

1 **Molecular Architecture of Early Dissemination**  
2 **and Massive Second Wave of the SARS-CoV-2 Virus in a**  
3 **Major Metropolitan Area**

4  
5 **Running Title: Two waves of COVID-19 disease in Houston, Texas**

6  
7 **S. Wesley Long,<sup>a,b,1</sup> Randall J. Olsen,<sup>a,b,1</sup> Paul A. Christensen,<sup>a,1</sup> David W.**  
8 **Bernard,<sup>a,b</sup> James J. Davis,<sup>c,d</sup> Maulik Shukla,<sup>c,d</sup> Marcus Nguyen,<sup>c,d</sup> Matthew**  
9 **Ojeda Saavedra,<sup>a</sup> Prasanti Yerramilli,<sup>a</sup> Layne Pruitt,<sup>a</sup> Sishir Subedi,<sup>a</sup> Hung-**  
10 **Che Kuo,<sup>e</sup> Heather Hendrickson,<sup>a</sup> Ghazaleh Eskandari,<sup>a</sup> Hoang A. T.**  
11 **Nguyen,<sup>a</sup> J. Hunter Long,<sup>a</sup> Muthiah Kumaraswami,<sup>a</sup> Jule Goike,<sup>e</sup> Daniel**  
12 **Boutz,<sup>f</sup> Jimmy Gollihar,<sup>a,f</sup> Jason S. McLellan,<sup>e</sup> Chia-Wei Chou,<sup>e</sup> Kamyab**  
13 **Javanmardi,<sup>e</sup> Ilya J. Finkelstein,<sup>e,g</sup>, and James M. Musser<sup>a,b,#</sup>**

14  
15 <sup>a</sup>Center for Molecular and Translational Human Infectious Diseases Research,  
16 Department of Pathology and Genomic Medicine, Houston Methodist Research  
17 Institute and Houston Methodist Hospital, 6565 Fannin Street, Houston, Texas  
18 77030

19 <sup>b</sup>Departments of Pathology and Laboratory Medicine, and Microbiology and  
20 Immunology, Weill Cornell Medical College, 1300 York Avenue, New York, New  
21 York 10065

22 <sup>c</sup>Consortium for Advanced Science and Engineering, University of Chicago, 5801  
23 South Ellis Avenue, Chicago, Illinois, 60637

24 <sup>d</sup>Computing, Environment and Life Sciences, Argonne National Laboratory, 9700  
25 South Cass Avenue, Lemont, Illinois 60439

26 <sup>e</sup>Department of Molecular Biosciences and Institute of Molecular Biosciences,  
27 The University of Texas at Austin, Austin, Texas 78712

28 <sup>f</sup>CCDC Army Research Laboratory-South, University of Texas, Austin, Texas 78712

29 <sup>g</sup>Center for Systems and Synthetic Biology, University of Texas at Austin, Austin,  
30 Texas 78712

31

32 <sup>1</sup>S.W.L., R.J.O., and P.A.C. contributed equally to this article. The order of co-first  
33 authors was determined by discussion and mutual agreement between the three  
34 co-first authors.

35

36 <sup>#</sup>Address correspondence to: James M. Musser, M.D., Ph.D., Department of  
37 Pathology and Genomic Medicine, Houston Methodist Research Institute, 6565  
38 Fannin Street, Suite B490, Houston, Texas 77030. Tel: 713.441.5890, E-mail:  
39 [jmmusser@houstonmethodist.org](mailto:jmmusser@houstonmethodist.org).

40

41 This article is a direct contribution from James M. Musser, a Fellow of the  
42 American Academy of Microbiology, who arranged for and secured reviews by  
43 Barry N. Kreiswirth, Center for Discovery and Innovation, Hackensack Meridian

44 Health, New Jersey; and David M. Morens, National Institute of Allergy and  
45 Infectious Diseases, National Institutes of Health, Maryland.

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66 **ABSTRACT** We sequenced the genomes of 5,085 SARS-CoV-2 strains  
67 causing two COVID-19 disease waves in metropolitan Houston, Texas, an  
68 ethnically diverse region with seven million residents. The genomes were  
69 from viruses recovered in the earliest recognized phase of the pandemic in  
70 Houston, and an ongoing massive second wave of infections. The virus  
71 was originally introduced into Houston many times independently. Virtually  
72 all strains in the second wave have a Gly614 amino acid replacement in the  
73 spike protein, a polymorphism that has been linked to increased  
74 transmission and infectivity. Patients infected with the Gly614 variant  
75 strains had significantly higher virus loads in the nasopharynx on initial  
76 diagnosis. We found little evidence of a significant relationship between  
77 virus genotypes and altered virulence, stressing the linkage between  
78 disease severity, underlying medical conditions, and host genetics. Some  
79 regions of the spike protein - the primary target of global vaccine efforts -  
80 are replete with amino acid replacements, perhaps indicating the action of  
81 selection. We exploited the genomic data to generate defined single amino  
82 acid replacements in the receptor binding domain of spike protein that,  
83 importantly, produced decreased recognition by the neutralizing  
84 monoclonal antibody CR30022. Our study is the first analysis of the  
85 molecular architecture of SARS-CoV-2 in two infection waves in a major  
86 metropolitan region. The findings will help us to understand the origin,  
87 composition, and trajectory of future infection waves, and the potential

88 **effect of the host immune response and therapeutic maneuvers on SARS-**  
89 **CoV-2 evolution.**

90

91 **IMPORTANCE** There is concern about second and subsequent waves of  
92 **COVID-19 caused by the SARS-CoV-2 coronavirus occurring in**  
93 **communities globally that had an initial disease wave. Metropolitan**  
94 **Houston, Texas, with a population of 7 million, is experiencing a massive**  
95 **second disease wave that began in late May 2020. To understand SARS-**  
96 **CoV-2 molecular population genomic architecture, evolution, and**  
97 **relationship between virus genotypes and patient features, we sequenced**  
98 **the genomes of 5,085 SARS-CoV-2 strains from these two waves. Our study**  
99 **provides the first molecular characterization of SARS-CoV-2 strains**  
100 **causing two distinct COVID-19 disease waves.**

101

102 **KEYWORDS:** SARS-CoV-2, COVID-19 disease, genome sequencing, molecular  
103 population genomics, evolution

104

105 [Introduction]

106 **P**andemic disease caused by the severe acute respiratory syndrome  
107 coronavirus 2 (SARS-CoV-2) virus is now responsible for massive human  
108 morbidity and mortality worldwide (1-5). The virus was first documented to cause  
109 severe respiratory infections in Wuhan, China, beginning in late December 2019  
110 (6-9). Global dissemination occurred extremely rapidly and has affected major  
111 population centers on most continents (10, 11). In the United States, the Seattle  
112 and the New York City (NYC) regions have been especially important centers of  
113 COVID-19 disease caused by SARS-CoV-2. For example, as of August 19,  
114 2020, there were 227,419 confirmed SARS-CoV-2 cases in NYC, causing 56,831  
115 hospitalizations and 19,005 confirmed fatalities and 4,638 probable fatalities (12).  
116 Similarly, in Seattle and King County, 17,989 positive patients and 696 deaths  
117 have been reported as of August 18, 2020 (13).

118       The Houston metropolitan area is the fourth largest and most ethnically  
119 diverse city in the United States, with a population of approximately 7 million  
120 (14, 15). The 2,400-bed Houston Methodist health system has seven hospitals  
121 and serves a large, multiethnic, and socioeconomically diverse patient population  
122 throughout greater Houston (13, 14). The first COVID-19 case in metropolitan  
123 Houston was reported on March 5, 2020 with community spread occurring one  
124 week later (16). Many of the first cases in our region were associated with  
125 national or international travel in areas known to have SARS-CoV-2 virus  
126 outbreaks (16). A central molecular diagnostic laboratory serving all Houston  
127 Methodist hospitals and our very early adoption of a molecular test for the SARS-

128 CoV-2 virus permitted us to rapidly identify positive patients and interrogate  
129 genomic variation among strains causing early infections in the greater Houston  
130 area. Our analysis of SARS-CoV-2 genomes causing disease in Houston has  
131 continued unabated since early March and is ongoing. Genome sequencing and  
132 related efforts were expanded extensively in late May as we recognized that a  
133 prominent second wave was underway (**Figure 1**).

134 Here, we report that SARS-CoV-2 was introduced to the Houston area  
135 many times, independently, from diverse geographic regions, with virus  
136 genotypes representing genetic clades causing disease in Europe, Asia, South  
137 America and elsewhere in the United States. There was widespread community  
138 dissemination soon after COVID-19 cases were reported in Houston. Strains with  
139 a Gly614 amino acid replacement in the spike protein, a polymorphism that has  
140 been linked to increased transmission and *in vitro* cell infectivity, increased  
141 significantly over time and caused virtually all COVID-19 cases in the massive  
142 second disease wave. Patients infected with strains with the Gly614 variant had  
143 significantly higher virus loads in the nasopharynx on initial diagnosis. Some  
144 naturally occurring single amino acid replacements in the receptor binding  
145 domain (RBD) of spike protein resulted in decreased reactivity with a neutralizing  
146 monoclonal antibody, consistent with the idea that some virus variants arise due  
147 to host immune pressure.

148

## 149 **RESULTS**

150           **Description of metropolitan Houston.** Houston, Texas, is located in the  
151 southwestern United States, 50 miles inland from the Gulf of Mexico. It is the  
152 most ethnically diverse city in the United States (14). Metropolitan Houston is  
153 comprised predominantly of Harris County plus parts of eight contiguous  
154 surrounding counties. In the aggregate, the metropolitan area includes 9,444  
155 square miles. The estimated population size of metropolitan Houston is 7 million  
156 (<https://www.houston.org/houston-data>).

157           **Epidemic curve characteristics over two disease waves.** The first  
158 confirmed case of COVID-19 in the Houston metropolitan region was reported on  
159 March 5, 2020 (16), and the first confirmed case diagnosed in Houston Methodist  
160 hospitals was reported on March 6, 2020. The epidemic curve indicated a first  
161 wave of COVID-19 cases that peaked around April 11-15, followed by a decline  
162 in cases until May 11. Soon thereafter, the slope of the case curve increased with  
163 a very sharp uptick in confirmed cases beginning on June 12 (**Figure 1B**). We  
164 consider May 11 as the transition between waves, as this date is the inflection  
165 point of the cumulative new cases curve and had the absolute lowest number of  
166 new cases in the mid-May time period. Thus, for the data presented herein, wave  
167 1 is defined as March 5 through May 11, 2020, and wave 2 is defined as May 12  
168 through July 7, 2020. Epidemiologic trends within the Houston Methodist Hospital  
169 population were mirrored by data from Harris County and the greater  
170 metropolitan Houston region (**Figure 1A**). Through the 7<sup>th</sup> of July, 25,366  
171 COVID-19 cases were reported in Houston, 37,776 cases in Harris County, and  
172 53,330 in metropolitan Houston, including 9,823 cases in Houston Methodist



173 facilities (inpatients and outpatients) ([https://www.tmc.edu/coronavirus-](https://www.tmc.edu/coronavirus-updates/infection-rate-in-the-greater-houston-area/)  
174 [updates/infection-rate-in-the-greater-houston-area/](https://www.tmc.edu/coronavirus-updates/infection-rate-in-the-greater-houston-area/) and  
175 <https://harriscounty.maps.arcgis.com/apps/opsdashboard/index.html#/c0de71f8e>  
176 [a484b85bb5efcb7c07c6914](https://harriscounty.maps.arcgis.com/apps/opsdashboard/index.html#/c0de71f8ea484b85bb5efcb7c07c6914)).

177 During the first wave (early March through May 11), 11,476 COVID-19  
178 cases were reported in Houston, including 1,729 cases in the Houston Methodist  
179 Hospital system. Early in the first wave (from March 5 through March 30, 2020),  
180 we tested 3,080 patient specimens. Of these, 406 (13.2%) samples were positive  
181 for SARS-CoV-2, representing 40% (358/898) of all confirmed cases in  
182 metropolitan Houston during that time period. As our laboratory was the first  
183 hospital-based facility to have molecular testing capacity for SARS-CoV-2  
184 available on site, our strain samples are likely representative of COVID-19  
185 infections during the first wave.

186 For the entire study period (March 5 through July 7, 2020), we tested  
187 68,418 specimens from 55,800 patients. Of these, 9,121 patients (16.4%) had a  
188 positive test result, representing 17.1% (9,121/53,300) of all confirmed cases in  
189 metropolitan Houston. Thus, our strain samples are also representative of those  
190 responsible for COVID-19 infections in the massive second wave.

191 To test the hypothesis that, on average, the two waves affected different  
192 groups of patients, we analyzed individual patient characteristics (hospitalized  
193 and non-hospitalized) in each wave. Consistent with this hypothesis, we found  
194 significant differences in the COVID-19 patients in each wave (**Table S1**). For  
195 example, patients in the second wave were significantly younger, had fewer

196 comorbidities, were more likely to be Hispanic/Latino (by self-report), and lived in  
197 zip codes with lower median incomes (**Table S1**). A detailed analysis of the  
198 characteristics of patients hospitalized in Houston Methodist facilities in the two  
199 waves has recently been published (17).

200 **SARS-CoV-2 genome sequencing and phylogenetic analysis.** To  
201 investigate the genomic architecture of the virus across the two waves, we  
202 sequenced the genomes of 5,085 SARS-CoV-2 strains dating to the earliest time  
203 of confirmed COVID-19 cases in Houston. Analysis of SARS-CoV-2 strains  
204 causing disease in the first wave (March 5 through May 11) identified the  
205 presence of many diverse virus genomes that, in the aggregate, represent the  
206 major clades identified globally to date (**Figure 1B**). Clades G, GH, GR, and S  
207 were the four most abundantly represented phylogenetic groups (**Figure 1B**).  
208 Strains with the Gly614 amino acid variant in spike protein represented 82% of  
209 the SARS-CoV-2 strains in wave 1, and 99.9% in wave 2 ( $p < 0.0001$ ; Fisher's  
210 exact test) (**Figure 1B**). This spike protein variant is characteristic of clades G,  
211 GH, and GR. Importantly, strains with the Gly614 variant represented only 71%  
212 of the specimens sequenced in March, the early part of wave 1 (**Figure 1B**). We  
213 attribute the decrease in strains with this variant observed in the first two weeks  
214 of March (**Figure 1B**) to fluctuation caused by the relatively fewer COVID-19  
215 cases occurring during this period.

216 **Relating spatiotemporal genome analysis with virus genotypes over**  
217 **two disease waves.** We examined the spatial and temporal mapping of genomic  
218 data to investigate community spread during wave 1 (**Figure 2**). Rapid and

219 widespread community dissemination occurred soon after the initial COVID-19  
220 cases were reported in Houston. The heterogenous virus genotypes present very  
221 early in wave 1 indicate that multiple strains independently entered metropolitan  
222 Houston, rather than introduction and spread of a single strain. An important  
223 observation was that strains of most of the individual subclades were distributed  
224 over broad geographic areas (**Figure S1**). These findings are consistent with the  
225 known ability of SARS-CoV-2 to spread very rapidly from person to person.

226 **Relationship between virus clades, clinical characteristics of infected**  
227 **patients, and additional metadata.** It is possible that SARS-CoV-2 genome  
228 subtypes have different clinical characteristics, analogous to what is believed to  
229 have occurred with Ebola virus (18-20) and known to occur for other pathogenic  
230 microbes (21). As an initial examination of this issue in SARS-CoV-2, we tested  
231 the hypothesis that patients with disease severe enough to warrant  
232 hospitalization were infected with a non-random subset of virus genotypes. We  
233 also examined the association between virus clades and disease severity based  
234 on overall mortality, highest level of required care (intensive care unit,  
235 intermediate care unit, inpatient or outpatient), need for mechanical ventilation,  
236 and length of stay. There was no simple relationship between virus clades and  
237 disease severity using these four indicators. Similarly, there was no simple  
238 relationship between virus clades and other metadata, such as sex, age, or  
239 ethnicity (**Figure S2**).

240 **Machine learning analysis.** Machine learning models can be used to  
241 identify complex relationships not revealed by statistical analyses. We built

242 machine learning models to test the hypothesis that virus genome sequence can  
243 predict patient outcomes including mortality, length of stay, level of care, ICU  
244 admission, supplemental oxygen use, and mechanical ventilation. Models to  
245 predict outcomes based on virus genome sequence alone resulted in low F1  
246 scores less than 50% (0.41 – 0.49) and regression models showed similarly low  
247  $R^2$  values (-0.01 – -0.20) (**Table S2**). F1 scores near 50% are indicative of  
248 classifiers that are performing similarly to random chance. The use of patient  
249 metadata alone to predict patient outcome improved the model's F1 scores by 5-  
250 10% (0.51 – 0.56) overall. The inclusion of patient metadata with virus genome  
251 sequence data improved most predictions of outcomes, compared to genome  
252 sequence alone, to 50% to 55% F1 overall (0.42 – 0.55) in the models (**Table**  
253 **S2**). The findings are indicative of two possibilities that are not mutually  
254 exclusive. First, patient metadata, such as age and sex, may provide more signal  
255 for the model to use and thus result in better accuracies. Second, the model's  
256 use of single nucleotide polymorphisms (SNPs) may have resulted in overfitting.  
257 Most importantly, no SNP predicted a significant difference in outcome. A table of  
258 classifier accuracy scores and performance information is provided in **Table S2**.

259 **Patient outcome and metadata correlations.** Overall, very few metadata  
260 categories correlated with patient outcomes (**Table S3**). Mortality was  
261 independently correlated with increasing age, with a Pearson correlation  
262 coefficient (PCC) equal to 0.27. This means that 27% of the variation in mortality  
263 can be predicted from patient age. Length of stay correlated independently with

264 increasing age (PCC=0.20). All other patient metadata correlations to outcomes  
265 had PCC less than 0.20 (**Table S3**).

266 We further analyzed outcomes correlated to isolates from wave 1 and 2,  
267 and the presence of the Gly614 variant in spike protein. Being in wave 1 was  
268 independently correlated with mechanical ventilation days, overall length of stay,  
269 and ICU length of stay, with PCC equal to 0.20, 0.18, and 0.14, respectively.  
270 Importantly, the presence of the Gly614 variant did not correlate with patient  
271 outcomes (**Table S3**).

272 **Analysis of the *nsp12* polymerase gene.** The SARS-CoV-2 genome  
273 encodes an RNA-dependent RNA polymerase (RdRp, also referred to as Nsp12)  
274 used in virus replication (22-25). Two amino acid substitutions (Phe479Leu and  
275 Val556Leu) in RdRp each confer significant resistance *in vitro* to remdesivir, an  
276 adenosine analog (26). Remdesivir is inserted into RNA chains by RdRp during  
277 replication, resulting in premature termination of RNA synthesis and inhibition of  
278 virus replication. This compound has shown prophylactic and therapeutic benefit  
279 against MERS-CoV and SARS-CoV-2 experimental infection in rhesus  
280 macaques (27, 28). Recent reports indicate that remdesivir has therapeutic  
281 benefit in some COVID-19 hospitalized patients (29-33), leading it to be now  
282 widely used in patients worldwide. Thus, it may be important to understand  
283 variation in RdRp in large strain samples.

284 To acquire data about allelic variation in the *nsp12* gene, we analyzed our  
285 5,085 virus genomes. The analysis identified 265 SNPs, including 140  
286 nonsynonymous (amino acid-altering) SNPs, resulting in amino acid

287 replacements throughout the protein (**Table 1, Figure 3, Figure 4, Figure S3,**  
288 **and Figure S4**). The most common amino acid change was Pro322Leu,  
289 identified in 4,893 of the 5,085 (96%) patient isolates. This amino acid  
290 replacement is common in genomes from clades G, GH, and GR, which are  
291 distinguished from other SARS-CoV-2 clades by the presence of the Gly614  
292 amino acid change in the spike protein. Most of the other amino acid changes in  
293 RdRp were present in relatively small numbers of strains, and some have been  
294 identified in other isolates in a publicly available database (34). Five prominent  
295 exceptions included amino acid replacements: Ala15Val in 138 strains, Met462Ile  
296 in 59 strains, Met600Ile in 75 strains, Thr907Ile in 45 strains, and Pro917Ser in  
297 80 strains. All 75 Met600Ile strains were phylogenetically closely related  
298 members of clade G, and also had the Pro322Leu amino acid replacement  
299 characteristic of this clade (**Figure S3**). These data indicate that the Met600Ile  
300 change is likely the evolved state, derived from a precursor strain with the  
301 Pro322Leu replacement. Similarly, we investigated phylogenetic relationships  
302 among strains with the other four amino acid changes noted above. In all cases,  
303 the vast majority of strains with each amino acid replacement were found among  
304 individual subclades of strains (**Figure S3**).

305       Importantly, none of the observed amino acid polymorphisms in RdRp  
306 were located precisely at two sites known to cause *in vitro* resistance to  
307 remdesivir (26). Most of the amino acid changes are located distantly from the  
308 RNA-binding and catalytic sites (**Figure S4 and Table 1**). However,  
309 replacements at six amino acid residues (Ala442Val, Ala448Val, Ala553Pro/Val,

310 Gly682Arg, Ser758Pro, and Cys812Phe) may potentially interfere with either  
311 remdesivir binding or RNA synthesis. Four (Ala442Val, Ala448Val,  
312 Ala553Pro/Val, and Gly682Arg) of the six substitution sites are located  
313 immediately above the nucleotide-binding site, that is comprised of Lys544,  
314 Arg552, and Arg554 residues as shown by structural studies (**Figure 4**). The  
315 positions of these four variant amino acid sites are comparable to Val556 (**Figure**  
316 **4**), for which a Val556Leu mutation in SARS-CoV was identified to confer  
317 resistance to remdesivir *in vitro* (26). The other two substitutions (Ser758Pro and  
318 Cys812Phe) are inferred to be located either at, or in the immediate proximity of,  
319 the catalytic active site, that is comprised of three contiguous residues (Ser758,  
320 Asp759, and Asp760). A proline substitution we identified at Ser758 (Ser758Pro)  
321 is likely to negatively impact RNA synthesis. Although Cys812 is not directly  
322 involved in the catalysis of RNA synthesis, it is only 3.5 Å away from Asp760.  
323 The introduction of the bulkier phenylalanine substitution at Cys812 (Cys812Phe)  
324 may impair RNA synthesis. Consequently, these two substitutions are expected  
325 to detrimentally affect virus replication or fitness.

326 **Analysis of the gene encoding the spike protein.** The densely glycosylated  
327 spike protein of SARS-CoV-2 and its close coronavirus relatives binds directly to  
328 host-cell angiotensin-converting enzyme 2 (ACE2) receptors to enter host cells  
329 (35-37). Thus, the spike protein is a major translational research target, including  
330 intensive vaccine and therapeutic antibody (35-64). Analysis of the gene  
331 encoding the spike protein identified 470 SNPs, including 285 that produce  
332 amino acid changes (**Table 2, Figure 5**). Forty-nine of these replacements

333 (V11A, T51A, W64C, I119T, E156Q, S205A, D228G, L229W, P230T, N234D,  
334 I235T, T274A, A288V, E324Q, E324V, S325P, S349F, S371P, S373P, T385I,  
335 A419V, C480F, Y495S, L517F, K528R, Q628E, T632I, S708P, T719I, P728L,  
336 S746P, E748K, G757V, V772A, K814R, D843N, S884A, M902I, I909V, E918Q,  
337 S982L, M1029I, Q1142K, K1157M, Q1180R, D1199A, C1241F, C1247G, and  
338 V1268A) are not represented in a publicly available database (34) as of August  
339 19, 2020. Interestingly, 25 amino acid sites have three distinct variants (that is,  
340 the reference amino acid plus two additional variant amino acids), and five amino  
341 acid sites (amino acid positions 21, 27, 228, 936, and 1050) have four distinct  
342 variants represented in our sample of 5,085 genomes (**Table 2, Figure 5**).

343 We mapped the location of amino acid replacements onto a model of the  
344 full-length spike protein (35, 65) and observed that the substitutions are found in  
345 each subunit and domain of the spike (**Figure 6**). However, the distribution of  
346 amino acid changes is not uniform throughout the protein regions. For example,  
347 compared to some other regions of the spike protein, the RBD has relatively few  
348 amino acid changes, and the frequency of strains with these substitutions is low,  
349 each occurring in fewer than 10 isolates. This finding is consistent with the  
350 functional constraints on RBD to mediate interaction with ACE2. In contrast, the  
351 periphery of the S1 subunit NTD contains a dense cluster of substituted residues,  
352 with some single amino acid replacements found in 10–20 isolates (**Table 2,**  
353 **Figure 5, Figure 6**). Clustering of amino acid changes in a distinct region of the  
354 spike protein may be a signal of positive selection. Inasmuch as infected patients  
355 make antibodies against the NTD, we favor the idea that host immune selection



356 is one force contributing to some of the amino acid variation in this region. One  
357 NTD substitution, H49Y, was found in 142 isolates. This position is not well  
358 exposed on the surface of the NTD and is likely not a result of immune pressure.  
359 The same is true for another highly represented substitution, F1052L. This  
360 substitution was observed in 167 isolates, and F1052 is buried within the core of  
361 the S2 subunit. The substitution observed most frequently in the spike protein in  
362 our sample is D614G, a change observed in 4,895 of the isolates. As noted  
363 above, strains with the Gly614 variant significantly increased in wave 2 compared  
364 to wave 1.

365 As observed with RdRp, the majority of strains with each single amino  
366 acid change in the spike protein were found on a distinct phylogenetic lineage  
367 (**Figure S5**), indicating identity by descent. A prominent exception is the  
368 Leu5Phe replacement that is present in all major clades, suggesting that this  
369 amino acid change arose multiple times independently or very early in the course  
370 of SARS-CoV-2 evolution. Finally, we note that examination of the phylogenetic  
371 distribution of strains with multiple distinct amino acid replacements at the same  
372 site (e.g., Arg21Ile/Lys/Thr, Ala27Ser/Thr/Val, etc.) revealed that they were  
373 commonly found in different genetic branches, consistent with independent origin  
374 (**Figure S5**).

375 **Cycle threshold (Ct) comparison of SARS-CoV-2 strains with either**  
376 **the Asp614 or Gly614 amino acid replacements in spike protein.** It has been  
377 reported that patients infected with strains having spike protein Gly614 variant  
378 have, on average, higher virus loads on initial diagnosis (66-70). To determine if

379 this is the case in Houston strains, we examined the cycle threshold (Ct) for  
380 every sequenced strain that was detected from a patient specimen using the  
381 SARS-CoV-2 Assay done by the Hologic Panther instrument. We identified a  
382 significant difference ( $p < 0.0001$ ) between the mean Ct value for strains with an  
383 Asp614 ( $n=102$ ) or Gly614 ( $n=812$ ) variant of the spike protein (**Figure 7**).  
384 Strains with Gly614 had a Ct value significantly lower than strains with the  
385 Asp614 variant, indicating that patients infected with the Gly614 strains had, on  
386 average, higher virus loads on initial diagnosis than patients infected by strains  
387 with the Asp614 variant (**Figure 7**). This observation is consistent with the  
388 conjecture that, on average, strains with the Gly614 variant are better able to  
389 disseminate.

390 **Characterization of recombinant proteins with single amino acid**  
391 **replacements in the receptor binding domain region of spike protein.** The  
392 RBD of spike protein binds the ACE2 surface receptor and is also targeted by  
393 neutralizing (36, 37, 41, 43-46, 48-62, 71). Thus, single amino acid replacements  
394 in this domain may have functional consequences that enhance virus fitness. To  
395 begin to test this idea, we expressed spike variants with the Asp614Gly  
396 replacement and 13 clinical RBD variants identified in our genome sequencing  
397 studies (**Figure 8, Table S4A, B**). All RBD variants were cloned into an  
398 engineered spike protein construct that stabilizes the prefusion state and  
399 increases overall expression yield (spike-6P, here referred to as spike) (64).

400 We first assessed the biophysical properties of spike-Asp614Gly, an  
401 amino acid polymorphism that is common globally and increased significantly in  
402 our wave 2 strain isolates. Pseudotyped viruses expressing spike-Gly614 have  
403 higher infectivity for host cells *in vitro* than spike-Asp614 (66, 67, 69, 72, 73). The  
404 higher infectivity of spike-Gly614 is correlated with increased stability and  
405 incorporation of the spike protein into the pseudovirion (73). We observed a  
406 higher expression level (**Figure 8A, B**) and increased thermostability for the  
407 spike protein construct containing this variant (**Figure 8C, D**). The size exclusion  
408 chromatography (SEC) elution profile of spike-Asp614 was indistinguishable from  
409 spike-Gly614, consistent with a trimeric conformation (**Figure 8A**). These results  
410 are broadly consistent with higher-resolution structural analyses of both spike  
411 variants.

412 Next, we purified and biophysically characterized 13 RBD mutants that  
413 each contain Gly614 and one additional single amino acid replacement we  
414 identified by genome sequencing our clinical samples (**Table S4C**). All variants  
415 eluted as trimers, indicating the global structure, remained intact (**Figure 8** and  
416 **Figure S6**). However, several variants had reduced expression levels and  
417 virtually all had decreased thermostability relative to the variant that had only a  
418 D614G single amino acid replacement (**Figure 8D**). The A419V and A522V  
419 mutations were especially deleterious, reducing yield and precluding further  
420 downstream analysis (**Figure 8B**). We next assayed the affinity of the 11 highest-  
421 expressing spike variants for ACE2 and the neutralizing monoclonal antibody  
422 CR3022 via enzyme-linked immunosorbent assays (ELISAs) (**Figure 8E-G** and

423 **Table S4C**). Most variants retained high affinity for the ACE2 surface receptor.  
424 However, importantly, three RBD variants (F338L, S373P, and R408T) had  
425 substantially reduced affinity for CR3022, a monoclonal antibody that disrupts the  
426 spike protein homotrimerization interface (63, 74). Notably, the S373P mutation  
427 is one amino acid away from the epitope recognized by CR3022. These results  
428 are consistent with the interpretation that some RBD mutants arising in COVID-  
429 19 patients may have increased ability to escape humoral immune pressure, but  
430 otherwise retain strong ACE2 binding affinity.

431

## 432 **DISCUSSION**

433 In this work we analyzed the molecular population genomics, sociodemographic,  
434 and medical features of two waves of COVID-19 disease occurring in  
435 metropolitan Houston, Texas, between early March and early July 2020. We also  
436 studied the biophysical and immunologic properties of some naturally occurring  
437 single amino acid changes in the spike protein RBD identified by sequencing the  
438 5,085 genomes. We discovered that the first COVID-19 wave was caused by a  
439 heterogenous array of virus genotypes assigned to several different clades. The  
440 majority of cases in the first wave are related to strains that caused widespread  
441 disease in European and Asian countries, as well as other localities. We  
442 conclude that the SARS-CoV-2 virus was introduced into Houston many times  
443 independently, likely by individuals who had traveled to or from different parts of  
444 the world, including other communities in the United States. In support of this  
445 conclusion, the first cases in metropolitan Houston were associated with a travel

446 history to a known COVID-19 region (16). The data are consistent with the fact  
447 that Houston is a large international city characterized by a multi-ethnic  
448 population and is a prominent transport hub with direct flights to major cities  
449 globally.

450 The second wave of COVID-19 cases also is characterized by SARS-  
451 CoV-2 strains with diverse genotypes. Virtually all cases in the second and  
452 ongoing disease wave were caused by strains with the Gly614 variant of spike  
453 protein (**Figure 1B**). Our data unambiguously demonstrate that strains with the  
454 Gly614 variant increased significantly in frequency in wave 2 relative to wave 1 in  
455 the Houston metropolitan region. This shift occurred very rapidly in a matter of  
456 just a few months. Amino acid residue Asp614 is located in subdomain 2 (SD-2)  
457 of the spike protein and forms a hydrogen bond and electrostatic interaction with  
458 two residues in the S2 subunit of a neighboring protomer. Replacement of  
459 aspartate with glycine would eliminate both interactions, thereby substantively  
460 weakening the contact between the S1 and S2 subunits. We previously  
461 speculated (75) that this weakening produces a more fusogenic spike protein, as  
462 S1 must first dissociate from S2 before S2 can refold and mediate fusion of virus  
463 and cell membranes. Stated another way, virus strains with the Gly614 variant  
464 may be better able to enter host cells, potentially resulting in enhanced spread.  
465 Consistent with this idea, Korber et al. (66) showed that the Gly614 variant grows  
466 to higher titer as pseudotyped virions. On initial diagnosis infected individuals had  
467 lower RT-PCR cycle thresholds suggesting higher upper respiratory tract viral  
468 loads. Our data (**Figure 7**) are fully consistent with that finding Zhang et al. (73)

469 reported that pseudovirus with the 614Gly variant infected ACE2-expressing cells  
470 more efficiently than the 614Asp. Similar results have been described by Hu et  
471 al. (67) and Lorenzo-Redondo et al. (68). Plante et al. (76) recently studied  
472 isogenic mutant SARS-CoV-2 strains with either the 614Asp or 614Gly variant  
473 and found that the 614Gly variant virus had significantly increased replication in  
474 human lung epithelial cells *in vitro* and increased infectious titers in nasal and  
475 trachea washes obtained from experimentally infected hamsters. These results  
476 are consistent with the idea that the 614Gly variant bestows increased virus  
477 fitness in the upper respiratory tract (76).

478 Additional work is needed to investigate the potential biomedical relevance  
479 and public health importance of the Asp614Gly polymorphism, including but not  
480 limited to virus dissemination, overall fitness, impact on clinical course and  
481 virulence, and development of vaccines and therapeutics. Although it is possible  
482 that stochastic processes alone may account for the rapid increase in COVID-19  
483 disease frequency caused by viruses containing the Gly614 variant, we do not  
484 favor that interpretation in part because of the cumulative weight of the  
485 epidemiologic, human RT-PCR diagnostics data, *in vitro* experimental findings,  
486 and animal infection studies using isogenic mutant virus strains. In addition, if  
487 stochastic processes solely are responsible, we believe it is difficult to explain  
488 essentially simultaneous increase in frequency of the Gly614 variant in  
489 genetically diverse viruses in three distinct clades (G, GH, and GR) in a  
490 geographically large metropolitan area with 7 million ethnically diverse people.  
491 Regardless, more research on this important topic is warranted.

492           The diversity present in our 1,026 virus genomes from the first disease  
493 wave contrasts somewhat with data reported by Gonzalez-Reiche et al., who  
494 studied 84 SARS-CoV-2 isolates causing disease in patients in the New York  
495 City region (11). Those investigators concluded that the vast majority of disease  
496 was caused by progeny of strains imported from Europe. Similarly, Bedford et al.  
497 (10) reported that much of the COVID-19 disease in the Seattle, Washington  
498 area was caused by strains that are progeny of a virus strain recently introduced  
499 from China. Some aspects of our findings are similar to those reported recently  
500 by Lemieux et al. based on analysis of strains causing disease in the Boston  
501 area (81). Our findings, like theirs, highlight the importance of multiple  
502 importation events of genetically diverse strains in the epidemiology of COVID-19  
503 disease in this pandemic. Similarly, Icelandic and Brazilian investigators  
504 documented that SARS-CoV-2 was imported by individuals traveling to or from  
505 many European and other countries (82, 83).

506           The virus genome diversity and large sample size in our study permitted  
507 us to test the hypothesis that distinct virus clades were nonrandomly associated  
508 with hospitalized COVID-19 patients or disease severity. We did not find  
509 evidence to support this hypothesis, but our continuing study of COVID-19 cases  
510 accruing in the second wave will further improve statistical stratification.

511           We used machine learning classifiers to identify if any SNPs contribute to  
512 increased infection severity or otherwise affect virus-host outcome. The models  
513 could not be trained to accurately predict these outcomes from the available virus  
514 genome sequence data. This may be due to sample size or class imbalance.

515 However, we do not favor this interpretation. Rather, we think that the inability to  
516 identify particular virus SNPs predictive of disease severity or infection outcome  
517 likely reflects the substantial heterogeneity in underlying medical conditions and  
518 treatment regimens among COVID-19 patients studied herein. An alternative but  
519 not mutually exclusive hypothesis is that patient genotypes play an important role  
520 in determining virus-human interactions and resulting pathology. Although some  
521 evidence has been presented in support of this idea (84, 85), available data  
522 suggest that in the aggregate, host genetics does not play an overwhelming role  
523 in determining outcome in the great majority of adult patients, once virus infection  
524 is established.

525 Remdesivir is a nucleoside analog reported to have activity against  
526 MERS-CoV, a coronavirus related to SARS-CoV-2. Recently, several studies  
527 have reported that remdesivir shows promise in treating COVID-19 patients (29-  
528 33), leading the FDA to issue an emergency use authorization. Because *in vitro*  
529 resistance of SARS-CoV to remdesivir has been reported to be caused by either  
530 of two amino acid replacements in RdRp (Phe479Leu or Val556Leu), we  
531 interrogated our data for polymorphisms in the *nsp12* gene. Although we  
532 identified 140 different inferred amino acid replacements in RdRp in the 5,085  
533 genomes analyzed, none of these were located precisely at the two positions  
534 associated with *in vitro* resistance to remdesivir. Inasmuch as remdesivir is now  
535 being deployed widely to treat COVID-19 patients in Houston and elsewhere, our  
536 findings suggest that the majority of SARS-CoV-2 strains currently circulating in  
537 our region should be susceptible to this drug.



538           The amino acid replacements Ala442Val, Ala448Val, Ala553Pro/Val, and  
539 Gly682Arg that we identified occur at sites that, intriguingly, are located directly  
540 above the nucleotide substrate entry channel and nucleotide binding residues  
541 Lys544, Arg552, and Arg554 (22, 23) (**Figure 4**). One possibility is that  
542 substitution of the smaller alanine or glycine residues with the bulkier side chains  
543 of Val/Pro/Arg may impose structural constraints for the modified nucleotide  
544 analog to bind, and thereby disfavor remdesivir binding. This, in turn, may lead to  
545 reduced incorporation of remdesivir into the nascent RNA, increased fidelity of  
546 RNA synthesis, and ultimately drug resistance. A similar mechanism has been  
547 proposed for a Val556Leu change (23).

548           We also identified one strain with a Lys477Asn replacement in RdRp. This  
549 substitution is located close to a Phe479Leu replacement reported to produce  
550 partial resistance to remdesivir *in vitro* in SARS-CoV patients from 2004,  
551 although the amino acid positions are numbered differently in SARS-CoV and  
552 SARS-CoV-2. Structural studies have suggested that this amino acid is surface-  
553 exposed, and distant from known key functional elements. Our observed  
554 Lys477Asn change is also located in a conserved motif described as a finger  
555 domain of RdRp (**Figure 3 and 4**). One speculative possibility is that Lys477 is  
556 involved in binding a yet unidentified cofactor such as Nsp7 or Nsp8, an  
557 interaction that could modify nucleotide binding and/or fidelity at a distance.  
558 These data warrant additional study in larger patient cohorts, especially in  
559 individuals treated with remdesivir.

560 Analysis of the gene encoding the spike protein identified 285 polymorphic  
561 amino acid sites relative to the reference genome, including 49 inferred amino  
562 acid replacements not present in available databases as of August 19, 2020.  
563 Importantly, 30 amino acid sites in the spike protein had two or three distinct  
564 replacements relative to the reference strain. The occurrence of multiple variants  
565 at the same amino acid site is one characteristic that may suggest functional  
566 consequences. These data, coupled with structural information available for  
567 spike protein, raise the possibility that some of the amino acid variants have  
568 functional consequences, for example including altered serologic reactivity and  
569 shown here. These data permit generation of many biomedically relevant  
570 hypotheses now under study.

571 A recent study reported that RBD amino acid changes could be selected  
572 *in vitro* using a pseudovirus neutralization assay and sera obtained from  
573 convalescent plasma or monoclonal antibodies (86). The amino acid sites  
574 included positions V445 and E484 in the RBD. Important to note, variants G446V  
575 and E484Q were present in our patient samples. However, these mutations  
576 retain high affinity to CR3022 (**Figure 8F, G**). The high-resolution structure of the  
577 RBD/CR3022 complex shows that CR3022 makes contacts to residues 369-386,  
578 380-392, and 427-430 of RBD (74). Although there is no overlap between  
579 CR3022 and ACE2 epitopes, CR3022 is able to neutralize the virus through an  
580 allosteric effect. We found that the Ser373Pro change, which is located within the  
581 CR3022 epitope, has reduced affinity to CR3022 (**Figure 8F, G**). The F338L and  
582 R408T mutations, although not found directly within the interacting epitope, also

583 display reduced binding to CR3022. Other investigators (86) using *in vitro*  
584 antibody selection identified a change at amino acid site S151 in the N-terminal  
585 domain, and we found mutations S151N and S151I in our patient samples. We  
586 also note that two variant amino acids (Gly446Val and Phe456Leu) we identified  
587 are located in a linear epitope found to be critical for a neutralizing monoclonal  
588 antibody described recently by Li et al. (87).

589 In the aggregate, these findings suggest that mutations emerging within  
590 the spike protein at positions within and proximal to known neutralization  
591 epitopes may result in escape from antibodies and other therapeutics currently  
592 under development. Importantly, our study did not reveal that these mutant  
593 strains had disproportionately increased over time. The findings may also bear  
594 on the occurrence of multiple amino acid substitutions at the same amino acid  
595 site that we identified in this study, commonly a signal of selection. In the  
596 aggregate, the data support a multifaceted approach to serological monitoring  
597 and biologics development, including the use of monoclonal antibody cocktails  
598 (46, 47, 88).

599

## 600 **CONCLUDING STATEMENT**

601 Our work represents analysis of the largest sample to date of SARS-CoV-2  
602 genome sequences from patients in one metropolitan region in the United States.  
603 The investigation was facilitated by the fact that we had rapidly assessed a  
604 SARS-CoV-2 molecular diagnostic test in January 2020, more than a month  
605 before the first COVID-19 patient was diagnosed in Houston. In addition, our

606 large healthcare system has seven hospitals and many facilities (e.g., outpatient  
607 care centers, emergency departments) located in geographically diverse areas of  
608 the city. We also provide reference laboratory services for other healthcare  
609 entities in the Houston area. Together, our facilities serve patients of diverse  
610 ethnicities and socioeconomic status. Thus, the data presented here likely reflect  
611 a broad overview of virus diversity causing COVID-19 infections throughout  
612 metropolitan Houston. We previously exploited these features to study influenza  
613 and *Klebsiella pneumoniae* dissemination in metropolitan Houston (89, 90). We  
614 acknowledge that every “twig” of the SARS-CoV-2 evolutionary tree in Houston is  
615 not represented in these data. The samples studied are not comprehensive for  
616 the entire metropolitan region. For example, it is possible that our strain samples  
617 are not fully representative of individuals who are indigent, homeless, or of very  
618 low socioeconomic groups. In addition, although the strain sample size is  
619 relatively large compared to other studies, the sample represents only about 10%  
620 of all COVID-19 cases in metropolitan Houston documented in the study period.  
621 In addition, some patient samples contain relatively small amounts of virus  
622 nucleic acid and do not yield adequate sequence data for high-quality genome  
623 analysis. Thus, our data likely underestimate the extent of genome diversity  
624 present among SARS-CoV-2 causing COVID-19 and will not identify all amino  
625 acid replacements in the virus in this geographic region. It will be important to  
626 sequence and analyze the genomes of additional SARS-CoV-2 strains causing  
627 COVID-19 cases in the ongoing second massive disease wave in metropolitan  
628 Houston, and these studies are underway. Data of this type will be especially

629 important to have if a third and subsequent waves were to occur in metropolitan  
630 Houston, as it could provide insight into molecular and epidemiologic events  
631 contributing to them.

632         The genomes reported here are an important data resource that will  
633 underpin our ongoing study of SARS-CoV-2 molecular evolution, dissemination,  
634 and medical features of COVID-19 in Houston. As of August 19, 2020, there  
635 were 135,866 reported cases of COVID-19 in metropolitan Houston, and the  
636 number of cases is increasing daily. Although the full array of factors contributing  
637 to the massive second wave in Houston is not known, it is possible that the  
638 potential for increased transmissibility of SARS-CoV-2 with the Gly614 may have  
639 played a role, as well as changes in behavior associated with the Memorial Day  
640 and July 4<sup>th</sup> holidays, and relaxation of some of the social constraints imposed  
641 during the first wave. The availability of extensive virus genome data dating from  
642 the earliest reported cases of COVID-19 in metropolitan Houston, coupled with  
643 the database we have now constructed, may provide critical insights into the  
644 origin of new infection spikes and waves occurring as public health constraints  
645 are further relaxed, schools and colleges re-open, holidays occur, commercial air  
646 travel increases, and individuals change their behavior because of COVID-19  
647 “fatigue.” The genome data will also be useful in assessing ongoing molecular  
648 evolution in spike and other proteins as baseline herd immunity is generated,  
649 either by natural exposure to SARS-CoV-2 or by vaccination. The signal of  
650 potential selection contributing to some spike protein diversity and identification

651 of naturally occurring mutant RBD variants with altered serologic recognition  
652 warrant close attention and expanded study.

653

## 654 **MATERIALS AND METHODS**

655 **Patient specimens.** All specimens were obtained from individuals who  
656 were registered patients at Houston Methodist hospitals, associated facilities  
657 (e.g., urgent care centers), or institutions in the greater Houston metropolitan  
658 region that use our laboratory services. Virtually all individuals met the criteria  
659 specified by the Centers for Disease Control and Prevention to be classified as a  
660 person under investigation.

661

662 **SARS-CoV-2 molecular diagnostic testing.** Specimens obtained from  
663 symptomatic patients with a high degree of suspicion for COVID-19 disease were  
664 tested in the Molecular Diagnostics Laboratory at Houston Methodist Hospital  
665 using an assay granted Emergency Use Authorization (EUA) from the FDA  
666 ([https://www.fda.gov/medical-devices/emergency-situations-medical-](https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-diagnostic-testing-sars-cov-2#offeringtests)  
667 [devices/faqs-diagnostic-testing-sars-cov-2#offeringtests](https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-diagnostic-testing-sars-cov-2#offeringtests)). Multiple testing  
668 platforms were used, including an assay that follows the protocol published by  
669 the WHO ([https://www.who.int/docs/default-source/coronaviruse/protocol-v2-](https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf)  
670 [1.pdf](https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf)) using the EZ1 virus extraction kit and EZ1 Advanced XL instrument or  
671 QIASymphony DSP Virus kit and QIASymphony instrument for nucleic acid  
672 extraction and ABI 7500 Fast Dx instrument with 7500 SDS software for reverse  
673 transcription RT-PCR, the COVID-19 test using BioFire Film Array 2.0

674 instruments, the Xpert Xpress SARS-CoV-2 test using Cepheid GeneXpert  
675 Infinity or Cepheid GeneXpert Xpress IV instruments, the SARS-CoV-2 Assay  
676 using the Hologic Panther instrument, and the Aptima SARS-CoV-2 Assay using  
677 the Hologic Panther Fusion system. All assays were performed according to the  
678 manufacturer's instructions. Testing was performed on material obtained from  
679 nasopharyngeal or oropharyngeal swabs immersed in universal transport media  
680 (UTM), bronchoalveolar lavage fluid, or sputum treated with dithiothreitol (DTT).  
681 To standardize specimen collection, an instructional video was created for  
682 Houston Methodist healthcare workers  
683 (<https://vimeo.com/396996468/2228335d56>).

684

685 **Epidemiologic curve.** The number of confirmed COVID-19 positive cases  
686 was obtained from USAFacts.org ([https://usafacts.org/visualizations/coronavirus-](https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/)  
687 [covid-19-spread-map/](https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/)) for Austin, Brazoria, Chambers, Fort Bend, Galveston,  
688 Harris, Liberty, Montgomery, and Waller counties. Positive cases for Houston  
689 Methodist Hospital patients were obtained from our Laboratory Information  
690 System and plotted using the documented collection time.

691

692 **SARS-CoV-2 genome sequencing.** Libraries for whole virus genome  
693 sequencing were prepared according to version 1 or version 3 of the ARTIC  
694 nCoV-2019 sequencing protocol (<https://artic.network/ncov-2019>). Long reads  
695 were generated with the LSK-109 sequencing kit, 24 native barcodes (NBD104

696 and NBD114 kits), and a GridION instrument (Oxford Nanopore). Short reads  
697 were generated with the NexteraXT kit and NextSeq 550 instrument (Illumina).

698

699 **SARS-CoV-2 genome sequence analysis.** Consensus virus genome  
700 sequences from the Houston area isolates were generated using the ARTIC  
701 nCoV-2019 bioinformatics pipeline. Publicly available genomes and metadata  
702 were acquired through GISAID on August 19, 2020. GISAID sequences  
703 containing greater than 1% N characters, and Houston sequences with greater  
704 than 5% N characters were removed from consideration. Identical GISAID  
705 sequences originating from the same geographic location with the same  
706 collection date were also removed from consideration to reduce redundancy.  
707 Nucleotide sequence alignments for the combined Houston and GISAID strains  
708 were generated using MAFFT version 7.130b with default parameters (91).  
709 Sequences were manually curated in JalView (92) to trim the ends and to  
710 remove sequences containing spurious inserts. Phylogenetic trees were  
711 generated using FastTree with the generalized time-reversible model for  
712 nucleotide sequences (93). CLC Genomics Workbench (QIAGEN) was used to  
713 generate the phylogenetic tree figures.

714

715 **Geospatial mapping.** The home address zip code for all SARS-CoV-2  
716 positive patients was used to generate the geospatial maps. To examine  
717 geographic relatedness among genetically similar isolates, geospatial maps were  
718 filtered to isolates containing specific amino acid changes.



719

720           **Time series.** Geospatial data were filtered into wave 1 (3/5/2020-  
721 5/11/2020) and wave 2 (5/12/2020-7/7/2020) time intervals to illustrate the  
722 spread of confirmed SARS-CoV-2 positive patients identified over time.

723

724           **Machine learning.** Virus genome alignments and patient metadata were  
725 used to build models to predict patient metadata and outcomes using both  
726 classification models and regression. Metadata considered for prediction in the  
727 classification models included age, ABO and Rh blood type, ethnic group,  
728 ethnicity, sex, ICU admission, IMU admission, supplemental oxygen use, and  
729 ventilator use. Metadata considered for prediction in regression analysis included  
730 ICU length of stay, IMU length of stay, total length of stay, supplemental oxygen  
731 use, and ventilator use. Because sex, blood type, Rh factor, age, age decade,  
732 ethnicity, and ethnic group are features in the patient features and combined  
733 feature sets, models were not trained for these labels using patient and  
734 combined feature sets. Additionally, age, length of stay, IMU length of stay, ICU  
735 length of stay, mechanical ventilation days, and supplemental oxygen days were  
736 treated as regression problems and XGBoost regressors were built while the rest  
737 were treated as classification problems and XGBoost classifiers were built.

738           Three types of features were considered for training the XGBoost  
739 classifiers: alignment features, patient features, and the combination of alignment  
740 and patient features. Alignment features were generated from the consensus  
741 genome alignment such that columns containing ambiguous nucleotide bases

742 were removed to ensure the models did not learn patterns from areas of low  
743 coverage. These alignments were then one-hot encoded to form the alignment  
744 features. Patient metadata values were one-hot encoded with the exception of  
745 age, which remained as a raw integer value, to create the patient features.  
746 These metadata values consisted of age, ABO, Rh blood type, ethnic group,  
747 ethnicity, and sex. All three types of feature sets were used to train models that  
748 predict ICU length of stay, IMU length of stay, overall length of stay, days of  
749 supplemental oxygen therapy, and days of ventilator usage while only alignment  
750 features were used to train models that predict age, ABO, Rh blood type, ethnic  
751 group, ethnicity, and sex.

752 A ten-fold cross validation was used to train XGBoost models (94) as  
753 described previously (95, 96). Depths of 4, 8, 16, 32, and 64 were used to tune  
754 the models, but accuracies plateaued after a depth of 16. SciKit-Learn's (97)  
755 classification report and r2 score were then used to assess overall accuracy of  
756 the classification and regression models, respectively.

757

758 **Patient metadata correlations.** We encoded values into multiple columns  
759 for each metadata field for patients if metadata was available. For example, the  
760 ABO column was divided into four columns for A, B, AB, and O blood type.  
761 Those columns were encoded with a 1 for the patients' ABO type, with all other  
762 columns encoded with 0. This was repeated for all non-outcome metadata fields.  
763 Age, however, was not re-encoded, as the raw integer values were used. Each  
764 column was then correlated to the various outcome values for each patient

765 (deceased, ICU length, IMU length, length of stay, supplemental oxygen length,  
766 and ventilator length) to obtain a Pearson coefficient correlation value for each  
767 metadata label and outcome.

768

769 **Analysis of the *nsp12* polymerase and S protein genes.** The *nsp12*  
770 virus polymerase and S protein genes were analyzed by plotting SNP density in  
771 the consensus alignment using Python (Python v3.4.3, Biopython Package  
772 v1.72). The frequency of SNPs in the Houston isolates was assessed, along with  
773 amino acid changes for nonsynonymous SNPs.

774

775 **Cycle threshold (Ct) comparison of SARS-CoV-2 strains with either**  
776 **Asp614 or Gly614 amino acid replacements in the spike protein.** The cycle  
777 threshold (Ct) for every sequenced strain that was detected from a patient  
778 specimen using the SARS-CoV-2 Assay on the Hologic Panther instrument was  
779 retrieved from the Houston Methodist Hospital Laboratory Information System.  
780 Statistical significance between the mean Ct value for strains with an aspartate  
781 ( $n=102$ ) or glycine ( $n=812$ ) amino acid at position 614 of the spike protein was  
782 determined with the Mann-Whitney test (GraphPad PRISM 8).

783

784 **Creation and characterization of spike protein RBD variants.** Spike  
785 RBD variants were cloned into the spike-6P (HexaPro; F817P, A892P, A899P,  
786 A942P, K986P, V987P) base construct that also includes the D614G substitution  
787 (pIF638). Briefly, a segment of the gene encoding the RBD was excised with

788 EcoRI and NheI, mutagenized by PCR, and assembled with a HiFi DNA  
789 Assembly Cloning Kit (NEB).

790 FreeStyle 293-F cells (Thermo Fisher Scientific) were cultured and  
791 maintained in a humidified atmosphere of 37°C and 8% CO<sub>2</sub> while shaking at  
792 110-125rpm. Cells were transfected with plasmids encoding spike protein  
793 variants using polyethylenimine. Three hours post-transfection, 5µM kifunensine  
794 was added to each culture. Cells were harvested four days after transfection and  
795 the protein containing supernatant was separated from the cells by two  
796 centrifugation steps: 10 min at 500rcf and 20 min at 10,000rcf. Supernatants  
797 were kept at 4°C throughout. Clarified supernatant was loaded on a Poly-Prep  
798 chromatography column (Bio-Rad) containing Strep-Tactin Superflow resin (IBA),  
799 washed with five column volumes (CV) of wash buffer (100mM Tris-HCl pH 8.0,  
800 150mM NaCl; 1mM EDTA), and eluted with four CV of elution buffer (100mM  
801 Tris-HCl pH 8.0, 150mM NaCl, 1mM EDTA, 2.5mM d-Desthiobiotin). The eluate  
802 was spin-concentrated (Amicon Ultra-15) to 600µL and further purified via size-  
803 exclusion chromatography (SEC) using a Superose 6 Increase 10/300 column  
804 (G.E.) in SEC buffer (2mM Tris pH 8.0, 200mM NaCl and 0.02% NaN<sub>3</sub>). Proteins  
805 were concentrated to 300µL and stored in SEC buffer.

806 The RBD spike mutants chosen for analysis were all RBD amino acid  
807 mutants identified by our genome sequencing study as of June 15, 2020. We  
808 note that the exact boundaries of the RBD domain varies depending on the paper  
809 used as reference. We used the boundaries demarcated in Figure 1A of Cai et al.  
810 Science paper 21 July) (98) that have K528R located at the RBD-CTD1 interface.

811

812           **Differential scanning fluorimetry.** Recombinant spike proteins were  
813 diluted to a final concentration of 0.05mg/mL with 5X SYPRO orange (Sigma) in  
814 a 96-well qPCR plate. Continuous fluorescence measurements ( $\lambda_{ex}=465\text{nm}$ ,  
815  $\lambda_{em}=580\text{nm}$ ) were collected with a Roche LightCycler 480 II. The temperature  
816 was increased from 22°C to 95°C at a rate of 4.4°C/min. We report the first  
817 melting transition.

818

819           **Enzyme-linked immunosorbent assays.** ELISAs were performed to  
820 characterize binding of S6P, S6P D614G, and S6P D614G-RBD variants to  
821 human ACE2 and the RBD-binding monoclonal antibody CR3022. The ACE2-  
822 hFc chimera was obtained from GenScript (Z03484), and the CR3022 antibody  
823 was purchased from Abcam (Ab273073). Corning 96-well high-binding plates  
824 (CLS9018BC) were coated with spike variants at 2 $\mu\text{g/mL}$  overnight at 4°C. After  
825 washing four times with phosphate buffered saline + 0.1% Tween20 (PBST;  
826 300 $\mu\text{L}$ /well), plates were blocked with PBS+2% milk (PBSM) for 2 h at room  
827 temperature and again washed four times with PBST. These were serially diluted  
828 in PBSM 1:3 seven times in triplicate. After 1 h incubation at room temperature,  
829 plates were washed four times in PBST, labeled with 50 $\mu\text{L}$  mouse anti-human  
830 IgG1 Fc-HRP (SouthernBlots, 9054-05) for 45 min in PBSM, and washed again  
831 in PBST before addition of 50 $\mu\text{L}$  1-step Ultra TMB-ELISA substrate (Thermo  
832 Scientific, 34028). Reactions were developed for 15 min and stopped by addition  
833 of 50 $\mu\text{L}$  4M H<sub>2</sub>SO<sub>4</sub>. Absorbance intensity (450nm) was normalized within a plate

834 and EC<sub>50</sub> values were calculated through 4-parameter logistic curve (4PL)  
835 analysis using GraphPad PRISM 8.4.3.

836

### 837 **ACKNOWLEDGMENTS**

838 We thank Dr. Steven Hinrichs and colleagues at the Nebraska Public Health  
839 Laboratory, and Dr. David Persse and colleagues at the Houston Health  
840 Department for providing samples used to validate our initial SARS-CoV-2  
841 molecular assay. We thank Drs. Jessica Thomas and Zejuan Li, Erika Walker,  
842 Concepcion C. Cantu, the very talented and dedicated molecular technologists,  
843 and the many labor pool volunteers in the Molecular Diagnostics Laboratory for  
844 their dedication to patient care. We also thank Brandi Robinson, Harrold Cano,  
845 Cory Romero, Brooke Burns, and Hayder Mahmood for technical assistance. We  
846 are indebted to Drs. Marc Boom and Dirk Sostman for their support, and to many  
847 very generous Houston philanthropists for their tremendous support of this  
848 ongoing project, including but not limited to anonymous, Ann and John Bookout  
849 III, Carolyn and John Bookout, Ting Tsung and Wei Fong Chao Foundation, Ann  
850 and Leslie Doggett, Freeport LNG, the Hearst Foundations, Jerold B. Katz  
851 Foundation, C. James and Carole Walter Looke, Diane and David Modesett, the  
852 Sherman Foundation, and Paula and Joseph C. “Rusty” Walter III. We gratefully  
853 acknowledge the originating and submitting laboratories of the SARS-CoV-2  
854 genome sequences from GISAID’s EpiFlu™ Database used in some of the work  
855 presented here. We also thank many colleagues for critical reading of the  
856 manuscript and suggesting improvements, and Sasha Pejerrey, Adrienne

857 Winston, Heather McConnell, and Kathryn Stockbauer for editorial contributions.  
858 We appreciate Dr. Stephen Schaffner for his helpful comments regarding the  
859 correlation analysis. We are especially indebted to Drs. Nancy Jenkins and Neal  
860 Copeland for their scholarly suggestions to improve an early version of the  
861 manuscript.

862

863 **Author contributions:** J.M.M. conceptualized and designed the project; S.W.L,  
864 R.J.O., P.A.C., D.W.B., J.J.D., M.S., M.N., M.O.S., C.C.C., P.Y., L.P., S.S., H.-C.  
865 K., H.H., G.E., H.A.T.N., J.H.L., M.K., J.G., D.B., J.G., J.S.M., C.-W.C., K.J., and  
866 I.F. performed research. All authors contributed to writing the manuscript.

867 Data and material availability: The spike-6P (“HexaPro”) plasmid is available from  
868 Addgene (ID: 154754) or from I.J.F. under a material transfer agreement with  
869 The University of Texas at Austin. Additional plasmids are available upon request  
870 from I.J.F.

871

872 This study was supported by the Fondren Foundation, Houston Methodist  
873 Hospital and Research Institute (to J.M.M.), NIH grant AI127521 (to J.S.M.), NIH  
874 grants GM120554 and GM124141 to I.J.F., the Welch Foundation (F-1808 to  
875 I.J.F.), and the National Science Foundation (1453358 to I.J.F.). I.J.F. is a CPRIT  
876 Scholar in Cancer Research. J.J.D., M.S., and M.N. are supported by the NIAID  
877 Bacterial and Viral Bioinformatics resource center award (contract number  
878 75N93019C00076).

879

880 **[Figure Legends]**

881 **FIG 1** (A) Confirmed COVID-19 cases in the Greater Houston Metropolitan  
882 region. Cumulative number of COVID-19 patients over time through July 7, 2020.  
883 Counties include Austin, Brazoria, Chambers, Fort Bend, Galveston, Harris,  
884 Liberty, Montgomery, and Waller. The shaded area represents the time period  
885 during which virus genomes characterized in this study were recovered from  
886 COVID-19 patients. The red line represents the number of COVID-19 patients  
887 diagnosed in the Houston Methodist Hospital Molecular Diagnostic Laboratory.  
888 (B) Distribution of strains with either the Asp614 or Gly614 amino acid variant in  
889 spike protein among the two waves of COVID-19 patients diagnosed in the  
890 Houston Methodist Hospital Molecular Diagnostic Laboratory. The large inset  
891 shows major clade frequency for the time frame studied.

892

893 **FIG 2** Sequential time-series heatmaps for all COVID-19 Houston Methodist  
894 patients during the study period. Geospatial distribution of COVID-19 patients is  
895 based on zip code. Panel A (left) shows geospatial distribution of sequenced  
896 SARS-CoV-2 strains in wave 1 and panel B (right) shows wave 2 distribution.  
897 The collection dates are shown at the bottom of each panel. The insets refer to  
898 numbers of strains in the color spectrum used. Note difference in numbers of  
899 strains used in panel A and panel B insets.

900



901 **FIG 3** Location of amino acid replacements in RNA-dependent RNA polymerase  
902 (RdRp/Nsp12) among the 5,085 genomes of SARS-CoV-2 sequenced. The  
903 various RdRp domains are color-coded. The numbers refer to amino acid site.  
904 Note that several amino acid sites have multiple variants identified.

905

906 **FIG 4** Amino acid changes identified in Nsp12 (RdRp) in this study that may  
907 influence interaction with remdesivir. The schematic at the top shows the domain  
908 architecture of Nsp12. (Left) Ribbon representation of the crystal structure of  
909 Nsp12-remdesivir monophosphate-RNA complex (PDB code: 7BV2). The  
910 structure in the right panel shows a magnified view of the boxed area in the left  
911 panel. The Nsp12 domains are colored as in the schematic at the top. The  
912 catalytic site in Nsp12 is marked by a black circle in the right panel. The side  
913 chains of amino acids comprising the catalytic site of RdRp (Ser758, Asp759,  
914 and Asp760) are shown as balls and stick and colored yellow. The nucleotide  
915 binding site is boxed in the right panel. The side chains of amino acids  
916 participating in nucleotide binding (Lys544, Arg552, and Arg554) are shown as  
917 balls and sticks and colored light blue. Remdesivir molecule incorporated into the  
918 nascent RNA is shown as balls and sticks and colored light pink. The RNA is  
919 shown as a blue cartoon and bases are shown as sticks. The positions of C $\alpha$   
920 atoms of amino acids identified in this study are shown as red and green spheres  
921 and labeled. The amino acids that are shown as red spheres are located above  
922 the nucleotide binding site, whereas Cys812 located at the catalytic site is shown

923 as a green sphere. The side chain of active site residue Ser758 is shown as ball  
924 and sticks and colored yellow. The location of C $\alpha$  atoms of remdesivir resistance  
925 conferring amino acid Val556 is shown as blue sphere and labeled.

926

927 **FIG 5** Location of amino acid replacements in spike protein among the 5,085  
928 genomes of SARS-CoV-2 sequenced. The various spike protein domains are  
929 color-coded. The numbers refer to amino acid site. Note that many amino acid  
930 sites have multiple variants identified.

931

932 **FIG 6** Location of amino acid substitutions mapped on the SARS-CoV-2 spike  
933 protein. Model of the SARS-CoV-2 spike protein with one protomer shown as  
934 ribbons and the other two protomers shown as a molecular surface. The C $\alpha$  atom  
935 of residues found to be substituted in one or more virus isolates identified in this  
936 study is shown as a sphere on the ribbon representation. Residues found to be  
937 substituted in 1–9 isolates are colored tan, 10–99 isolates yellow, 100–999  
938 isolates colored red (H49Y and F1052L), and >1000 isolates purple (D614G).  
939 The surface of the aminoterminal domain (NTD) that is distal to the trimeric axis  
940 has a high density of substituted residues. RBD, receptor binding domain.

941

942 **FIG 7** Cycle threshold (Ct) for every SARS-CoV-2 patient sample tested using  
943 the Hologic Panther assay. Data are presented as mean +/- standard error of the

944 mean for strains with an aspartate (D614,  $n=102$  strains, blue) or glycine  
945 (G614,  $n=812$  strains, red) at amino acid 614 of the spike protein. Mann-Whitney  
946 test,  $*P<0.0001$ .

947

948 **FIG 8** Biochemical characterization of spike RBD variants. (A) Size-exclusion  
949 chromatography (SEC) traces of the indicated spike-RBD variants. Dashed line  
950 indicates the elution peak of spike-6P. (B) The relative expression of all RBD  
951 variants as determined by the area under the SEC traces. All expression levels  
952 are normalized relative to spike-6P. (C) Thermostability analysis of RBD variants  
953 by differential scanning fluorimetry. Each sample had three replicates and only  
954 mean values were plotted. Black vertical dashed line indicates the first melting  
955 temperature of 6P-D614G and orange vertical dashed line indicates the first  
956 melting temperature of the least stable variant (spike-G446V). (D) First apparent  
957 melting temperature of all RBD variants. (E) ELISA-based binding affinities for  
958 ACE2 and (F) the neutralizing antibody CR3022 to the indicated RBD variants.  
959 (G) Summary of EC50s for all measured RBD variants.

960

961

962 **[Supplemental Figure Legends]**

963 **Supplemental FIG 1** Geographic distribution of representative SARS-CoV-2  
964 subclades in the Houston metropolitan region. Blue shaded areas denote zip  
965 codes containing COVID-19 cases with the designated subclade.

966

967 **Supplemental FIG 2** Cladograms showing distribution of patient metadata,  
968 including (A) age (in decade), (B) sex, (C) ethnicity/ethnic group, (D) wave, (E)  
969 level of care, (F) mechanical ventilation, (G) length of stay, and (H) mortality.

970

971 **Supplemental FIG 3** Distribution of subclades characterized by particular amino  
972 acid replacements in Nsp12 (RdRp).

973

974 **Supplemental FIG 4** Mapping the location of amino acid replacements on  
975 Nsp12 (RdRp) from COVID-19 virus. The schematic on the top shows the  
976 domain architecture of Nsp12. The individual domains of Nsp12 are color-coded  
977 and labeled. Ribbon representation of the crystal structure of Nsp12-remdesivir  
978 monophosphate-RNA complex is shown (PDB code: 7BV2). The structure in the  
979 right panel is obtained by rotating the left panel 180° along the y-axis. The Nsp12  
980 domains are colored as in the schematic at the top. The positions of C $\alpha$  atoms of  
981 the surface-exposed amino acids identified in this study are shown as yellow  
982 spheres, whereas the positions of C $\alpha$  atoms of the buried amino acids are  
983 depicted as cyan spheres. The catalytic site in RdRp is marked by a black circle  
984 in the right panel. The side chains of amino acids comprising the catalytic site of

985 RdRp are shown as balls and sticks and colored yellow. The nucleotide binding  
986 site is boxed and labeled in the right panel. The side chains of amino acids  
987 participating in nucleotide binding (Lys545, Arg553, and Arg555) are shown as  
988 balls and sticks. Remdesivir molecule incorporated into the nascent RNA is  
989 shown as balls and sticks and colored light pink. The RNA is shown as blue  
990 cartoon and bases are shown as sticks. The positions of C $\alpha$  atoms of amino  
991 acids that are predicted to influence remdesivir binding are shown as red  
992 spheres. The amino acid Cys812 located at the catalytic site is shown as green  
993 sphere. The location of C $\alpha$  atoms of remdesivir resistance conferring amino acid  
994 Val556 is shown as blue sphere and labeled.

995

996 **Supplemental FIG 5** Distribution of subclades characterized by particular amino  
997 acid replacements in spike protein.

998

999 **Supplemental FIG 6** Biochemical characterization of single amino acid variants  
1000 of spike protein RBD. (A, B) Size-exclusion chromatography (SEC) traces of the  
1001 indicated spike-RBD variants. Dashed line indicates the elution peak of spike-6P.  
1002 (C) Thermostability analysis of RBD variants. Each sample had three replicates  
1003 and only mean values were plotted. Black vertical dashed line indicates the first  
1004 melting temperature of 6P-D614G. (D) ELISA-based binding affinities for ACE2  
1005 and (E) neutralizing monoclonal antibody CR3022 to the indicated RBD variants.

1006

1007

1008 **[Supplemental Table Legends]**

1009 **Supplemental Table 1** Patient demographics in wave 1 and wave 2.

1010

1011 **Supplemental Table 2** Classifier accuracy scores and performance of machine  
1012 learning models.

1013

1014 **Supplemental Table 3** Pearson correlation coefficient data for correlation  
1015 analysis.

1016

1017 **Supplemental Table 4** Primers and plasmids used for the *in vitro*  
1018 characterization of recombinant proteins with single amino acid replacements in  
1019 the receptor binding domain (RBD) region of spike protein, and their biophysical  
1020 properties. To test the hypothesis that RBD amino acid changes enhance viral  
1021 fitness, we expressed spike variants with the Asp614Gly replacement and 13  
1022 clinical RBD variants identified in our genome sequencing studies. Table S4A  
1023 contains the primers used, Table S4B contains the plasmid construct information,  
1024 and Table S4C contains the biophysical properties of the resultant spike protein  
1025 variants.

## 1026 REFERENCES

- 1027 1. 2020. World Health Organization Coronavirus Disease 2019 (COVID-19) Situation Report.  
1028 [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200420-sitrep-91-](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200420-sitrep-91-covid-19.pdf?sfvrsn=fcf0670b_4)  
1029 [covid-19.pdf?sfvrsn=fcf0670b\\_4](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200420-sitrep-91-covid-19.pdf?sfvrsn=fcf0670b_4). Accessed April 21.
- 1030 2. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL,  
1031 Lauber C, Leontovich AM, Neuman BW, Penzar D, Perlman S, Poon LLM, Samborskiy DV,  
1032 Sidorov IA, Sola I, Ziebuhr J, Coronaviridae Study Group of the International Committee on  
1033 Taxonomy of V. 2020. The species Severe acute respiratory syndrome-related coronavirus:  
1034 classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* 5:536-544.
- 1035 3. Wang C, Horby PW, Hayden FG, Gao GF. 2020. A novel coronavirus outbreak of global health  
1036 concern. *Lancet* 395:470-473.
- 1037 4. Perlman S. 2020. Another Decade, Another Coronavirus. *New England Journal of Medicine*  
1038 382:760-762.
- 1039 5. Allel K, Tapia-Muñoz T, Morris W. 2020. Country-level factors associated with the early spread  
1040 of COVID-19 cases at 5, 10 and 15 days since the onset. *Glob Public Health*  
1041 doi:10.1080/17441692.2020.1814835:1-14.
- 1042 6. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia  
1043 J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R,  
1044 Gao Z, Jin Q, Wang J, Cao B. 2020. Clinical features of patients infected with 2019 novel  
1045 coronavirus in Wuhan, China. *Lancet* 395:497-506.
- 1046 7. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F,  
1047 Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A Novel Coronavirus from Patients with  
1048 Pneumonia in China, 2019. *New England Journal of Medicine* 382:727-733.
- 1049 8. Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, Xing F, Liu J, Yip CC, Poon RW, Tsoi HW,  
1050 Lo SK, Chan KH, Poon VK, Chan WM, Ip JD, Cai JP, Cheng VC, Chen H, Hui CK, Yuen KY.  
1051 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating  
1052 person-to-person transmission: a study of a family cluster. *Lancet* 395:514-523.
- 1053 9. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML,  
1054 Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new  
1055 coronavirus associated with human respiratory disease in China. *Nature* 579:265-269.
- 1056 10. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, Nalla A,  
1057 Pepper G, Reinhardt A, Xie H, Shrestha L, Nguyen TN, Adler A, Brandstetter E, Cho S, Giroux  
1058 D, Han PD, Fay K, Frazar CD, Ilcisin M, Lacombe K, Lee J, Kiavand A, Richardson M, Sibley  
1059 TR, Truong M, Wolf CR, Nickerson DA, Rieder MJ, Englund JA, Hadfield J, Hodcroft EB,  
1060 Huddleston J, Moncla LH, Müller NF, Neher RA, Deng X, Gu W, Federman S, Chiu C, Duchin J,  
1061 Gautom R, Melly G, Hiatt B, Dykema P, Lindquist S, Queen K, Tao Y, Uehara A, Tong S, et al.  
1062 2020. Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv*  
1063 doi:10.1101/2020.04.02.20051417:2020.04.02.20051417.
- 1064 11. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S,  
1065 Kleiner G, Polanco J, Khan Z, Albuquerque B, van de Guchte A, Dutta J, Francoeur N, Melo BS,  
1066 Oussenko I, Deikus G, Soto J, Sridhar SH, Wang Y-C, Twyman K, Kasarskis A, Altman DR,  
1067 Smith M, Sebra R, Aberg J, Krammer F, García-Sastre A, Luksza M, Patel G, Paniz-Mondolfi A,  
1068 Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and early spread of SARS-  
1069 CoV-2 in the New York City area. *Science* 369:297-301.
- 1070 12. Health N. 2020. COVID-19 Data. <https://www1.nyc.gov/site/doh/covid/covid-19-data.page>.  
1071 Accessed August 19.
- 1072 13. County K. 2020. Daily COVID-19 outbreak summary.  
1073 <https://www.kingcounty.gov/depts/health/covid-19/data/daily-summary.aspx>. Accessed August  
1074 18.
- 1075 14. Cline M, Emerson M, bratter j, howell j, Jeanty P. 2012. Houston Region Grows More  
1076 Racially/Ethnically Diverse, With Small Declines in Segregation. A Joint Report Analyzing Census  
1077 Data from 1990, 2000, and 2010.
- 1078 15. Emerson M, Bratter J, Howell J, Jeanty P, Cline M. 2012. Houston Region Grows More  
1079 Racially/Ethnically Diverse, With Small Declines in Segregation. A Joint Report Analyzing



- 1080 Census Data from 1990, 2000, and 2010. Kinder Institute for Urban Research & the Hobby  
1081 Center for the Study of Texas,  
1082 16. Services THaH. 2020. Texas Health and Human Services. <https://hhs.texas.gov/>. Accessed  
1083 August 18.  
1084 17. Vahidy FS, Drews AL, Masud FN, Schwartz RL, Askary BB, Boom ML, Phillips RA. 2020.  
1085 Characteristics and Outcomes of COVID-19 Patients During Initial Peak and Resurgence in the  
1086 Houston Metropolitan Area. *Jama* doi:10.1001/jama.2020.15301.  
1087 18. Diehl WE, Lin AE, Grubaugh ND, Carvalho LM, Kim K, Kyawe PP, McCauley SM, Donnard E,  
1088 Kucukural A, McDonel P, Schaffner SF, Garber M, Rambaut A, Andersen KG, Sabeti PC, Luban  
1089 J. 2016. Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013-2016 Epidemic.  
1090 *Cell* 167:1088-1098.e6.  
1091 19. Urbanowicz RA, McClure CP, Sakuntabhai A, Sall AA, Kobinger G, Müller MA, Holmes EC,  
1092 Rey FA, Simon-Loriere E, Ball JK. 2016. Human Adaptation of Ebola Virus during the West  
1093 African Outbreak. *Cell* 167:1079-1087.e5.  
1094 20. Dietzel E, Schudt G, Krähling V, Matrosovich M, Becker S. 2017. Functional Characterization of  
1095 Adaptive Mutations during the West African Ebola Virus Outbreak. *J Virol* 91.  
1096 21. Kachroo P, Eraso JM, Beres SB, Olsen RJ, Zhu L, Nasser W, Bernard PE, Cantu CC, Saavedra  
1097 MO, Arredondo MJ, Strobe B, Do H, Kumaraswami M, Vuopio J, Grondahl-Yli-Hannuksela K,  
1098 Kristinsson KG, Gottfredsson M, Pesonen M, Pensar J, Davenport ER, Clark AG, Corander J,  
1099 Caugant DA, Gaini S, Magnussen MD, Kubiak SL, Nguyen HAT, Long SW, Porter AR, DeLeo  
1100 FR, Musser JM. 2019. Integrated analysis of population genomics, transcriptomics and virulence  
1101 provides novel insights into *Streptococcus pyogenes* pathogenesis. *Nat Genet* 51:548-559.  
1102 22. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng  
1103 L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B, Yang X, Li J, Yang H, Liu Z, Xu W,  
1104 Guddat LW, Wang Q, Lou Z, Rao Z. 2020. Structure of the RNA-dependent RNA polymerase  
1105 from COVID-19 virus. *Science* doi:10.1126/science.abb7498:eabb7498.  
1106 23. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, Wang X, Zhou F, Zhao W, Gao M, Chang S,  
1107 Xie Y-C, Tian G, Jiang H-W, Tao S-C, Shen J, Jiang Y, Jiang H, Xu Y, Zhang S, Zhang Y, Xu  
1108 HE. 2020. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-  
1109 CoV-2 by remdesivir. *Science* 368:1499-1504.  
1110 24. Shannon A, Le NT, Selisko B, Eydoux C, Alvarez K, Guillemot JC, Decroly E, Peersen O, Ferron  
1111 F, Canard B. 2020. Remdesivir and SARS-CoV-2: Structural requirements at both nsp12 RdRp  
1112 and nsp14 Exonuclease active-sites. *Antiviral Res* 178:104793.  
1113 25. Gordon CJ, Tchesnokov EP, Woolner E, Perry JK, Feng JY, Porter DP, Gotte M. 2020.  
1114 Remdesivir is a direct-acting antiviral that inhibits RNA-dependent RNA polymerase from severe  
1115 acute respiratory syndrome coronavirus 2 with high potency. *J Biol Chem*  
1116 doi:10.1074/jbc.RA120.013679.  
1117 26. Agostini ML, Andres EL, Sims AC, Graham RL, Sheahan TP, Lu X, Smith EC, Case JB, Feng  
1118 JY, Jordan R, Ray AS, Cihlar T, Siegel D, Mackman RL, Clarke MO, Baric RS, Denison MR.  
1119 2018. Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral  
1120 Polymerase and the Proofreading Exoribonuclease. *mBio* 9.  
1121 27. de Wit E, Feldmann F, Cronin J, Jordan R, Okumura A, Thomas T, Scott D, Cihlar T, Feldmann  
1122 H. 2020. Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque  
1123 model of MERS-CoV infection. *Proc Natl Acad Sci U S A* 117:6771-6776.  
1124 28. Williamson BN, Feldmann F, Schwarz B, Meade-White K, Porter DP, Schulz J, van Doremalen  
1125 N, Leighton I, Yinda CK, Pérez-Pérez L, Okumura A, Lovaglio J, Hanley PW, Saturday G, Bosio  
1126 CM, Anzick S, Barbian K, Cihlar T, Martens C, Scott DP, Munster VJ, de Wit E. 2020. Clinical  
1127 benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. *Nature* 585:273-276.  
1128 29. Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, Feldt T, Green G, Green ML,  
1129 Lescure FX, Nicastri E, Oda R, Yo K, Quiros-Roldan E, Studemeister A, Redinski J, Ahmed S,  
1130 Bennett J, Chelliah D, Chen D, Chihara S, Cohen SH, Cunningham J, D'Arminio Monforte A,  
1131 Ismail S, Kato H, Lapadula G, L'Her E, Maeno T, Majumder S, Massari M, Mora-Rillo M, Mutoh  
1132 Y, Nguyen D, Verweij E, Zoufaly A, Osinusi AO, DeZure A, Zhao Y, Zhong L, Chokkalingam A,  
1133 Elboudwarej E, Telep L, Timbs L, Henne I, Sellers S, Cao H, Tan SK, Winterbourne L, Desai P,  
1134 et al. 2020. Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N Engl J Med*  
1135 doi:10.1056/NEJMoa2007016.



- 1136 30. Goldman JD, Lye DCB, Hui DS, Marks KM, Bruno R, Montejano R, Spinner CD, Galli M, Ahn  
1137 MY, Nahass RG, Chen YS, SenGupta D, Hyland RH, Osinusi AO, Cao H, Blair C, Wei X,  
1138 Gaggar A, Brainard DM, Towner WJ, Muñoz J, Mullane KM, Marty FM, Tashima KT, Diaz G,  
1139 Subramanian A. 2020. Remdesivir for 5 or 10 Days in Patients with Severe Covid-19. *N Engl J*  
1140 *Med* doi:10.1056/NEJMoa2015301.
- 1141 31. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, Hohmann E, Chu HY,  
1142 Luetkemeyer A, Kline S, Lopez de Castilla D, Finberg RW, Dierberg K, Tapson V, Hsieh L,  
1143 Patterson TF, Paredes R, Sweeney DA, Short WR, Touloumi G, Lye DC, Ohmagari N, Oh MD,  
1144 Ruiz-Palacios GM, Benfield T, Fätkenheuer G, Kortepeter MG, Atmar RL, Creech CB, Lundgren  
1145 J, Babiker AG, Pett S, Neaton JD, Burgess TH, Bonnett T, Green M, Makowski M, Osinusi A,  
1146 Nayak S, Lane HC. 2020. Remdesivir for the Treatment of Covid-19 - Preliminary Report. *N Engl*  
1147 *J Med* doi:10.1056/NEJMoa2007764.
- 1148 32. Spinner CD, Gottlieb RL, Criner GJ, Arribas López JR, Cattelan AM, Soriano Viladomiu A,  
1149 Ogbuagu O, Malhotra P, Mullane KM, Castagna A, Chai LYA, Roestenberg M, Tsang OTY,  
1150 Bernasconi E, Le Turnier P, Chang SC, SenGupta D, Hyland RH, Osinusi AO, Cao H, Blair C,  
1151 Wang H, Gaggar A, Brainard DM, McPhail MJ, Bhagani S, Ahn MY, Sanyal AJ, Huhn G, Marty  
1152 FM. 2020. Effect of Remdesivir vs Standard Care on Clinical Status at 11 Days in Patients With  
1153 Moderate COVID-19: A Randomized Clinical Trial. *Jama* doi:10.1001/jama.2020.16349.
- 1154 33. Olender SA, Perez KK, Go AS, Balani B, Price-Haywood EG, Shah NS, Wang S, Walunas TL,  
1155 Swaminathan S, Slim J, Chin B, De Wit S, Ali SM, Soriano Viladomiu A, Robinson P, Gottlieb  
1156 RL, Tsang TYO, Lee IH, Haubrich RH, Chokkalingam AP, Lin L, Zhong L, Bekele BN, Mera-  
1157 Giler R, Gallant J, Smith LE, Osinusi AO, Brainard DM, Hu H, Phulpin C, Edgar H, Diaz-Cuervo  
1158 H, Bernardino JI. 2020. Remdesivir for Severe COVID-19 versus a Cohort Receiving Standard of  
1159 Care. *Clin Infect Dis* doi:10.1093/cid/ciaa1041.
- 1160 34. (CNCB) CNCfB. 2020. 2019 Novel Coronavirus Resource (2019nCoV).  
1161 <https://bigd.big.ac.cn/ncov/about?lang=en>. Accessed August 19.
- 1162 35. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS.  
1163 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*  
1164 367:1260-1263.
- 1165 36. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, Function,  
1166 and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181:281-292.e6.
- 1167 37. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, Lu G, Qiao C, Hu Y, Yuen KY, Wang Q,  
1168 Zhou H, Yan J, Qi J. 2020. Structural and Functional Basis of SARS-CoV-2 Entry by Using  
1169 Human ACE2. *Cell* doi:10.1016/j.cell.2020.03.045.
- 1170 38. Jackson LA, Anderson EJ, Roush NG, Roberts PC, Makhene M, Coler RN, McCullough MP,  
1171 Chappell JD, Denison MR, Stevens LJ, Pruijssers AJ, McDermott A, Flach B, Doria-Rose NA,  
1172 Corbett KS, Morabito KM, O'Dell S, Schmidt SD, Swanson PA, 2nd, Padilla M, Mascola JR,  
1173 Neuzil KM, Bennett H, Sun W, Peters E, Makowski M, Albert J, Cross K, Buchanan W, Pikaart-  
1174 Tautges R, Ledgerwood JE, Graham BS, Beigel JH. 2020. An mRNA Vaccine against SARS-  
1175 CoV-2 - Preliminary Report. *N Engl J Med* doi:10.1056/NEJMoa2022483.
- 1176 39. Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, Bellamy D, Bibi S,  
1177 Bittaye M, Clutterbuck EA, Dold C, Faust SN, Finn A, Flaxman AL, Hallis B, Heath P, Jenkin D,  
1178 Lazarus R, Makinson R, Minassian AM, Pollock KM, Ramasamy M, Robinson H, Snape M,  
1179 Tarrant R, Voysey M, Green C, Douglas AD, Hill AVS, Lambe T, Gilbert SC, Pollard AJ. 2020.  
1180 Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a  
1181 preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* 396:467-478.
- 1182 40. Zhu FC, Guan XH, Li YH, Huang JY, Jiang T, Hou LH, Li JX, Yang BF, Wang L, Wang WJ, Wu  
1183 SP, Wang Z, Wu XH, Xu JJ, Zhang Z, Jia SY, Wang BS, Hu Y, Liu JJ, Zhang J, Qian XA, Li Q,  
1184 Pan HX, Jiang HD, Deng P, Gou JB, Wang XW, Wang XH, Chen W. 2020. Immunogenicity and  
1185 safety of a recombinant adenovirus type-5-vectored COVID-19 vaccine in healthy adults aged 18  
1186 years or older: a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet* 396:479-488.
- 1187 41. Brouwer PJM, Caniels TG, van der Straten K, Snitselaar JL, Aldon Y, Bangaru S, Torres JL, Okba  
1188 NMA, Claireaux M, Kerster G, Bentlage AEH, van Haaren MM, Guerra D, Burger JA, Schermer  
1189 EE, Verheul KD, van der Velde N, van der Kooi A, van Schooten J, van Breemen MJ, Bijl TPL,  
1190 Slieden K, Aartse A, Derking R, Bontjer I, Kootstra NA, Wiersinga WJ, Vidarsson G, Haagmans

- 1191 BL, Ward AB, de Bree GJ, Sanders RW, van Gils MJ. 2020. Potent neutralizing antibodies from  
1192 COVID-19 patients define multiple targets of vulnerability. *Science* 369:643-650.
- 1193 42. Chi X, Yan R, Zhang J, Zhang G, Zhang Y, Hao M, Zhang Z, Fan P, Dong Y, Yang Y, Chen Z,  
1194 Guo Y, Zhang J, Li Y, Song X, Chen Y, Xia L, Fu L, Hou L, Xu J, Yu C, Li J, Zhou Q, Chen W.  
1195 2020. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of  
1196 SARS-CoV-2. *Science* 369:650-655.
- 1197 43. Wec AZ, Wrapp D, Herbert AS, Maurer DP, Haslwanter D, Sakharkar M, Jangra RK, Dieterle  
1198 ME, Lilov A, Huang D, Tse LV, Johnson NV, Hsieh C-L, Wang N, Nett JH, Champney E,  
1199 Burnina I, Brown M, Lin S, Sinclair M, Johnson C, Pudi S, Bortz R, Wirchnianski AS,  
1200 Laudermilch E, Florez C, Fels JM, O'Brien CM, Graham BS, Nemazee D, Burton DR, Baric RS,  
1201 Voss JE, Chandran K, Dye JM, McLellan JS, Walker LM. 2020. Broad neutralization of SARS-  
1202 related viruses by human monoclonal antibodies. *Science* 369:731-736.
- 1203 44. Zost SJ, Gilchuk P, Case JB, Binshtein E, Chen RE, Nkolola JP, Schäfer A, Reidy JX, Trivette A,  
1204 Nargi RS, Sutton RE, Suryadevara N, Martinez DR, Williamson LE, Chen EC, Jones T, Day S,  
1205 Myers L, Hassan AO, Kafai NM, Winkler ES, Fox JM, Shrihari S, Mueller BK, Meiler J,  
1206 Chandrashekar A, Mercado NB, Steinhardt JJ, Ren K, Loo YM, Kallewaard NL, McCune BT,  
1207 Keeler SP, Holtzman MJ, Barouch DH, Gralinski LE, Baric RS, Thackray LB, Diamond MS,  
1208 Carnahan RH, Crowe JE, Jr. 2020. Potently neutralizing and protective human antibodies against  
1209 SARS-CoV-2. *Nature* 584:443-449.
- 1210 45. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J, Eguia  
1211 R, Crawford KH, Dingens AS, Nargi RS, Sutton RE, Suryadevara N, Rothlauf PW, Liu Z, Whelan  
1212 SP, Carnahan RH, Crowe JE, Bloom JD. 2020. Complete mapping of mutations to the SARS-  
1213 CoV-2 spike receptor-binding domain that escape antibody recognition. *bioRxiv*  
1214 doi:10.1101/2020.09.10.292078:2020.09.10.292078.
- 1215 46. Baum A, Copin R, Ajithdoss D, Zhou A, Lanza K, Negron N, Ni M, Wei Y, Atwal GS, Oyejide  
1216 A, Goetz-Gazi Y, Dutton J, Clemmons E, Staples HM, Bartley C, Klaffke B, Alfson K, Gazi M,  
1217 Gonzales O, Dick E, Carrion R, Pessaint L, Porto M, Cook A, Brown R, Ali V, Greenhouse J,  
1218 Taylor T, Andersen H, Lewis MG, Stahl N, Murphy AJ, Yancopoulos GD, Kyratsous CA. 2020.  
1219 REGN-COV2 antibody cocktail prevents and treats SARS-CoV-2 infection in rhesus macaques  
1220 and hamsters. *bioRxiv* doi:10.1101/2020.08.02.233320:2020.08.02.233320.
- 1221 47. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, Giordano S, Lanza K, Negron N, Ni  
1222 M, Wei Y, Atwal GS, Murphy AJ, Stahl N, Yancopoulos GD, Kyratsous CA. 2020. Antibody  
1223 cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual  
1224 antibodies. *Science* 369:1014-1018.
- 1225 48. Barnes CO, West AP, Jr., Huey-Tubman KE, Hoffmann MAG, Sharaf NG, Hoffman PR, Koranda  
1226 N, Gristick HB, Gaebler C, Muecksch F, Lorenzi JCC, Finkin S, Hägglöf T, Hurley A, Millard  
1227 KG, Weisblum Y, Schmidt F, Hatziioannou T, Bieniasz PD, Caskey M, Robbani DF,  
1228 Nussenzweig MC, Bjorkman PJ. 2020. Structures of Human Antibodies Bound to SARS-CoV-2  
1229 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* 182:828-842.e16.
- 1230 49. Alsoussi WB, Turner JS, Case JB, Zhao H, Schmitz AJ, Zhou JQ, Chen RE, Lei T, Rizk AA,  
1231 McIntire KM, Winkler ES, Fox JM, Kafai NM, Thackray LB, Hassan AO, Amanat F, Krammer F,  
1232 Watson CT, Kleinstein SH, Fremont DH, Diamond MS, Ellebedy AH. 2020. A Potently  
1233 Neutralizing Antibody Protects Mice against SARS-CoV-2 Infection. *J Immunol* 205:915-922.
- 1234 50. Salazar E, Kuchipudi SV, Christensen PA, Eagar T, Yi X, Zhao P, Jin Z, Long SW, Olsen RJ,  
1235 Chen J, Castillo B, Leveque C, Towers D, Lavinder JJ, Gollihar J, Cardona JA, Ippolito GC,  
1236 Nissly RH, Bird I, Greenawalt D, Rossi RM, Gontu A, Srinivasan S, Poojary I, Cattadori IM,  
1237 Hudson P, Josleyn NM, Prugar L, Huie KE, Herbert AS, Bernard DW, Dye JM, Kapur V, Musser  
1238 JM. 2020. Convalescent plasma anti-SARS-CoV-2 spike protein ectodomain and receptor binding  
1239 domain IgG correlate with virus neutralization. *The Journal of Clinical Investigation*  
1240 doi:10.1172/JCI141206.
- 1241 51. Salazar E, Christensen PA, Graviss EA, Nguyen DT, Castillo B, Chen J, Lopez BV, Eagar TN, Yi  
1242 X, Zhao P, Rogers J, Shehabeldin A, Joseph D, Leveque C, Olsen RJ, Bernard DW, Gollihar J,  
1243 Musser JM. 2020. Treatment of COVID-19 Patients with Convalescent Plasma Reveals a Signal  
1244 of Significantly Decreased Mortality. *Am J Pathol* doi:10.1016/j.ajpath.2020.08.001.
- 1245 52. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE,  
1246 Tortorici MA, Walls AC, King NP, Velesler D, Bloom JD. 2020. Deep Mutational Scanning of

- 1247 SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*  
1248 doi:10.1016/j.cell.2020.08.012.
- 1249 53. Steffen TL, Stone ET, Hassert M, Geerling E, Grimberg BT, Espino AM, Pantoja P, Climent C,  
1250 Hoft DF, George SL, Sariol CA, Pinto AK, Brien JD. 2020. The receptor binding domain of  
1251 SARS-CoV-2 spike is the key target of neutralizing antibody in human polyclonal sera. *bioRxiv*  
1252 doi:10.1101/2020.08.21.261727:2020.08.21.261727.
- 1253 54. Corbett KS, Flynn B, Foulds KE, Francica JR, Boyoglu-Barnum S, Werner AP, Flach B,  
1254 O'Connell S, Bock KW, Minai M, Nagata BM, Andersen H, Martinez DR, Noe AT, Douek N,  
1255 Donaldson MM, Nji NN, Alvarado GS, Edwards DK, Flebbe DR, Lamb E, Doria-Rose NA, Lin  
1256 BC, Louder MK, O'Dell S, Schmidt SD, Phung E, Chang LA, Yap C, Todd J-PM, Pessaint L, Van  
1257 Ry A, Browne S, Greenhouse J, Putman-Taylor T, Strasbaugh A, Campbell T-A, Cook A, Dodson  
1258 A, Steingrebe K, Shi W, Zhang Y, Abiona OM, Wang L, Pegu A, Yang ES, Leung K, Zhou T,  
1259 Teng I-T, Widge A, et al. 2020. Evaluation of the mRNA-1273 Vaccine against SARS-CoV-2 in  
1260 Nonhuman Primates. *New England Journal of Medicine* doi:10.1056/NEJMoa2024671.
- 1261 55. van Doremalen N, Lambe T, Spencer A, Belij-Rammerstorfer S, Purushotham JN, Port JR,  
1262 Avanzato VA, Bushmaker T, Flaxman A, Ulaszewska M, Feldmann F, Allen ER, Sharpe H,  
1263 Schulz J, Holbrook M, Okumura A, Meade-White K, Pérez-Pérez L, Edwards NJ, Wright D,  
1264 Bissett C, Gilbride C, Williamson BN, Rosenke R, Long D, Ishwarbhai A, Kailath R, Rose L,  
1265 Morris S, Powers C, Lovaglio J, Hanley PW, Scott D, Saturday G, de Wit E, Gilbert SC, Munster  
1266 VJ. 2020. ChAdOx1 nCoV-19 vaccine prevents SARS-CoV-2 pneumonia in rhesus macaques.  
1267 *Nature* doi:10.1038/s41586-020-2608-y.
- 1268 56. Wang C, Li W, Drabek D, Okba NMA, van Haperen R, Osterhaus A, van Kuppeveld FJM,  
1269 Haagmans BL, Grosveld F, Bosch BJ. 2020. A human monoclonal antibody blocking SARS-CoV-  
1270 2 infection. *Nat Commun* 11:2251.
- 1271 57. Ju B, Zhang Q, Ge J, Wang R, Sun J, Ge X, Yu J, Shan S, Zhou B, Song S, Tang X, Yu J, Lan J,  
1272 Yuan J, Wang H, Zhao J, Zhang S, Wang Y, Shi X, Liu L, Zhao J, Wang X, Zhang Z, Zhang L.  
1273 2020. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* 584:115-119.
- 1274 58. Liu L, Wang P, Nair MS, Yu J, Rapp M, Wang Q, Luo Y, Chan JF, Sahi V, Figueroa A, Guo XV,  
1275 Cerutti G, Bimela J, Gorman J, Zhou T, Chen Z, Yuen KY, Kwong PD, Sodroski JG, Yin MT,  
1276 Sheng Z, Huang Y, Shapiro L, Ho DD. 2020. Potent neutralizing antibodies against multiple  
1277 epitopes on SARS-CoV-2 spike. *Nature* 584:450-456.
- 1278 59. Rogers TF, Zhao F, Huang D, Beutler N, Burns A, He W-t, Limbo O, Smith C, Song G, Woehl J,  
1279 Yang L, Abbott RK, Callaghan S, Garcia E, Hurtado J, Parren M, Peng L, Ramirez S, Ricketts J,  
1280 Ricciardi MJ, Rawlings SA, Wu NC, Yuan M, Smith DM, Nemazee D, Teijaro JR, Voss JE,  
1281 Wilson IA, Andrabi R, Briney B, Landais E, Sok D, Jardine JG, Burton DR. 2020. Isolation of  
1282 potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model.  
1283 *Science* 369:956-963.
- 1284 60. Hassan AO, Case JB, Winkler ES, Thackray LB, Kafai NM, Bailey AL, McCune BT, Fox JM,  
1285 Chen RE, Alsoussi WB, Turner JS, Schmitz AJ, Lei T, Shrihari S, Keeler SP, Fremont DH, Greco  
1286 S, McCray PB, Jr., Perlman S, Holtzman MJ, Ellebedy AH, Diamond MS. 2020. A SARS-CoV-2  
1287 Infection Model in Mice Demonstrates Protection by Neutralizing Antibodies. *Cell* 182:744-  
1288 753.e4.
- 1289 61. Chandrashekar A, Liu J, Martinot AJ, McMahan K, Mercado NB, Peter L, Tostanoski LH, Yu J,  
1290 Maliga Z, Nekorchuk M, Busman-Sahay K, Terry M, Wrijil LM, Ducat S, Martinez DR, Atyeo C,  
1291 Fischinger S, Burke JS, Slein MD, Pessaint L, Van Ry A, Greenhouse J, Taylor T, Blade K, Cook  
1292 A, Finneyfrock B, Brown R, Teow E, Velasco J, Zahn R, Wegmann F, Abbink P, Bondzie EA,  
1293 Dagotto G, Gebre MS, He X, Jacob-Dolan C, Kordana N, Li Z, Lifton MA, Mahrokhian SH,  
1294 Maxfield LF, Nityanandam R, Nkolola JP, Schmidt AG, Miller AD, Baric RS, Alter G, Sorger  
1295 PK, Estes JD, et al. 2020. SARS-CoV-2 infection protects against rechallenge in rhesus macaques.  
1296 *Science* 369:812-817.
- 1297 62. Mercado NB, Zahn R, Wegmann F, Loos C, Chandrashekar A, Yu J, Liu J, Peter L, McMahan K,  
1298 Tostanoski LH, He X, Martinez DR, Rutten L, Bos R, van Manen D, Vellinga J, Custers J,  
1299 Langedijk JP, Kwaks T, Bakkens MJG, Zuijdgeest D, Huber SKR, Atyeo C, Fischinger S, Burke  
1300 JS, Feldman J, Hauser BM, Caradonna TM, Bondzie EA, Dagotto G, Gebre MS, Hoffman E,  
1301 Jacob-Dolan C, Kirilova M, Li Z, Lin Z, Mahrokhian SH, Maxfield LF, Nampanya F,  
1302 Nityanandam R, Nkolola JP, Patel S, Ventura JD, Verrington K, Wan H, Pessaint L, Ry AV,



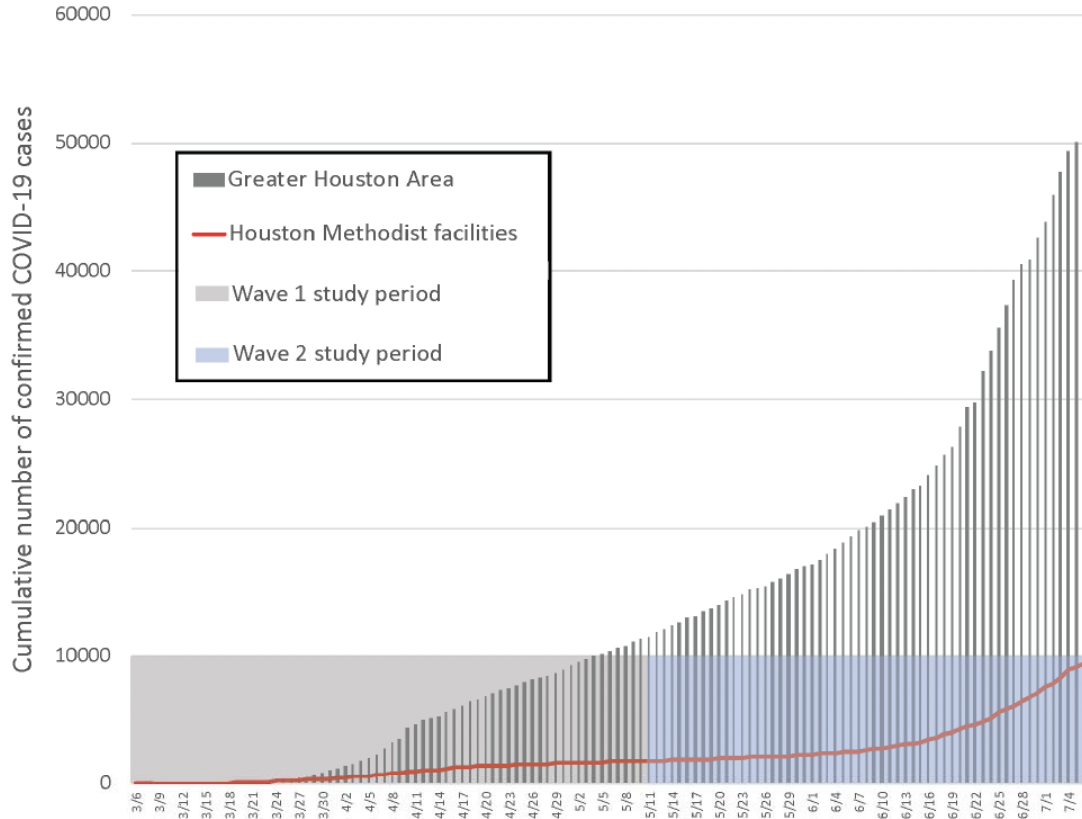
- 1303 Blade K, Strasbaugh A, Cabus M, et al. 2020. Single-shot Ad26 vaccine protects against SARS-  
1304 CoV-2 in rhesus macaques. *Nature* doi:10.1038/s41586-020-2607-z.
- 1305 63. Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, Mok CKP, Wilson IA. 2020. A highly  
1306 conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV.  
1307 *Science* 368:630-633.
- 1308 64. Hsieh C-L, Goldsmith JA, Schaub JM, DiVenere AM, Kuo H-C, Javanmardi K, Le KC, Wrapp D,  
1309 Lee AG, Liu Y, Chou C-W, Byrne PO, Hjorth CK, Johnson NV, Ludes-Meyers J, Nguyen AW,  
1310 Park J, Wang N, Amengor D, Lavinder JJ, Ippolito GC, Maynard JA, Finkelstein IJ, McLellan JS.  
1311 2020. Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* 369:1501-  
1312 1505.
- 1313 65. Woo H, Park SJ, Choi YK, Park T, Tanveer M, Cao Y, Kern NR, Lee J, Yeom MS, Croll TI, Seok  
1314 C, Im W. 2020. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model  
1315 in a Viral Membrane. *J Phys Chem B* 124:7128-7137.
- 1316 66. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi  
1317 EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de  
1318 Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire  
1319 EO, Montefiori DC. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G  
1320 Increases Infectivity of the COVID-19 Virus. *Cell* 182:812-827.e19.
- 1321 67. Hu J, He C-L, Gao Q-Z, Zhang G-J, Cao X-X, Long Q-X, Deng H-J, Huang L-Y, Chen J, Wang  
1322 K, Tang N, Huang A-L. 2020. D614G mutation of SARS-CoV-2 spike protein enhances viral  
1323 infectivity. *bioRxiv* doi:10.1101/2020.06.20.161323:2020.06.20.161323.
- 1324 68. Lorenzo-Redondo R, Nam HH, Roberts SC, Simons LM, Jennings LJ, Qi C, Achenbach CJ,  
1325 Hauser AR, Ison MG, Hultquist JF, Ozer EA. 2020. A Unique Clade of SARS-CoV-2 Viruses is  
1326 Associated with Lower Viral Loads in Patient Upper Airways. *medRxiv*  
1327 doi:10.1101/2020.05.19.20107144:2020.05.19.20107144.
- 1328 69. Cassia Wagner PR, Chris D. Frazar, Jover Lee, Nicola F. Müller, Louise H. Moncla, James  
1329 Hadfield, Emma B. Hodcroft, Benjamin Pelle, Matthew Richardson, Caitlin Behrens, Meei-Li  
1330 Huang, Patrick Mathias, Gregory Pepper, Lasata Shrestha, Hong Xie, Amin Addetia, Truong  
1331 Nguyen, Victoria M Rachleff, Romesh Gautam, Geoff Melly, Brian Hiatt, Philip Dykema,  
1332 Amanda Adler, Elisabeth Brandstetter, Peter D. Han, Kairsten Fay, Misja Ilcisin, Kirsten  
1333 Lacombe, Thomas R. Sibley, Melissa Truong, Caitlin R. Wolf, Karen Cowgill, Stephanie Schrag,  
1334 Jeff Duchin, Michael Boeckh, Janet A. Englund, Michael Famulare, Barry R. Lutz, Mark J.  
1335 Rieder, Matthew Thompson, Richard A. Neher, Geoffrey S. Baird, Lea M. Starita, Helen Y. Chu,  
1336 Jay Shendure, Scott Lindquist, Deborah A. Nickerson, Alexander L. Greninger, Keith R. Jerome,  
1337 Trevor Bedford. 2020. Comparing viral load and clinical outcomes in Washington State across  
1338 D614G substitution in spike protein of SARS-CoV-2. <https://github.com/blab/ncov-wa-d614g>.  
1339 Accessed September 8.
- 1340 70. Volz EM, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole A, Southgate JA, Johnson R,  
1341 Jackson B, Nascimento FF, Rey SM, Nicholls SM, Colquhoun RM, da Silva Filipe A, Shepherd  
1342 JG, Pascall DJ, Shah R, Jesudason N, Li K, Jarrett R, Pacchiarini N, Bull M, Geidelberg L,  
1343 Siveroni I, Goodfellow IG, Loman NJ, Pybus O, Robertson DL, Thomson EC, Rambaut A,  
1344 Connor TR. 2020. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on  
1345 transmissibility and pathogenicity. *medRxiv*  
1346 doi:10.1101/2020.07.31.20166082:2020.07.31.20166082.
- 1347 71. Lv Z, Deng Y-Q, Ye Q, Cao L, Sun C-Y, Fan C, Huang W, Sun S, Sun Y, Zhu L, Chen Q, Wang  
1348 N, Nie J, Cui Z, Zhu D, Shaw N, Li X-F, Li Q, Xie L, Wang Y, Rao Z, Qin C-F, Wang X. 2020.  
1349 Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic  
1350 antibody. *Science* 369:1505-1509.
- 1351 72. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile T, Wang Y, Baum A, Diehl WE,  
1352 Dauphin A, Carbone C, Veinotte K, Egri SB, Schaffner SF, Lemieux JE, Munro J, Rafique A,  
1353 Barve A, Sabeti PC, Kyratsous CA, Dudkina N, Shen K, Luban J. 2020. Structural and Functional  
1354 Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *bioRxiv*  
1355 doi:10.1101/2020.07.04.187757:2020.07.04.187757.
- 1356 73. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, IZard T, Farzan M, Choe H. 2020. The  
1357 D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.  
1358 *bioRxiv* doi:10.1101/2020.06.12.148726:2020.06.12.148726.

- 1359 74. Huo J, Zhao Y, Ren J, Zhou D, Duyvesteyn HME, Ginn HM, Carrique L, Malinauskas T, Ruza  
1360 RR, Shah PNM, Tan TK, Rijal P, Coombes N, Bewley KR, Tree JA, Radecke J, Paterson NG,  
1361 Supasa P, Mongkolsapaya J, Sreaton GR, Carroll M, Townsend A, Fry EE, Owens RJ, Stuart DI.  
1362 2020. Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host Microbe*  
1363 doi:10.1016/j.chom.2020.06.010.
- 1364 75. Long SW, Olsen RJ, Christensen PA, Bernard DW, Davis JR, Shukla M, Nguyen M, Ojeda  
1365 Saavedra M, Cantu CC, Yerramilli P, Pruitt L, Subedi S, Hendrickson H, Eskandari G,  
1366 Kumaraswami M, McLellan JS, Musser JM. 2020. Molecular Architecture of Early Dissemination  
1367 and Evolution of the SARS-CoV-2 Virus in Metropolitan Houston, Texas. *bioRxiv*  
1368 doi:10.1101/2020.05.01.072652:2020.05.01.072652.
- 1369 76. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J,  
1370 Fontes-Garfias CR, Mirchandani D, Scharton D, Bilello JP, Ku Z, An Z, Kalveram B, Freiberg  
1371 AN, Menachery VD, Xie X, Plante KS, Weaver SC, Shi P-Y. 2020. Spike mutation D614G alters  
1372 SARS-CoV-2 fitness and neutralization susceptibility. *bioRxiv*  
1373 doi:10.1101/2020.09.01.278689:2020.09.01.278689.
- 1374 77. Latz CA, DeCarlo C, Boitano L, Png CYM, Patell R, Conrad MF, Eagleton M, Dua A. 2020.  
1375 Blood type and outcomes in patients with COVID-19. *Ann Hematol* 99:2113-2118.
- 1376 78. Wu BB, Gu DZ, Yu JN, Yang J, Shen WQ. 2020. Association between ABO blood groups and  
1377 COVID-19 infection, severity and demise: A systematic review and meta-analysis. *Infect Genet*  
1378 *Evol* 84:104485.
- 1379 79. Zhao J, Yang Y, Huang H, Li D, Gu D, Lu X, Zhang Z, Liu L, Liu T, Liu Y, He Y, Sun B, Wei M,  
1380 Yang G, Wang X, Zhang L, Zhou X, Xing M, Wang PG. 2020. Relationship between the ABO  
1381 Blood Group and the COVID-19 Susceptibility. *Clin Infect Dis* doi:10.1093/cid/ciaa1150.
- 1382 80. Zietz M, Tatonetti NP. 2020. Testing the association between blood type and COVID-19 infection,  
1383 intubation, and death. *medRxiv* doi:10.1101/2020.04.08.20058073.
- 1384 81. Lemieux J, Siddle KJ, Shaw BM, Loreth C, Schaffner S, Gladden-Young A, Adams G, Fink T,  
1385 Tomkins-Tinch CH, Krasilnikova LA, Deruff KC, Rudy M, Bauer MR, Lagerborg KA,  
1386 Normandin E, Chapman SB, Reilly SK, Anahtar MN, Lin AE, Carter A, Myhrvold C, Kembal M,  
1387 Chaluvadi S, Cusick C, Flowers K, Neumann A, Cerrato F, Farhat M, Slater D, Harris JB, Branda  
1388 J, Hooper D, Gaeta JM, Baggett TP, O'Connell J, Gnirke A, Lieberman TD, Philippakis A, Burns  
1389 M, Brown C, Luban J, Ryan ET, Turbett SE, LaRocque RC, Hanage WP, Gallagher G, Madoff  
1390 LC, Smole S, Pierce VM, Rosenberg ES, et al. 2020. Phylogenetic analysis of SARS-CoV-2 in the  
1391 Boston area highlights the role of recurrent importation and superspreading events. *medRxiv*  
1392 doi:10.1101/2020.08.23.20178236:2020.08.23.20178236.
- 1393 82. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL,  
1394 Saemundsdottir J, Sigurdsson A, Sulem P, Agustsdottir AB, Eiriksdottir B, Fridriksdottir R,  
1395 Gardarsdottir EE, Georgsson G, Gretarsdottir OS, Gudmundsson KR, Gunnarsdottir TR, Gylfason  
1396 A, Holm H, Jenson BO, Jonasdottir A, Jonsson F, Josefsdottir KS, Kristjansson T, Magnusdottir  
1397 DN, le Roux L, Sigmundsdottir G, Sveinbjornsson G, Sveinsdottir KE, Sveinsdottir M,  
1398 Thorarensen EA, Thorbjornsson B, Löve A, Masson G, Jonsdottir I, Möller AD, Gudnason T,  
1399 Kristinsson KG, Thorsteinsdottir U, Stefansson K. 2020. Spread of SARS-CoV-2 in the Icelandic  
1400 Population. *N Engl J Med* doi:10.1056/NEJMoa2006100.
- 1401 83. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, Mellan TA, du  
1402 Plessis L, Pereira RHM, Sales FCS, Manuli ER, Thézé J, Almeida L, Menezes MT, Voloch CM,  
1403 Fumagalli MJ, Coletti TM, da Silva CAM, Ramundo MS, Amorim MR, Hoeltgebaum HH, Mishra  
1404 S, Gill MS, Carvalho LM, Buss LF, Prete CA, Ashworth J, Nakaya HI, Peixoto PS, Brady OJ,  
1405 Nicholls SM, Tanuri A, Rossi ÁD, Braga CKV, Gerber AL, de C. Guimarães AP, Gaburo N,  
1406 Alencar CS, Ferreira ACS, Lima CX, Levi JE, Granato C, Ferreira GM, Francisco RS, Granja F,  
1407 Garcia MT, Moretti ML, Perroud MW, Castiñeiras TMPP, Lazari CS, et al. 2020. Evolution and  
1408 epidemic spread of SARS-CoV-2 in Brazil. *Science* 369:1255-1260.
- 1409 84. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernández J, Prati D,  
1410 Baselli G, Asselta R, Grimsrud MM, Milani C, Aziz F, Kässens J, May S, Wendorff M,  
1411 Wienbrandt L, Uellendahl-Werth F, Zheng T, Yi X, de Pablo R, Chercoles AG, Palom A, Garcia-  
1412 Fernandez AE, Rodriguez-Frias F, Zanella A, Bandera A, Protti A, Aghemo A, Lleo A, Biondi A,  
1413 Caballero-Garralda A, Gori A, Tanck A, Carreras Nolla A, Latiano A, Fracanzani AL, Peschuck  
1414 A, Julià A, Pesenti A, Voza A, Jiménez D, Mateos B, Nafria Jimenez B, Quereda C, Paccapelo C,

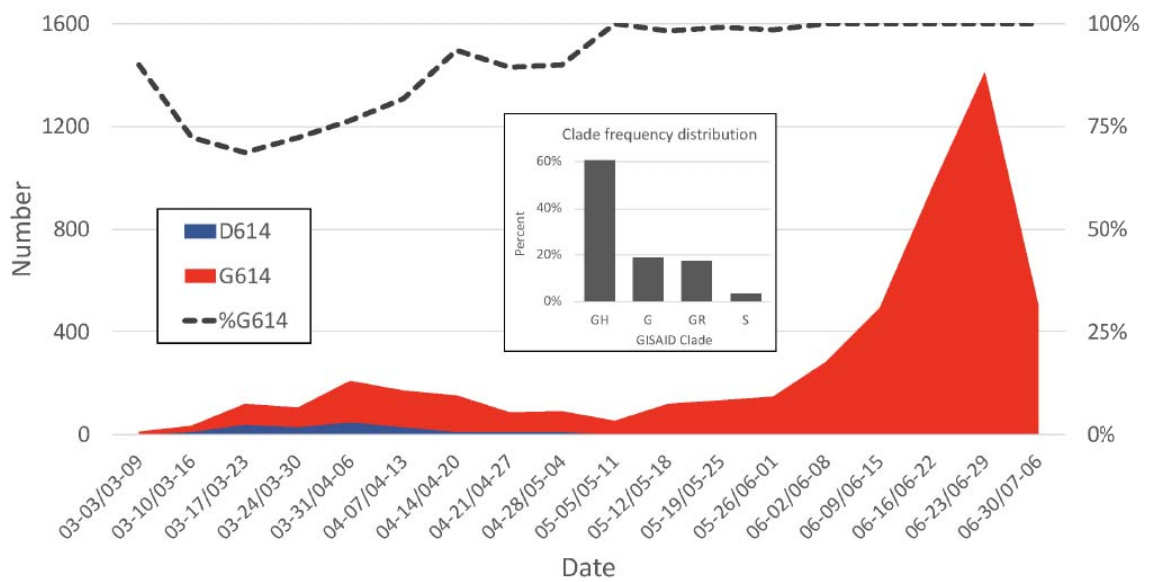
- 1415 Gassner C, Angelini C, Cea C, Solier A, et al. 2020. Genomewide Association Study of Severe  
1416 Covid-19 with Respiratory Failure. *N Engl J Med* doi:10.1056/NEJMoa2020283.
- 1417 85. 2020. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host  
1418 genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum*  
1419 *Genet* 28:715-718.
- 1420 86. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, Muecksch F, Rutkowska M,  
1421 Hoffmann H-H, Michailidis E, Gaebler C, Agudelo M, Cho A, Wang Z, Gazumyan A, Cipolla M,  
1422 Luchsinger L, Hillyer CD, Caskey M, Robbiani DF, Rice CM, Nussenzweig MC, Hatzioannou T,  
1423 Bieniasz PD. 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants.  
1424 *bioRxiv* doi:10.1101/2020.07.21.214759:2020.07.21.214759.
- 1425 87. Li T, Han X, Wang Y, Gu C, Wang J, Hu C, Li S, Wang K, Luo F, Huang J, Long Y, Song S,  
1426 Wang W, Hu J, Wu R, Mu S, Hao Y, Chen Q, Gao F, Shen M, Long S, Gong F, Li L, Wu Y, Xu  
1427 W, Cai X, Qu D, Yuan Z, Gao Q, Zhang G, He C, Nai Y, Deng K, Du L, Tang N, Xie Y, Huang  
1428 A, Jin A. 2020. A key linear epitope for a potent neutralizing antibody to SARS-CoV-2 S-RBD.  
1429 *bioRxiv* doi:10.1101/2020.09.11.292631:2020.09.11.292631.
- 1430 88. Hansen J, Baum A, Pascal KE, Russo V, Giordano S, Wloga E, Fulton BO, Yan Y, Koon K, Patel  
1431 K, Chung KM, Hermann A, Ullman E, Cruz J, Rafique A, Huang T, Fairhurst J, Libertiny C,  
1432 Malbec M, Lee W-y, Welsh R, Farr G, Pennington S, Deshpande D, Cheng J, Watty A, Bouffard  
1433 P, Babb R, Levenkova N, Chen C, Zhang B, Romero Hernandez A, Saotome K, Zhou Y, Franklin  
1434 M, Sivapalasingam S, Lye DC, Weston S, Logue J, Haupt R, Frieman M, Chen G, Olson W,  
1435 Murphy AJ, Stahl N, Yancopoulos GD, Kyratsous CA. 2020. Studies in humanized mice and  
1436 convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* 369:1010-1014.
- 1437 89. Long SW, Olsen RJ, Eagar TN, Beres SB, Zhao P, Davis JJ, Brettin T, Xia F, Musser JM. 2017.  
1438 Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing  
1439 *Klebsiella pneumoniae* Isolates, Houston, Texas: Unexpected Abundance of Clonal  
1440 Group 307. *mBio* 8:e00489-17.
- 1441 90. Stucker KM, Schobel SA, Olsen RJ, Hodges HL, Lin X, Halpin RA, Fedorova N, Stockwell TB,  
1442 Tovchigrechko A, Das SR, Wentworth DE, Musser JM. 2015. Haemagglutinin mutations and  
1443 glycosylation changes shaped the 2012/13 influenza A(H3N2) epidemic, Houston, Texas. *Euro*  
1444 *Surveill* 20.
- 1445 91. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:  
1446 improvements in performance and usability. *Mol Biol Evol* 30:772-80.
- 1447 92. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2--a  
1448 multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-91.
- 1449 93. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for  
1450 large alignments. *PLoS One* 5:e9490.
- 1451 94. Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System, abstr Proceedings of the  
1452 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San  
1453 Francisco, California, USA, Association for Computing Machinery,
- 1454 95. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, Shukla M, Stevens RL, Xia F,  
1455 Yoo H. 2018. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella*  
1456 *pneumoniae*. *Scientific reports* 8:421.
- 1457 96. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davis  
1458 JJ. 2019. Using machine learning to predict antimicrobial MICs and associated genomic features  
1459 for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology* 57:e01260-18.
- 1460 97. Pedregosa F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,  
1461 Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,  
1462 Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of*  
1463 *Machine Learning Research* 12 (2011) 2825-2830.
- 1464 98. Cai Y, Zhang J, Xiao T, Peng H, Sterling SM, Walsh RM, Jr., Rawson S, Rits-Volloch S, Chen B.  
1465 2020. Distinct conformational states of SARS-CoV-2 spike protein. *Science*  
1466 doi:10.1126/science.abd4251.
- 1467

1468 **FIG 1**

**A**

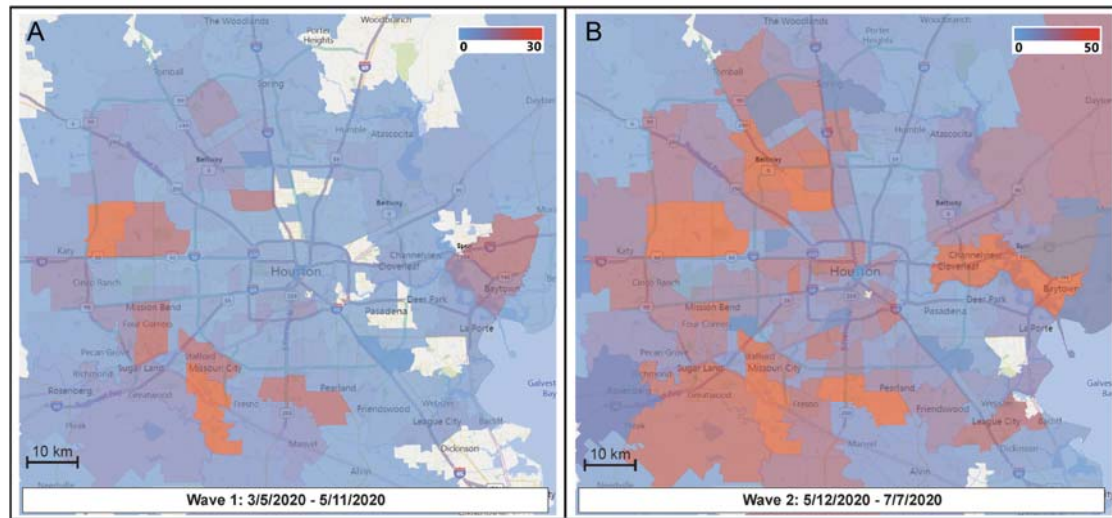


**B**



1469  
1470

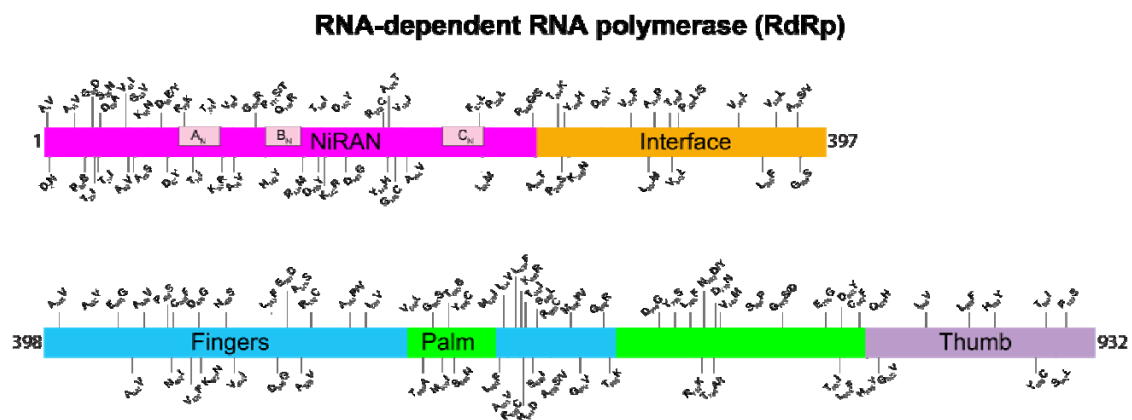
1471 **FIG 2**



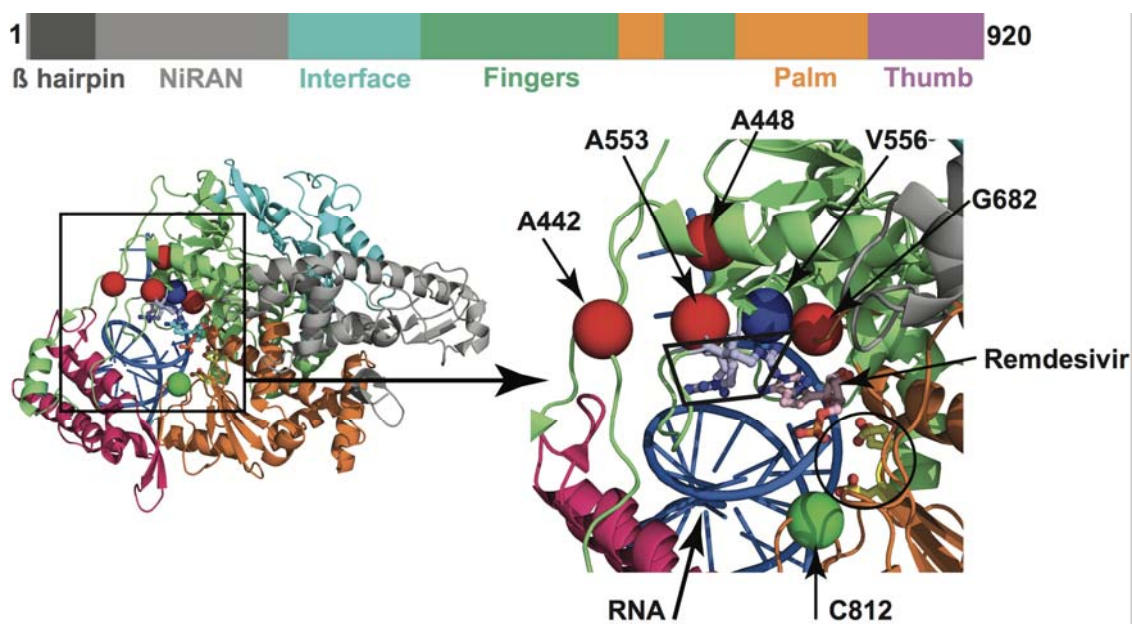
1472



1473 **FIG 3**



1475 **FIG 4**



1476

1477

1478 FIG 5



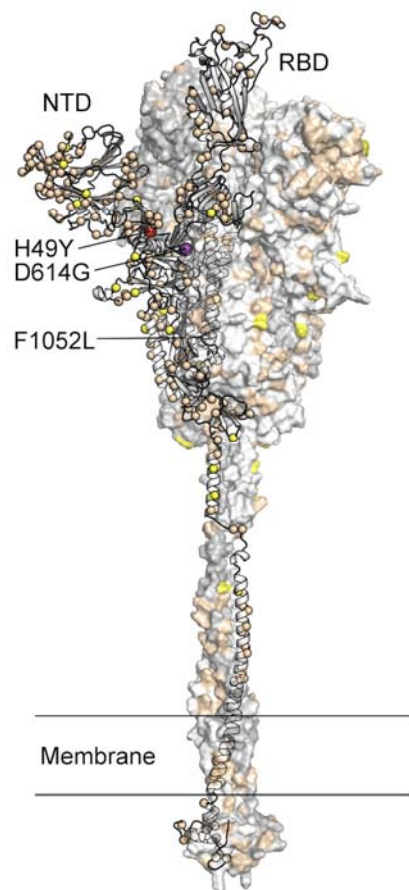
1479

1480

1481

1482

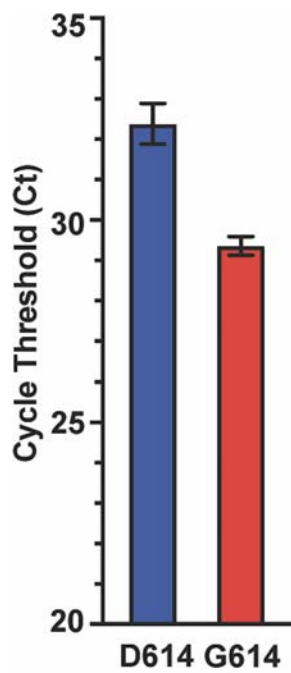
1483 FIG 6



1484

1485

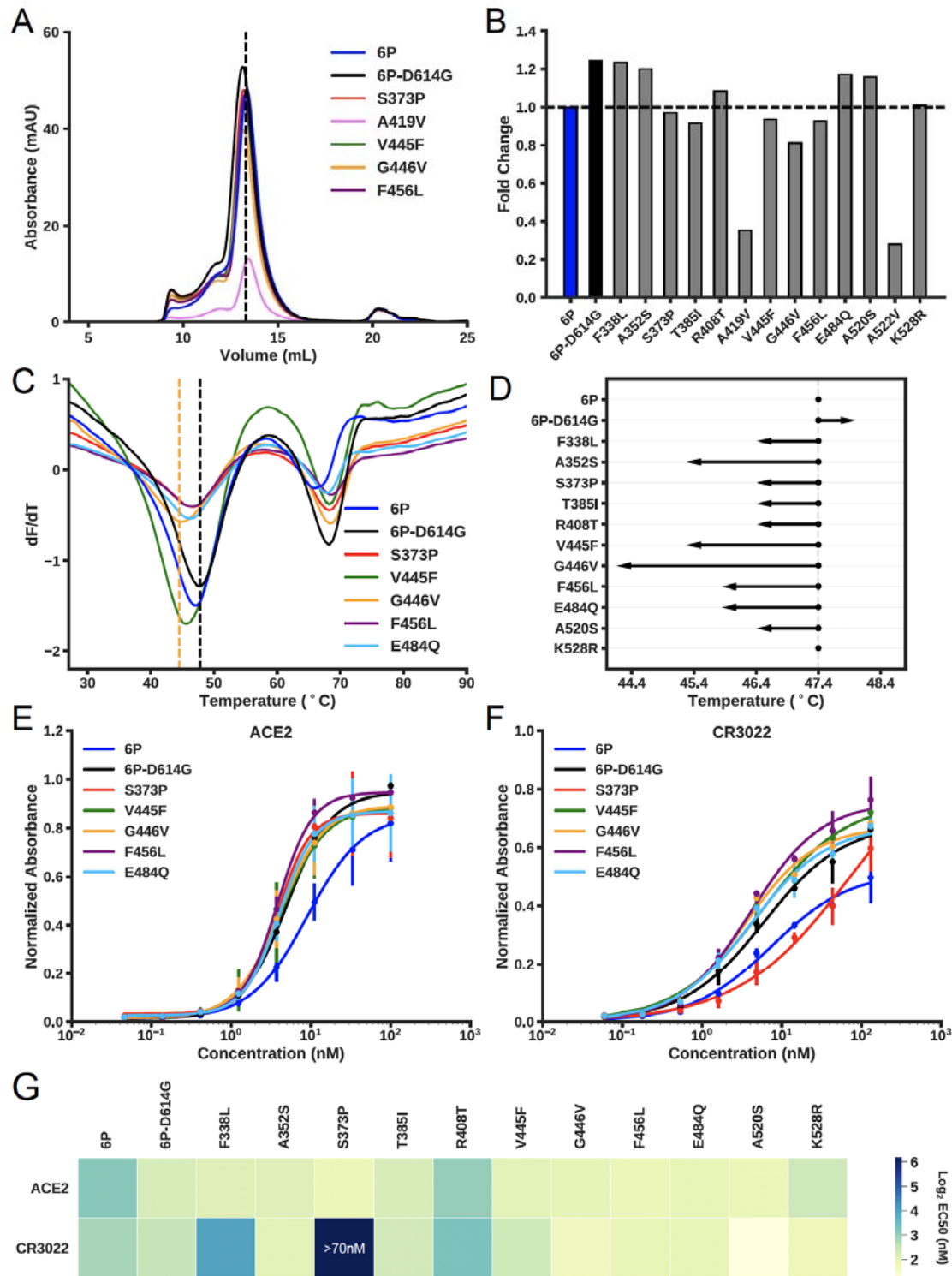
1486 **FIG 7**



1487

1488

1489 **FIG 8**



1490

1491

1492

1493 **Table 1.** Nonsynonymous SNPs of SARS-CoV-2 *nsp12*.

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
13446	C3T	A1V	N-terminus		2	2
13448	G5A	D2N	N-terminus	1		1
13487	C44T	A15V	N-terminus		138	138
13501	C58T	P20S	N-terminus		1	1
13514	G71A	G24D	N-terminus		3	3
13517	C74T	T25I	N-terminus		4	4
13520	G77A	S26N	N-terminus		1	1
13523	C80T	T27I	N-terminus		1	1
13526	A83C	D28A	N-terminus		1	1
13564	G121A	V41I	B hairpin		1	1
13568	C125T	A42V	B hairpin	1		1
13571	G128T	G43V	B hairpin	1		1
13576	G133T	A45S	B hairpin		12	12
13617	G174T	K58N	NiRAN		1	1
13618	G175T	D59Y	NiRAN		24	24
13620	C177G	D59E	NiRAN		1	1
13627	G184T	D62Y	NiRAN	1		1
13661	G218A	R73K	NiRAN		1	1
13667	C224T	T75I	NiRAN		2	2
13694	C251T	T84I	NiRAN		1	1
13712	A269G	K90R	NiRAN		1	1
13726	G283A	V95I	NiRAN		1	1
13730	C287T	A96V	NiRAN	2	2	4
13762	G319C	G107R	NiRAN		1	1
13774	C331A	P111T	NiRAN		1	1
13774	C331T	P111S	NiRAN		15	15
13777	C334T	H112Y	NiRAN		1	1
13790	A347G	Q116R	NiRAN		2	2
13835	G392T	R131M	NiRAN		1	1
13858	G415T	D139Y	NiRAN		3	3
13862	C419T	T140I	NiRAN	1	5	6
13868	A425G	K142R	NiRAN	1		1
13897	G454T	D152Y	NiRAN		4	4
13901	A458G	D153G	NiRAN		2	2
13957	C514T	R172C	NiRAN		2	2
13963	T520C	Y174H	NiRAN		1	1
13966	G523A	A175T	NiRAN		1	1
13975	G532T	G178C	NiRAN		4	4

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
13984	G541A	V181I	NiRAN		1	1
13994	C551T	A184V	NiRAN		8	8
14104	T661C	F221L	NiRAN	2		2
14109	A666G	I222M	NiRAN	1		1
14120	C677T	P226L	NiRAN		2	2
14185	A742G	R248G	NiRAN		1	1
14187	G744T	R248S	NiRAN		1	1
14188	G745A	A249T	NiRAN		1	1
14225	C782A	T261K	Interface		4	4
14230	C787T	P263S	Interface		1	1
14233	T790C	Y264H	Interface		1	1
14241	G798T	K266N	Interface		1	1
14290	G847T	D283Y	Interface		1	1
14335	G892T	V298F	Interface		8	8
14362	C919A	L307M	Interface		2	2
14371	G928C	A310P	Interface	1		1
14396	C953T	T318I	Interface		1	1
14398	G955T	V319L	Interface		1	1
14407	C964T	P322S	Interface		2	2
14408	C965T	P322L	Interface	843	4050	4893
14500	G1057T	V353L	Interface		5	5
14536	C1093T	L365F	Interface		1	1
14557	G1114T	V372L	Fingers		4	4
14584	G1141T	A381S			1	1
14585	C1142T	A381V	Fingers		10	10
14593	G1150A	G384S	Fingers	1		1
14657	C1214T	A405V	Fingers		1	1
14708	C1265T	A422V			1	1
14747	A1304G	E435G	Fingers		2	2
14768	C1325T	A442V	Fingers		21	21
14786	C1343T	A448V	Fingers	3	6	9
14821	C1378T	P460S			1	1
14829	G1386T	M462I	Fingers		59	59
14831	G1388T	C463F	Fingers		3	3
14857	G1414T	V472F			1	1
14870	A1427G	D476G	Fingers		5	5
14874	G1431T	K477N	Fingers	1		1
14912	A1469G	N490S	Fingers	1	1	2
14923	G1480A	V494I	Fingers		2	2



<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
14980	C1537T	L513F	Fingers	1	1	2
14990	A1547G	D516G			1	1
15006	G1563C	E521D	Fingers	2	3	5
15016	G1573T	A525S	Fingers		3	3
15026	C1583T	A528V	Fingers	5	1	6
15037	C1594T	R532C	Fingers		1	1
15100	G1657C	A553P	Fingers		1	1
15101	C1658T	A553V	Fingers		1	1
15124	A1681G	I561V	Fingers		2	2
15202	G1759C	V587L	Palm		7	7
15211	A1768G	T590A			1	1
15226	G1783A	G595S	Palm		1	1
15243	G1800T	M600I	Palm	71	4	75
15251	C1808G	T603S	Palm		1	1
15257	A1814G	Y605C			1	1
15260	G1817A	S606N	Palm		1	1
15327	G1884T	M628I	Fingers	3	1	4
15328	C1885T	L629F	Fingers	1		1
15334	A1891G	I631V	Fingers		1	1
15341	C1898T	A633V	Fingers		1	1
15352	C1909T	L637F	Fingers		1	1
15358	C1915T	R639C	Fingers		1	1
15362	A1919G	K640R	Fingers		1	1
15364	C1921G	H641D	Fingers	1		1
15368	C1925T	T642I	Fingers	1		1
15380	G1937T	S646I	Fingers	1		1
15386	C1943T	S648L	Fingers		2	2
15391	C1948T	R650C	Fingers		1	1
15406	G1963T	A655S	Fingers		3	3
15407	C1964T	A655V	Fingers	1		1
15436	A1993G	M665V	Fingers		2	2
15438	G1995T	M665I	Fingers		24	24
15452	G2009T	G670V	Fingers		28	28
15487	G2044C	G682R	Palm		1	1
15497	C2054A	T685K	Palm		1	1
15572	A2129G	D710G	Palm		1	1
15596	A2153G	Y718S	Palm	2		2
15619	C2176T	L726F	Palm	1		1
15638	G2195A	R732K	Palm	1		1

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
15640	A2197G	N733D	Palm	1		1
15640	A2197T	N733Y	Palm	1		1
15655	A2212G	T738A	Palm		2	2
15656	C2213T	T738I	Palm		2	2
15658	G2215A	D739N	Palm		2	2
15664	G2221A	V741M	Palm		1	1
15715	T2272C	S758P	Palm		1	1
15760	G2317A	G773S	Palm	1		1
15761	G2318A	G773D	Palm		1	1
15827	A2384G	E795G	Palm	1		1
15848	C2405T	T802I	Palm		1	1
15850	G2407T	D803Y	Palm		1	1
15853	C2410T	L804F	Palm		2	2
15878	G2435T	C812F	Palm		1	1
15886	C2443T	H815Y	Palm		1	1
15906	G2463T	Q821H	Thumb	1	1	2
15908	G2465T	G822V	Thumb		1	1
15979	A2536G	I846V	Thumb	4		4
16045	C2602T	L868F	Thumb		1	1
16084	C2641T	H881Y	Thumb		1	1
16148	A2705G	Y902C	Thumb		1	1
16163	C2720T	T907I	Thumb		45	45
16178	C2735T	S912L	Thumb		2	2
16192	C2749T	P917S	Thumb		80	80

1494

1495

**Table 2.** Nonsynonymous SNPs in SARS-CoV-2 spike protein.

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
21575	C13T	L5F	S1	11	25	36
21578	G16T	V6F	S1		1	1
21587	C25T	P9S	S1	2		2
21588	C26T	P9L	S1	1	1	2
21594	T32C	V11A	S1		1	1
21597	C35T	S12F	S1		6	6
21604	G42T	Q14H	S1		1	1
21614	C52T	L18F	S1 - NTD	1	11	12
21618	C56T	T19I	S1 - NTD	1	1	2
21621	C59T	T20I	S1 - NTD		1	1
21624	G62T	R21I	S1 - NTD		6	6
21624	G62A	R21K	S1 - NTD		1	1
21624	G62C	R21T	S1 - NTD		3	3
21627	C65T	T22I	S1 - NTD	2	4	6
21638	C76T	P26S	S1 - NTD		17	17
21641	G79T	A27S	S1 - NTD	1	1	2
21641	G79A	A27T	S1 - NTD	1		1
21642	C80T	A27V	S1 - NTD		1	1
21648	C86T	T29I	S1 - NTD	1	4	5
21707	C145T	H49Y	S1 - NTD		142	142
21713	A151G	T51A	S1 - NTD		1	1
21724	G162T	L54F	S1 - NTD		11	11
21754	G192T	W64C	S1 - NTD		1	1
21767	C205T	H69Y	S1 - NTD	1	7	8
21770	G208A	V70I	S1 - NTD		1	1
21770	G208T	V70F	S1 - NTD	1		1
21774	C212T	S71F	S1 - NTD		1	1
21784	T222A	N74K	S1 - NTD	1		1
21785	G223C	G75R	S1 - NTD		1	1
21793	G231T	K77N	S1 - NTD		1	1
21824	G262A	D88N	S1 - NTD		1	1
21834	A272T	Y91F	S1 - NTD		1	1
21846	C284T	T95I	S1 - NTD	1	10	11
21852	A290G	K97R	S1 - NTD		1	1
21855	C293T	S98F	S1 - NTD	1	2	3
21861	T299C	I100T	S1 - NTD		2	2

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
21918	T356C	I119T	S1 - NTD	1		1
21930	C368T	A123V	S1 - NTD		1	1
21941	G379T	V127F	S1 - NTD		1	1
21942	T380C	V127A	S1 - NTD		4	4
21974	G412T	D138Y	S1 - NTD	2		2
21985	G423T	L141F	S1 - NTD		1	1
21986	G424A	G142S	S1 - NTD		2	2
21993	A431G	Y144C	S1 - NTD	1		1
21995	T433C	Y145H	S1 - NTD	2		2
21998	C436T	H146Y	S1 - NTD	1	2	3
22014	G452A	S151N	S1 - NTD		1	1
22014	G452T	S151I	S1 - NTD		2	2
22017	G455T	W152L	S1 - NTD	1	1	2
22021	G459T	M153I	S1 - NTD		1	1
22021	G459A	M153I	S1 - NTD		1	1
22022	G460A	E154K	S1 - NTD		1	1
22028	G466C	E156Q	S1 - NTD	2		2
22037	G475A	V159I	S1 - NTD	1		1
22097	C535T	L179F	S1 - NTD		1	1
22104	G542T	G181V	S1 - NTD		1	1
22107	A545G	K182R	S1 - NTD		1	1
22135	A573T	E191D	S1 - NTD		1	1
22139	G577T	V193L	S1 - NTD		1	1
22150	T588G	N196K	S1 - NTD	1		1
22175	T613G	S205A	S1 - NTD		1	1
22205	G643T	D215Y	S1 - NTD		1	1
22206	A644G	D215G	S1 - NTD		2	2
22214	C652G	Q218E	S1 - NTD		1	1
22227	C665T	A222V	S1 - NTD		1	1
22241	G679A	V227I	S1 - NTD		2	2
22242	T680C	V227A	S1 - NTD	1		1
22244	G682C	D228H	S1 - NTD		2	2
22245	A683G	D228G	S1 - NTD	1		1
22246	T684G	D228E	S1 - NTD	2		2
22248	T686G	L229W	S1 - NTD	1		1
22250	C688A	P230T	S1 - NTD	1		1
22253	A691G	I231V	S1 - NTD	1		1

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
22254	T692C	I231T	S1 - NTD	1		1
22259	A697G	I233V	S1 - NTD	1		1
22260	T698C	I233T	S1 - NTD	1		1
22262	A700G	N234D	S1 - NTD	1		1
22266	T704C	I235T	S1 - NTD	1		1
22281	C719T	T240I	S1 - NTD		5	5
22286	C724T	L242F	S1 - NTD		1	1
22295	C733T	H245Y	S1 - NTD		2	2
22304	T742C	Y248H	S1 - NTD		3	3
22311	C749T	T250I	S1 - NTD	1	4	5
22313	C751T	P251S	S1 - NTD		2	2
22320	A758G	D253G	S1 - NTD		2	2
22320	A758C	D253A	S1 - NTD	1		1
22323	C761T	S254F	S1 - NTD		3	3
22329	C767T	S256L	S1 - NTD	1		1
22335	G773T	W258L	S1 - NTD	1		1
22344	G782T	G261V	S1 - NTD	3		3
22346	G784T	A262S	S1 - NTD		4	4
22350	C788T	A263V	S1 - NTD	1		1
22382	A820G	T274A	S1 - NTD		1	1
22398	A836T	Y279F	S1 - NTD	1		1
22408	T846G	N282K	S1 - NTD		1	1
22425	C863T	A288V	S1 - NTD		1	1
22430	G868T	D290Y	S1 - NTD	1		1
22484	G922T	V308L	S1		3	3
22487	G925C	E309Q	S1		1	1
22532	G970C	E324Q	S1		1	1
22533	A971T	E324V	S1		1	1
22535	T973C	S325P	S1		1	1
22536	C974T	S325F	S1		1	1
22550	C988T	P330S	S1 - RBD		2	2
22574	T1012C	F338L	S1 - RBD	1		1
22608	C1046T	S349F	S1 - RBD		1	1
22616	G1054T	A352S	S1 - RBD		7	7
22661	G1099T	V367F	S1 - RBD		1	1
22673	T1111C	S371P	S1 - RBD		3	3
22679	T1117C	S373P	S1 - RBD		1	1

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
22712	C1150T	P384S	S1 - RBD		1	1
22716	C1154T	T385I	S1 - RBD	3		3
22785	G1223C	R408T	S1 - RBD		1	1
22793	G1231T	A411S	S1 - RBD		1	1
22818	C1256T	A419V	S1 - RBD	1		1
22895	G1333T	V445F	S1 - RBD		1	1
22899	G1337T	G446V	S1 - RBD	2		2
22928	T1366C	F456L	S1 - RBD	1		1
23001	G1439T	C480F	S1 - RBD		1	1
23012	G1450C	E484Q	S1 - RBD	1		1
23046	A1484C	Y495S	S1 - RBD		1	1
23111	C1549T	L517F	S1 - RBD		1	1
23120	G1558T	A520S	S1 - RBD	1	6	7
23121	C1559T	A520V	S1 - RBD		1	1
23127	C1565T	A522V	S1 - RBD	1	1	2
23145	A1583G	K528R	S1 - RBD		2	2
23149	G1587T	K529N	S1		1	1
23170	C1608A	N536K	S1		1	1
23202	C1640A	T547K	S1		2	2
23202	C1640T	T547I	S1		1	1
23223	A1661T	E554V	S1		2	2
23224	G1662T	E554D	S1	4	31	35
23270	G1708T	A570S	S1		3	3
23277	C1715T	T572I	S1	5	5	10
23282	G1720T	D574Y	S1		1	1
23292	G1730T	R577L	S1	1		1
23311	G1749T	E583D	S1		6	6
23312	A1750G	I584V	S1		1	1
23315	C1753T	L585F	S1	1	7	8
23349	G1787A	S596N	S1		1	1
23373	C1811T	T604I	S1		2	2
23380	C1818A	N606K	S1		2	2
23403	A1841G	D614G	S1	841	4054	4895
23426	G1864T	V622F	S1		2	2
23426	G1864C	V622L	S1		2	2
23435	C1873T	H625Y	S1		1	1
23439	C1877T	A626V	S1		1	1

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
23444	C1882G	Q628E	S1		7	7
23453	C1891T	P631S	S1	1		1
23457	C1895T	T632I	S1		1	1
23481	C1919T	S640F	S1	1	42	43
23486	G1924T	V642F	S1		1	1
23502	C1940T	A647V	S1		1	1
23536	C1974A	N658K	S1		4	4
23564	G2002T	A668S	S1		1	1
23586	A2024G	Q675R	S1		14	14
23587	G2025C	Q675H	S1		1	1
23587	G2025T	Q675H	S1		4	4
23589	C2027T	T676I	S1	1	2	3
23593	G2031T	Q677H	S1	1	1	2
23595	C2033T	T678I	S1	1		1
23624	G2062T	A688S	S2		4	4
23625	C2063T	A688V	S2		16	16
23655	C2093T	S698L	S2		1	1
23664	C2102T	A701V	S2		21	21
23670	A2108G	N703S	S2		1	1
23679	C2117T	A706V	S2		1	1
23684	T2122C	S708P	S2		1	1
23709	C2147T	T716I	S2		1	1
23718	C2156T	T719I	S2		1	1
23745	C2183T	P728L	S2	1		1
23755	G2193T	M731I	S2	3	1	4
23798	T2236C	S746P	S2		1	1
23802	C2240T	T747I	S2		1	1
23804	G2242A	E748K	S2		1	1
23832	G2270T	G757V	S2		1	1
23856	G2294T	R765L	S2		1	1
23868	G2306T	G769V	S2		3	3
23873	G2311T	A771S	S2		8	8
23877	T2315C	V772A	S2		1	1
23895	C2333T	T778I	S2		1	1
23900	G2338C	E780Q	S2		1	1
23936	C2374T	P792S	S2		1	1

<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
23948	G2386T	D796Y	S2		2	2
23955	G2393T	G798V	S2	1		1
23987	C2425T	P809S	S2		2	2
23988	C2426T	P809L	S2		1	1
23997	C2435T	P812L	S2		1	1
24003	A2441G	K814R	S2		1	1
24014	A2452G	I818V	S2 - FP		5	5
24026	C2464T	L822F	S2 - FP		97	97
24041	A2479T	T827S	S2 - FP		4	4
24077	G2515T	D839Y	S2	2		2
24089	G2527A	D843N	S2	1	1	2
24095	G2533T	A845S	S2		5	5
24099	C2537T	A846V	S2		1	1
24129	A2567G	N856S	S2		7	7
24138	C2576T	T859I	S2		5	5
24141	T2579C	V860A	S2		1	1
24170	A2608G	I870V	S2		3	3
24188	G2626T	A876S	S2		1	1
24197	G2635T	A879S	S2		31	31
24198	C2636T	A879V	S2		1	1
24212	T2650G	S884A	S2		11	11
24237	C2675T	A892V	S2		1	1
24240	C2678T	A893V	S2	1		1
24268	G2706T	M902I	S2		1	1
24287	A2725G	I909V	S2 - HR1		2	2
24314	G2752C	E918Q	S2 - HR1	1		1
24328	G2766C	L922F	S2 - HR1		2	2
24348	G2786T	S929I	S2 - HR1		1	1
24356	G2794T	G932C	S2 - HR1		1	1
24357	G2795T	G932V	S2 - HR1		1	1
24368	G2806A	D936N	S2 - HR1		3	3
24368	G2806C	D936H	S2 - HR1		1	1
24368	G2806T	D936Y	S2 - HR1	3	4	7
24374	C2812T	L938F	S2 - HR1		3	3
24378	C2816T	S939F	S2 - HR1		4	4
24380	T2818G	S940A	S2 - HR1		5	5
24389	A2827G	S943G	S2 - HR1		6	6



<b>Genomic Locus</b>	<b>Gene Locus</b>	<b>Amino Acid Change</b>	<b>Domain</b>	<b>Wave 1 (n=1026)</b>	<b>Wave 2 (n=4059)</b>	<b>Total (n=5085)</b>
24463	C2901A	S967R	S2 - HR1	2		2
24507	C2945T	S982L	S2 - HR1		1	1
24579	C3017T	T1006I	S2 - CH		1	1
24588	C3026G	T1009S	S2 - CH		1	1
24621	C3059T	A1020V	S2 - CH	1		1
24638	G3076T	A1026S	S2 - CH		2	2
24642	C3080T	T1027I	S2 - CH	5		5
24649	G3087T	M1029I	S2 - CH		1	1
24710	A3148T	M1050L	S2		1	1
24710	A3148G	M1050V	S2	1	1	2
24712	G3150T	M1050I	S2		2	2
24718	C3156A	F1052L	S2	1	166	167
24770	G3208T	A1070S	S2		2	2
24794	G3232T	A1078S	S2 - CD	3	2	5
24812	G3250T	D1084Y	S2 - CD	1	29	30
24834	G3272T	R1091L	S2 - CD	1		1
24867	G3305T	W1102L	S2 - CD		1	1
24872	G3310T	V1104L	S2 - CD		1	1
24893	G3331C	E1111Q	S2 - CD		2	2
24897	C3335T	P1112L	S2 - CD	2	2	4
24912	C3350T	T1117I	S2 - CD		1	1
24923	T3361C	F1121L	S2 - CD		2	2
24933	G3371T	G1124V	S2 - CD	1	2	3
24959	G3397T	V1133F	S2 - CD		1	1
24977	G3415T	D1139Y	S2 - CD		1	1
24986	C3424A	Q1142K	S2	1		1
24998	G3436T	D1146Y	S2		4	4
24998	G3436C	D1146H	S2		13	13
25019	G3457T	D1153Y	S2		11	11
25032	A3470T	K1157M	S2	1		1
25046	C3484T	P1162S	S2		5	5
25047	C3485T	P1162L	S2		3	3
25050	A3488T	D1163V	S2		2	2
25088	G3526T	V1176F	S2		18	18
25101	A3539G	Q1180R	S2		1	1
25104	A3542G	K1181R	S2		4	4
25116	G3554A	R1185H	S2		2	2

Genomic Locus	Gene Locus	Amino Acid Change	Domain	Wave 1 (n=1026)	Wave 2 (n=4059)	Total (n=5085)
25121	A3559T	N1187Y	S2		1	1
25135	G3573T	K1191N	S2		1	1
25137	A3575C	N1192T	S2	1		1
25158	A3596C	D1199A	S2		1	1
25160	C3598T	L1200F	S2		1	1
25163	C3601A	Q1201K	S2		1	1
25169	C3607T	L1203F	S2	1		1
25183	G3621T	E1207D	S2		1	1
25186	G3624T	Q1208H	S2	1		1
25217	G3655T	G1219C	S2	1	3	4
25234	G3672T	L1224F	S2		1	1
25241	A3679G	I1227V	S2	1		1
25244	G3682T	V1228L	S2		2	2
25249	G3687T	M1229I	S2		1	1
25249	G3687C	M1229I	S2		2	2
25250	G3688A	V1230M	S2		1	1
25266	G3704T	C1235F	S2		4	4
25273	G3711T	M1237I	S2		2	2
25284	G3722T	C1241F	S2		1	1
25287	G3725T	S1242I	S2		4	4
25297	G3735T	K1245N	S2		1	1
25301	T3739G	C1247G	S2		1	1
25302	G3740T	C1247F	S2		4	4
25305	G3743T	C1248F	S2		2	2
25317	C3755T	S1252F	S2		1	1
25340	G3778T	D1260Y	S2		2	2
25350	C3788T	P1263L	S2	1	2	3
25352	G3790T	V1264L	S2		1	1
25365	T3803C	V1268A	S2		1	1

1497

1498 The domain region of RBD is based on structural information found in Cai et al.  
1499 2020 (98).

1500

1501

1502 Forty-nine of these amino acid replacements (V11A, T51A, W64C, I119T,  
1503 E156Q, S205A, D228G, L229W, P230T, N234D, I235T, T274A, A288V, E324Q,  
1504 E324V, S325P, S349F, S371P, S373P, T385I, A419V, C480F, Y495S, L517F,  
1505 K528R, Q628E, T632I, S708P, T719I, P728L, S746P, E748K, G757V, V772A,

1506 K814R, D843N, S884A, M902I, I909V, E918Q, S982L, M1029I, Q1142K,  
1507 K1157M, Q1180R, D1199A, C1241F, C1247G, and V1268A) were not  
1508 represented in a publicly available database (34) as of August 19, 2020.

1509