Comparative Analysis of Emerging B.1.1.7+E484K SARS-CoV-2 isolates from Pennsylvania

Ahmed M. Moustafa[1,2], Colleen Bianco[1], Lidiya Denu[1], Azad Ahmed[3], 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], Paul J. Planet[1,7,8*]

1. Division of Pediatric Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

2. Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

3. Department of Microbiology and Immunology, Center for Genomic Sciences, Drexel University College of Medicine, Philadelphia, PA 19129, USA.

4. Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

5. Hospital of the University of Pennsylvania, Philadelphia, PA 19104, USA.

6. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

7. Department of Pediatrics, Perelman College of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

8. Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA.

Phone

PJP: +1 215-590-1169

***Corresponding Author**

Emails

AMM: moustafaam@chop.edu

PJP: planetp@chop.edu

Abstract

Rapid whole genome sequencing of SARS-CoV-2 has presented the ability to detect new emerging variants of concern in near real time. Here we report the genome of a virus isolated in Pennsylvania in March 2021 that was identified as lineage B.1.1.7 (VOC-202012/01) that also harbors the E484K spike mutation, which has been shown to promote “escape” from neutralizing antibodies *in vitro*. We compare this sequence to the only 5 other B.1.1.7+E484K genomes from Pennsylvania, all of which were isolated in mid March. Beginning in February 2021, only a small number (n=60) of isolates with this profile have been detected in the US, and only a total of 253 have been reported globally (first in the UK in December 2020). Comparative genomics of all currently available high coverage B.1.1.7+E484K genomes (n=235) available on GISAID suggested the existence of 7 distinct groups or clonal complexes (CC; as defined by GNUVID) bearing the E484K mutation raising the possibility of 7 independent acquisitions of the E484K spike mutation in each background. Phylogenetic analysis suggested the presence of at least 3 distinct clades of B.1.1.7+E484K circulating in the US, with the Pennsylvanian isolates belonging to two distinct clades. Increased genomic surveillance will be crucial for detection of emerging variants of concern that can escape natural and vaccine induced immunity.

During the past six 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 UK (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 harbor 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 harbor this mutation. However, in February 2021 Public Health England (PHE) published a concerning report of eleven 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, one B.1.1.7 isolate 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 SARS-CoV-2 genomes, we compared it to all available B.1.1.7+E484K high coverage genomes available on GISAID[13] (n=235). 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 USA)[13] (as of 04/17/2021).

A temporal plot of the number of B.1.1.7+E484K isolates collected between December 2020 to March 2021 (2-week window) is shown in **Figure 1B**. The first isolate of the 60 US isolates available on GISAID was collected on 02/06/2021 from Oregon (OR). 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), 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. This increase raises the concern that more B.1.1.7+E484K sequences may be emerging even as herd immunity increases by natural immunity and vaccines.

Although all 236 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 and 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 raises the possibility that the E484K mutation was acquired independently in each of these CCs in independent events.

Phylogenetic analysis of the 235 B.1.1.7+E484K GISAID isolates showed that US isolates are found in at least 3 different clades. The genome presented here falls in a well-supported clade of 28 isolates, 6 of which were from the US (CT, FL, OR, PA and NY), 18 from Sweden, 2 from Poland and 1 from Germany (**Figure 2A**)**.** The only other 4 isolates reported from PA, were in a large clade containing the majority of US genomes, and were located in a well-supported subclade with genomes from the nearby state of West Virginia.

Analysis of SNPs in the 236 isolates compared to the reference MN908947.3[16] (**Figure 2B and Supplementary Figure 1**) 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 with 9 other US isolates a stop mutation (A28095T) in ORF8 (**Figure 2B**).

Here we present a comparative analysis of the first SARS-CoV-2 B.1.1.7 isolates detected in PA that harbor the E484K spike mutation, a mutation that could be associated with reduced efficacy of both vaccine-induced and natural immunity. Our analysis suggests that multiple lineages of B.1.1.7+E484K are circulating in the US, and that these lineages may have acquired E484K independently.

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 QIAamp Viral RNA Mini (Qiagen). Whole genome sequencing was done by The Genomics Core Facility at Drexel University. Briefly, WGS of extracted viral RNA was performed as previously described using Paragon Genomics CleanPlex SARS-CoV-2 Research and Surveillance NGS Panel[17,18]. Libraries were quantified using the Qubit dsDNA HS (High Sensitivity) Assay Kit (Invitrogen) with the Qubit Fluorometer (Invitrogen). Library quality was assessed using Agilent High Sensitivity DNA Kit and the 2100 Bioanalyzer instrument (Agilent). Libraries were then normalized to 5nM and pooled in equimolar concentrations. The resulting pool was quantified again using the Qubit dsDNA HS (High Sensitivity) Assay Kit (Invitrogen) and diluted to a final concentration of 4nM; libraries were denatured and diluted according to Illumina protocols and loaded on the MiSeq at 10pM. Paired-end and dual-indexed 2x150bp sequencing was done using MiSeq Reagent Kits v3 (300 cycles). Sequences were demultiplexed and basecalls were converted to FASTQ using bcl2fastq2 v2.20. 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[19] 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 acknowledgement 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 v2.2 database (version January 6th 2021)[15]. Temporal plots were plotted in GraphPad Prism v7.0a.

To show the relationship amongst 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[16] using MAFFT's FFT-NS-2 algorithm [20] (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[21] was then estimated using the GTR+F+I model of nucleotide substitution[22], default heuristic search options, and ultrafast bootstrapping with 1000 replicates[23]. 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).

The sample was obtained by as part of routine clinical care, solely for non-research purposes, carrying minimal risk, and were therefore granted a waiver of informed consent as reviewed under protocol number under IRB 21-018478.

Availability of data and material

The sequence has been uploaded to GISAID with accession number EPI_ISL_1629709.

Conflict of interest

172 The authors declare that they have no competing interests.

173

174 **Acknowledgements**

175 We would like to thank the Global Initiative on Sharing All Influenza Data (GISAID) and
176 thousands of contributing laboratories for making the genomes publicly available. A full
177 acknowledgements table is available in Supplementary Table 3. We would like to
178 acknowledge the staff members of the Drexel Genomics Core Facility at the Drexel
179 University College of Medicine for processing and sequencing the isolates. P.J.P and
180 A.M.M are supported by 1R01AI137526-01 and 1R21AI144561-01A1 (A.M.M. and
181 P.J.P.), and R01NR015639 (P.J.P.).

182

183

Figure Legends

Figure 1. Diversity of SARS-CoV-2 in Philadelphia and global diversity of sequenced B.1.1.7+E484K genomes. A. Stacked bar plot showing the diversity of random genomes sequenced by our laboratory at Children's Hospital of Philadelphia during January, February and March 2021. Ten lineages that were represented by only one 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 isolate that is B.1.526.1 was counted with the parent B.1.526 for easier visualization. **B.** Bar plot showing number of GISAID genomes (n=250) that are 20I/501Y.V1 and have the E484K spike mutation over time in the US and globally. **C.** Diversity of 236 isolates according to GNUVID. Bar plot showing relative abundance of circulating clonal complexes (CC) 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.

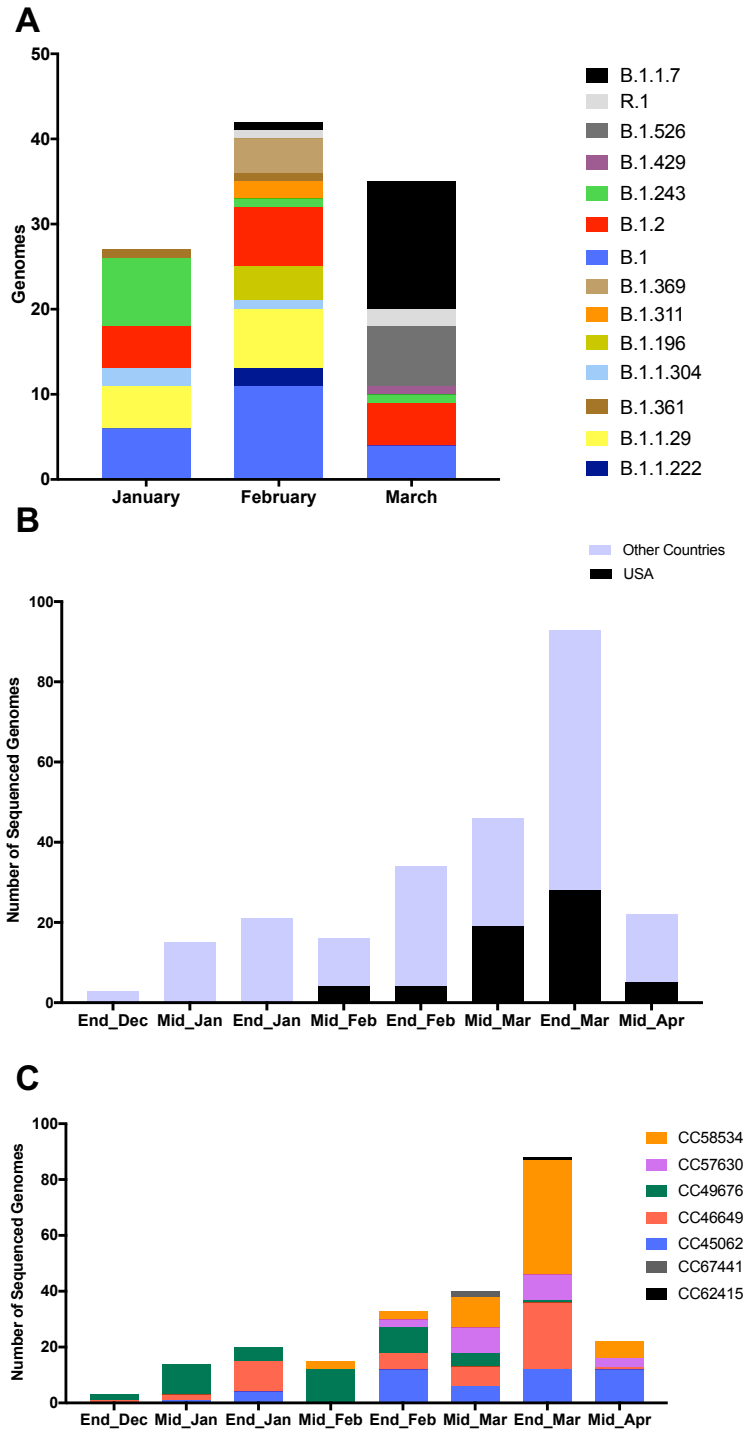
Figure 2. SNP-based Phylogeny and variations of the B.1.1.7+E484K isolates. A. Maximum likelihood tree of the B.1.1.7+E484K isolates. US isolates are in red. For the CHOP_204 isolate the alternative allele was called as consensus if its frequency was at least 0.75. The tree was rooted with MN908947.3. Bootstrap values are shown on the branches. **B.** SNP patterns in the 53 US isolates compared to MN908947.3. SNP variations in the 236 isolates are shown in Supplementary Figure 1. Mutations identified in CHOP_204 are available in Supplementary Table 2. Seven US isolates were excluded from the plot as they had > 5% nucleotides designated "N" in the sequence.

207 An acknowledgement table of the submitting laboratories providing the SARS-CoV-2

208 genomes used in this study is in Supplemental Table 3.

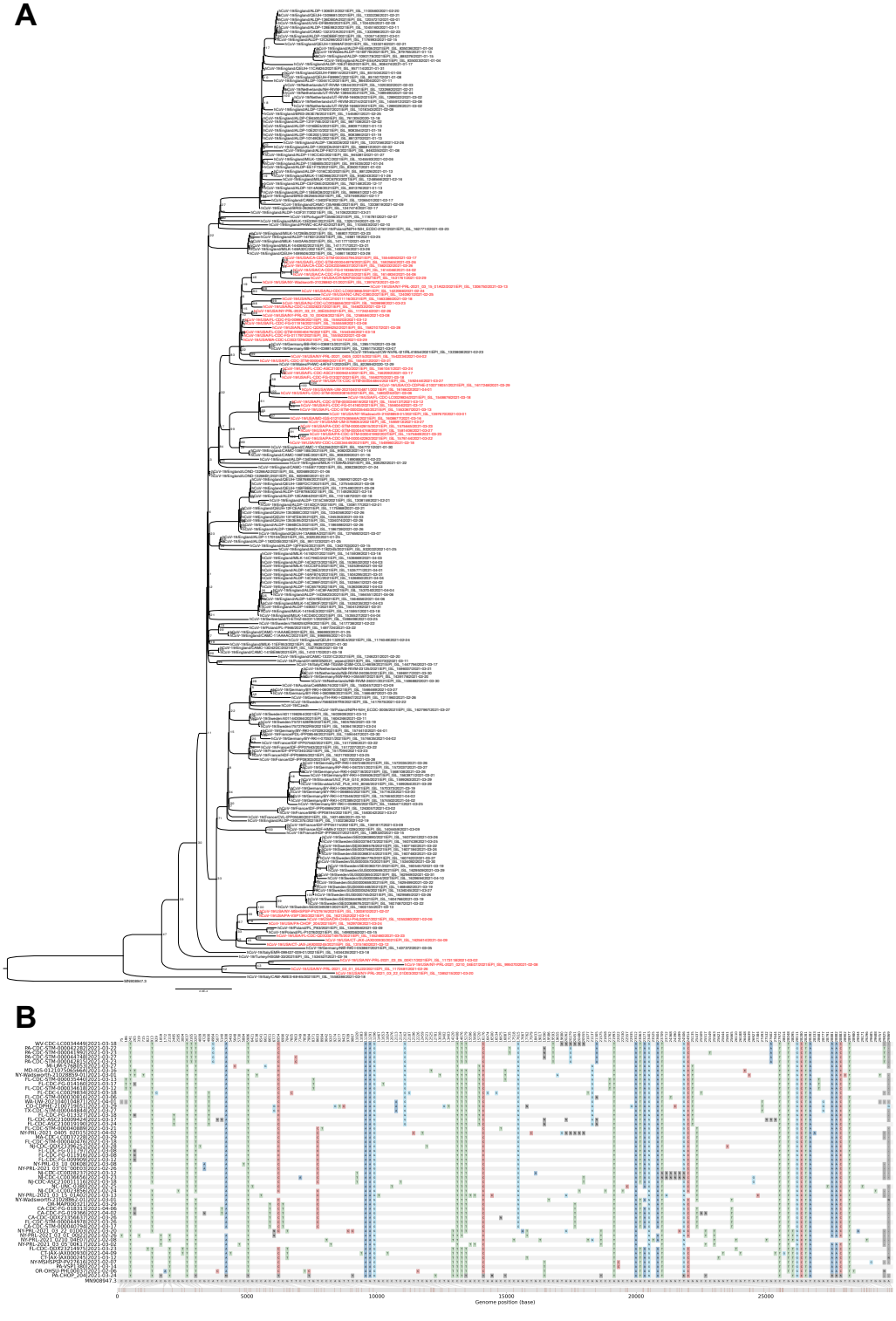
209

210 **Figure 1**



211

212 **Figure 2**



213

Supplementary Table 1. Excel Sheet of GNUVID results for the 236 isolates.

Supplementary Table 2. Mutations and deletions in CHOP_204 compared to MN908947.3.

Mutation	Protein	AA change	Frequency
C241T	-	-	1
C913T	ORF1ab	synonymous	0.92
C1059T	ORF1ab	T265I	0.33
C2110T	ORF1ab	synonymous	0.98
C3037T	ORF1ab	synonymous	1
C3267T	ORF1ab	T1001I	0.64
C4320T	ORF1ab	synonymous	0.38
C5388A	ORF1ab	A1708D	0.65
C5986T	ORF1ab	synonymous	0.62
T6954C	ORF1ab	I2230T	0.74
T7984C	ORF1ab	synonymous	0.65
T9867C	ORF1ab	L3201P	0.33
11288 (del-9)	ORF1ab	SGF3675-77 deletion	0.99
C12781T	ORF1ab	synonymous	0.96
C14120T	ORF1ab	Q4619*	0.95
C14408T	ORF1ab	synonymous	1
C14676T	ORF1ab	P4804L	0.96
C15279T	ORF1ab	T5005I	0.66
T16176C	ORF1ab	L5304P	0.72
A16500C	ORF1ab	K5412T	0.30
C16887T	ORF1ab	synonymous	0.31
C19390T	ORF1ab	synonymous	1
C21575T	S	L5F	0.35
21765 (del6)	S	HV69-70 deletion	0.99
21991 (del3)	S	Y144 deletion	0.98
G23012A	S	E484K	0.77
A23063T	S	N501Y	0.95
C23271A	S	A570D	1
A23403G	S	D614G	1

C23604A	S	P681H	0.98
C23664T	S	A701V	0.41
C23709T	S	T716I	0.99
T24506G	S	S982A	0.55
G24914C	S	D1118H	0.94
C25517T	ORF3a	P42L	0.36
C27972T	ORF8	Q27*	0.93
A28095T	ORF8	K68*	0.93
A28111G	ORF8	Y73C	0.96
A28271 (del1)	-	deletion	0.97
GAT28280CTA	N	D3L	0.97
C28869T	N	P199L	0.38
GGG28881AAC	N	R203K, G204R	0.55
C28977T	N	S235F	0.88
C29137T	N	synonymous	0.54

Supplementary Table 3. GISAID Acknowledgement Table.

Supplementary Figure 1. SNP variations in all available 20I/501Y.V1+E484K isolates.

References

1 Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, eabg3055, doi:10.1126/science.abg3055 (2021).

2 Tegally, H. *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*, 2020.2012.2021.20248640, doi:10.1101/2020.12.21.20248640 (2020).

3 Faria, N. R. *et al.* Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. *medRxiv*, 2021.2002.2026.21252554, doi:10.1101/2021.02.26.21252554 (2021).

4 Rambaut, A.*, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by anovel set of spike mutations.* https://virological.org/t/563 (2020).

5 Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nature Medicine*, doi:10.1038/s41591-021-01294-w (2021).

6 Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*, 2020.2012.2030.20249034, doi:10.1101/2020.12.30.20249034 (2021).

7 Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, e61312, doi:10.7554/eLife.61312 (2020).

8 Zhou, D. *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell*, doi:10.1016/j.cell.2021.02.037.

9 Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, doi:10.1016/j.cell.2021.03.013 (2021).

10 Wu, K. *et al.* Serum Neutralizing Activity Elicited by mRNA-1273 Vaccine. *N Engl J Med*, doi:10.1056/NEJMc2102179 (2021).

11 Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature*, doi:10.1038/s41586-021-03412-7 (2021).

12 *Public Health England. Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 (Technical briefing 5).* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf (2021).

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

14 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*, doi:10.1038/s41564-020-0770-5 (2020).

15 Moustafa, A. M. & Planet, P. J. Emerging SARS-CoV-2 diversity revealed by rapid whole genome sequence typing. *bioRxiv*, doi:10.1101/2020.12.28.424582 (2020).

16 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

17 Li, C. *et al.* Highly sensitive and full-genome interrogation of SARS-CoV-2 using multiplexed PCR enrichment followed by next-generation sequencing. *bioRxiv*, 2020.2003.2012.988246, doi:10.1101/2020.03.12.988246 (2020).

18 Pandey, U. *et al.* High Prevalence of SARS-CoV-2 Genetic Variation and D614G Mutation in Pediatric Patients with COVID-19. *Open Forum Infectious Diseases* (2020 (In Press)).

19 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8, doi:10.1186/s13059-018-1618-7 (2019).

20 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* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).

21 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

22 TavarÈ, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57-86 (1986).

23 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522, doi:10.1093/molbev/msx281 (2018).