medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

SARS-CoV-2 variants of concern dominate in Lahore, Pakistan in April 2021.

Muhammad Bilal Sarwar[1+], Muhammad Yasir[2+], Nabil-Fareed Alikhan*[2], Nadeem Afzal[1], Leonardo de Oliveira Martins[2], Thanh Le Viet[2], Alexander J Trotter[2], Sophie J Prosolek[2], Gemma L Kay[2], Ebenezer Foster-Nyarko[2], Steven Rudder[2], David J Baker[2], [1], Sidra-tul-muntaha[1,3], Muhammad Roman[1], Mark A Webber[2,4], Almina Shafiq[1],Balqees Shabir[5], Javed Akram[1], Andrew J Page*[2], Shah Jahan*[1]

[1]Department of Immunology, University of Health Sciences, Lahore, Pakistan.

[2]Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK.

[3]Central Diagnostic Facility, Mayo Hospital, Lahore Pakistan.

[4]University of East Anglia, Norwich, UK.

[5]Department of Medicine, East Medical ward, King Edward medical University Mayo Hospital Lahore Pakistan.

*Corresponding authors: Nabil-Fareed Alikhan <nabil-fareed.alikhan@quadram.ac.uk> & Shah Jahan <shahjahan@uhs.edu.pk>

+Joint first authors / authors contributed equally

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Background

The SARS-CoV-2 pandemic continues to expand globally, with case numbers rising in many areas of the world, including the Indian sub-continent. Pakistan has one of the worlds largest populations, of over 200 million people and is experiencing a severe third wave of infections caused by SARS-CoV-2 beginning in March 2021. Currently very few SARS-CoV-2 genomes collected in Pakistan are available, with just 12 covering the third wave, 9 of which are from Islamabad. This highlights the need for more genome sequencing to allow surveillance of variants in circulation. In fact more genomes are available for travellers with a travel history from Pakistan, than from within the country itself.

Methods

For an understanding of the circulating variants in Lahore and surrounding areas with a combined population of 11.1 million, 102 samples were sequenced, covering one week period from April 2021. The samples were randomly chosen from 2 hospitals with a diagnostic polymerase chain reaction (PCR) cutoff value of less than 25 cycles.

Results

Analysis of the lineages shows that B.1.1.7 (first identified in the UK, Alpha variant) dominates, accounting for 97.9% (97/99) of cases, with B.1.351 (first identified in South Africa, Beta variant) accounting for 2.0% (2/99) of cases. No other lineages were observed.

Discussion

In depth analysis of the B.1.1.7 lineages indicates multiple separate introductions and subsequent establishment within the region. Eight samples were identical to genomes observed in Europe (7 UK, 1 Switzerland), indicating recent transmission. Genomes of other samples show evidence that these have evolved, indicating sustained transmission over a period of time either within Pakistan or other countries with low density genome sequencing. Vaccines remain effective against B.1.1.7, however the low level of B.1.351 against which some vaccines are less effective demonstrates the requirement for continued prospective genomic surveillance.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

The COVID-19 pandemic has spread rapidly throughout the world and continues to expand in many regions. It began with an unknown case of pneumonia in the city of Wuhan, China (Huang *et al.*, 2020). The causative pathogen has since been named 'severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2)'. As of May 2021, there have been over 167 million reported cases and 3.4 million fatalities (Dong, Du and Gardner, 2020). Genomic surveillance has assisted the pandemic response providing information for outbreak investigations and detecting possible epitope changes that would allow the virus to escape vaccines. Multiple classification systems have been developed to quickly communicate SARS-CoV-2 variants circulating in a community (Alm *et al.*, 2020; Hodcroft *et al.*, 2020; Rambaut *et al.*, 2020). From these definitions, certain lineages have been designated as Variants of Concern (VOCs), which are defined as such due to indications of increased transmission patterns and/or possible resistance to vaccine and/or other treatments (Volz *et al.*, 2021). World Health Organisation (WHO) has introduced a new nomenclature of these VOCs and Variants of Interest (VOIs) based on Greek alphabets.

B.1.1.7 (Alpha) and B.1.351 (Beta) are two VOCs that have circulated globally. The SARS-CoV-2 lineage B.1.1.7, designated Variant of Concern 202012/01 (VOC) by Public Health England, was first identified in the UK in late Summer to early Autumn 2020 (Volz *et al.*, 2021). B.1.351 is another VOC identified in South Africa and defined by eight mutations in the spike protein including (K417N, E484K and N501Y) (Tegally *et al.*, 2021).

Pakistan is currently experiencing a severe third wave of infections caused by SARS-CoV-2 which began in March 2021. Vaccination rates nationally are under 2% (*National Command Operation Center*, 2021), leaving large segments of the community at risk of serious illness from COVID-19. Currently very few SARS-CoV-2 genomes collected in Pakistan are available, with just 12 covering the third wave, 9 of which are from one city, Islamabad. This highlights the need for more genome sequencing from Pakistan, particularly given the current situation and the very high population in order to allow surveillance of variants in circulation. Currently, more genomes are available for travellers with a travel history from Pakistan, than from within the country itself (Shu and McCauley, 2017).

We have amplicon sequenced 102 samples, randomly chosen from a 1 week period in April 2021 from Lahore and surrounding areas, to get a snapshot assessment of the circulating lineage in the region. This has identified the B.1.1.7 variant of concern (Alpha) as the primary lineage circulating, found in 97.9% of cases, with clear signals of repeated overseas introductions into the region.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results and Discussion

In this study, one hundred and two SARS-CoV-2 samples in viral transport medium (VTM) were randomly selected from SARS-CoV-2 diagnostic positives by the Mayo and Sheikh Zayed hospitals in Lahore, Pakistan, over a 6 day period from 2021-04-06 to 2021-04-11. All samples had a Ct < 25. These public hospitals serve a predominantly middle income population. The patients were aged between 21 and 91 with a broad distribution of ages, with 74% (n=74) male, 26% (n=26) female and 2 unknown. Four samples were from patients who died (all aged over 60), with all others recovering. Patients had no association with international travel, with 77 patients explicitly reporting they and their household had not recently travelled overseas (Supplementary Table S1).

Viral RNA was amplified using the ARTIC protocol (Quick, 2020) with sequencing libraries prepared using CoronaHiT (Baker *et al.*, 2021). Resulting sequenced reads were used to generate consensus sequences with the ARTIC bioinformatics protocol (See methods).

Analysis of the PANGO lineages shows that B.1.1.7 (first identified in the UK) dominates, accounting for 97.9% (97/99) of cases, with B.1.351 (first identified in South Africa) accounting for 2.0% (2/99) of cases. No other lineages were observed.

SARS-CoV-2 was first identified in Pakistan on 26th February 2020 (Javed *et al.*, 2020). The first B.1.1.7 genome was submitted to GISAID on 2020-12-25. There were previously 268 SARS-CoV-2 genomes available on GISAID where Pakistan was listed as the country of exposure with 90 assigned as B.1.1.7 or B.1.351 (Table 1; Supplementary Table S2). Most of these samples were associated with international travel from the country and were collected outside of Pakistan (Supplementary Table S2). A list of submitting authors of these data can be found in Supplementary Table S3. We combined all B.1.1.7 and B.1.351 public data with the genomes presented in this study to provide a snapshot of B.1.1.7 and B.1.351 dissemination (Supplementary Table S4).

Table 1: PANGO lineages of GISAID data where Pakistan is the country of exposure

Lineage	Count	Lineage	Count	Lineage	Count
B.1.1.7	87	C.23	5	B.1.525	1
B.1.36	35	B.1.36.24	3	C.36	1
B.1	30	A	3	B.1.260	1
B.1.471	24	B	3	B.1.36.8	1
B.1.1.1	18	B.1.351	3	B.1.523	1
B.1.36.31	17	C.11	2	B.1.459	1
B.6	12	B.1.1.370	1	B.1.562	1
B.1.1	10	B.1.36.22	1	B.1.617.1	1
B.1.36.34	5	B.1.1.214	1		

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

B.1.1.7 was likely introduced into Pakistan as the lineage emerged in the United Kingdom. B.1.1.7 samples are found in multiple clades spanning the entire phylogeny of B.1.1.7, suggesting the time to most recent common ancestor of all B.1.1.7 and B.1.1.7 in Pakistan were similar, which was calculated here as October 2020 (Figure 1). The date of emergence of B.1.1.7 is before this date and has been calculated as September 2020 elsewhere (Galloway *et al.*, 2021). These early introduction dates into Pakistan are plausible as the first confirmed B.1.1.7 genomes associated with Pakistan date back as early as late December 2020 (Figure 1).

By reconstructing the exposure history over the phylogenetic tree, we estimate the number of B.1.1.7 importations as 127 using the dated tree (Figure 1). More than half the introductions into the country happened before March 2021, despite a constant rate of recent importations. Genomes from Pakistan were intermingled with genomes from elsewhere, as shown in Figure 2 where we calculated the patristic distance of genomes in this study to their closest neighbour and found that they were generally closer to samples from abroad. The number of substitutions to the nearest neighbour also varied (0-8 substitutions)(Figure 2). Considering that all samples in this study were sourced from patients with no travel history, this suggests that the current data does not fully capture transmission within Pakistan. The virus is detected and sequenced from the source (usually the United Kingdom) and then in some cases sampled in Pakistan immediately or after several months, where the number of substitutions were lower or higher respectively.

B.1.351 was likely introduced into Pakistan recently. B.1.351 samples are found in four clades in the phylogeny of B.1.351, suggesting at least 4 separate introductions into the country (Figure 3). The time to the most recent common ancestor of B.1.351 globally was calculated here as during September 2020, whereas the clades containing B.1.351 genomes from Pakistan date no earlier than February 2021 (Figure 3). The date of emergence of B.1.351 is before these dates and has been calculated as early August 2020 elsewhere (Tegally *et al.*, 2021).

It has been demonstrated (Abu-Raddad, Chemaitelly and Butt, 2021) that vaccines, in particular BNT162b2, are effective against B.1.1.7. Therefore, if this remains the dominant lineage in the region, public health vaccination policy can be implemented accordingly. However, whilst the prevalence of B.1.351, which has been linked to lower vaccine efficacy, is very low, this reinforces the need for prospective surveillance of SARS-CoV-2 using genome sequencing to inform public health interventions in a continual manner.

The most regular source of genomic sequencing is from travellers from Pakistan being sequenced by their destination countries, and being annotated as such in the public databases (GISAID). Japan is the largest contributor of genomes from

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

travellers originating from Pakistan (111 out of 269). This indirect surveillance is useful but unreliable as travel restrictions and pre-flight testing can bias the results.

This report shows the critical importance of whole-genome sequencing of SARS-CoV-2 to determine the prevalence and changing epidemiology of different variants of the virus. This data is crucial to inform public health decision making as well as to allow global epidemiology to be understood. Building capacity for sequencing and analysis of genomes in countries with high infection rates will be crucial for the global response to CoVID-19.

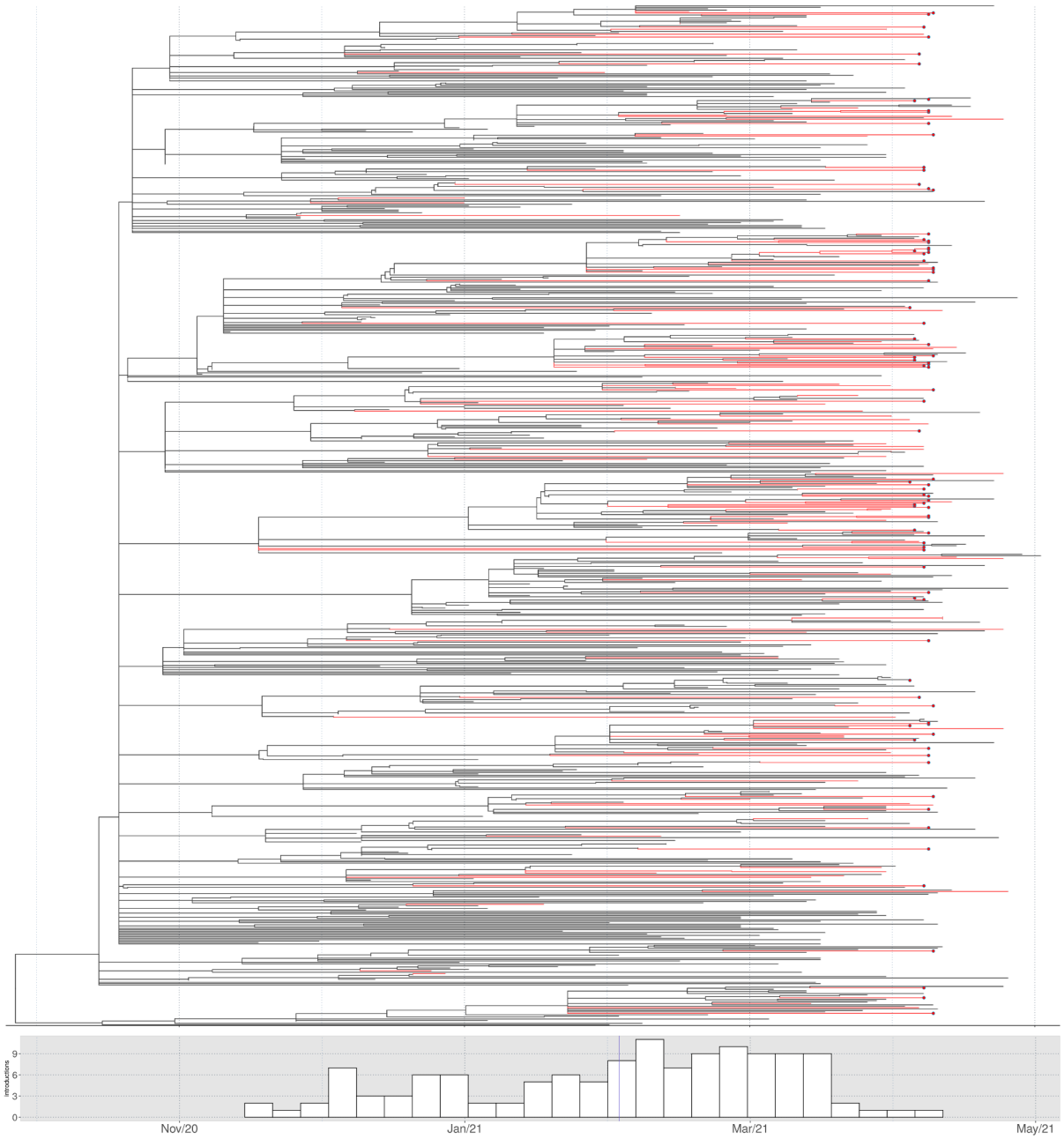


Figure 1: Phylogeny of B.1.1.7 including genomes from Pakistan

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

At the top, the maximum likelihood dated tree of lineage B.1.1.7, showing only samples close to sequences related to Pakistan. Branches are coloured red when all descendant tips have Pakistan as the country of exposure. Samples sequenced in the present study are highlighted as red dots. At the bottom we have the histogram of introductions into Pakistan over time, as estimated by ancestral state reconstruction, with the average time represented by a blue vertical line. Phylogenetic tree was estimated with IQTREE2 followed by divergence times estimation usingTreeTime after excluding outliers. The figure was plotted with ggtree and introduction events were estimated with castor.

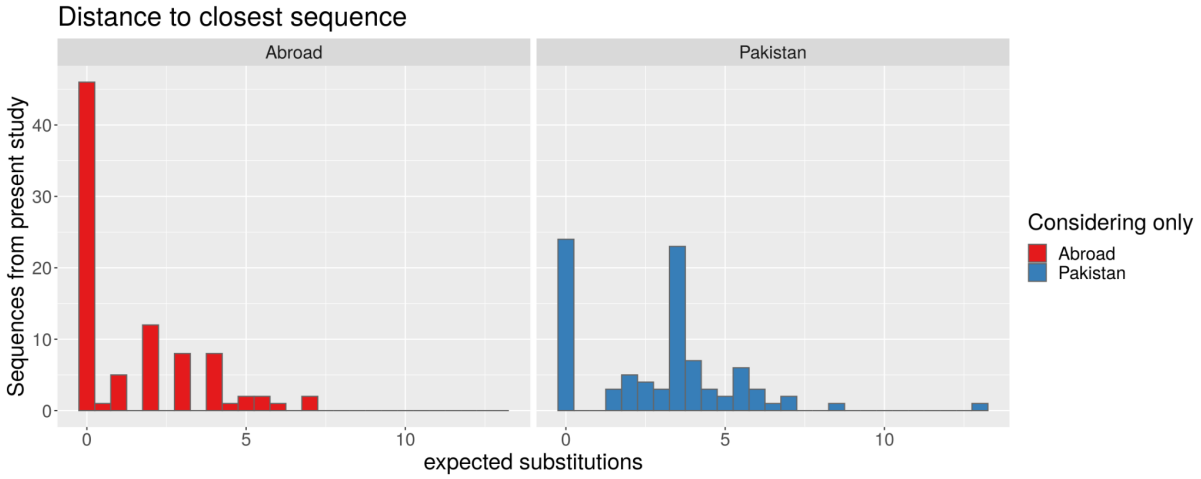


Figure 2: Distribution of distances to closest neighbours from Pakistan and abroad, for B.1.1.7 sequences.

The number of expected substitutions is calculated from the patristic distances between leaves over the maximum likelihood tree, transformed from substitutions per site by multiplying by genome length. For each of the 88 UHS-PAK sequences, we find the closest distance considering only Pakistan (blue) or international (red) tips on the tree.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

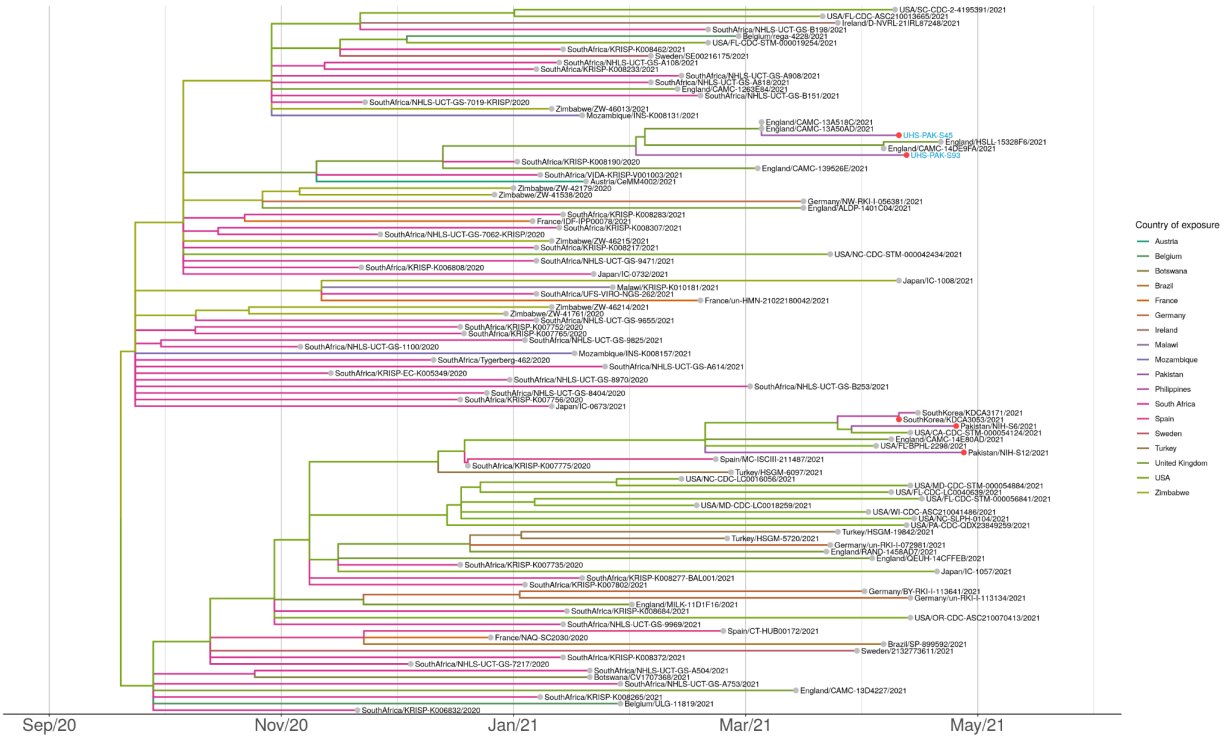


Figure 3: Maximum likelihood dated tree of lineage B.1.351 focused on samples sequenced in this study.
Branches are coloured according to country of exposure, with Pakistan highlighted with a red circle. The two samples from the current study, UHS-PAK-S45 and UHS-PAK-S93, are labelled in blue. Phylogenetic tree was estimated with IQTREE2 (Minh *et al.*, 2020) followed by divergence times estimation usingTreeTime after excluding outliers. The figure was plotted with ggtree (Yu *et al.*, 2017).

Supplementary data

Supplementary table S1 - Samples sequenced in this study

Supplementary table S2 - GISAID metadata of samples associated with Pakistan

Supplementary Table S3 - GISAID Submitting author information

Supplementary Table S4 - GISAID metadata of B.1.351 and B.1.1.7

Methods

Genome sequencing and analysis

RNA was extracted using Viral RNA Extraction kit, FavorPrep™ and Polymerase Chain Reaction (RNA) was carried out using GenomeCoV19 Detection Kit by abm (cat: 628) on IQ 5 bioRad in diagnostic laboratories of Lahore Pakistan. Positive

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

samples with a CT < 25 were randomly selected for genome sequencing. Viral RNA was converted in cDNA and was amplified using the ARTIC protocol v3 (LoCost) (Quick, 2020) with sequencing libraries prepared using CoronaHiT (Baker *et al.*, 2021). We carried out genome sequencing using the Illumina NextSeq 500 platform.

The raw reads were demultiplexed using bcl2fastq (v2.20). The reads were used to generate a consensus sequence using the ARTIC bioinformatic pipeline (*Artic Network*, no date). Briefly, the reads had adapters trimmed with TrimGalore (Krueger, 2020) and were aligned to the WuhanHu-1 reference genome (accession MN908947.3) using BWA-MEM (v0.7.17) (Li, 2013); the ARTIC amplicons were trimmed and a consensus built using iVAR (v.1.2.3) (Grubaugh *et al.*, 2019).

PANGO lineages assigned using Pangolin v2.4.2 and PangoLEARN model dated 2021-05-12 (Rambaut *et al.*, 2020).

Phylogenetic analysis

For the phylogenetic analysis, all sequences from GISAID where Pakistan is the country of exposure were downloaded and added to the sequences from the current study. All remaining sequences from GISAID were then compared to this data set where we kept the closest ones —for each Pakistan sequence, the four closest from non-Pakistan were kept for subsequent analysis to provide context. The alignment and neighbour search were done with uvaia (de Oliveira Martins, Leonardo, 2021), and problematic (homoplasic or difficult to sequence) sites were masked from the alignment (Turakhia *et al.*, 2020). For the B.1.1.7 and B.1.351 dated phylogenetic inference, we further enriched each alignment with more distant neighbours which maximised the phylogenetic diversity (Minh, Klaere and von Haeseler, 2006), based on the neighbour-joining tree (Simonsen and Pedersen, 2011) of all sequences close to each Pakistan sequence, for each lineage. Sequences with more than 10% N, or with incomplete date information were excluded from analysis. Clusters distant and unrelated to Pakistan sequences were reduced or removed by visual inspection of maximum likelihood trees.

From the 102 UHS-PAK sequences, 90 samples were included in the phylogenetic analysis: 88 from B.1.1.7 and 2 samples from B.1.351; the final alignments have 723 and 107 sequences, respectively. The maximum likelihood trees were inferred with IQTREE2 v2.1.2 (Minh *et al.*, 2020) and the divergence times were estimated by marginalisation under a strict clock using TreeTime. Ancestral state reconstruction of the country of exposure was done with castor (Louca and Doebeli, 2018) and trees were plotted with ggtree (Yu *et al.*, 2017), both for R. The number of introductions into Pakistan was estimated by counting edges where the reconstructed probability of being exposed in Pakistan increased to one, weighting or not by the number of most parsimonious scenarios. Due to unequal sampling and sequencing, the number of introductions is an underestimate and their dates are subject to selection bias (e.g.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

previous introductions were not sequenced due to regional differences or severity of infection).

Declaration of Interests

None declared

Funding

The Quadram Institute authors gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); their research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10352, also Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). The University of Health Sciences authors acknowledge the support provided by the Higher Education Commission (HEC) of Pakistan under the project RRG-211.

Ethics

This project was conducted under approval number UHS/REG-20/ERC/1758 from the University of Health Sciences Lahore Ethical Review Committee

Data availability

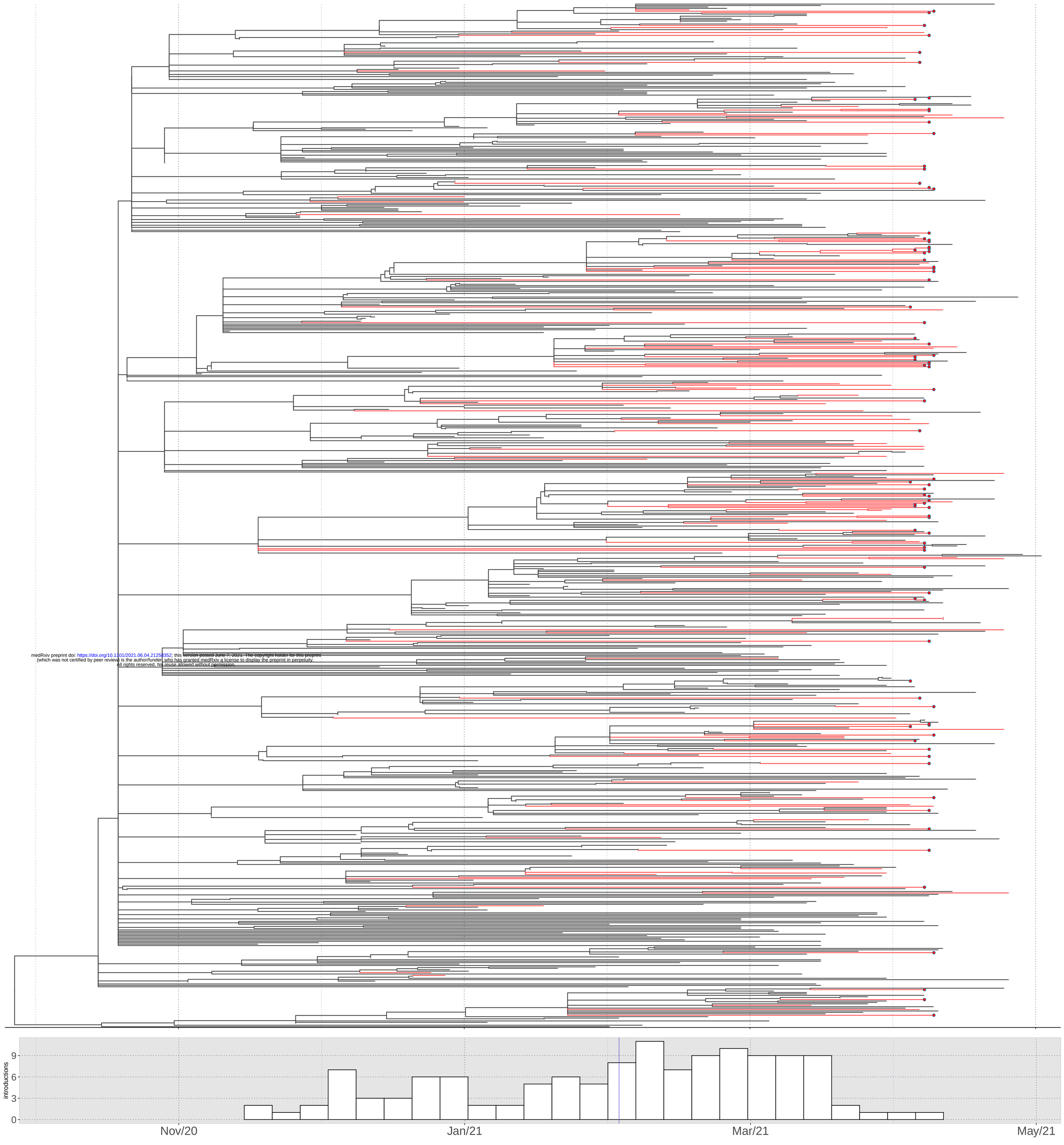
Assembled/consensus genomes are available from GISAID (Shu and McCauley, 2017) subject to minimum quality control criteria. Raw reads are available from European Nucleotide Archive (ENA) in Bioproject PRJEB45462.

References

Abu-Raddad, L. J., Chemaitelly, H. and Butt, A. A. (2021) 'Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants', *New England Journal of Medicine*, p. NEJMc2104974. doi: 10.1056/NEJMc2104974.
Alm, E. *et al.* (2020) 'Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020', *Eurosurveillance*, 25(32), p. 2001410. doi: 10.2807/1560-7917.ES.2020.25.32.2001410.
Artic Network (no date). Available at: https://artic.network/ (Accessed: 6 July 2020).
Baker, D. J. *et al.* (2021) 'CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes', *Genome Medicine*, 13(1), p. 21. doi: 10.1186/s13073-021-00839-5.
de Oliveira Martins, Leonardo (2021) *Uvaia*. Available at: https://github.com/quadram-institute-bioscience/uvaia.
Dong, E., Du, H. and Gardner, L. (2020) 'An interactive web-based dashboard to track COVID-19 in real time', *The Lancet Infectious Diseases*, 20(5), pp. 533–534. doi: 10.1016/S1473-3099(20)30120-1.
Galloway, S. E. *et al.* (2021) 'Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States,

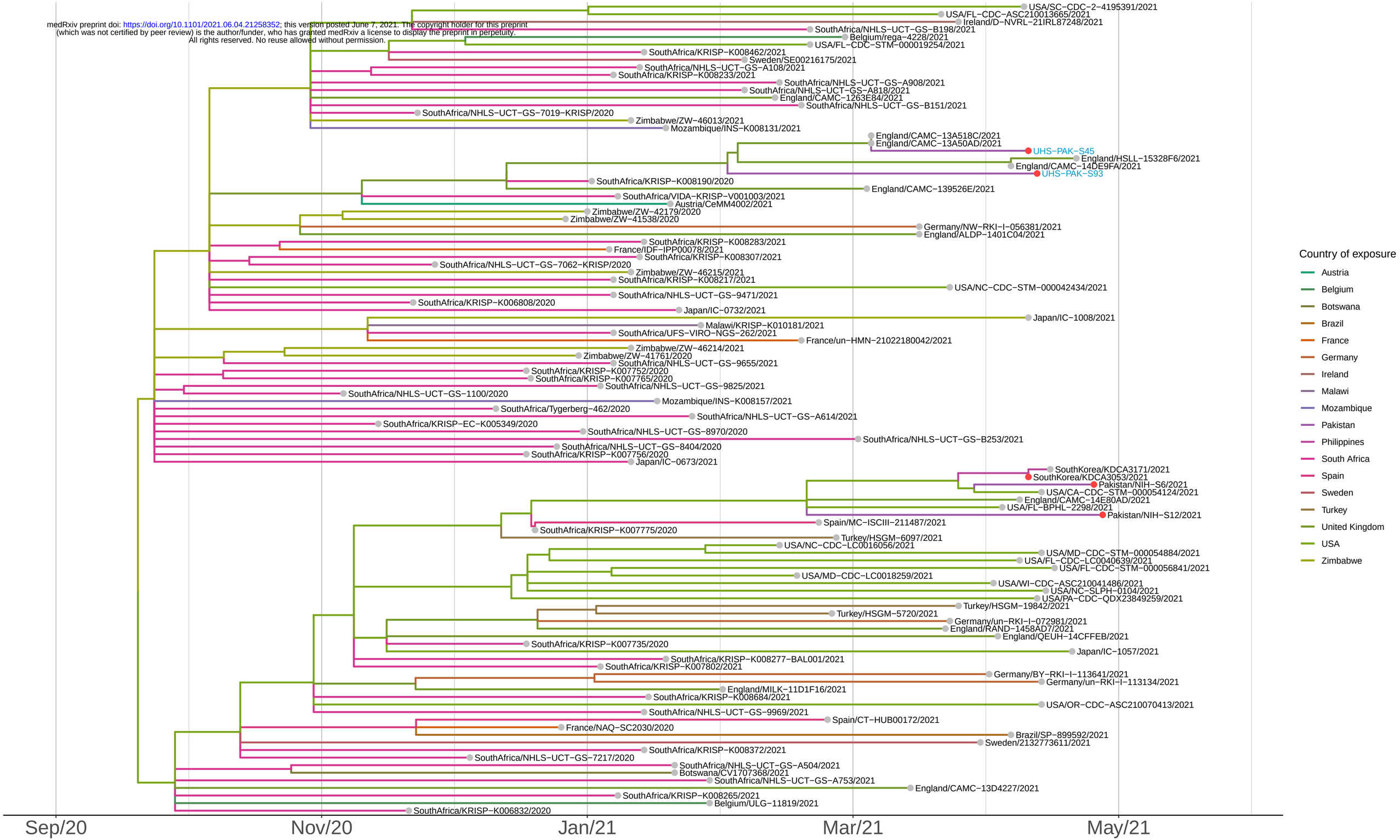
medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

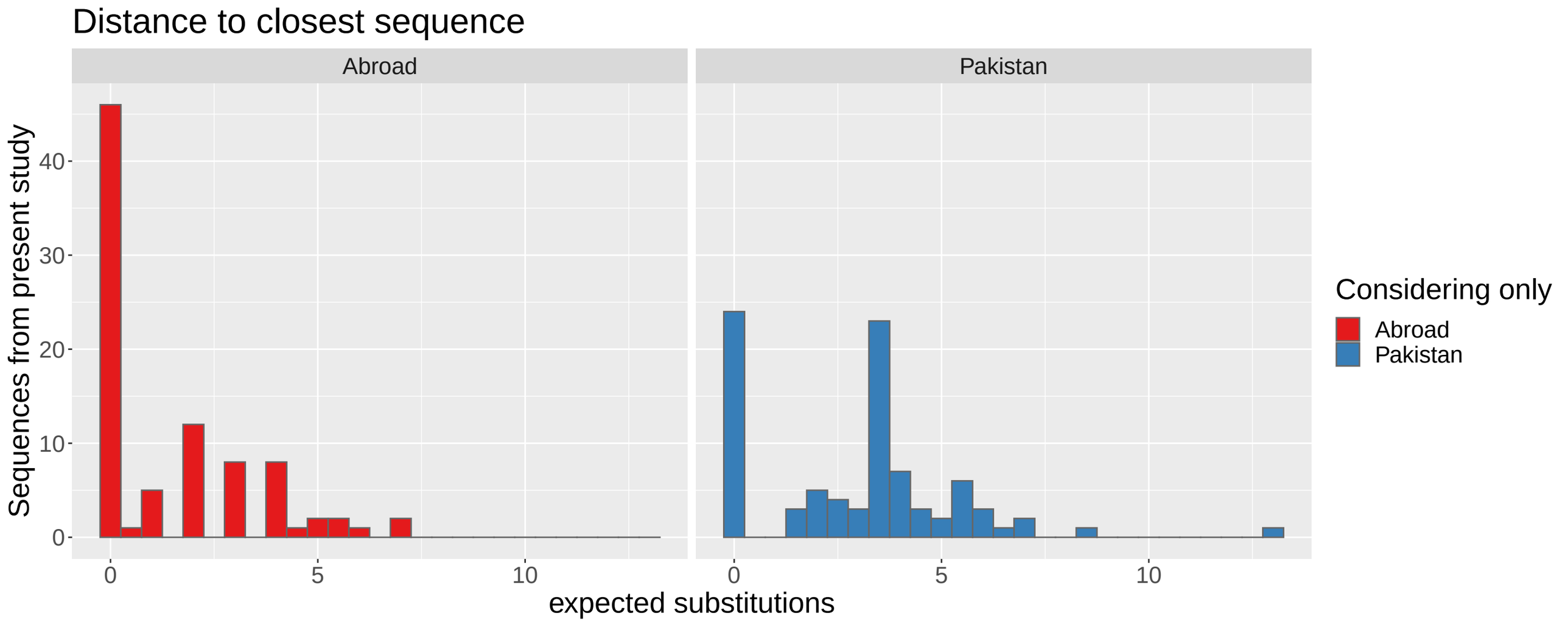
December 29, 2020–January 12, 2021', *MMWR. Morbidity and Mortality Weekly Report*, 70(3), pp. 95–99. doi: 10.15585/mmwr.mm7003e2.
Grubaugh, N. D. *et al.* (2019) 'An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar', *Genome Biology*, 20(1), p. 8. doi: 10.1186/s13059-018-1618-7.
Hodcroft, E. B. *et al.* (2020) *Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org*. Available at: https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming.
Huang, C. *et al.* (2020) 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', *The Lancet*, 395(10223), pp. 497–506. doi: 10.1016/S0140-6736(20)30183-5.
Javed, B. *et al.* (2020) 'Is Pakistan's Response to Coronavirus (SARS-CoV-2) Adequate to Prevent an Outbreak?', *Frontiers in Medicine*, 7, p. 158. doi: 10.3389/fmed.2020.00158.
Krueger, F. (2020) *FelixKrueger/TrimGalore*. Available at: https://github.com/FelixKrueger/TrimGalore (Accessed: 6 July 2020).
Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv:1303.3997 [q-bio]*. Available at: http://arxiv.org/abs/1303.3997 (Accessed: 26 July 2017).
Louca, S. and Doebeli, M. (2018) 'Efficient comparative phylogenetics on large trees', *Bioinformatics*. Edited by A. Valencia, 34(6), pp. 1053–1055. doi: 10.1093/bioinformatics/btx701.
Minh, B. Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*. Edited by E. Teeling, 37(5), pp. 1530–1534. doi: 10.1093/molbev/msaa015.
Minh, B. Q., Klaere, S. and von Haeseler, A. (2006) 'Phylogenetic Diversity within Seconds', *Systematic Biology*. Edited by M. Steel, 55(5), pp. 769–773. doi: 10.1080/10635150600981604.
National Command Operation Center (2021). Available at: https://covid.gov.pk/stats/pakistan (Accessed: 25 May 2021).
Quick, J. (2020) 'nCoV-2019 sequencing protocol v2 v1 (protocols.io.bdp7i5rn)'. doi: 10.17504/protocols.io.bdp7i5rn.
Rambaut, A. *et al.* (2020) 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature Microbiology*, 5(11), pp. 1403–1407. doi: 10.1038/s41564-020-0770-5.
Shu, Y. and McCauley, J. (2017) 'GISAID: Global initiative on sharing all influenza data – from vision to reality', *Eurosurveillance*, 22(13), p. 30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494.
Simonsen, M. and Pedersen, C. N. S. (2011) 'Rapid computation of distance estimators from nucleotide and amino acid alignments', in *Proceedings of the 2011 ACM Symposium on Applied Computing - SAC '11*. *the 2011 ACM Symposium*, TaiChung, Taiwan: ACM Press, p. 89. doi: 10.1145/1982185.1982208.
Tegally, H. *et al.* (2021) 'Detection of a SARS-CoV-2 variant of concern in South Africa', *Nature*, 592(7854), pp. 438–443. doi: 10.1038/s41586-021-03402-9.
Turakhia, Y. *et al.* (2020) 'Stability of SARS-CoV-2 phylogenies', *PLOS Genetics*. Edited by G. S. Barsh, 16(11), p. e1009175. doi: 10.1371/journal.pgen.1009175.
Volz, E. *et al.* (2021) *Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data*. preprint. Infectious Diseases (except HIV/AIDS). doi: 10.1101/2020.12.30.20249034.
Yu, G. *et al.* (2017) 'ggtree : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data', *Methods in Ecology and Evolution*. Edited by G. McInerny, 8(1), pp. 28–36. doi: 10.1111/2041-210X.12628.



medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.04.21258352; this version posted June 7, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.




Distance to closest sequence
Abroad
Pakistan
Sequences from present study
0
10
20
30
40
0
5
10
expected substitutions
Considering only
Abroad
Pakistan