# Comparative Analysis of Emerging B.1.1.7+E484K SARS-CoV-2 Isolates

**Ahmed M. Moustafa,[1,2] Colleen Bianco,[1] Lidiya Denu,[1] Azad Ahmed,[3] Susan E. Coffin,[1] Brandy Neide,[1] John Everett,[4] Shantan Reddy,[4] Emilie Rabut,[5] Jasmine Deseignora,[5] Michael D. Feldman,[6] Kyle G. Rodino,[6] Frederic Bushman,[4] Rebecca M. Harris,[6,7] Josh Chang Mell,[3] and Paul J. Planet[1,7,8]**

[1]Division of Pediatric Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, [2]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, [3]Department of Microbiology and Immunology, Center for Genomic Sciences, Drexel University College of Medicine, Philadelphia, Pennsylvania, USA, [4]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [5]Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, USA, [6]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [7]Department of Pediatrics, Perelman College of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and [8]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York, USA

We report the genome of a B.1.1.7+E484K severe acute respiratory syndrome coronavirus 2 from Southeastern Pennsylvania and compare it with all high-coverage B.1.1.7+E484K genomes (n = 235) available. Analyses showed the existence of at least 4 distinct clades of this variant circulating in the United States and the possibility of at least 59 independent acquisitions of the E484K mutation.

**Keywords.** B.1.1.7+E484K; convergent; vaccine; escape mutation; variant of concern.

During the past 6 months of the pandemic, several variants of concern (VOC), each represented by a constellation of specific mutations thought to enhance viral fitness, have emerged in viral lineages from the United Kingdom (20I/501Y.V1; B.1.1.7), South Africa (20H/501Y.V2; B.1.351), and Brazil (20J/501Y.V3; P.1). These lineages were concerning due to likely increased transmission rates [1–6]. Two of these lineages, B.1.351 and P.1, were of specific concern because they have the mutation E484K, which has been shown to enhance escape from neutralizing antibody inhibition in vitro [7] and may be associated with reduced efficacy of the vaccine [8–11]. In general, viruses from the B.1.1.7 lineage do not have this mutation. However, in February 2021 Public Health England (PHE) published a concerning report of 11 B.1.1.7 genomes that had acquired the E484K spike mutation [12].

Here we report a B.1.1.7 isolate with the E484K spike mutation isolated in Southeastern Pennsylvania (PA). Our laboratory at the Children's Hospital of Philadelphia performed sequencing on randomly selected isolates collected since January 2021. Figure 1A shows the diversity of 114 randomly sequenced genomes. Lineages B.1.1.7, B.1.429 (California), B.1.526 (New York), and R.1 (international lineage with the E484K mutation) accounted for 69% of the sequenced genomes in March. There was a massive increase in lineage B.1.1.7 from 2% (1/47) in February to 42% in March (15/36). Interestingly, 1 B.1.1.7 isolate from this surveillance, collected on 3/24/2021 from a 52-year-old male, carried the E484K spike mutation that is present in the South African and Brazilian lineages.

To better understand the relationship between this isolate and publicly available severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes, we compared it with all available B.1.1.7+E484K high-coverage genomes available on the Global Initiative on Sharing All Influenza Data (GISAID; n = 235) [13]. Since the first report by PHE in February, a total of 253 B.1.1.7+E484K genomes have been uploaded to GISAID from England and 14 other countries (Germany, France, Italy, Poland, Sweden, Ireland, Netherlands, Portugal, Wales, Turkey, Slovakia, Austria, Czech Republic, and the United States) [13].

A temporal plot of the number of global B.1.1.7+E484K isolates collected between December 2020 and March 2021 (2-week window) is shown in Figure 1B (as of 04/17/2021). The first isolate of the 60 US isolates available on GISAID [13] was collected on 02/06/2021 from Oregon. Isolates were also reported from 15 other states (New York, North Carolina, Connecticut, Georgia, New Jersey, Maryland, Florida, West Virginia, California, Pennsylvania, Michigan, Texas, Massachusetts, Washington, and Colorado). Of these isolates, 48% were from Florida (n = 17) and New York (n = 12) and 28% were from New Jersey (n = 7), California (n = 4), and Pennsylvania (n = 6). Two isolates were from Oregon (OR), Connecticut (CT), and Maryland (MD), and single isolates are recorded from Georgia (GA), Texas (TX), Massachusetts (MA), Washington (WA), Colorado (CO), West Virginia (WV), Michigan (MI), and North Carolina (NC). The number of US isolates in March (n = 47 including the PA isolates) was nearly 6 times the number of the isolates reported in February. Since this analysis was completed and at the time of this submission (05/26/2021), there are 1400 B.1.1.7+E484K genomes on GISAID [13], which raises the concern that more B.1.1.7+E484K sequences may be emerging even as herd immunity increases by natural immunity and vaccines.

Received 25 April 2021; editorial decision 29 May 2021; accepted 3 June 2021.

Correspondence: Paul Planet, MD, PhD, Division of Infectious Disease, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Suite 1202, Philadelphia, PA 19104, USA (planetp@chop.edu).

**Open Forum Infectious Diseases®2021**

DOI: 10.1093/ofid/ofab300

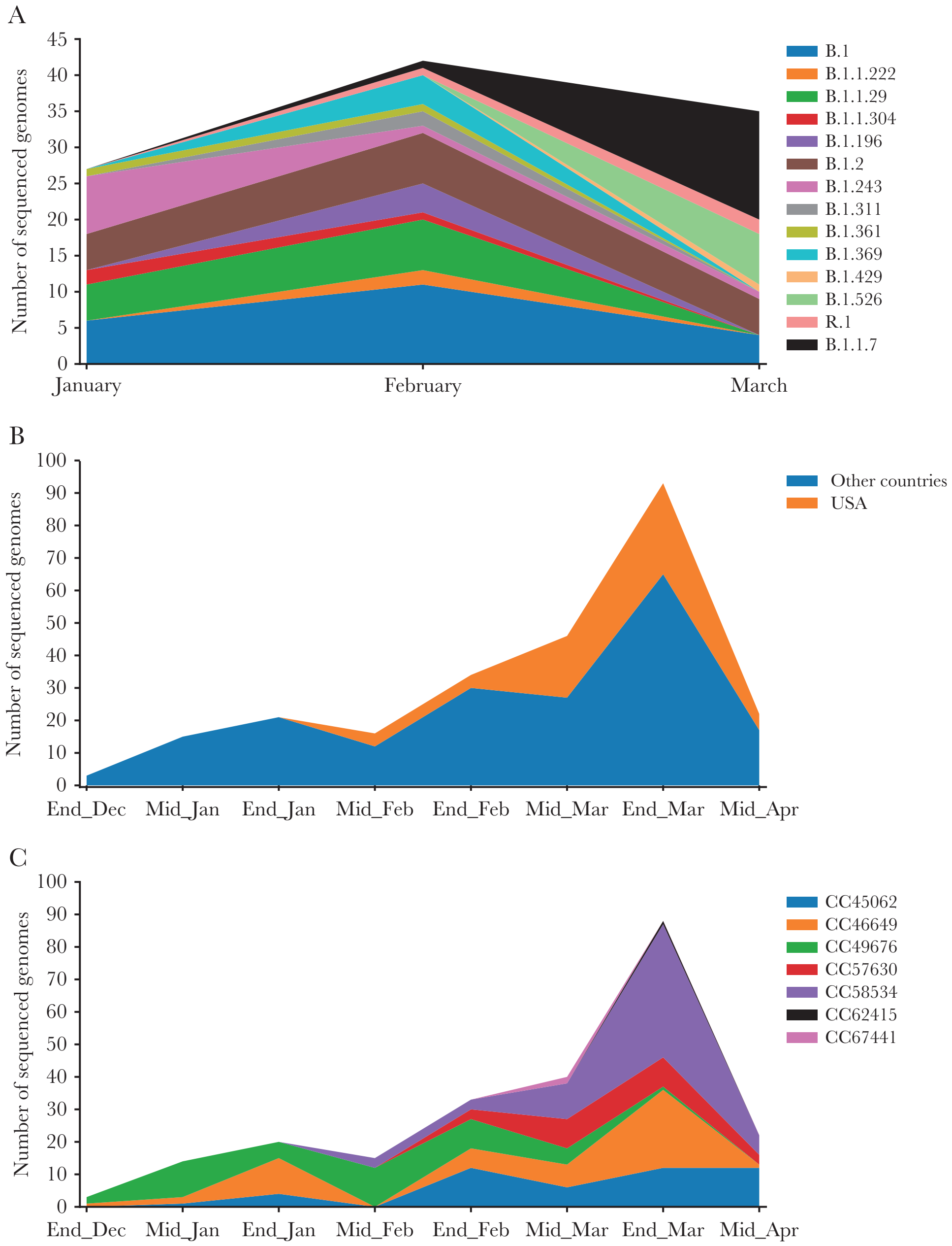

**Figure 1.** Diversity of SARS-CoV-2 in Philadelphia and global diversity of sequenced B.1.1.7+E484K genomes. A, Stacked area plot showing the diversity of random genomes sequenced by our laboratory at Children's Hospital of Philadelphia from January, February, and March 2021. Ten lineages that were represented by only 1 genome (B.1.1, B.1.1.106, B.1.1.129, B.1.1.197, B.1.1.281, B.1.1.296, B.1.119, B.1.234, B.1.350, B.1.409) were excluded from the plot. One B.1.526.1 isolate was counted with the parent B.1.526 for easier visualization. B, Stacked area plot showing the number of GISAID genomes (n = 250) that are B.1.1.7 (20I/501Y.V1) and have the E484K spike mutation over time in the United States and globally. C, Diversity of 236 isolates according to GNUVID. Stacked area plot showing relative abundance of circulating CCs for the 236 B.1.1.7+E484K isolates (typed by GNUVID). The bar plot shows that the isolates belong to 7 different CCs. Isolate EPI_ISL_1385215 was not assigned to any of the 7 CCs (CC255). Fourteen isolates were excluded from the plot as they had >5% nucleotides designated "N" in the sequence. Abbreviations: CCs, clonal complexes; GISAID, Global Initiative on Sharing All Influenza Data; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Although all 236 B.1.1.7+E484K genomes were typed as B.1.1.7 using Pangolin [14], a more granular view using our typing tool "GNUVID" [15] shows that they belong to 7 different clonal complexes (CCs; 45062, 46649, 49676, 57630, 58534, 62415, and 67441) (Figure 1C; Supplementary Table 1). In the GNUVID typing system, these correspond to 7 of 10 CCs in the B.1.1.7 lineage. For each of these CCs, representative sequences without the E484K mutation have been circulating since at least November 2020, predating the first E484K in each CC. This raised the possibility that the E484K mutation was acquired independently in each of these CCs in independent events.

To test the hypothesis of multiple acquisitions in a rigorous phylogenetic framework, we queried the GISAID database for closely related genomes for each of the 236 high-quality

B.1.1.7+E484K genomes and retrieved 354 closely related B.1.1.7 genomes. Using the phylogeny reconstructed from the combined data set, we performed an ancestral reconstruction using Fitch optimization [16], which revealed 59 de novo acquisitions of the E484K mutation in the B.1.1.7 lineage (Figure 2). In the phylogeny, 6 E484K clades were found in >1 country and 4 E484K clades in >1 US state, suggesting widespread dissemination of these viruses. The genome from PA presented here falls in a well-supported clade of 77 B.1.1.7 isolates, 48 of which have the E484K mutation. Seven of the 48 isolates were from the United States (CT, FL, OR, PA, and NY), 33 from Sweden, 4 from Denmark, 2 from Poland, 1 from the Netherlands, and 1 from Germany (Figure 2). The 9 other B.1.1.7+E484K isolates reported from PA were in a large clade containing US genomes (from CA, CO, FL, MA, MD, MI, NC, NE, NJ, NY, OR, TX, WA, WV). This large well-supported clade also contained isolates from England. Another large clade of 75 B.1.1.7 isolates had 34 isolates that carry the E484K mutation; 30 of them are from the United States (CA, CT, FL, MA, NC, NH, NJ, NY, OR). This clade also contained isolates from Germany, Wales, and Ireland. Together these analyses support sustained transmission of these 3 clades within the United States and also globally.

Phylogenetic and single nucleotide polymorphism (SNP) analysis in the 236 isolates compared with the reference MN908947.3 (Supplementary Figures 1 and 2) [17] showed that the isolate presented here had 12/17 of the B.1.1.7-defining SNPs (Supplementary Table 2), while the other Pennsylvanian isolate in the same clade had 17/17 of the SNPs. It also shared

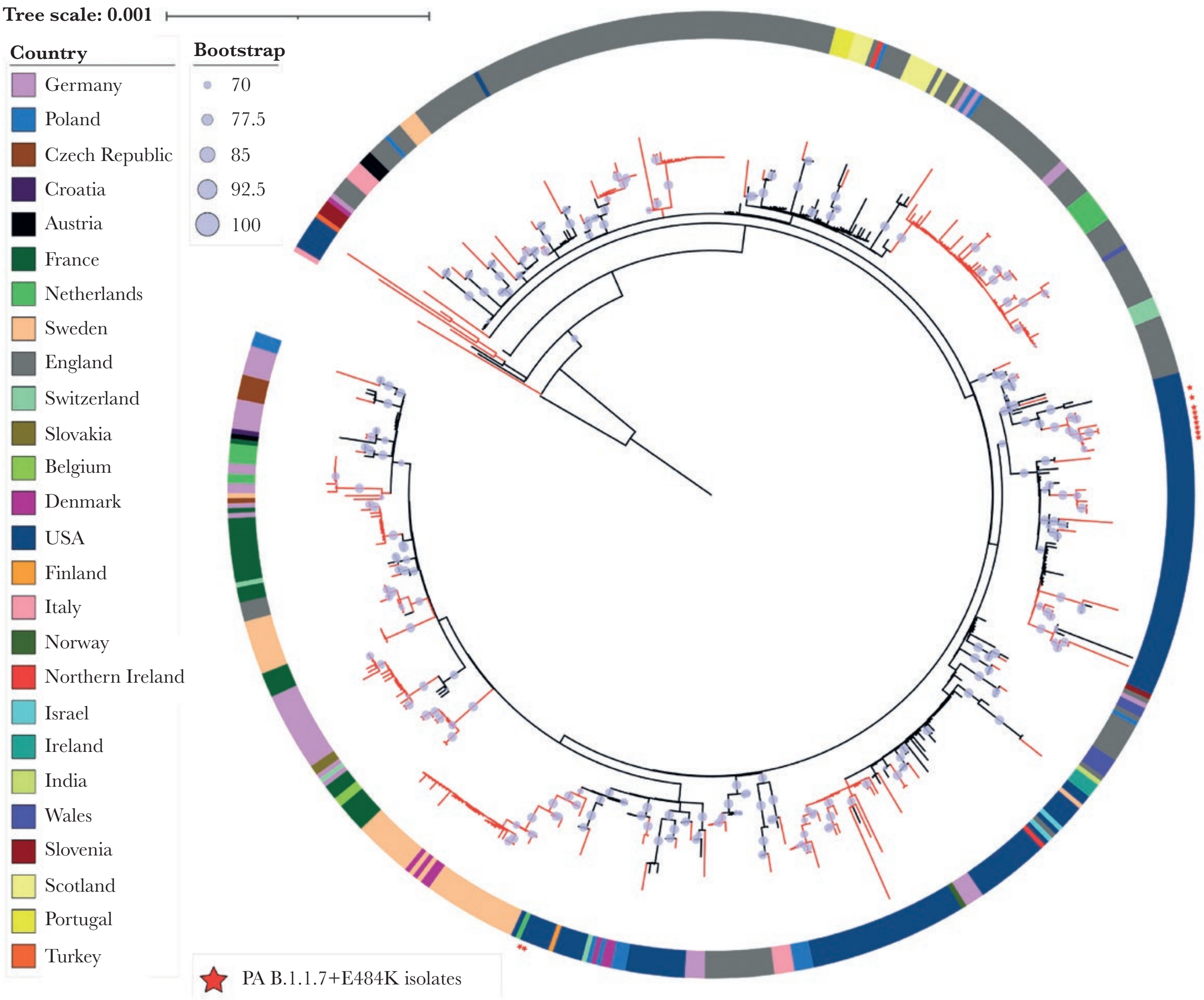

**Figure 2.** SNP-based phylogeny showing the independent acquisitions of E484K in the B.1.1.7 lineage. Maximum likelihood tree of the B.1.1.7+E484K isolates. US isolates are in red. For the CHOP_204 isolate, the alternative allele was called a consensus if its frequency was at least 0.75. The tree was rooted with MN908947.3. The countries of the isolates are shown as a ring. The PA B.1.1.7+E484K isolates are represented with red stars. The red branches represent ancestral reconstruction of the E484K mutation in the B.1.1.7 lineage. Bootstrap values >70 are shown on the branches. Abbreviations: SNP, single nucleotide polymorphism; PA, Pennsylvania.

with 9 other US isolates a stop mutation (A28095T) in ORF8 (Supplementary Figure 2).

Here we present a comparative analysis of the first SARS-CoV-2 B.1.1.7 isolates detected in PA that carry the E484K spike mutation, a mutation that could be associated with reduced efficacy of both vaccine-induced and natural immunity. Our analysis shows that multiple lineages of B.1.1.7+E484K are circulating in the United States and globally and that these lineages have acquired the E484K mutation independently, which argues for strong selective pressure for this mutation.

## METHODS

A nasopharyngeal swab sample that had residual volume after initial laboratory processing, positive PCR testing for SARS-CoV-2, was obtained for this study. RNA was extracted from nasopharyngeal swab samples using the QIAamp Viral RNA Mini (Qiagen). Whole-genome sequencing was done by the Genomics Core Facility at Drexel University. Briefly, whole-genome sequencing of extracted viral RNA was performed as previously described using the Paragon Genomics CleanPlex SARS-CoV-2 Research and Surveillance NGS Panel [18, 19]. Libraries were quantified using the Qubit dsDNA HS (High Sensitivity) Assay Kit (Invitrogen) with the Qubit Fluorometer (Invitrogen). Library quality was assessed using the Agilent High Sensitivity DNA Kit and the 2100 Bioanalyzer instrument (Agilent). Libraries were then normalized to 5 nM and pooled in equimolar concentrations. The resulting pool was quantified again using the Qubit dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) and diluted to a final concentration of 4 nM; libraries were denatured and diluted according to Illumina protocols and loaded on the MiSeq at 10 pM. Paired-end and dual-indexed 2×150-bp sequencing was done using MiSeq Reagent Kits, version 3 (300 cycles). Sequences were demultiplexed and basecalls were converted to FASTQ using bcl2fastq2, version 2.20. The FASTQ reads were then processed to consensus sequence, and variants were identified using the ncov2019-artic-nf pipeline (https://github.com/connor-lab/ncov2019-artic-nf). Briefly, the pipeline uses iVar [20] for primer trimming and consensus sequence-making (options: --ivarFreqThreshold 0.75). A bed file for the Paragon kit primers was used in the pipeline.

All 253 SARS-CoV-2 genomes that were assigned to Pango lineage [14] B.1.1.7 and possessing the E484K spike mutation (including the study isolate CHOP_204) were downloaded from GISAID [13] on 04/17/2021. An acknowledgments table of the submitting laboratories providing the SARS-CoV-2 genomes used in this study is in Supplemental Table 3. Seventeen sequences were excluded for lower coverage (>5% Ns; n = 14) and missing collection date (n = 3). All the high-coverage SARS-CoV-2 genomes (n = 236) were assigned a clonal complex using the GNUVID, version 2.2, database [15]. Temporal plots were plotted using a custom script.

To show the relationship among the genomes of the 236 isolates, a maximum likelihood tree was constructed. Briefly, consensus SARS-CoV-2 sequences for the 236 isolates were aligned to MN908947.3 [17] using MAFFT's FFT-NS-2 algorithm [21] (options: --add --keeplength)). The 5' and 3' untranslated regions were masked in the alignment file using a custom script. A maximum likelihood tree using IQ-TREE 2 [22] was then estimated using the GTR+F+I model of nucleotide substitution [23], default heuristic search options, and ultrafast bootstrapping with 1000 replicates [24]. The tree was rooted to MN908947.3. The snipit tool was then used to summarize the SNPs in the 236 isolates relative to MN908947.3 (https://github.com/aineniamh/snipit). To investigate the number of independent acquisitions of the E484K mutation, a maximum likelihood tree was constructed that has both the B.1.1.7 and B.1.1.7+E484K genomes. Briefly, all available genomes were downloaded from GISAID. A Mash [25] database of all available GISAID SARS-CoV-2 genomes was sketched (options: -i -p 32 -k 32 -s 10000). As there are hundreds of thousands of B.1.1.7 genomes available, we identified the 3 closest genomes to each of the 236 B.1.1.7+E484K genomes by sorting the GISAID genomes using the Mash distance (options: -i -p 32 -d 0.00055). This process identified 354 nonredundant B.1.1.7 and B.1.1.7+E484K genomes close to the 236 study genomes. A maximum likelihood tree of the 590 (354 + 236) genomes was then produced as discussed. The trees were visualized in iTOL [26].

### Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

### Acknowledgments

We would like to thank the Global Initiative on Sharing All Influenza Data (GISAID) and thousands of contributing laboratories for making the genomes publicly available. A full acknowledgments table is available in Supplementary Table 3. We would like to acknowledge the staff members of the Drexel Genomics Core Facility at the Drexel University College of Medicine for processing and sequencing the isolates.

***Financial support.*** P.J.P. and A.M.M. are supported by CDC200-2021-11102 from the Centers for Disease Control and Prevention (CDC) and funding from the Women's Committee of the Children's Hospital of Philadelphia (CHOP).

***Potential conflicts of interest.*** All authors: no reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

***Author contributions.*** A.M.M. and P.J.P. designed and conceptualized the study. A.M.M., B.C., L.D., A.A., B.N., J.E., S.R., E.R., J.D., J.C.M., and P.J.P. contributed to the data collection, data analysis, data interpretation, and whole-genome sequencing. S.E.C., M.D.F., K.G.R., F.B., R.M.H., J.C.M., and P.J.P. supervised the study and contributed to the data collection, data analysis, and data interpretation. A.M.M. and P.J.P. wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

***Patient consent.*** The samples were obtained as part of routine clinical care, solely for nonresearch purposes, carrying minimal risk, and were therefore granted a waiver of informed consent as reviewed under protocol number IRB 21-018726 by the Internal Review Board at the Children's Hospital of Philadelphia.

***Availability of data and material.*** The sequence has been uploaded to GISAID with accession number EPI_ISL_1629709.

## References

1. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science **2021**; eabg3055.
2. Tegally H, Wilkinson E, Giovanetti M, et al. 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 2020.12.21.20248640 [Preprint]. 22 December **2020**. Available at: https://doi.org/10.1101/2020.12.21.20248640. Accessed 25 May 2021.
3. Faria NR, Mellan, TA, Whittaker C, et al. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. Science **2021**; 372:815–21.
4. Rambaut A, Loman N, Pybus O, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. **2020**. Available at: https://virological.org/t/563. Accessed 25 May 2021.
5. Chen RE, Zhang X, Case JB, et al. Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. Nat Med **2021**; 27:717–26.
6. Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: insights from linking epidemiological and genetic data. medRxiv 2020.12.30.20249034 [Preprint]. 4 January **2021**. Available at: https://doi.org/10.1101/2020.12.30.20249034. Accessed 25 May 2021.
7. Weisblum Y, Schmidt F, Zhang F, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife **2020**; 9:e61312.
8. Zhou D, Dejnirattisai W, Supasa P, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell **2021;** 184:2348–61.e6.
9. Garcia-Beltran WF, Lam EC, St Denis K, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell **2021**; 184:2372-2383.e9.
10. Wu K, Werner AP, Koch M, et al. Serum neutralizing activity elicited by mRNA-1273 vaccine. N Engl J Med **2021**; 384:1468–70.
11. Collier DA, De Marco A, Ferreira IATM, et al; CITIID-NIHR BioResource COVID-19 Collaboration; COVID-19 Genomics UK (COG-UK) Consortium. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature **2021**; 593:136–41.
12. Public Health England. Investigation of novel SARS-CoV-2 variant: variant of concern 202012/01 (technical briefing 5). **2021**. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf. Accessed 25 May 2021.
13. Shu Y, McCauley J. GISAID: Global Initiative on Sharing All Influenza Data - from vision to reality. Euro Surveill **2017**; 22:30494.
14. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol **2020**; 5:1403–7.
15. Moustafa AM, Planet PJ. Emerging SARS-CoV-2 diversity revealed by rapid whole genome sequence typing. bioRxiv 2020.12.28.424582 [Preprint]. **2020**. Available at: https://doi.org/10.1101/2020.12.28.424582. Accessed 25 May 2021.
16. Fitch WM. Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool **1971**; 20:406–16.
17. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature **2020**; 579:265–9.
18. Li C, Debruyne DN, Spencer J, et al. Highly sensitive and full-genome interrogation of SARS-CoV-2 using multiplexed PCR enrichment followed by next-generation sequencing. bioRxiv 2020.03.12.988246 [Preprint]. **2020**. Available at: https://doi.org/10.1101/2020.03.12.988246. Accessed 25 May 2021.
19. Pandey U, Yee R, Shen L, et al. High prevalence of SARS-CoV-2 genetic variation and D614G mutation in pediatric patients with COVID-19. Open Forum Infect Dis **2021**; 8:ofaa551.
20. Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol **2019**; 20:8.
21. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **2002**; 30:3059–66.
22. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol **2020**; 37:1530–4.
23. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci **1986**; 17:57–86.
24. Hoang DT, Chernomor O, von Haeseler A, et al. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol **2018**; 35:518–22.
25. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol **2016**; 17:132.
26. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res **2019**; 47:W256–9.