

Novel SARS-CoV-2 spike variant identified in Washington, D.C.

1

1 **Novel SARS-CoV-2 spike variant identified through viral genome sequencing of the**  
2 **pediatric Washington D.C. COVID-19 outbreak.**

3

4 Jonathan LoTempio<sup>1,2</sup>, Erik Billings<sup>3</sup>, Kyah Draper<sup>3</sup>, Christal Ralph<sup>3</sup>, Mahdi Moshgriz<sup>3</sup>, Nhat  
5 Duong<sup>3</sup>, Jennifer Dien Bard<sup>4,5</sup>, Xiaowu Gai<sup>4,5</sup>, David Wessel<sup>6,7,8</sup>, Roberta L. DeBiasi<sup>8,10,11</sup>,  
6 Joseph M. Campos<sup>3,8,9,10</sup>, Eric Vilain<sup>1,2</sup>, Meghan Delaney<sup>3,8,10</sup>, Drew G. Michael<sup>1,2,3,8,#</sup>

7

8 1. Children's National Hospital, Center for Genetic Medicine Research

9 2. George Washington University School of Medicine and Health Sciences, Department of  
10 Genomics and Precision Medicine

11 3. Children's National Hospital Division of Pathology and Laboratory Medicine Department

12 4. Children's Hospital of Los Angeles, Department of Pathology and Laboratory Medicine

13 5. Keck School of Medicine, University of Southern California

14 6. Children's National Hospital, Division of Cardiology

15 7. Children's National Hospital, Division of Critical Care Medicine

16 8. George Washington University School of Medicine and Health Sciences, Department of  
17 Pediatrics

18 9. George Washington University School of Medicine and Health Sciences, Department of  
19 Microbiology, Immunology and Tropical Medicine

20 10. George Washington University School of Medicine and Health Sciences, Department of  
21 Pathology

22 11. Children's National Hospital, Division of Pediatric Infectious Diseases

23 # Address correspondence to Drew G. Michael [dgmichael@childrensnational.org](mailto:dgmichael@childrensnational.org)

24

25 Running head: Novel SARS-CoV-2 spike variant identified in Washington, D.C.

26 Abstract word count: 197 | Importance word count: 148 | Main paper word count: 4230

27 **ABSTRACT**

28 The SARS-CoV-2 virus has emerged as a global pandemic, severely impacting everyday life.  
29 Significant resources have been dedicated towards profiling the viral genome in the adult  
30 population. We present an analysis of viral genomes acquired from pediatric patients presenting  
31 to Children's National Hospital in Washington D.C, including 24 with primary SARS CoV2  
32 infection and 3 with Multisystem Inflammatory Syndrome in Children (MIS-C) undergoing  
33 treatment at our facility. Viral genome analysis using next generation sequencing indicated that  
34 approximately 81% of the analyzed strains were of the GH clade, 7% of the cases belonged to  
35 the GR clade, and 12% of the cases belonged to S, V, or G clades. One sample, acquired from  
36 a neonatal patient, presented with the highest viral RNA load of all patients evaluated at our  
37 center. Viral sequencing of this sample identified a SARS-CoV-2 spike variant, S:N679S.  
38 Analysis of data deposited in the GISAID global database of viral sequences shows the  
39 S:N679S variant is present in eight other sequenced samples within the US mid-Atlantic region.  
40 The similarity of the regional sequences suggests transmission and persistence of the SARS-  
41 CoV-2 variant within the Capitol region, raising the importance of increasing the frequency of  
42 SARS-CoV-2 genomic surveillance.

43

44 **IMPORTANCE** A variant in the SARS-CoV-2 spike protein was identified in a febrile neonate  
45 who was hospitalized with COVID-19. This patient exhibited the highest viral RNA load of any  
46 COVID-19 patient tested at our center. Viral sequencing identified a spike protein variant,  
47 S:N679S, which is proximal to the cleavage site at residue 681. The SARS-CoV-2 surface spike  
48 is a protein trimer (three subunits) which serves as the key target for antibody therapies and  
49 vaccine development. Study of viral sequences from the GISAID database revealed eight  
50 related sequences from the US mid-Atlantic region. The identification of this variant in a very  
51 young patient, its critical location in the spike polyprotein, and the evidence that it has been  
52 detected in other patients in our region underscores the need for increased viral sequencing to

53 monitor variant prevalence and emergence, which may have a direct impact on recommended  
54 public health measures and vaccination strategies.

55

## 56 **INTRODUCTION**

57 SARS-CoV-2, a positive-sense single-stranded RNA virus, is the causative agent of the ongoing  
58 COVID-19 pandemic (1–3). Reports in early 2020 suggested that children were spared the  
59 harshest manifestations of disease, with the majority of patients reported as asymptomatic (4).  
60 The first wave of outbreaks across Europe and the Americas demonstrated that this was not the  
61 case (5, 6), and in addition to the classical array of COVID-19 symptomatology, children were  
62 shown to be susceptible to a novel disease presentation, Multi-system Inflammatory Syndrome  
63 in Children (MIS-C) (7, 8).

64

65 Children’s National Hospital (CNH) in Washington, D.C., has evaluated and treated large  
66 numbers of pediatric and young adult patients with both primary SARS-CoV2 infection as well  
67 as MIS-C. The volume of patients seeking care has mirrored trends seen on the US East  
68 Coast, with an early spring 2020 peak, summer plateau, and an ongoing wave of infections  
69 which began in late November. As of January 12th, 2021, at least 821 Washingtonians have  
70 died of COVID-19, representing over 0.1% of the District of Columbia (D.C.) population (9),  
71 although none of these deaths have occurred in children at CNH, despite hundreds requiring  
72 hospitalization. Accordingly, our institution has allocated significant resources to understand the  
73 multi-factorial nature of SARS-CoV-2 primary infection and MIS-C.

74

75 The diagnosis of suspected COVID-19 patients at CNH, which includes D.C., Maryland,  
76 Virginia, West Virginia, and Delaware in its hospital catchment area, is routinely performed  
77 using semi-quantitative RT-PCR commercial platforms with EUA approved tests designed to  
78 assess multiple loci within the SARS-CoV-2 genome. When a test returns as positive, that

Novel SARS-CoV-2 spike variant identified in Washington, D.C.

4

79 patient enters treatment following a protocol designed to optimize care and resources. For  
80 selected samples, virus samples underwent viral sequencing to characterize the nature of the  
81 viral strains in our region.

82

83 Analysis of one viral isolate from a neonatal patient with a high viral RNA load resulted in the  
84 identification of a novel spike protein variant representing the earliest known sample of an  
85 emergent SARS-CoV-2 lineage in the US mid-Atlantic region. As the vaccine rollout for SARS-  
86 CoV-2 continues, the identification of this variant in the viral spike protein highlights the need for  
87 consistent, comprehensive molecular monitoring, transparency in reporting, and rapid response  
88 at all levels of healthcare and pandemic response systems.

89

## 90 **RESULTS**

91

92 **Genomic profiling of pediatric COVID-19 SARS-CoV-2 virus in the Washington D.C. area**  
93 **identifies the GH clade as the predominant local viral clade.** Seventy-six samples from  
94 COVID-19 patients presenting at CNH were selected for viral sequencing. Of these, twenty-  
95 seven samples met genetic sequencing coverage criteria for further analysis (>1,000 mean  
96 coverage), allowing for consensus sequences to be built and used for phylogenetic and variant  
97 analysis.

98

99 Viral sequences were analyzed via multiple sequence alignment (MSA) using ClustalW  
100 alongside representative samples from each viral clade as annotated in the Global Initiative on  
101 Sharing All Influenza Data (GISAID) repository and summarized in **S1**. The phylogenetic  
102 relationship between our viral samples **Fig. 1A**, allows for comparison to samples deposited  
103 from other laboratories in the D.C. area **Fig. 1B**. These data indicate that the pediatric COVID-  
104 19 outbreak within the D.C. area followed trends observed in the adult population, as

105 established by the clades of viruses deposited in the Global Initiative on Sharing All Influenza  
106 Data (GISAID) database and annotated for our geographic locale (**Fig. 1B**). This finding  
107 supports the hypothesis that the viral strains propagating in adults are similar to those in  
108 children.

109

110 The D.C. outbreak mirrored the New York State (NYS) outbreak in terms of early predominance  
111 of sequences assigned to the GH clade. Genetic differences in outbreaks over time can be seen  
112 when examining key US West Coast outbreaks, as well as the early Chinese outbreak, where  
113 different clades of the virus reached equilibrium as the pandemic progressed (**Fig. 1B**). Analysis  
114 of the distribution of deposited viral sequences over time allows for the coarse analysis of  
115 predominant strains in each area. In the D.C. area, along with NYS, Washington State, and  
116 California. The ratio of clades present in each location changes with an increase in deposited  
117 viral sequences, demonstrating the evolving and dynamic nature of the SARS-CoV-2 variant  
118 profile across the pandemic.

119

120 To assess the possible geographic origin of viral strains propagating locally, we utilized the  
121 Phylogenetic Assignment of Named Global Outbreak LINEages, or PANGOLIN tool. PANGOLIN  
122 leverages data deposited at GISAID to assign lineages and find near global genetic neighbors  
123 based on their internal variant-based lineage assignment, which allows for sequence level  
124 analysis of global relationships. The countries which had the highest number of sequence  
125 submissions with the same lineages as our 27 samples were the US, United Kingdom (UK), and  
126 France, with minor relationships to sequences deposited from Portugal, Australia, Israel, and  
127 New Zealand (**Fig. 2**).

128

129 The results point toward a European origin for the virus propagating in the US Capitol region  
130 pediatric population. The observation of the US and UK as major locations of genetic similarity

131 is consistent with domestic population flux across US state borders, which remain porous, as  
132 well as international population flux, which was restricted with China on January 31, 2020, but  
133 unrestricted with Europe generally until March 11, 2020 (10, 11). Considering this, the  
134 incomplete nature of early travel restrictions does not appear to have had a protective effect on  
135 the US population. The UK is an outsized-GISAID contributor, responsible for sharing over  
136 104,000 of the 263,000 high-quality, complete genomes in GISAID as of January 12,  
137 contributing to its weight in comparison to the samples sequenced in D.C. The third highest  
138 genetic contributor, France, was likely an early contributor to our outbreak in the D.C. area as  
139 determined by the presence of variants specific to the GH lineage samples found both in the  
140 Washington Metro region and first identified in France in February 2020 (DeBiasi et al,  
141 unpublished data under review). One limitation of the use of the PANGOLIN tool is that, since it  
142 is tied to the ever-expanding GISAID resource, relationships over time may change based on  
143 newly-deposited viral sequences.

144

145 **Viral variant profile of three pediatric MIS-C patients highlights the complexity of COVID-**  
146 **19 genotype-phenotype associations.** To attempt correlation of SARS-CoV-2 variants with  
147 phenotypic outcomes, we identified consensus variants within the viral genome and overlaid the  
148 distribution of viral genetic variants with patient MIS-C clinical status. These analyses identified  
149 multiple variants present within each viral sample. Of note, 92% (25/27) of the viral genomes  
150 within the sampled cases contained the S:D614G spike variant associated with increased viral  
151 transmissibility (12).

152

153 Interestingly, five patients were identified with identical viral variant profiles. Four of these  
154 patients presented with primary COVID-19 disease and one presented with MIS-C. These  
155 samples are distinguished from other sequences assigned to the GH clade by the presence of  
156 variants in both the nucleoprotein (N:S193I) and non-structural protein 2 (NSP2:T371I), (**Fig. 3**).

Novel SARS-CoV-2 spike variant identified in Washington, D.C.

7

157 In the GISAID database, 1198 sequences have N:S193I, while 277 have NSP2:T371I - only  
158 seven have both (GISAID accessions: EPI\_ISL\_516718, EPI\_ISL\_683413, EPI\_ISL\_745300,  
159 EPI\_ISL\_676627, EPI\_ISL\_683422, EPI\_ISL\_710051, EPI\_ISL\_683418). Six of those seven  
160 were generated from samples originating in Texas in late July, while the final sequence was  
161 generated from a sample originating in Virginia collected on April 30, 2020.

162

163 All patients presenting with MIS-C shared two GH clade-specific variants (NSP2:T85I and  
164 ORF3a:Q57H) with other samples from patients who presented with primary COVID-19. One,  
165 which was identical to other previously-discussed samples, had nucleocapsid variants N:S193I  
166 and NSP2:T371I. The second had the additional variant NSP1:L16I, while the third had  
167 NSP3:T424I and NSP8:A45V.

168

169 The observation of five identical viral genomic profiles with different disease outcomes, and two  
170 clinically similar MIS-C cases with different viral genotypes highlights the difficulties inherent in  
171 forming viral genotype-human phenotype correlations in this disease (**Fig. 3**). This observation  
172 points toward the importance of additional variables, such as initial viral dosage, environmental  
173 factors, and host genetic predispositions as key elements of COVID-19 progression and  
174 possibly underlying MIS-C presentation.

175

176 **High viral RNA load on presentation in neonatal patient leads to identification of novel**  
177 **spike protein S:N679S variant.** In early September 2020, a febrile neonatal patient was  
178 hospitalized at CNH with COVID-19 symptoms. SARS-CoV-2 RT-PCR testing detected a high  
179 diagnostic viral load as measured by the polymerase chain reaction (PCR) cycle threshold (Ct)  
180 of 6.45 using an assay which targets ORF (Ct 6.4) and spike (Ct 6.5) (Simplexa COVID-19  
181 Direct, Diasorin, CA, USA). In comparison, we observed a median Ct of 22.1 for both genes  
182 during 499 previous positive tests. The observed Ct for this patient represented an approximate

## Novel SARS-CoV-2 spike variant identified in Washington, D.C.

8

183 51,418-fold increase in viral load compared to the median presenting viral load in previously  
184 seen patients. Supplemental testing of this patient's sample on a different platform yielded a Ct  
185 of 6.5 (Allplex™ 2019-nCoV Assay, Seegene, S Korea) confirming the extremely high viral load.  
186 As this patient represented an outlier in viral load, we selected this sample for NGS profiling of  
187 the SARS-CoV-2 genome.

188

189 Short read sequencing analysis of the SARS-CoV-2 genome identified a novel variant which  
190 resulted in the nonsynonymous amino acid substitution S:N679S, which is responsible for viral  
191 entry. The presence of >9,000 individual reads was strong support for a spike protein variant,  
192 rather than a sequencing error or artifact of PCR (**Fig. 4B**). In order to confirm with an  
193 orthogonal sequencing technology, the sample was sent for confirmation of the variant by  
194 Sanger sequencing, which confirmed the presence of the S:N679S variant (**Fig. 4C**). The  
195 Sanger data analysis also confirmed the presence of the S:D614G variant in the viral genome,  
196 which is present in the majority of global samples due in part to its hypothesized greater  
197 infective potential (12). This is significant, as association of the S:N679S with S:D614G may  
198 contribute to the persistence of the S:N679S variant.

199

200 To assess where the S:N679S variant is present in the community, we queried the GISAID  
201 database. At the time of the initial query, the GISAID database contained six high-quality  
202 complete genomes containing the S:N679S variant. All six of these genomes were deposited by  
203 labs in Maryland and Virginia. In mid-December, 2020 re-query identified an additional four  
204 samples from Australia and Japan, with a third re-query revealing a sequence from Brazil.  
205 Finally, a fourth query on January 12, 2021 revealed two more high-quality sequences from  
206 Delaware. Phylogenetic analysis (**Fig. 5A**) showed that this novel spike variant had emerged in  
207 two distinct viral clades which are geographically and genetically independent of each other.  
208 Variant profiles and lineage assignments suggest the singleton samples from Brazil and Japan,



209 and the samples from Australia, while all in clade GR, are genetically distinct in lineages  
210 B.1.1.33, B.1.1.284, and B.1.1.162, respectively. Conversely, all samples collected in the US  
211 were assigned to the lineage B.1.189 by PANGOLIN, further supporting their similarity within the  
212 large G clade.

213

214 To probe the similarity between the reported samples of US origin, we constructed a maximum-  
215 likelihood (ML) phylogeny with international samples for context (**Fig. 5A**). The tree topology  
216 among US mid-Atlantic regional samples is consistent with time-of-sampling metadata, which is  
217 available for all US samples. The patient treated at CNH, who exists in a branch distinct from  
218 the other six samples from the D.C. area, was sampled on 17 days before the next earliest  
219 sample on September 23rd. Both samples from Maryland collected on October 20, as well as  
220 one from October 13 are genetically identical with branch length zero. It is remarkable that  
221 samples collected on the 20th and 21st of October in Delaware share a common ancestor  
222 predating the samples found locally (including our sample from September 7, 2020). This  
223 suggests both the persistence of the variant, and earlier emergence than was previously  
224 appreciated with samples from the District of Columbia, Maryland, and Virginia.

225

226 **Spike protein residue 679 is proximal to the S1/S2 cleavage site and may impact function.**

227 We queried GISAID for all amino acid substitutions at spike polyprotein position 679, the results  
228 of which are presented in **S2 a**. The most common variants observed so far include substitution  
229 with lysine or deletion, both of which represent 703 of 731 variants described in all human  
230 sequences in GISAID. Substitution with serine, the variant observed at CNH, represented 16  
231 human sequences as of January 12, 2021. Five substitutions to histidine, two to isoleucine, two  
232 to tyrosine, and one to threonine were also observed. There are over 370,000 sequences  
233 containing the wild-type asparagine.

234

235 The relatively high number of lysine variants can be partially attributed to codon wobble - both  
236 have tRNA three letter codes AAX, with asparagine coded by AAU and AAC, while lysine is  
237 coded by AAA and AAG. The serine, threonine, and tyrosine require transition or transversion at  
238 the second position of the codon, while transversion of the first codon position is required to  
239 result in histidine. A multifasta (**S2\_b**) with each of these amino acid substitutions, and a  
240 deletion of the residue at 679 was constructed and analyzed with the ProP tool for predicting  
241 cleavage sites (13). With the exception of the S:N679Y variant, each substitution scored higher  
242 than the wild type threshold of 0.62. Higher ProP scores indicate a higher likelihood the site will  
243 function as a site for cleavage. This suggests our observed S:N679S substitution at this residue  
244 does not preclude the variant protein from activation. A robust and rapid pipeline for functional  
245 classification of SARS-CoV-2 variants will be an important public health tool as the global  
246 community continues to navigate the COVID-19 pandemic.

247

248 In 2011, it was shown that SARS-CoV spike protein is cleaved and activated by human airway  
249 trypsin-like protease (HAT)(14). Furthermore, it has been suggested that the insertion of the  
250 RRAR amino acid sequence at the S1/S2 cleavage site in SARS-CoV-2, relative to SARS-CoV,  
251 represents a gain of function as a target of human peptidase furin, which is known to cleave  
252 basic amino acids at the motif R-X-[R/K]-R. This motif is indeed observed between S1 and S2  
253 spike subunits in SARS-CoV-2. *In silico* and benchtop experimental evidence has been  
254 presented for SARS-CoV-2 variants with and without this motif showing SARS-CoV-2 variants  
255 lacking the RRAR motif are unable to fuse with HEK293 cells in the absence of trypsin or HAT,  
256 while the wild type SARS-CoV-2 spike can undergo cellular fusion without HAT or trypsin (15).

257

258 This is strong evidence for a cell fusion gain of function through the RRAR protein motif – but  
259 not for an increase in virulence or infectivity. Indeed, SARS-CoV-2 variants where the furin  
260 cleavage site has been deleted show increased infectivity (16, 17). Whether preservation of the

Novel SARS-CoV-2 spike variant identified in Washington, D.C.

11

261 furin cleavage site and its gain of function for cell fusion relative to SARS-CoV is outweighed by  
262 the increase in infectivity in its absence (as in SARS-CoV-2) should be functionally assessed in  
263 cell and animal experiments.

264

265 **Analysis of global SARS-CoV-2 data provides evidence for mid-Atlantic circulation of the**  
266 **spike variant S:N679S.** As of January 12, 2021, there are 161 sequences in GISAID which are  
267 annotated as B.1.189 (including low quality sequences) which represent the nearest genetic  
268 neighbors of our sample of interest. Of these sequences, 149 are from the US, nine are from  
269 Mexico, two are from Canada, and one is from Japan. The earliest available sequences were  
270 from samples collected in March in California and Mexico City. Two samples from Manitoba,  
271 Canada, were collected in May, and the single sample from Japan was collected in July.  
272 Examination of the temporal and geographic distribution of this lineage further suggests  
273 circulation in North America, from coast to coast, and across the Mexican-American, and  
274 Canadian-American borders (**Fig. 5 B**). Variant analysis with NextClade (18) shows only two  
275 silent mutations present inside and outside the D.C. area - A20268G and C4582T - suggesting  
276 the possibility of many missing representatives of the lineage.

277

278 The S:N679S-containing sample detected at CNH has silent mutation C2710T, which is absent  
279 in all other local samples. All D.C. area samples contain two missense mutations,  
280 ORF1a:A2994V and ORF1a:P3359S, as well as silent mutation T11408C. The difference  
281 between these samples suggests there is likely an earlier case which has the additional variants  
282 common to all seven samples including missense mutations ORF1a:K564Q, ORF1a:P2018L,  
283 ORF1a:S3885F, ORF1b:H2583Y, and ORF3a:S177I, as well as silent mutations C4276T,  
284 T6160C, C16293T, C16887T, C26222T. This predicted most recent common ancestor, likely  
285 appeared in summer 2020. Additional sequencing of archived samples will be required to  
286 assess this hypothesis.

287

288 When considering that there are two high-quality sequences in the database are from samples  
289 acquired in Delaware, the timeline can be revised to include longer circulation and wider spread.  
290 Those samples contain the coding variants ORF1a:S3885F, ORF1b:P314L, ORF1b:H2583Y,  
291 ORF3a:S177I, S:D614G, S:N679S, which are common to the samples from the D.C. area. One  
292 sample, EPI\_ISL\_803105 collected on October 20, also has ORF1a:K564Q which is common to  
293 the D.C. region samples, while the other, EPI\_ISL\_803106 collected one day later, does not  
294 have that variant which possibly represents a loss of the variant through back mutation to the  
295 reference nucleotide, A, at position 1955. These sequences also contain further coding variants  
296 ORF10:Q29\* and ORF1a:C1114F, but lack all other ORF1a variants.

297

298 Taken together, these data support the hypothesis wherein a SARS-CoV-2 virus containing the  
299 variant S:N679S is circulating within the US mid-Atlantic region. Based on these analyses, the  
300 ancestral state is hypothesized to include the following variants: ORF1a:K564Q,  
301 ORF1a:S3885F, ORF1b:P314L, ORF1b:H2583Y, ORF3a:S177I, S:D614G, S:N679S. All other  
302 variants are lineage specific. All variant profiles of these sequences can be found in S1.

303

304 **Low sequencing rates in the US hinder surveillance of emerging viral variation.** As of  
305 January 12, the US has passed 25 million reported cases and 433,000 reported deaths; the  
306 range of excess deaths in the US reported by the CDC and attributed to COVID-19 was given a  
307 lower bound of 346,750 and an upper bound of 469,831 (19). Concurrently, approximately  
308 71,000 sequences deposited to GISAID are from the US, representing ~0.3% of confirmed  
309 cases. We leverage data from Australia (59.6% of confirmed cases, or 17,081/28,650 cases  
310 sequenced), Japan (3.3% or 9,885/298,000 cases), and Brazil (0.02% or 2,102/8.2m cases).  
311 This differs from the UK, the world's largest COVID-19 viral genome contributor with 157,626

## Novel SARS-CoV-2 spike variant identified in Washington, D.C.

13

312 sequences deposited from 3.16 million cases (~4.99%). All data are based on GISAID  
313 submissions and the Johns Hopkins' interactive dashboard (20).

314

315 Our analysis of the S:N679S containing viral genome from sample CNH-B-40 began in early  
316 December; other local high-quality sequences with the variant had been deposited just weeks  
317 prior. In the intervening four weeks, the sequences from Australia, Japan, and Brazil were  
318 deposited, along with another low-quality sequence from Delaware in the US. This brings the  
319 total number of high-quality sequences from human samples in GISAID with S:N679S to twelve,  
320 including CNH-B-40 which is yet to be assigned a GISAID accession number.

321

322 The highly-related samples with S:N679S were all collected over a large, interconnected 4-  
323 state/district geographic area in a relatively short 7-week timeframe with no established person-  
324 to-person transmission network. Given the high likelihood of community circulation of this  
325 variant of SARS-CoV-2, it is possible that additional cases of this variant strain are present in  
326 the mid-Atlantic region but have yet to be detected due to low rates of viral sequencing and  
327 deposition to GISAID.

328

329 As of January 12, Delaware has approximately 67,000 confirmed cases and 345 sequences  
330 deposited (0.5%), Maryland has 315,000 confirmed cases and 802 sequences deposited  
331 (0.25%). D.C. has 33,000 confirmed cases and 130 sequences deposited (plus our 27) (0.4%,  
332 0.47% combined), and Virginia has 413,000 confirmed cases and 1,685 sequences deposited  
333 (0.4%). Taken together, in states from this region, there are at least 828,000 cases, but only  
334 2,962 sequences. This represents a sequencing rate of 0.36%, just over the national average of  
335 0.3%, but considerably lower than the 4.99% of samples sequenced in the UK, where a critical  
336 new variant was first identified (21, 22).

337

338 **DISCUSSION**

339 A key goal of precision medicine is to target care to those who need it most. An ideal COVID-19  
340 diagnostic and triage algorithm would integrate features such as viral genomic profile, host  
341 innate errors in immunity and viral load over the course of infection to inform care. Toward this  
342 end, we established systems to profile the SARS-CoV-2 genome and link viral genotype to  
343 pediatric disease outcomes. This report contains the initial analysis of twenty-seven patients  
344 and highlights the complexity of generating effective viral genotype-human phenotype  
345 correlations. Given the complex, multifactorial nature of COVID-19, larger pediatric studies  
346 which link to phenotypic outcomes will be required.

347

348 We note the parallel between the emergence of antibiotic resistance traits in bacteria (23, 24)  
349 and viral evolution. Physical and social distancing measures have presented a transmission and  
350 genetic bottleneck to SARS-CoV-2, with data supporting a different mix of virus clades in each  
351 successive wave. The wide-scale development and deployment of vaccines for SARS-CoV-2  
352 will present further bottlenecks to the virus. These measures, while demonstrably effective, have  
353 the potential to drive the evolution of strains capable of bypassing defenses. The emergence of  
354 the D614G variant and the increased infectivity of the UK variant strain B.1.1.7 / UK VUI  
355 202012/01 strain highlight the potential for stochastic genomic variants to generate vaccine  
356 escaping viral strains and to exhibit altered SARS-CoV-2 fitness (12, 22).

357

358 At present, there is high-quality, genomic evidence which supports the biological tolerance or  
359 persistence of the S:N679S variant in at least two distinct clades. This includes the variant  
360 S:N679S in the GR clade found in Brazil, Japan, and Australia, and the variant found in G clade  
361 in September and October of 2020 in the US mid-Atlantic region. While the sequences from  
362 Brazil, Japan, and Australia have been assigned the GR clade, they do not belong to the same  
363 PANGOLIN lineage due to variation in their genomes, which we can see recapitulated in the

364 tree topology of **Fig. 5a**. This is strong evidence for four different evolutionary events which are  
365 tolerated in GR and G SARS-CoV-2 clades.

366

367 The UK presents a strong case for widespread use of high throughput sequencing technologies  
368 and rapid data sharing so that discoveries like these are more rapidly made, confirmed, and  
369 acted upon. Under current funding levels, the US CDC plans to fund the sequencing and  
370 release of at least 6,000 viruses per week as reported on January 4, 2021 (25, 26). This  
371 represents approximately 0.4% of daily cases based on current daily caseloads in excess of  
372 200,000. This is still well below the UK's level of sequencing (4.99%), but is a commendable  
373 step in the right direction, assuming that samples that are sequenced will be representative of  
374 the population, allowing for surveillance. Inclusion of data on disease severity, patient age and  
375 ethnicity, co-morbidities, and other relevant contextual data will help researchers ascertain the  
376 generalizability of sequences in hand, or adjust for confounding factors as needed. Further  
377 funds should be allocated to promote a collaborative effort to sequence biobanked samples in  
378 an effort to understand viral evolution and transmission paths.

379

380 Our analyses identified the S:N679S variant within a neonatal patient with a high observed viral  
381 load at presentation. This single case observation currently represents insufficient evidence to  
382 propose a causal relationship between the S:N679S variant and increased viral loads or  
383 presentation at a very young age. Analysis of the GISAID data for pediatric enrichment was not  
384 possible due to the lack of patient metadata for records with this variant. The observation that  
385 this variant strain of SARS-CoV-2 is currently undergoing community transmission in the US  
386 mid-Atlantic area warrants continuous and rigorous monitoring. The SARS-CoV-2 spike protein  
387 not only moderates viral infectivity and cellular uptake, but is also a target for vaccine and  
388 monoclonal antibody therapeutic development. While vaccines are designed to elicit a  
389 polyclonal immune response, the primary target of the vaccine response is currently a

390 monogenic antigen, and monoclonal antibody cocktails like those produced earlier this year run  
391 the risk of selection for spike variants that escape those treatments (27).

392

393 In conclusion, a new protein coding spike variant, S:N679S, has been discovered in a clinically-  
394 unique patient in Washington, D.C. Analysis of global data shows that presence of the variant  
395 was tolerated by SARS-CoV-2 in at least four different viral lineages, one of which shows  
396 evidence of transmission of the variant to recipients in the US mid-Atlantic region. One of these  
397 lineages presented within the CNH catchment area, showing viral sequence similarity to our  
398 case. Variation in emerging infectious disease is expected, but this finding underscores the  
399 need for larger-scale whole viral genome monitoring to ensure that responses can be quickly  
400 adapted in the face of viral evolution. Until then, variants will continue to circulate undetected,  
401 hindering the global response to SARS-CoV-2. Progress toward the defeat of the COVID-19  
402 pandemic will require effective application of public health measures and wide-scale vaccination  
403 alongside careful monitoring for emergence of viral escape variants.

404

## 405 **MATERIALS AND METHODS**

406 **Initial COVID-19 Screening.** Initial COVID-19 screening was performed at CNH using either  
407 the Diasorin Liaison® MDX-Simplexa™ COVID-19 Direct real-time PCR assay or the Seegene  
408 Allplex™ 2019-nCoV real-time PCR assay. Both assays are qualitative and can be performed  
409 using nasopharyngeal swab, oropharyngeal swabs, or saliva collections. The Diasorin Molecular  
410 Simplexa COVID-19 assay allows for direct amplification of the ORF1ab region and the S gene  
411 (Spike glycoprotein) present in the viral sample. The Seegene Allplex COVID-19 assay requires  
412 viral RNA extraction prior to amplification and uses the Bio-Rad CFX96 real-time PCR  
413 thermocycler for detection. This assay uses primers designed to detect the presence of the E



414 gene (an envelope protein), RdRP (RNA-dependent RNA polymerase), and the N gene (a  
415 nucleoprotein).

416

417 **RNA extraction and cDNA generation.** Viral RNA was extracted with the Qiagen EZ1® DSP  
418 Virus Kit, a magnetic bead-based kit. The concentration of the viral RNA was assessed using  
419 the NanoDrop One. To perform Sanger confirmation on targeted variants, the viral RNA  
420 underwent reverse transcriptase cDNA generation using the High Capacity cDNA Reverse  
421 Transcription kit manufactured by Applied Biosystems. The RT master-mix was prepared  
422 according to manufacturer recommendations excluding the 10X RT Random Primers, as we  
423 had designed our own primers based on the sequencing data. To avoid potential primer bias,  
424 cDNA was synthesized using three different initiation sites in the viral genome

425

426 **Viral sequencing.** Sequence data was generated on the Illumina MiSeq platform as a part of a  
427 collaborative effort between CNH and Children's Hospital of Los Angeles with institutional  
428 review board approval. Following previously described methodology (28), the CleanPlex SARS-  
429 CoV-2 NGS Panel kit (Paragon Genomics, Inc., Hayward, CA) was used to produce cDNA and  
430 then amplify the SARS-CoV-2 genome in two multiplex PCR reactions containing 171 and 172  
431 primer pairs each, producing amplicons ~150bp in length, and covering nearly the entire viral  
432 genome once assembled (~100bp missing on each end). Assembly was performed using the  
433 NovoAlign v4.02.00 algorithm with NC\_045512 as the SARS-CoV-2 reference sequence.

434

435 **Sanger confirmation.** At completion of reverse transcription, Sanger sequencing of selected  
436 variants identified by NGS was performed. We utilized the AccuPrime™ Taq DNA Polymerase  
437 System manufactured by Invitrogen™. A total of 10 PCR reactions containing different forward  
438 and reverse primer pairings were prepared to confirm the observed variants. Thermocycler

439 conditions: denature 94°C 30s, anneal 55°C 30s, extend 68°C 60s, for 30 cycles. The final PCR  
440 products were run on a 2% E-Gel to assess for presence of bands and fragment sizes. The  
441 PCR products were Sanger sequenced by GeneWiz (Frederick, MD).

442

443 **RT primers:**

444 23902r\* TAAATTTGTTTGACTTGTGCAAAAAC

445 24527r CCTGTGATCAACCTATCAATTTGCACTTCAGC

446 24609r AAGATTAGCAGAAGCTCTGATTTCTG

447

448 **PCR primers:**

449 23320\_M13F\* CATTACACCATGTTCTTTTGGTGGTGT

450 23727\_M13R\* TAGACACTGGTAGAATTTCTGTGGTAA

451 23381\_M13F CAGGTTGCTGTTCTTTATCAGGGT

452 23902\_M13R TAAATTTGTTTGACTTGTGCAAAAAC

453 23564\_M13F GCAGGTATATGCGCT

454 24038\_M13R GCCAGCATCTGCAAGTGTCAC

455 23902\_M13F TTTTGCACAAGTCAAACAAATTTA

456 24570\_M13R GTTGAGTCACATATGTCTGCAAAC

457 \*used to generate the representative chromatogram shown in **Fig 4**. Primer

458 sequences shown 5' to 3' for their respective strands. M13 sequencing tags have

459 been removed for clarity.

460

461 **Data access.** GISAID is presently the leader in viral sequence data sharing, having rapidly  
462 expanded their influenza data sharing capabilities to suit the COVID-19 pandemic (29). All data  
463 from samples outside of CNH were accessed from the GISAID database in accordance with  
464 their data sharing agreement (30). As of submission, sequences from CNH have not been  
465 assigned GISAID accession IDs, but will be available in that repository.

466

467 **Phylogenetics.** Consensus sequences were generated from variant calls with bcf tools on the  
468 SARS-CoV-2 reference sequence from Wuhan, China (NC\_045512) (31). MSA were generated  
469 from consensus sequences from CNH samples and from GISAID through ClustalW alignment  
470 on default settings (32). An ML phylogenetic tree was generated on the MSA with MEGA 7.0  
471 (33, 34). All tree files were visualized with Interactive Tree of Life (35).

472

473 **Variant analysis.** The NextClade and CoV-Glue tools were used for variant analysis (18, 36)  
474 and compared to clade, lineage, and global distribution of related sequences via the PANGOLIN  
475 (37) tool. NextClade assesses variation by aligning a query sequence to the SARS-CoV-2  
476 reference sequence by scanning 21-mers across the genome followed by a banded Smith-  
477 Waterman alignment with affine gap penalty, while CoV-GLUE leverages MAFFT (38) and  
478 RAxML (39), with detailed methods described in their preprint. PANGOLIN leverages MAFFT  
479 (38) and IQ-TREE 2 (40), with methods previously described. All relevant files are available as

480 **S3.**

481

482 **Proteinase cleavage site prediction.** A multifasta of S1/S2 cleavage site was uploaded to  
483 ProP1.0 for analysis and prediction of cleavage sites (41). Scores were exported and organized  
484 by of GISAID variants. **S2.**

485

486 **Acknowledgements:**

487 We acknowledge the critical efforts of Kristen Kocher, Melissa Andrew, and Miguel Almalvez  
488 who made viral transport media and sample collection kits during the shortage which occurred  
489 in spring and summer 2020.

490

Novel SARS-CoV-2 spike variant identified in Washington, D.C.

20

491 We acknowledge Tara Workman, Hasani Malcolm, Farzana Siddiqui, Katy Draper, and Ashenafi  
492 Melkamu who ran SARS-CoV-2 PCR tests; Eric Freeman, William Suslovic, and Aszia Burrel  
493 who coordinated the clinical research program; Michael Evangelista and Joyce Granados who  
494 contributed from the microbiology department.

495

496 We acknowledge the GISAID consortium for their comprehensive work. We acknowledge the  
497 specific originating labs: Maryland Public Health Laboratory, Virginia Division of Consolidated  
498 Laboratory Services, Royal Hobart Hospital, Australia, Pathogen Genomics Center at the Japan  
499 National Institute of Infectious Disease, Laboratorio Central de Saude Publica do Estado do Rio  
500 Grande do Sul, and Respiratory Viruses Branch, US CDC. (**S4**)

501

502 Our funders included the Ikaria Fund, and the A. James Clark Distinguished Professorship,  
503 CNH and the Saban Research Institute at CHLA intramural support for COVID-19 Directed  
504 Research.

505

## 506 REFERENCES

- 507 1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P,  
508 Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A Novel Coronavirus from  
509 Patients with Pneumonia in China, 2019. *N Engl J Med* 382:727–733.
- 510 2. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J, Queen K, Tao Y, Paden C,  
511 Zhang J, Li Y, Uehara A, Wang H, Goldsmith C, Bullock H, Wang L, Whitaker B, Lynch B,  
512 Gautam R, Schindewolf C, Lokugamage K, Scharton D, Plante J, Mirchandani D, Widen  
513 S, Narayanan K, Makino S, Ksiazek T, Plante K, Weaver S, Lindstrom S, Tong S,  
514 Menachery V, Thornburg N. 2020. Isolation and characterization of SARS-CoV-2 from the  
515 first US COVID-19 patient. *bioRxiv Prepr Serv Biol* 2020.03.02.972935.
- 516 3. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu

- 517 J. 2020. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* 7:1012–  
518 1023.
- 519 4. Bialek S, Gierke R, Hughes M, McNamara LA, Pilishvili T, Skoff T. 2020. Coronavirus  
520 Disease 2019 in Children — United States, February 12–April 2, 2020. *MMWR Morb*  
521 *Mortal Wkly Rep* 69:422–426.
- 522 5. Götzinger F, Santiago-García B, Noguera-Julián A, Lanaspá M, Lancelli L, Calò  
523 Carducci FI, Gabrovská N, Velizarova S, Prunk P, Osterman V, Krivec U, Lo Vecchio A,  
524 Shingadia D, Soriano-Arandes A, Melendo S, Lanari M, Pierantoni L, Wagner N, L’Huillier  
525 AG, Heininger U, Ritz N, Bandi S, Krajcar N, Roglić S, Santos M, Christiaens C, Creuven  
526 M, Buonsenso D, Welch SB, Bogyi M, Brinkmann F, Tebruegge M, Pfefferle J,  
527 Zacharasiewicz A, Berger A, Berger R, Strenger V, Kohlfürst DS, Zschocke A, Bernar B,  
528 Simma B, Haberlandt E, Thir C, Biebl A, Vanden Driessche K, Boiy T, Van Brusselen D,  
529 Bael A, Debulpaep S, Schelstraete P, Pavic I, Nygaard U, Glenthøj JP, Heilmann  
530 Jensen L, Lind I, Tistsenko M, Uustalu Ü, Buchtala L, Thee S, Kobbe R, Rau C, Schwerk  
531 N, Barker M, Tsolia M, Eleftheriou I, Gavin P, Kozdoba O, Zsigmond B, Valentini P,  
532 Ivaškevičienė I, Ivaškevičius R, Vilc V, Schölvinck E, Rojahn A, Smyrniaios A,  
533 Klingenberg C, Carvalho I, Ribeiro A, Starshinova A, Solovic I, Falcón L, Neth O, Minguell  
534 L, Bustillo M, Gutiérrez-Sánchez AM, Guarch Ibáñez B, Ripoll F, Soto B, Kötz K,  
535 Zimmermann P, Schmid H, Zucol F, Niederer A, Buettcher M, Cetin BS, Bilogortseva O,  
536 Chechenyeva V, Demirjian A, Shackley F, McFetridge L, Speirs L, Doherty C, Jones L,  
537 McMaster P, Murray C, Child F, Beuvink Y, Makwana N, Whittaker E, Williams A, Fidler  
538 K, Bernatoniene J, Song R, Oliver Z, Riordan A. 2020. COVID-19 in children and  
539 adolescents in Europe: a multinational, multicentre cohort study. *Lancet Child Adolesc*  
540 *Heal* 4:653–661.
- 541 6. DeBiasi RL, Song X, Delaney M, Bell M, Smith K, Pershad J, Ansusinha E, Hahn A,  
542 Hamdy R, Harik N, Hanisch B, Jantusch B, Koay A, Steinhorn R, Newman K, Wessel D.

- 543 2020. Severe Coronavirus Disease-2019 in Children and Young Adults in the  
544 Washington, DC, Metropolitan Region. *J Pediatr* 223:199-203.e1.
- 545 7. Feldstein LR, Rose EB, Horwitz SM, Collins JP, Newhams MM, Son MBF, Newburger  
546 JW, Kleinman LC, Heidemann SM, Martin AA, Singh AR, Li S, Tarquinio KM, Jaggi P,  
547 Oster ME, Zackai SP, Gillen J, Ratner AJ, Walsh RF, Fitzgerald JC, Keenaghan MA,  
548 Alharash H, Doymaz S, Clouser KN, Giuliano JS, Gupta A, Parker RM, Maddux AB,  
549 Havalad V, Ramsingh S, Bukulmez H, Bradford TT, Smith LS, Tenforde MW, Carroll CL,  
550 Riggs BJ, Gertz SJ, Daube A, Lansell A, Coronado Munoz A, Hobbs C V., Marohn KL,  
551 Halasa NB, Patel MM, Randolph AG. 2020. Multisystem Inflammatory Syndrome in U.S.  
552 Children and Adolescents. *N Engl J Med* 383:334–346.
- 553 8. Godfred-Cato S, Bryant B, Leung J, Oster ME, Conklin L, Abrams J, Roguski K, Wallace  
554 B, Prezzato E, Koumans EH, Lee EH, Geevarughese A, Lash MK, Reilly KH, Pulver WP,  
555 Thomas D, Feder KA, Hsu KK, Plipat N, Richardson G, Reid H, Lim S, Schmitz A, Pierce  
556 T, Hrapcak S, Datta D, Morris SB, Clarke K, Belay E. 2020. COVID-19–Associated  
557 Multisystem Inflammatory Syndrome in Children — United States, March–July 2020.  
558 *MMWR Morb Mortal Wkly Rep* 69:1074–1080.
- 559 9. COVID-19 Surveillance. Washington, DC Gov. <https://coronavirus.dc.gov/data>
- 560 10. The White House. 2020. Suspension of Entry as Immigrants and Nonimmigrants of  
561 Certain Additional Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus.  
562 *Fed Regist.* [https://www.federalregister.gov/documents/2020/02/05/2020-](https://www.federalregister.gov/documents/2020/02/05/2020-02424/suspension-of-entry-as-immigrants-and-nonimmigrants-of-persons-who-pose-a-risk-of-transmitting-2019)  
563 [02424/suspension-of-entry-as-immigrants-and-nonimmigrants-of-persons-who-pose-a-](https://www.federalregister.gov/documents/2020/02/05/2020-02424/suspension-of-entry-as-immigrants-and-nonimmigrants-of-persons-who-pose-a-risk-of-transmitting-2019)  
564 [risk-of-transmitting-2019](https://www.federalregister.gov/documents/2020/02/05/2020-02424/suspension-of-entry-as-immigrants-and-nonimmigrants-of-persons-who-pose-a-risk-of-transmitting-2019)
- 565 11. The White House. 2020. Suspension of Entry as Immigrants and Nonimmigrants of  
566 Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus and Other  
567 Appropriate Measures To Address This Risk. *Fed*  
568 *Regist.* <https://www.federalregister.gov/documents/2020/03/16/2020-05578/suspension->

569 of-entry-as-immigrants-and-nonimmigrants-of-certain-additional-persons-who-pose-a-

570 risk-of

571 12. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou

572 J, Fontes-Garfias CR, Mirchandani D, Scharton D, Bilello JP, Ku Z, An Z, Kalveram B,

573 Freiberg AN, Menachery VD, Xie X, Plante KS, Weaver SC, Shi PY. 2020. Spike

574 mutation D614G alters SARS-CoV-2 fitness. *Nature* 1–6.

575 13. Duckert P, Brunak S, Blom N. 2004. Prediction of proprotein convertase cleavage sites.

576 *Protein Eng Des Sel* 17:107–112.

577 14. Bertram S, Glowacka I, Muller MA, Lavender H, Gnirss K, Nehlmeier I, Niemeyer D, He

578 Y, Simmons G, Drosten C, Soilleux EJ, Jahn O, Steffen I, Pohlmann S. 2011. Cleavage

579 and Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by

580 Human Airway Trypsin-Like Protease. *J Virol* 85:13363–13372.

581 15. Xia S, Lan Q, Su S, Wang X, Xu W, Liu Z, Zhu Y, Wang Q, Lu L, Jiang S. 2020. The role

582 of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the

583 presence or absence of trypsin. *Signal Transduct Target Ther*. Springer Nature.

584 16. Klimstra W, Tilston-Lunel N, Nambulli S, Boslett J, McMillen C, Gilliland T, Dunn M, Sun

585 C, Wheeler S, Wells A, Hartman A, McElroy A, Reed D, Rennick L, Duprex WP. 2020.

586 SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from

587 acutely infected, hospitalized COVID-19 patients. *bioRxiv Prepr Serv Biol*

588 <https://doi.org/10.1101/2020.06.19.154930>.

589 17. Nguyen HT, Zhang S, Wang Q, Anang S, Wang J, Ding H, Kappes JC, Sodroski J,

590 Sodroski JG. 2020. Spike glycoprotein and host cell determinants of SARS-CoV-2 entry

591 and cytopathic effects Downloaded from <https://doi.org/10.1128/JVI.02304-20>.

592 18. Hodcroft E, Aksamentov I, Stroud N, Neher R, Sibley T. 2020. Nextclade.

593 19. Excess Deaths Associated with COVID-19. US

594 CDC. [https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess\\_deaths.htm](https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm)

- 595 20. 2021. COVID-19 Map. Johns Hopkins Coronavirus Resour Cent.
- 596 21. Kirby T. 2021. New variant of SARS-CoV-2 in UK causes surge of COVID-19. *Lancet*  
597 *Respir Med* [https://doi.org/10.1016/S2213-2600\(21\)00005-9](https://doi.org/10.1016/S2213-2600(21)00005-9).
- 598 22. 2020. SARS-CoV-2 Variant – United Kingdom of Great Britain and Northern Ireland.  
599 WHO. World Health Organization. [http://www.who.int/csr/don/21-december-2020-sars-](http://www.who.int/csr/don/21-december-2020-sars-cov2-variant-united-kingdom/en/)  
600 [cov2-variant-united-kingdom/en/](http://www.who.int/csr/don/21-december-2020-sars-cov2-variant-united-kingdom/en/)
- 601 23. Ventola CL. 2015. The antibiotic resistance crisis: causes and threats. *P T J* 40:277–83.
- 602 24. Windels EM, Michiels JE, van den Bergh B, Fauvart M, Michiels J. 2019. Antibiotics:  
603 Combatting tolerance to stop resistance. *MBio* 10.
- 604 25. Emerging SARS-CoV-2 Variants. US CDC.
- 605 26. Cohen E. 2021. CDC hopes to check more samples for new Covid strain. CNN.
- 606 27. Coronavirus (COVID-19) Update: FDA Authorizes Monoclonal Antibodies for Treatment  
607 of COVID-19. US FDA. [https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-](https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html)  
608 [research/scientific-brief-emerging-variants.html](https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html)
- 609 28. Pandey U, Yee R, Shen L, Judkins AR, Bootwalla M, Ryutov A, Maglinte DT, Ostrow D,  
610 Precit M, Biegel JA, Bender JM, Gai X, Bard JD. 2020. High Prevalence of SARS-CoV-2  
611 Genetic Variation and D614G Mutation in Pediatric Patients with COVID-19. *Open Forum*  
612 *Infect Dis* <https://doi.org/10.1093/ofid/ofaa551>.
- 613 29. LoTempio J, Spencer D, Yarvitz R, Delot-Vilan A, Vilain E, Delot E. 2020. We Can Do  
614 Better: Lessons Learned on Data Sharing in COVID-19 Pandemic Can Inform Future  
615 Outbreak Preparedness and Response. *Sci Dipl.*
- 616 30. GISAID Data Access Terms of Use. GISAID  
617 Consort. <https://www.gisaid.org/registration/terms-of-use/>
- 618 31. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane  
619 T, McCarthy SA, Davies RM, Li H. 2020. Twelve years of SAMtools and BCFtools.
- 620 32. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of



- 621 progressive multiple sequence alignment through sequence weighting, position-specific  
622 gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
- 623 33. Hall BG. 2013. Building Phylogenetic Trees from Molecular Data with MEGA. *Mol Biol*  
624 *Evol* 30:1229–1235.
- 625 34. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis  
626 Version 7.0 for Bigger Datasets. *Mol Biol Evol* 33:1870–1874.
- 627 35. Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: Recent updates and new  
628 developments. *Nucleic Acids Res* 47:W256–W259.
- 629 36. Singer JB, Gifford RJ, Cotten M, Robertson DL. 2020. CoV-GLUE: A Web Application for  
630 Tracking SARS-CoV-2 Genomic Variation  
631 <https://doi.org/10.20944/preprints202006.0225.v1>.
- 632 37. Rambaut A, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG.  
633 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic  
634 epidemiology. *Nat Microbiol* 5:1403–1407.
- 635 38. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:  
636 Improvements in performance and usability. *Mol Biol Evol* 30:772–780.
- 637 39. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis  
638 of large phylogenies. *Bioinformatics* 30:1312–1313.
- 639 40. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,  
640 Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic  
641 Inference in the Genomic Era. *Mol Biol Evol* 37:1530–1534.
- 642 41. Duckert P, Brunak S, Blom N. 2004. Prediction of proprotein convertase cleavage sites.  
643 *Protein Eng Des Sel* 17:107–112.
- 644
- 645
- 646

647 **FIGURES**

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

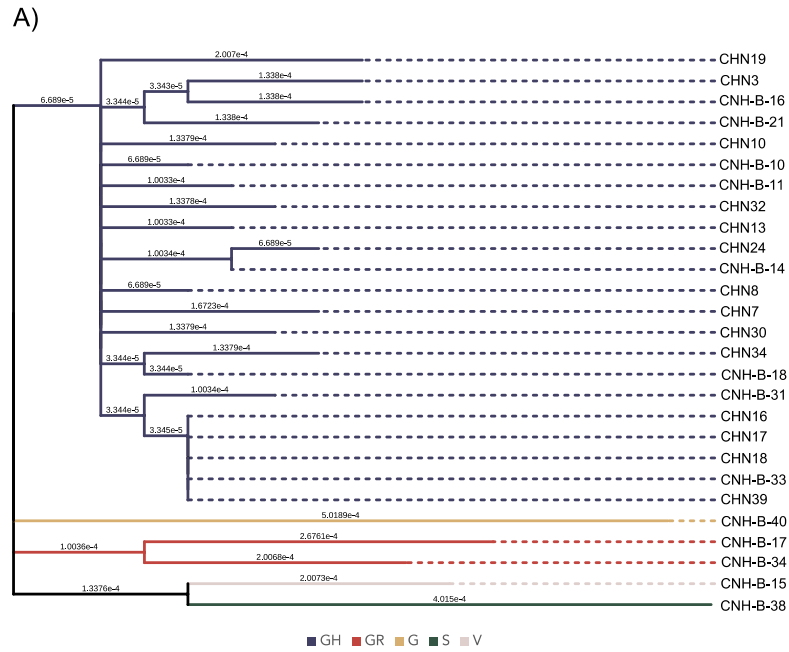
671

672

673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698

**Fig 1. A) Maximum-likelihood phylogenetic tree of 27 SARS-CoV-2 positive patients.** This tree was constructed from a ClustalW multiple sequence alignment and visualized with the Interactive Tree of Life GUI. Solid branches represent divergence in substitutions per nucleotide, while dashed lines are added for ease of interpretation. Colors represent the GISAID clade to which each sample belongs. **B) Time series plot of the GISAID clades of GISAID submissions over time in select cities.** Sequence metadata were accessed from GISAID to establish coarse differences in viral population diversity over time.

699

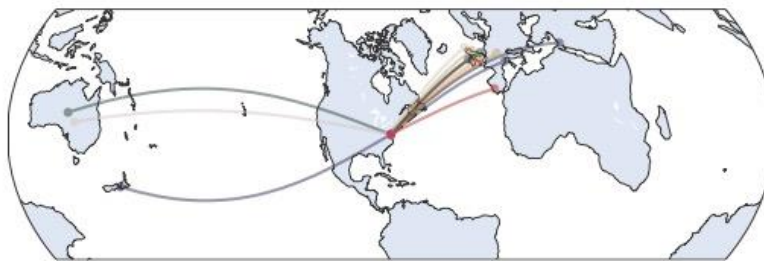
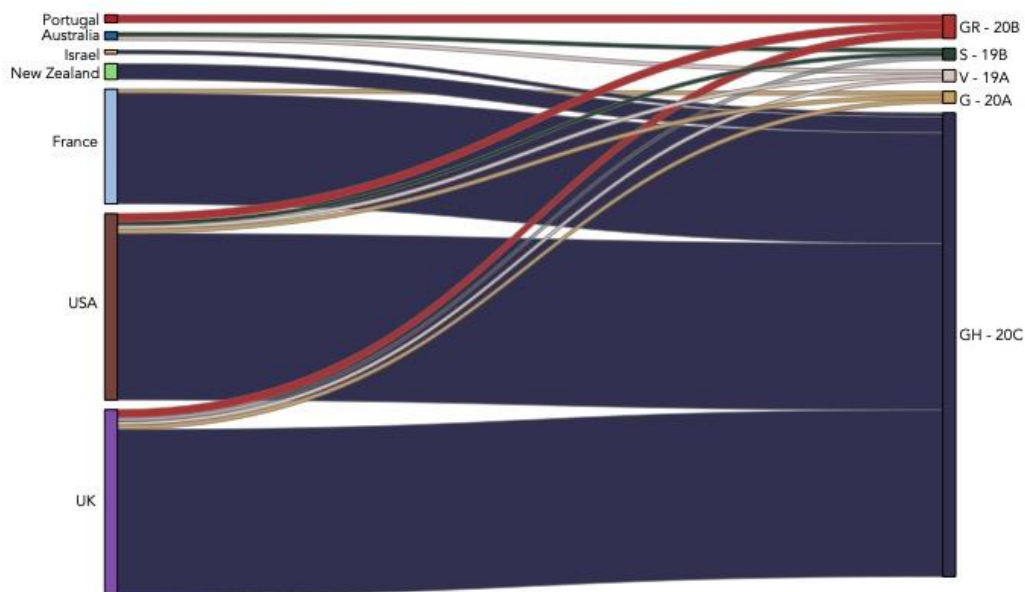


700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724

**Fig 2. Global network of SARS-CoV-2.** Representation of the locations where the highest number of sequences assigned to a given PANGOLIN lineage can be found in relation to DC sequences

## Novel SARS-CoV-2 spike variant identified in Washington, D.C.

2



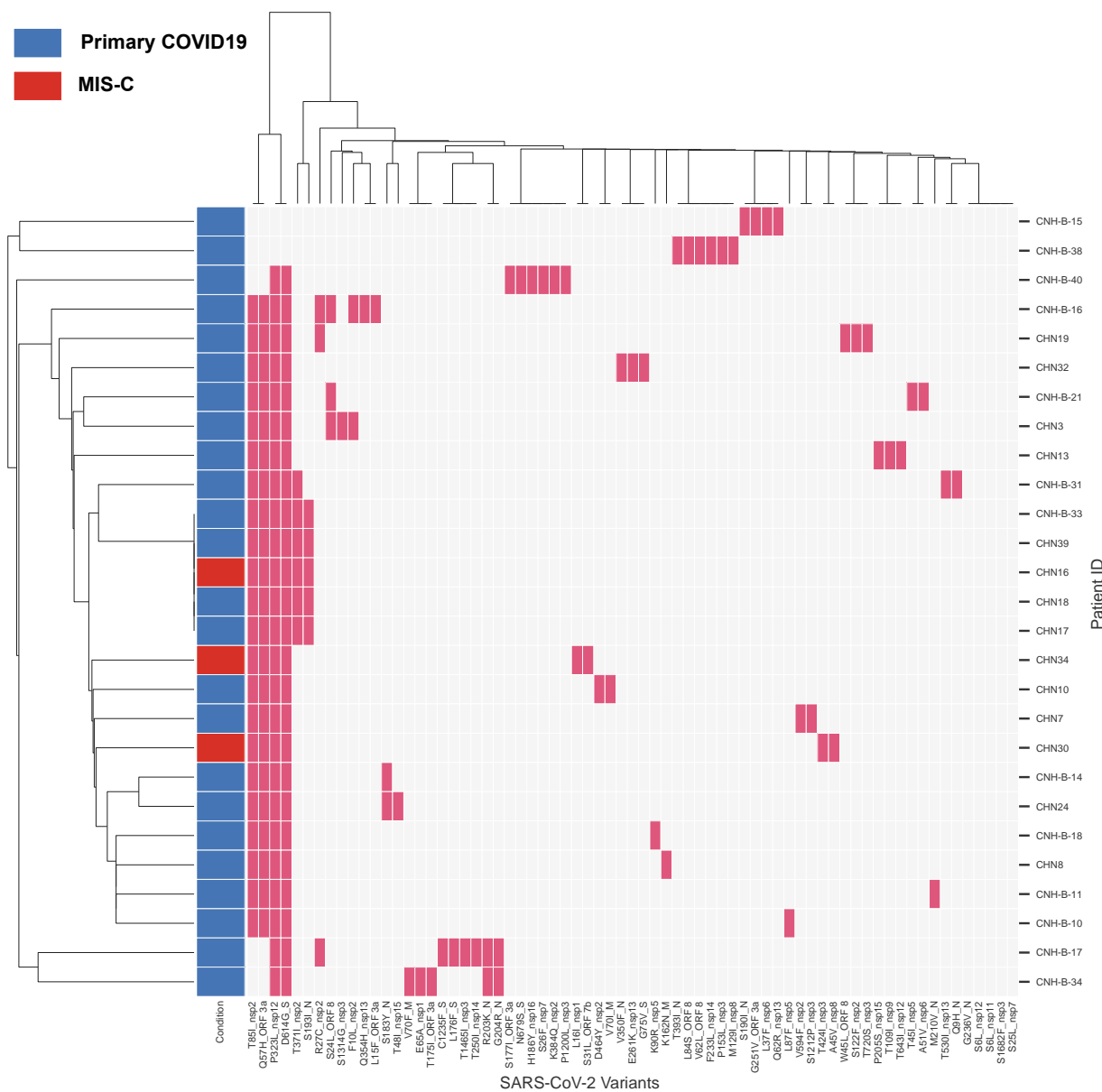
725

726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750

**Fig 3. Clustering of detected SARS-CoV-2 viral genome variants with disease outcome.**

Bi-clustering of the binary variant matrix of SARS reveals a diverse cluster of variants and five patients with identical viral genomic profiles. The condition of each patient is depicted indicated primary COVID-19 (blue) and MIS-C (red). Variants identified across the cohort were aggregated and a binary matrix was generated for each patient and variant. Red colored boxes indicate the detection of a variant within that case.

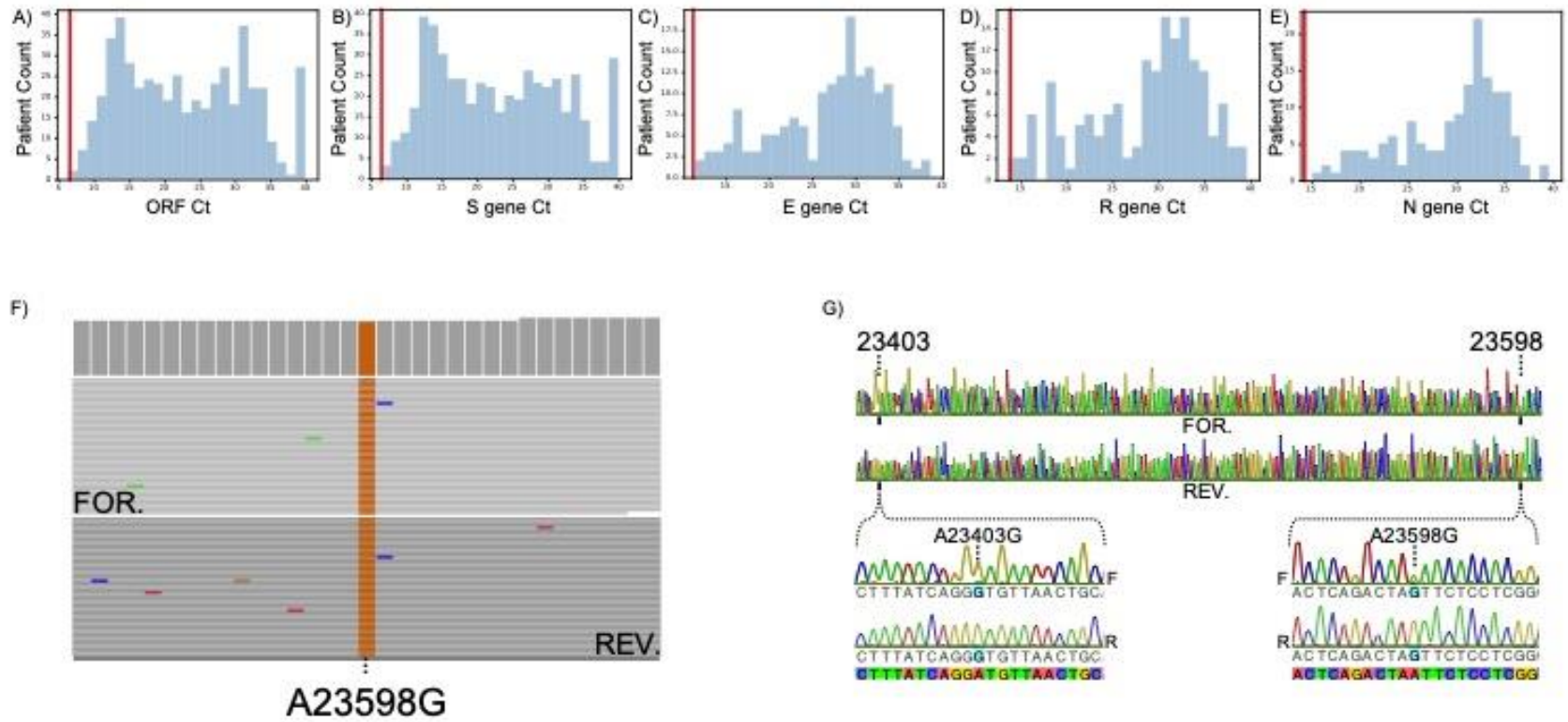
Novel SARS-CoV-2 spike variant identified in Washington, D.C.





752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776

**Fig 4. A-E) Histograms of viral PCR test cycle threshold values.** Vertical red line indicates the detected RT-PCR of the S:N679S case. **A)** Diasorin ORF Ct distribution **B)** Diasorin S gene Ct distribution. **C)** Seegene RT-PCR E gene Ct distribution. **D)** Seegeen R gene Ct distribution. **E)** Seegene N gene Ct distribution. **F)** Short-read pileup at the coding region for spike protein variant of interest, S:N679S. The read pileup shows strong homozygous signal (>9,000x coverage) for a non-synonymous variant in the viral spike protein. **G)** Sanger sequence confirmation of variant of interest, S:N679S. Chromatogram showing a contiguous span containing the previously characterized D614G variant and linkage to the S:N679S variant of interest.



778

779

780

781

782

783

784

785

786

787 **Fig 5 A) Maximum likelihood phylogenetic tree of all high-quality sequences in GISAID**

788 **with variant S:N679S in the spike protein.** A non-reference sample from Wuhan collected in

789 December 2019 was used as the outgroup in this tree to draw contrast between clades. Clade

790 G in yellow shows the high-quality sequences in lineage B.1.189 where spike variant S:N679S

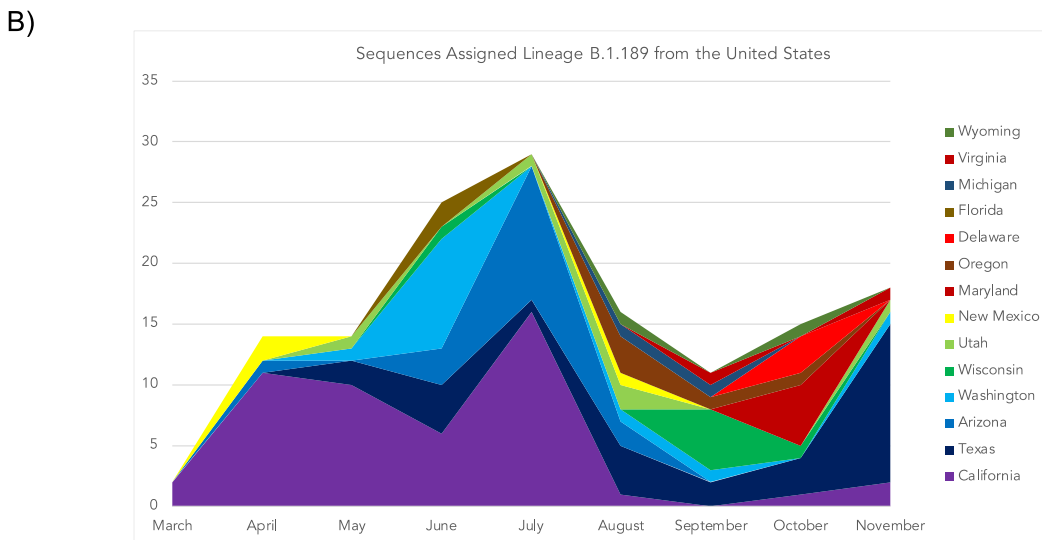
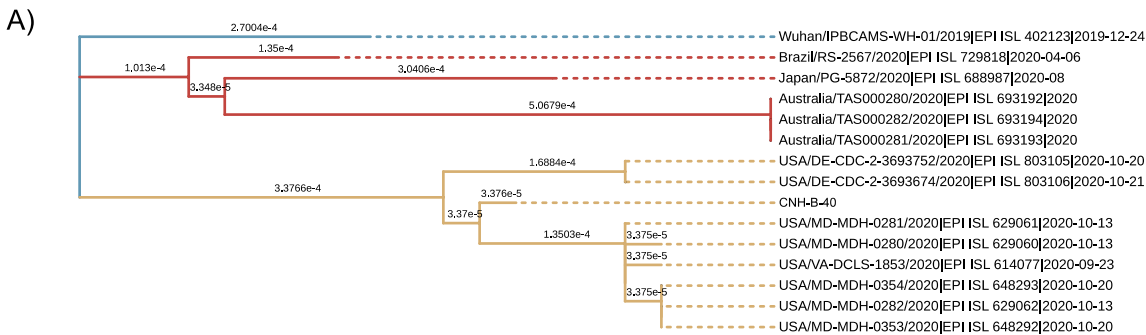
791 has emerged. Clade GR in yellow shows three separate evolutionary events resulting in spike

792 variant S:N679S. **B) Time series plot of the count, by location, of sequences in GISAID**

793 **belonging to PANGOLIN lineage B.1.189.** Sequence metadata were accessed from GISAID

794 to establish coarse differences in viral population diversity over time.

Novel SARS-CoV-2 spike variant identified in Washington, D.C.



Novel SARS-CoV-2 spike variant identified in Washington, D.C.

9

796 **SUPPLEMENT LEGEND**

797

798 **S1: List of case IDs, clades, and disease status**

799 S1\_Case\_clade\_and\_disease\_status\_12\_21\_2020.csv

800 **S2\_A: ProP scores**

801 S2\_A\_GISAID\_Spike\_residue\_679\_variant\_scores\_15\_jan\_2021.pdf

802 **S2\_B: Amino acid substitution multifasta**

803 S2\_B\_679\_aa\_sub.fa

804 **S3\_A: Coding and non-coding variants in sequences with S:N679S**

805 S3\_A\_B1189\_N679S\_variant\_summary.pdf

806 **S3\_B: Variants called via NextClade**

807 S3\_B\_NextClade\_variants.xlsx

808 **S4: Acknowledgements from originating labs, GISAID**

809 S4\_gisaid\_hcov-19\_acknowledgement\_table\_high\_quality\_spike\_n679S.pdf