for pragmatic resource allocation for high throughput sequencing techniques that systematically target cases, instead of focusing only on a few or rare lineages which would jeopardize the characterization of the true genetic landscape.

Genomic surveillance of COVID-19 patient samples allows the identification of polymorphisms such as deletions, synonymous and missense mutations in circulating strains that may contribute to increased transmissibility or pathogenicity of viral lineages creating SARS-CoV-2 variants of concern. Our analysis stresses the need to build capacity for real-time genomic surveillance and epidemiology in Pakistan and other low-resource nations to assist the public health response to COVID-19 pandemic or future outbreaks. These measures would help identify predominant and rare lineages while tracking their community transmission, providing important information for public health decision makers.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.105003.

Author contributions

ST and NST contributed to conceptualization, analysis, methodology, project administration, visualization. ST, NST, and PT drafted, reviewed and edited the manuscript. PT, TM and BM participated in data curation, formal analysis, investigation, methodology, visualization and supervision. MMA, MU, NB, AK and NM performed formal analysis, data curation, methodology, writing, review and editing. AK, MS were involved in funding acquisition, resources, supervision, writing, review and editing.

Ethical approval

This study was approved by the Institutional Review Committee of National Institute of Health, Islamabad, Pakistan.

Declaration of Competing Interest

The authors have no conflict of interest to declare.

Acknowledgements

We gratefully acknowledge the Fogarty International Center at the National Institutes of Health (NIH/FIC) and the Johns Hopkins University Applied Physics Laboratory (JHU/APL) for developing in-country capacity for whole genome sequencing of SARS-CoV-2 and providing technical guidance. The opinions expressed in this article are those of the authors and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

References

Ayres, D.L., Cummings, M.P., et al., 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical Phylogenetics. Syst. Biol. 68 (6), 1052–1061.
Bahl, J., Nelson, M.I., et al., 2011. Temporally structured metapopulation dynamics and persistence of influenza a H3N2 virus in humans. Proc. Natl. Acad. Sci. U. S. A. 108 (48), 19359–19364.
Baillie, G.J., Galiano, M., et al., 2012. Evolutionary dynamics of local pandemic H1N1/2009 influenza virus lineages revealed by whole-genome analysis. J. Virol. 86 (1), 11–18.
Bedford, T., Greninger, A.L., et al., 2020. Cryptic transmission of SARS-CoV-2 in Washington state. Science 370 (6516), 571–575.
Dudas, G., Carvalho, L.M., et al., 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544 (7650), 309–315.
Edwards, C.J., Suchard, M.A., et al., 2011. Ancient hybridization and an Irish origin for the modern polar bear matriline. Curr. Biol. 21 (15), 1251–1258.
Faye, O., Boëlle, P.-Y., et al., 2015. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect. Dis. 15 (3), 320–326.
Grubaugh, N.D., Ladner, J.T., et al., 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 546 (7658), 401–405.
Hadfield, J., Megill, C., et al., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34 (23), 4121–4123.
Huang, C., Wang, Y., et al., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395 (10223), 497–506.
Katoh, K., Rozewicki, J., et al., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20 (4), 1160–1166.
Kim, K., Tandi, T., et al., 2017. Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak in South Korea, 2015: epidemiology, characteristics and public health implications. J. Hosp. Infect. 95 (2), 207–213.
Korber, B., Fischer, W.M., et al., 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182 (4), 812–827 e819.
Lemey, P., Rambaut, A., et al., 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5 (9), e1000520.
Leo, Y., Chen, M., et al., 2003. Severe acute respiratory syndrome-Singapore, 2003. MMWR: Morbidity & Mortality Weekly Report 52 (18), 405.
Leung, Kathy, Pei, Yao, Leung, Gabriel M., Lam, Tommy T.Y., Wu, Joseph T., 2020. Empirical Transmission Advantage Of The D614G Mutant Strain Of SARS-Cov-2. medRxiv preprint. https://doi.org/10.1101/2020.09.22.20199810.
Motayo, B.O., Oluwasemowo, O.O., et al., 2021. Evolution and genetic diversity of SARS-CoV-2 in Africa using whole genome sequences. Int. J. Infect. Dis. 103, 282–287.
Nabeshima, T., Takazono, T., et al., 2021. COVID-19 cryptic transmission and genetic information blackouts: need for effective surveillance policy to better understand disease burden. Lancet Reg Health West Pac 7, 100104.
Nguyen, L.T., Schmidt, H.A., et al., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32 (1), 268–274.
Quick, J., 2020. nCoV-2019 Sequencing Protocol v1 (Protocols.Io.bbmuik6w).
Rambaut, A., Lam, T.T., et al., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2 (1) vew007.
Rambaut, A., Drummond, A.J., et al., 2018. Posterior summarization in Bayesian Phylogenetics using tracer 1.7. Syst. Biol. 67 (5), 901–904.
Rambaut, A., Holmes, E.C., et al., 2020a. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5 (11), 1403–1407.
Rambaut, A.L.N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, L.D., Volz, E., 2020b. Preliminary genomic characterization

of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. SARS-CoV-2 coronavirusnCoV-2019 Genomic Epidemiology.
Shakeel, Muhammad, Irfan, Muhammad, Ahmed, Zaib-Un-Nisa, Rashid, Muhammad, Ansari, Sabeeta Kanwal, Khan, IshtiaqAhmad, 2021. Surveillance of genetic diversity and evolution in locally transmitted SARS-CoV-2 in Pakistan during the first wave of the COVID-19 pandemic. bioRxiv. https://doi.org/10.1101/2021.01.13.426548.
Suchard, M.A., Lemey, P., et al., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4 (1) vey016.
Tegally, H.W.E., Giovanetti, M., et al., 2020. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. MedRxiv. https://doi.org/10.1101/2020.12.21.20248640. Corpus ID: 229348551.
World Health Organization, 2020. World Health Organization Coronavirus disease (COVID-2019) situation report – 126 Saf. Risk Pharmacother **8**, 3–8.
Wu, J., Li, J., et al., 2020. Clinical features of maintenance hemodialysis patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. Clin. J. Am. Soc. Nephrol. 15 (8), 1139–1145.