Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus

Highlights

- A SARS-CoV-2 variant with Spike G614 has replaced D614 as the dominant pandemic form
- The consistent increase of G614 at regional levels may indicate a fitness advantage
- G614 is associated with lower RT PCR Cts, suggestive of higher viral loads in patients
- The G614 variant grows to higher titers as pseudotyped virions

Authors

Bette Korber, Will M. Fischer, Sandrasegaram Gnanakaran, ..., Celia C. LaBranche, Erica O. Saphire, David C. Montefiori

Correspondence

btk@lanl.gov

In Brief

Korber et al. present evidence that there are now more SARS-CoV-2 viruses circulating in the human population globally that have the G614 form of the Spike protein versus the D614 form that was originally identified from the first human cases in Wuhan, China. Follow-up studies show that patients infected with G614 shed more viral nucleic acid compared with those with D614, and G614-bearing viruses show significantly higher infectious titers in vitro than their D614 counterparts.
Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus

Bette Korber,1,2,10,* Will M. Fischer,1 Sandrasegaram Gnanakaran,1 Hyejin Yoon,1 James Theiler,1 Werner Abfalterer,1 Nick Hengartner,1 Elena E. Giorgi,1 Tanmoy Bhattacharya,1 Brian Foley,1 Kathryn M. Hastie,3 Matthew D. Parker,4 David G. Partridge,5 Cariad M. Evans,5 Timothy M. Freeman,4 Thushan I. de Silva,5,6 on behalf of the Sheffield COVID-19 Genomics Group, Charlene McDanal,7 Lautaro G. Perez,7 Haili Tang,7 Alex Moon-Walker,3,8,9 Sean P. Whelan,9 Celia C. LaBranche,2 Erica O. Saphire,3 and David C. Montefiori7

1Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
2New Mexico Consortium, Los Alamos, NM 87545, USA
3La Jolla Institute for Immunology, La Jolla, CA 92037, USA
4Sheffield Biomedical Research Centre & Sheffield Bioinformatics Core, University of Sheffield, Sheffield S10 2HQ, UK
5Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield S10 2JF, UK
6Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield S10 2RX, UK
7Duke Human Vaccine Institute & Department of Surgery, Durham, NC 27710, USA
8Program in Virology, Harvard University, Boston, MA 02115, USA
9Department of Molecular Microbiology, Washington University in Saint Louis, St. Louis, MO 63130, USA
10Lead Contact
*Correspondence: btk@lanl.gov
https://doi.org/10.1016/j.cell.2020.06.043

SUMMARY

A SARS-CoV-2 variant carrying the Spike protein amino acid change D614G has become the most prevalent form in the global pandemic. Dynamic tracking of variant frequencies revealed a recurrent pattern of G614 increase at multiple geographic levels: national, regional, and municipal. The shift occurred even in local epidemics where the original D614 form was well established prior to introduction of the G614 variant. The consistency of this pattern was highly statistically significant, suggesting that the G614 variant may have a fitness advantage. We found that the G614 variant grows to a higher titer as pseudotyped virions. In infected individuals, G614 is associated with lower RT-PCR cycle thresholds, suggestive of higher upper respiratory tract viral loads, but not with increased disease severity. These findings illuminate changes important for a mechanistic understanding of the virus and support continuing surveillance of Spike mutations to aid with development of immunological interventions.

INTRODUCTION

The past two decades have seen three major pathogenic zoonotic disease outbreaks caused by betacoronaviruses (Cui et al., 2019; de Wit et al., 2016; Liu et al., 2020; Wu et al., 2020). Severe acute respiratory syndrome coronavirus (SARS-CoV) emerged in 2002, infecting ~8,000 people with a 10% mortality. Middle East respiratory syndrome coronavirus (MERS-CoV) emerged in 2012 with ~2,300 cases and 35% mortality (Graham and Baric, 2010). The third, SARS-CoV-2, causes the severe respiratory disease coronavirus disease 2019 (COVID-19) (Gorbalenyaya et al., 2020). First reported in China in December 2019 (Zhou et al., 2020), it rapidly became a pandemic with devastating effects. The June 21, 2020 World Health Organization (WHO) Situation Report records over 8.7 million COVID-19 cases and 460,000 deaths, numbers that increase daily. Humans have no direct immunological experience with SARS-CoV-2, leaving us vulnerable to infection and disease. SARS-CoV-2 is highly transmissible: basic reproduction number, $R_0$, estimates vary between 2.2 and 3.9 (Lv et al., 2020). Estimates of mortality vary regionally between 0.8% and 14.5% (mortality analyses, Johns Hopkins University of Medicine).

Coronaviruses have genetic proofreading mechanisms (Sevajol et al., 2014; Smith et al., 2013), and SARS-CoV-2 sequence diversity is very low (Fauver et al., 2020). Still, natural selection can act upon rare but favorable mutations. By analogy, antigenic drift results in gradual accumulation of mutations by the influenza virus during flu season, and the complex interplay between immunological resistance mutations and the fitness landscape enables antibody resistance to develop across populations (Boni et al., 2006). Although SARS-CoV-2 shows evidence of some seasonal waning (Sehra
et al., 2020), the persistence of the pandemic may enable accumulation of immunologically relevant mutations in the population even as vaccines are developed. Antigenic drift is seen among the common cold coronaviruses OC43 (Ren et al., 2015; Vijgen et al., 2005) and 229E (Chibo and Birch, 2006) and in SARS-CoV-1 (Quan et al., 2003; Song et al., 2005). Notably, a single SARS-CoV-1 amino acid change, Spike D480A/G in the receptor binding domain (RBD), arose in infected humans and civets and became the dominant variant among 2003/2004 viruses. D480A/G escapes neutralizing antibody 80R, and immune pressure from 80R in vitro could recapitulate emergence of the D480 mutation (Sui et al., 2008). Although there is no evidence yet of antigenic drift for SARS-CoV-2, with extended human-to-human transmission, SARS-CoV-2 could also acquire mutations with fitness advantages and immunological resistance. Attending to this risk now by identifying evolutionary transitions that may be relevant to the fitness or antigenic profile of the virus is important to ensure effectiveness of the vaccines and immunotherapeutic interventions as they advance to the clinic.

In response to the urgent need to develop effective vaccines and antibody-based therapeutic agents against SARS-CoV-2, over 90 vaccine and 50 antibody approaches are currently being explored (Cohen, 2020; Yu et al., 2020). Most target the trimeric Spike protein, which mediates host cell binding and entry and is the major target of neutralizing antibodies (Chen et al., 2020; Yuan et al., 2020). Spike monomers are comprised of an N-terminal S\textsubscript{1} subunit that mediates receptor binding and a membrane-proximal S\textsubscript{2} subunit that mediates membrane fusion (Hoffmann et al., 2020a; Walls et al., 2020; Wrapp et al., 2020). SARS-CoV-2 and SARS-CoV-1 share ~79% sequence identity (Lu et al., 2020), and both use angiotensin-converting enzyme 2 (ACE2) as their cellular receptor. Antibody responses to SARS-CoV-1 Spike are complex. In some patients with rapid and high neutralizing antibody responses, an early decline of these responses is associated with increased severity of disease and a higher risk of death (Ho et al., 2005; Liu et al., 2006; Temperton et al., 2005; Zhang et al., 2006). Some antibodies against SARS-CoV-1 Spike mediate antibody-dependent enhancement (ADE) of infection in vitro and exacerbate disease in animal models (Jaume et al., 2011; Wan et al., 2020; Wang et al., 2014; Yip et al., 2016).

Most current SARS-CoV-2 immunogens and testing reagents are based on the Spike protein sequence of the Wuhan reference sequence (Wang et al., 2020), and first-generation antibody therapeutic agents were discovered based on early pandemic infections and evaluated using the Wuhan reference sequence proteins. Alterations of the reference sequence as the virus propagates in human-to-human transmission could potentially alter the viral phenotype and/or the efficacy of immune-based interventions. Therefore, we designed bioinformatics tools to create an “early warning” strategy to evaluate Spike evolution during the pandemic to enable testing of mutations for phenotypic implications and generation of appropriate antibody breadth evaluation panels as vaccines and antibody-based therapeutic agents progress. Phylogenetic analysis of the global sampling of SARS-CoV-2 is being very capably addressed by the Global Initiative for Sharing All Influenza Data (GISAID) database (https://www.gisaid.org/; Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) and Nextstrain (https://nextstrain.org; Hadfield et al., 2018). However, in a setting of low genetic diversity like that of SARS-CoV-2, with very few de novo mutational events, phylogenetic methods that use homology to identify positive selection (Crispell et al., 2019) have limited statistical power. Additionally, recombination can add a confounding factor to phylogenetic reconstructions, and recombination is known to play a role in natural coronavirus evolution (Graham and Baric, 2010; Lau et al., 2011; Li et al., 2020; Oong et al., 2017; Rehman et al., 2020), and recombinant sequences (potential sequencing artifacts) have been found among SARS-CoV-2 sequences (De Maio et al., 2020). Given these issues, we developed an alternative indicator of potential positive selection by identifying variants that are recurrently becoming more prevalent in different geographic locations. If increases in relative frequency of a particular variant are observed repeatedly in distinct geographic regions, then that variant becomes a candidate for conferring a selective advantage.

Single amino acid changes are worth monitoring because they can be phenotypically relevant. Among coronavirus, point mutations have been demonstrated to confer resistance to neutralizing antibodies in MERS-CoV (Tang et al., 2014) and SARS-CoV-1 (Sui et al., 2008; ter Meulen et al., 2006). In the HIV envelope, single amino acid changes are known to alter host species susceptibility (Li et al., 2016), increase expression levels (Asmal et al., 2011), change the viral phenotype from tier 2 to tier 1, cause an overall change in neutralization sensitivity (Gao et al., 2014; LaBranche et al., 2019), and confer complete or nearly complete resistance to classes of neutralizing antibodies (Bricaut et al., 2019; Sadjadpour et al., 2013; Zhou et al., 2019).

We developed a bioinformatics pipeline to identify Spike amino acid variants that are increasing in frequency across many geographic regions by monitoring GISAID data. By early April 2020, it was clear that the Spike D614G mutation exhibited this behavior, and G614 has since become the dominant form in the pandemic. We present experimental evidence that the G614 variant is associated with greater infectivity as well as clinical evidence that it is associated with higher viral loads. We continue to monitor other mutations in Spike for frequency shifts at regional and global levels and provide regular updates at a public website (https://cov.lanl.gov/).

RESULTS

Website Overview

Our analysis pipeline to track SARS-CoV-2 mutations in the COVID-19 pandemic is based on regular updates from the GISAID SARS-CoV-2 sequence database (GISAID acknowledgments are in Table S1). GISAID sequences are generally linked to the location and date of sampling. Our website provides visualizations and summary data that allow regional tracking of SARS-CoV-2 mutations over time. Hundreds of new SARS-CoV-2 sequences are added to GISAID each day, so we have automated steps to create daily working alignments (Kurtz et al., 2004; Figure S1). The analysis presented here is based on a May 29, 2020 download of the GISAID data, when our Spike alignment included 28,576 sequences; updated versions of key figures can recreated at our website (https://cov.lanl.gov).
overall evolutionary rate for SARS-CoV-2 is very low, so we set a low threshold for a Spike mutation to be deemed “of interest,” and we track all sites in Spike where 0.3% of the sequences differ from the Wuhan reference sequence, monitoring them for increasing frequency over time in geographic regions as well as for recurrence in different geographic regions. Here we present results for the first amino acid variant to stand out by these metrics, D614G.

The D614G Variant

Increasing Frequency and Global Distribution

The Spike D614G amino acid change is caused by an A-to-G nucleotide mutation at position 23,403 in the Wuhan reference strain; it was the only site identified in our first Spike variation analysis in early March 2020 that met our threshold criterion. At that time, the G614 form was rare globally but gaining prominence in Europe, and GISAID was also tracking the clade carrying the D614G substitution, designating it the “G clade.” The D614G change is almost always accompanied by three other mutations: a C-to-T mutation in the 5’ UTR (position 241 relative to the Wuhan reference sequence), a silent C-to-T mutation at position 3,037, and a C-to-T mutation at position 14,408 that results in an amino acid change in RNA-dependent RNA polymerase (RdRp P323L). The haplotype comprising these 4 genetically linked mutations is now the globally dominant form. Prior to March 1, 2020, it was found in 10% of 997 global sequences; between March 1 and March 31, 2020, it represented 67% of 14,951 sequences; and between April 1 and May 18, 2020 (the last data point available in our May 29, 2020 sample), it represented 78% of 12,194 sequences. The transition from D614 to G614 occurred asynchronously in different regions throughout the world, beginning in Europe, followed by North America and Oceania and then Asia (Figures 1, 2, 3, S2, and S3).

We developed two statistical approaches to assess the consistency and significance of the D614-to-G614 transition. In general, to observe a significant change in the frequency of variants in a geographic region, three requirements must be met. First, both variants must at some point be co-circulating in the geographic area. Second, there must be sampling over an adequate duration to observe a change in frequency. Third, enough samples must be available for adequate statistical power to detect a difference. Both of our approaches enable us to systematically extract all GISAID local and regional data that meet these three requirements.

Our first approach requires that there be an “onset,” defined as the first day where the cumulative number of sequences reached 15 and both forms were represented at least 3 times; we further require that there be at least 15 sequences available at least 2 weeks after onset. Each geographic region that meets these criteria is extracted separately based on the hierarchical geographic/political levels designated in GISAID (Figure 1B). A two-sided Fisher’s exact test compares the counts in the pre-onset period with the counts after the 2-week delay period and provides a p value against the null hypothesis that the fraction of D614 versus G614 sequences did not change. All regions that met the above criteria and that showed a significant change in either direction (p < 0.05) are included. Almost all shifted toward increasing G614 frequencies: 5 of 5 continents, 16 of 17 countries (two-sided binomial p value of 0.00027), 16 of 16 regions (p = 0.00003), and 11 of 12 counties and cities (p = 0.0063).

In Figure 2 (Europe), Figure S2 (North America), and Figure S3 (Australia and Asia), we break down the relationships shown in Figure 1B in detail. The G614 variant increased in frequency even in regions where D614 was the clearly dominant form of a well-established local epidemic when G614 entered the population. Examples of this scenario include Wales, Nottingham, and Spain (Figure 2); Snohomish county and King county (Figure S2); and New South Wales, China, Japan, Hong Kong, and Thailand (Figure S3). Although introduction of a new variant might sometimes result in emergence of the new form because of stochastic effects or serial re-introductions or apparent emergence because of sampling biases, the consistency of the shift to G614 across regions is striking. The increase in G614 often continued after national stay-at-home orders were implemented and, in some cases, beyond the 2-week maximum incubation period.

We found two exceptions to the pattern of increasing G614 frequency in Figure 1B; details regarding these cases are shown in Figure S4. The first is Iceland. Changes in sampling strategy during a regional molecular epidemiology survey conducted through the month of March 2020 might explain this exception (Gudbjartsson et al., 2020). In early March 2020, only high-risk people were sampled, the majority being travelers from countries in Europe where G614 dominated. In mid-March 2020, screening began to include the local population; this coincided with the appearance of the D614 variant in the sequence dataset. The second exception is Santa Clara county, one of the most heavily sampled regions in California (Deng et al., 2020). The D614 variant dominates sequences from the Santa Clara County.

Figure 1. The Global Transition from the Original D614 Form to the G614 Variant

For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cell.2020.06.043#mmc4.

(A) Changes in the global distribution of the relative frequencies of the D614 (orange) and G614 (blue) variants in 2 time frames. Circle size indicates the relative sampling within each map. Through March 1, 2020, the G614 variant was rare outside of Europe, but by the end of March 2020 it had increased in frequency worldwide. These data are explored regionally in Figure 2 (Europe), Figure S2 (North America), and Figure S3 (Australia and Asia).

(B) Paired bar charts compare the fraction of sequences with D614 and with G614 for two time periods separated by a 2-week gap. The first time period (left bar) includes all sequences up to the onset day (see main text). The second time period (right bar) includes all sequences acquired at least 2 weeks after the onset date. All regions are shown that met the minimal threshold criterion for inclusion (see main text) with a significant shift in frequency (two-sided Fisher’s exact test, p < 0.05). Four hierarchical geographic levels are split out based on GISAID naming conventions.

(C) Running weekly average counts of sampled sequences exhibiting the D614 (orange) and G614 (blue) variants on different continents between January 12 and May 12, 2020. The measure of interest is the relative frequency over time. The shape of the overall curve just reflects sample availability; sequencing was more limited earlier in the epidemic (hence the left hand tail), and there is a time lag between viral sampling and sequence availability in GISAID (hence the right hand tail).

Weekly running count plots were generated with Python Matplotlib (Hunter, 2007); all elements of this figure are frequently updated at https://cov.lanl.gov/.
A

Prior to March 1

March 21-30

B

Weekly running counts

Italy

France

Scotland

England

Spain

Sheffield

United Kingdom

Nottingham

Iceland

Norwich

Date 2020

Date 2020

(legend on next page)
Department of Public Health (DPH) to date; the G614 variant was apparently not established in that community. In contrast, a smaller set of Santa Clara county sequences, sampled from mid-March to early April 2020, were specifically noted to be from Stanford; the Stanford samples had a mixture of both forms co-circulating (Figure S4), suggesting that the two communities in Santa Clara County are effectively distinct. A June 19, 2020 GISAID update for several California counties is provided in Figure S4C, and the G614 form is present in the most recent Santa Clara DPH samples.

Our second statistical approach to evaluating the significance of the D614-to-G614 transition (Figure 3) uses the time series data in GISAID more fully. Here we extracted all regional data from GISAID that had a minimum of 5 sequences representing each of the D614 and the G614 variants and at least 14 days of sampling. We then modeled the daily fraction of G614 as a function of time using isotonic regression, testing the null hypothesis that this fraction does not change over time (i.e., it remains roughly flat over time with equally likely random fluctuations of increase or decrease). We then separately tested the null against two alternative hypotheses: that the fraction of G614 increases or that it decreases. Figure 3A shows separate p values for all subcountries/states and counties/cities that met the minimal criteria. 30 of 31 subcountries/states with a significant change in frequency were increasing in G614; a binomial test indicates that G614 increases are highly significantly enriched (p = 2.98e–09). This was also found in 17 of 19 counties/cities (p = 0.0007). Figure 3B shows examples for 3 cities, plotting the daily fraction of G614 as a function of time. Country summaries (similar to Figure 3A) and plots for all regions (similar to Figure 3B) are included in Data S1.

Origins of the D614G 4-Base Haplotype

The earliest examples of sequences carrying parts of the 4-mutation haplotype that characterizes the D614G GISAID G clade were found in China and Germany in late January 2020, and they carried 3 of the 4 mutations that define the clade, lacking only the RdRp P323L substitution (Figure SSD). This may be an ancestral form of the G clade. One early Wuhan sequence and one early Thai sequence had the D614G change but not the other 3 mutations (Figure SSD); these may have arisen independently.

The earliest sequence we detected that carried all 4 mutations was found in China and Germany in late January 2020, and it had the D614G change but not the other 3 mutations (Figure SSD); these may have arisen independently. The earliest sequence we detected that carried all 4 mutations was found in China and Germany in late January 2020, and it had the D614G change but not the other 3 mutations (Figure SSD); these may have arisen independently. The earliest sequence we detected that carried all 4 mutations was found in China and Germany in late January 2020, and it had the D614G change but not the other 3 mutations (Figure SSD); these may have arisen independently.

Structural Implications of the Spike D614G Change

G614 is located on the surface of the Spike protein protomer, where it can form contacts with the neighboring protomer (Figure 4A). Cryoelectron microscopy (cryo-EM) structures (Walls et al., 2020; Wrapp et al., 2020) indicate that the side chains of D614 and T859 of the neighboring protomer (Figure 4B) form a between-protomer hydrogen bond, bringing together a residue from the S1 unit of one protomer and a residue of the S2 unit of the other protomer (Figure 4C). The change to G614 would eliminate this side-chain hydrogen bond, possibly increasing main-chain flexibility and altering between-protomer interactions. In addition, this substitution could modulate glycosylation at the nearby N616 site, which may influence the dynamics of the spatially proximal fusion peptide (Figure 4D) of the neighboring protomer, or have other effects.

G614 Is Associated with Potentially Higher Viral Loads in COVID-19 Patients but Not with Disease Severity

SARS-CoV-2 sequences from 999 individuals presenting with COVID-19 disease at the Sheffield Teaching Hospitals NHS Foundation Trust were available and linked to clinical data. The Sheffield data include age, sex, date of sampling, hospitalization status (defined as outpatient [OP], inpatient [IP], requiring hospitalization, or admittance to the intensive care unit [ICU]), and the cycle threshold (Ct) for a positive signal in E-gene based RT-PCR. The Ct is used here as a surrogate for relative viral loads; lower Ct values indicate higher viral loads (Corman et al., 2020), but not all viral nucleic acids represent infectious viral particles. RT-PCR methods changed during the course of the study because of limited availability of testing kits. The first method involved nucleic acid extraction and the second method heat treatment (Fomsgaard and Rosenstierne, 2020). A generalized linear model (GLM) used to predict the PCR Ct based on the RT-PCR method, sex, age, and D614G status showed only the RT-PCR method (p < 2e–16) and D614G status (p = 0.037) to be statistically significant (Figure 5A). Lower Ct values were observed in G614 infections. While our paper was in revision, G614-variant association with lower Ct values in vivo (Figure 5) was reported independently by two other groups (Lorenzo-Reondo et al., 2020; Wagner et al., 2020) in preprints that have not yet been peer reviewed.

We found no significant association between D614G status and disease severity as measured by hospitalization outcomes. A comparison of D614G status and hospitalization (combining IP and ICU) was not significant (p = 0.66, Fisher’s exact test), although comparing ICU admission with IP and OP did have borderline significance (p = 0.047) (Figure 5B). Regression analysis reinforced the result that G614 status was not associated with greater levels of hospitalization but that higher age (Dowd et al., 2020; Promislow, 2020), male sex (Conti and Younes, 2020), and hospitalization outcome were significant.
2020; Promislow, 2020) and higher Ct values (lower viral loads) were highly predictive of hospitalization. Further analysis showed that viral load was not masking a potential D614G status effect on hospitalization (STAR Methods). Univariate analysis also found highly significant associations between age and male sex and hospitalization (STAR Methods).

**G614 Is Associated with Higher Infectious Titers of Spike-Pseudotyped Virus**

We quantified the infectious titers of pseudotyped single-cycle vesicular stomatitis virus (VSV) and lentiviral particles displaying D614 or G614 SARS-CoV2 Spike protein. For the VSV and lentiviral pseudotypes, G614-bearing viruses had significantly higher infectious titers (2.6- to 9.3-fold increase) than their D614 counterparts; this was confirmed in multiple cell types (Figures 6A–6C). Similar results, reported recently in a preprint that has not yet been peer reviewed, also suggest that G614 increases Spike stability and membrane incorporation (Zhang et al., 2020).

TMPRSS2, a type-II transmembrane serine protease, cleaves the viral Spike after receptor binding to enhance entry of MERS-CoV, SARS-CoV, and SARS-CoV-2 (Hoffmann et al., 2020b; Kleine-Weber et al., 2018; Matsuyama et al., 2020; Millet and Whittaker, 2014; Park et al., 2016; Shulla et al., 2011; Zang et al., 2020). Spike 614 is in a pocket adjacent to the fusion peptide near the expected TMPRSS2 cleavage site, suggesting that there could be differences in the propensity and/or requirement for TMPRSS2 of the G614 variant. To test this hypothesis, we infected 293T cells stably expressing the ACE2 receptor in the presence or absence of TMPRSS2 and quantified the titer of infectious virus. We found similar fold changes in titers between D614 and G614 regardless of TMPRSS2 expression (Figure 6A). Hence, entry of G614-bearing viruses into 293T-ACE2 cells compared with D614-bearing viruses is not enhanced by TMPRSS2. Further studies are required to determine whether the G614 variant shows increased titers in lung cells, which may recapitulate native protease expression levels more faithfully, and to determine whether this variant increases the fitness of authentic SARS-CoV-2.

We also tested whether the D614G variations would be similarly neutralized by a polyclonal antibody. Convalescent sera of six San Diego residents, likely infected in early to mid-March 2020, when D614 and G614 were circulating, demonstrate equivalent or better neutralization of a G614-bearing pseudovirus compared with a D614-bearing pseudovirus (Figures 6D and 6E). Although we do not know with which virus each of these individuals were infected, these initial data suggest that, despite increased fitness in cell culture, G614-bearing virions are not intrinsically more resistant to neutralization by convalescent sera.

**DISCUSSION**

Our data show that, over the course of 1 month, the variant carrying the D614G Spike mutation became the globally dominant form of SARS-CoV-2. Phylogenetic tracking of SARS-CoV-2 variants at Nextstrain reveals complex webs of evolutionary and geographical relationships (https://nextstrain.org; Hadfield et al., 2018); travelers globally dispersed G614 variants and likely
would have introduced and reintroduced G614 variants into different locations. Still, D614 prevalent epidemics were very well established in many locations when G614 first began to appear (see Figure S2 for examples). The mutation that causes the D614G amino change is transmitted as part of a conserved haplotype defined by 4 mutations that almost always track together (Figures S5 and S6). The pattern of increasing G614 frequency within many different populations where D614 and G614 were co-circulating is highly significant, suggesting that G614 may be under positive selection (Figures 1B and 3). We also found G614 to be associated with higher levels of viral nucleic acid in the upper respiratory tract in human patients (Figure 5), suggestive of higher viral loads, and with higher infectivity in multiple pseudotyping assays (Figure 6).

Given that most G614 variants belong to the G clade lineage, phylogenetic methods that depend on recurrence of mutational events for their signal are poorly powered to resolve whether D614G is under positive selection. The GISAID data, however, suggest that D614G is associated with higher viral loads in human patients (Figure 5). The mutational pattern of G614 suggests that it may be under positive selection (Figures 1B and 3).

**Figure 4. Structural Mapping of Amino Acid Changes and Clusters of Variation in the Spike Protein**

(A) Sites including Spike 614 and those noted in Figure 7 mapped onto S1 and S2 units of the Spike protein (PDB: 6VSB). S1 and S2 are defined based on the furin cleavage site (protomer 1: S1, dark blue; S2, light blue; protomer 2: S1, dark green; S2, light green; protomer 3: gray). The RBD of protomer 1 is in the “up” position for engagement with the ACE2 receptor. Sites of interest are indicated by red balls, and variable clusters are labeled in red. The missing RBD residues at the ACE2 interface are shown in (D).

(B) Proximity of D614 (red) to an N linked glycosylation sequon of its own protomer (blue) and to residues T859 and Q853 of the neighboring protomer (green) are shown. Black dashed lines indicate possible hydrogen bond formation. Dotted lines indicate the structurally unresolved region of the fusion peptide connecting to Q853.

(C) Schematic representation of potential protomer protomer interactions shown in (B), in which D614 (red) from the S1 unit of one protomer (blue) is brought close to T859 from the S2 unit of the neighboring protomer.

(D) Sites of interest (red, residues 475 483) near the RBD (blue) ACE2 (yellow) binding interface. The interfacial region is shown as a molecular surface (PDB: 6M17).

(E) Variable cluster 936 940 (red) in the HR1 region of S2. These residues occur in a region that undergoes conformational transition during fusion; pre fusion (PDB: 6VSB) and post fusion (PDB: 6LXT) conformations of HR1 are shown (left and right).

**Figure 5. Clinical Status and D614G Associations Based on 999 Subjects with COVID-19 and Linked Sequence and Clinical Data Were Sampled in Sheffield, England**

(A) G614 was associated with a lower cycle threshold (Ct) required for detection; lower values are indicative of higher viral loads. The PCR method was changed partway through April 2020 because of shortages of nucleic acid extraction kits (Fomsgaard and Rosenstierne, 2020). The Ct levels for the two PCR methods (nucleic acid extraction versus simple heat inactivation) differ, and so we used a GLM to evaluate the statistical effect of D614G across methods.

(B) D614G status was not statistically associated with hospitalization status (outpatient [OP], inpatient [IP], or ICU) as a marker of disease severity, but age was highly correlated. The number of counts in each category is noted in the top right corner of each graph. See the main text and STAR Methods for statistical details.
Figure 6. Viral Infectivity and D614G Associations

(A) A recombinant VSV pseudotyped with the G614 Spike grows to a higher titer than D614 Spike in Vero, 293T ACE2, and 293T ACE2 TMPRSS2 cells, as measured in terms of focus forming units (ffu). ****p < 0.0001 by Student’s t test in pairwise comparisons. Experiments were repeated twice, each time in triplicate. Using a GLM to assess viral infectivity of the D614 and G614 variants across cell types and to account for repeat experiments, we found that the G614 variant had an average 3 fold higher infectious titer than D614 and that this difference was highly significant (p = 9 \times 10^{-11})(STAR Methods).

(B and C) Recombinant lentiviruses pseudotyped with the G614 Spike were more infectious than corresponding D614 S pseudotyped viruses in (B) 293T/ACE2 (6.5 fold increase) and (C) TZM-bi/ACE2 cells (2.8 fold increase, p < 0.0001). Relative luminescence units (RLUs) of Luc reporter gene expression (Naldini et al., 1996) were standardized to the p24 content of the pseudoviruses (p24 content of pseudoviruses for 293T/ACE2 cells: D614 = 269 ng/mL, G614 = 255 ng/mL; p24 content of pseudoviruses for TZM-bi/ACE2 cells, D614 = 680 ng/mL, G614 = 605 ng/mL). Background RLUs were measured in wells that received cells but no pseudovirus.

(D and E) Convalescent serum from six individuals in San Diego (Donors A-F) can neutralize D614 bearing (orange) and G614 bearing (blue) VSV pseudoviruses. Percent relative infection is plotted versus log polyclonal antibody concentration. Error bars indicate the mean ± SD of two biological replicates, each having two technical replicates. In (D), the sensitivity to each form is shown seperately for each sera. In (E) the responses to all sera are combined in one graph, and two negative control normal human sera are indicated in grey.
A

B

<table>
<thead>
<tr>
<th>Spike Mutation</th>
<th>Spike Region Possible impact</th>
<th>N=28,637 Count</th>
<th>Geographic Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>D614G</td>
<td>SARS-CoV epitope</td>
<td>20468 (71%)</td>
<td>Global</td>
</tr>
<tr>
<td>L5F</td>
<td>Signal Peptide</td>
<td>170 (0.6%)</td>
<td>Global</td>
</tr>
<tr>
<td>R211/K/T</td>
<td>S1 NTD domain</td>
<td>133 (0.5%)</td>
<td>Wales, England, others</td>
</tr>
<tr>
<td>A829T/S</td>
<td>Fusion Peptide</td>
<td>91 (0.3%)</td>
<td>Thailand</td>
</tr>
<tr>
<td>D839Y/N/E</td>
<td>Fusion Peptide</td>
<td>149 (0.5%)</td>
<td>Portugal, New Zealand, others</td>
</tr>
<tr>
<td>D936Y/H</td>
<td>Heptad Repeat 1(HR1)</td>
<td>257 (0.9%)</td>
<td>Sweden, UK, others</td>
</tr>
<tr>
<td>P1263L</td>
<td>Cytoplasmic Tail</td>
<td>201 (0.7%)</td>
<td>UK, others</td>
</tr>
</tbody>
</table>

C

Monitoring sites of interest in local epidemics

D

Higher Entropy Mutational clusters

<table>
<thead>
<tr>
<th>Count</th>
<th>Percent.</th>
<th>Local cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>28856</td>
<td>99.78</td>
<td>Portugal, New Zealand</td>
</tr>
<tr>
<td>149</td>
<td>0.51</td>
<td>Iceland</td>
</tr>
<tr>
<td>91</td>
<td>0.31</td>
<td>Thailand</td>
</tr>
<tr>
<td>32</td>
<td>0.11</td>
<td>Oxford</td>
</tr>
<tr>
<td>18</td>
<td>0.06</td>
<td>Cambridge</td>
</tr>
<tr>
<td>11</td>
<td>0.04</td>
<td>UK</td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>Cambridge</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>Virginia</td>
</tr>
<tr>
<td>936-940</td>
<td>HR1</td>
<td></td>
</tr>
</tbody>
</table>

(legend on next page)
provided the opportunity to look into the relationships among the SARS-CoV-2 variants in the context of time and geography, enabling us to track the increase in frequency of G614 as an early indicator of possible positive selection. This approach is potentially subject to founder effects and sampling biases, and so we generally view this strategy as simply an early indicator of an amino acid change that should be monitored further and tested. The G614 variant stood out, however, in our early detection framework for several reasons. First was the consistency of increase across geographic regions, which was highly significantly non-random (Figures 1B and 3). Second, if the two forms were equally likely to propagate, then one would expect the D614 form to persist in many locations where the G614 form was introduced into the ongoing well-established D614 epidemics. Instead, we found that, even in such cases, G614 increased (Figures 1, 2, 3, S2, and S3). Third, the increase in G614 frequency often continued well after national stay-at-home orders were in place, when serial reseeding from travelers was likely to be reduced significantly (Figures 2, S2, and S3).

Our global tracking data show that the G614 variant in Spike has spread faster than D614. We interpret this to mean that the virus is likely to be more infectious, a hypothesis consistent with the higher infectivity observed with G614 Spike-pseudotyped viruses we observed in vitro (Figure 6) and the G614 variant association with higher patient Ct values, indicative of potentially higher in vivo viral loads (Figure 5). Interestingly, we did not find evidence of G614 effects on disease severity; i.e., it was not significantly associated with hospitalization status. However, an association between the G614 variant and higher fatality rates has been reported in a comparison of mortality rates across countries, although this kind of analysis can be complicated by different availability of testing and care in different nations (Beceña-Flores and Cardozo, 2020).

Although higher infectiousness of the G614 variant may fully account for its rapid spread and persistence, other factors should also be considered. These include epidemiological factors because viral spread also depends on whom it infects, and epidemiological influences can also cause changes in genotype frequency to mimic evolutionary pressures. In all likelihood, a combination of evolutionary selection for G614 and the founder’s effects of being introduced into highly mobile and connected populations may have together contributed, in part, to its rise. The G-clade mutations in the 5’ UTR or in the RdRP protein might also have effects. In addition, there could be immunological consequences resulting from the G614 change in Spike. The G614 variant is sensitive to neutralization by polyclonal convalescent serum (Figure 5), which is encouraging in terms of immune interventions, but it will be important to determine whether the D614 and G614 forms of SARS-CoV-2 are differentially sensitive to neutralization by vaccine-elicited antibodies or by antibodies produced in response to infection with either form of the virus. Also, if the G614 variant is indeed more infectious than the D614 form (Figure 6), then it may require higher antibody levels for protection by vaccines or antibody therapeutic agents than the D614 form. Antibodies against an immunodominant linear epitope spanning Spike 614 in SARS-CoV-1 were associated with ADE activity (Wang et al., 2016), and so it is possible that this mutation may affect ADE.

Tracking mutations in the Spike gene has been our primary focus to date because of its relevance to vaccine and antibody-based therapy strategies currently under development. Such interventions take months to years to develop. For the sake of efficiency, contemporary variation should be factored in during development to ensure that the interventions will be effective against circulating variants when they are eventually deployed. To this end, we built a data analysis pipeline to enable exploration of potentially interesting mutations on SARS-CoV-2 sequences. The analysis is updated daily as the data become available through GISAID, enabling experimentalists to make use of the most current data available to inform vaccine development, reagents for evaluating antibody response, and experimental design. The speed with which the G614 variant became the dominant form globally suggests a need for continued vigilance.

**Limitations of Study**

Shifts in frequency toward the G614 variant in any given geographic region could, in principle, result from founder effects or sampling biases; it was the consistency of this pattern across regions where both forms of the virus were initially co-circulating that led us to suggest that the G614 form might be transmitted more readily because of an intrinsic fitness advantage; however,
systematic biases across many regions could affect the levels of significance we observed. The lack of association between G614 and hospitalization we report may miss effects on disease severity that are more subtle than we can detect. The experimental approach taken here to acquire laboratory evidence of increased fitness of the D614G mutation is based on two different pseudovirus models of infection in established cell lines. The extent to which this model faithfully recapitulates wild-type virus infection in natural target cells of the respiratory system is still being determined, and our laboratory experiments do not directly address the biology and mechanics of natural transmission. Infectiousness and transmissibility are not always synonymous, and more studies are needed to determine whether the D614G mutation actually led to an increase in number of infections and not just higher viral loads during infection. We encourage others to study this phenomenon in greater detail with a wild-type virus in natural infection and varied target cells (Hou et al., 2020) and in relevant animal models. Finally, the neutralization assays performed were based on sera from SARS-CoV-2 infected individuals with an unknown D614G status. Thus, although they show that the G614 variants are neutralization sensitive, more work is needed to resolve whether the potency of neutralization is affected when the variant that initiated the immune response differs from the test variant or when monoclonal antibodies are used.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human Subjects
- METHOD DETAILS
  - Detection and Sequencing of Sars-Cov-2 Isolates from Clinical Samples
  - Pseudotyped Virus Infectivity
  - Data Pipeline
  - Global Maps
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Systematic Regional Analysis of D614/G614 Frequencies
  - Clinical Data and Modeling
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.06.043.

CONSORTIA


ACKNOWLEDGMENTS

We thank Andrew McMichael, Sarah Rowland Jones, and Xiao Ning Xu for bringing together the clinical and theory teams. We thank Anthony West for pointing out the 5' UTR G clade mutation; George Ellison for suggestions regarding clinical data analyses; Barbara Imperiali for insights regarding the structural implications of the D614G change; and Rachael Mansbach, Sirupa Chakraborty, and Kien Nguyen for sharing preliminary MD data. We thank Da vey Smith and Stephen Rawlings of UCSD and Alessandro Sette, Jennifer M. Dan, and Shane Crotty of LJI for survivor sera and Sharon Schendel for manu script edits. We acknowledge Barney Graham, Kizzmekia Corbett, Nicole Doria Rose, Adrian McDermott, and John Masciola at the Vaccine Research Cen ter, NIH for reagents and assistance with the lentivirus based SARS CoV 2 infection assay and Elize Domin for technical support. The Sheffield COVID 19 Genomics Group is part of the COG UK CONSORTIUM, supported by the Medical Research Council (MRC), part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR), and Genome Research Limited, operating as the Wellcome Sanger Institute. T.I.d.S. is supported by a Wellcome Trust intermediate clinical fellowship (110058/Z/15/Z). M.P. was funded by the NIH Sheikh Biomedical Research Centre (BRC). B.K., E.E.G., T.B., N.H., W.M.F., H.Y., and W.A. were supported by LANL LDRD projects 20200554ECR and 20200706ER and through NIH NIAID, DHHSInteragency Agreement AA12007 001 00000. D.S. acknowledges support from the John and Mary Tu Foundation and the San Diego CFR AI036214. A.S. and S.C. acknowledge support from NIH NIAID AI42742 (Cooperative Centers for Human Immunology). E.O.S. acknowledges support from CoVic INV 006133 of the COVID 19 Therapeutics Accelerator, supported by the Bill and Melinda Gates Foundation, Mastercard, Wellcome and private philanthropic support, the Overton family, and a FastGrant from Emergent Ventures in aid of COVID 19 research. We gratefully acknowledge the team at GISAID for creating SARS CoV 2 global database and the many people who provided sequence data (Table S1).

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 29, 2020
Revised: June 10, 2020
Accepted: June 26, 2020
Published: July 3, 2020

REFERENCES

chronic infection impacts envelope processing and infectivity. PLoS ONE 6, e23673.


### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polyclonal human sera</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Bacterial and Virus Strains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VSV ΔG GFP</td>
<td>Karafast</td>
<td>Cat# EH1020</td>
</tr>
<tr>
<td>rVSV SARS CoV 2</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>rVSV SARS CoV 2 ΔD614G</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Chemicals, Peptides, and Recombinant Proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fugene 6</td>
<td>Promega</td>
<td>Cat# E2692</td>
</tr>
<tr>
<td>TransIT LT1</td>
<td>Mirus</td>
<td>Cat# MIR 2304</td>
</tr>
<tr>
<td>PFA</td>
<td>Electron Microscopy Sciences</td>
<td>Cat# 15710</td>
</tr>
<tr>
<td>Hoechst</td>
<td>ThermoFisher Scientific</td>
<td>Cat# 62249</td>
</tr>
<tr>
<td>SuperScript IV (50x)</td>
<td>ThermoFisher Scientific</td>
<td>Cat# 18090050</td>
</tr>
<tr>
<td>dNTP mix (10mM each)</td>
<td>ThermoFisher Scientific</td>
<td>Cat# R0192</td>
</tr>
<tr>
<td>Random hexamers (50uM)</td>
<td>ThermoFisher Scientific</td>
<td>Cat# N8080127</td>
</tr>
<tr>
<td>RNase OUT</td>
<td>ThermoFisher Scientific</td>
<td>Cat# 10777019</td>
</tr>
<tr>
<td>Q5 High fidelity polymerase</td>
<td>New England Biolabs</td>
<td>Cat# M0491S</td>
</tr>
<tr>
<td>Critical Commercial Assays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Promega Luciferase Assay System</td>
<td>Promega</td>
<td>Cat# E1501</td>
</tr>
<tr>
<td>Britelite Plus Reporter Gene Assay System</td>
<td>Perkin Elmer Part</td>
<td>Cat# 6066769</td>
</tr>
<tr>
<td>MagnaPure96 extraction platform</td>
<td>Roche Diagnostics Ltd, Burgess Hill, UK</td>
<td>Cat# 06 543 588 001</td>
</tr>
<tr>
<td>SensiFASTTM Probe No ROX One Step Real time PCR kit</td>
<td>Bioline</td>
<td>Cat# BIO 76001</td>
</tr>
<tr>
<td>Ligation sequencing kit</td>
<td>Oxford Nanopore</td>
<td>Cat# SQK LSK109</td>
</tr>
<tr>
<td>Native barcoding expansion kit 13 24</td>
<td>Oxford Nanopore</td>
<td>Cat# EXP NBD114</td>
</tr>
<tr>
<td>Native barcoding expansion kit 12</td>
<td>Oxford Nanopore</td>
<td>Cat# EXP NBD104</td>
</tr>
<tr>
<td>Flow cell priming kit</td>
<td>Oxford Nanopore</td>
<td>Cat# EXP FLP002</td>
</tr>
<tr>
<td>Flow cells R9.4.1.48p</td>
<td>Oxford Nanopore</td>
<td>Cat# FLO MIN106D</td>
</tr>
<tr>
<td>Flow cell wash kit</td>
<td>Oxford Nanopore</td>
<td>Cat# EXP WSH003</td>
</tr>
<tr>
<td>SFP Expansion kit</td>
<td>Oxford Nanopore</td>
<td>Cat# EXP SFB001</td>
</tr>
<tr>
<td>Next Ultra II library prep kit (illumina)</td>
<td>New England Biolabs</td>
<td>Cat# E7645L</td>
</tr>
<tr>
<td>Quick ligation module</td>
<td>New England Biolabs</td>
<td>Cat# E6056L</td>
</tr>
<tr>
<td>HIV 1 p24 ELISA</td>
<td>Perkin Elmer</td>
<td>Cat# NEK050B</td>
</tr>
<tr>
<td>Deposited Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GISAID</td>
<td>Elbe and Buckland Merrett, 2017; Shu and McCauley, 2017</td>
<td><a href="https://www.gisaid.org">https://www.gisaid.org</a></td>
</tr>
<tr>
<td>STAY AT HOME ORDERS IN EUROPE</td>
<td></td>
<td><a href="https://www.sidley.com/media/uploads/stay_at_home_tracker_europe.pdf?la=en">https://www.sidley.com/media/uploads/stay_at_home_tracker_europe.pdf?la=en</a></td>
</tr>
<tr>
<td>Experimental Models: Cell Lines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEK293T/17 cells</td>
<td>ATCC</td>
<td>Cat# CRL 11268</td>
</tr>
<tr>
<td>293T/ACE2.MF</td>
<td>Dr. Michael Farzan and Huihui Mu</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(Continued on next page)
<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>TZM bl/ACE2.MF</td>
<td>Dr. Michael Farzan and Huihui Mu</td>
<td>N/A</td>
</tr>
<tr>
<td>293T</td>
<td>ATCC</td>
<td>Cat# CRL 3216</td>
</tr>
<tr>
<td>Vero</td>
<td>ATCC</td>
<td>Cat# CCL 81</td>
</tr>
<tr>
<td>293T Ace2</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>293T Ace2 TMPRSS2</td>
<td>This study</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Oligonucleotides

| Primer pool1 and Primer Pool2 for sequencing (98 oligonucleotides) | Artic Network | https://artic.network/ncov 2019 |

### Recombinant DNA

<table>
<thead>
<tr>
<th>VRC7480 (Spike plasmid)</th>
<th>Drs. Barney Graham and Kizzmekia Corbett</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRC7480.D614G (Spike plasmid)</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pCMV ΔR8.2 (lentiviral backbone)</td>
<td>Drs. Barney Graham and Kizzmekia Corbett</td>
<td>N/A</td>
</tr>
<tr>
<td>pHRI CMV Luc (luciferase reporter)</td>
<td>Drs. Barney Graham and Kizzmekia Corbett; Naldini et al., 1996</td>
<td>N/A</td>
</tr>
<tr>
<td>pSG3&amp;Env</td>
<td>Drs. Beatrice Hahn and Feng Gao</td>
<td>N/A</td>
</tr>
<tr>
<td>Empty vector: phCMV3</td>
<td>Genlantis</td>
<td>Cat# P003300</td>
</tr>
<tr>
<td>pCAGGS VSV G</td>
<td>Kerfast</td>
<td>Cat# EH1017</td>
</tr>
<tr>
<td>phCMV3 SARS CoV 2</td>
<td>This study, Spike cloned from synthetic, codon optimized DNA</td>
<td>N/A</td>
</tr>
<tr>
<td>phCMV3 SARS CoV 2 D614G</td>
<td>This study, generated through site directed mutagenesis</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Software and Algorithms

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanopolish</td>
<td>© Ontario Institute for Cancer Research 2015 MPL license</td>
<td><a href="https://github.com/jts/nanopolish">https://github.com/jts/nanopolish</a></td>
</tr>
<tr>
<td>R packages: phangorn (version 2.5.5), ggplot2 (version 3.3.0), beeswarm (version 0.2.0), tidyverse (version 1.3.0), ape (version 5.3), lme4 (version 1.1.21)</td>
<td>The R Foundation for Statistical Computing</td>
<td><a href="https://cran.r">https://cran.r</a> project.org/</td>
</tr>
<tr>
<td>data.table (version 1.12.8)</td>
<td>Matt Dowle</td>
<td><a href="https://github.com/Rdatatable/data.table">https://github.com/Rdatatable/data.table</a></td>
</tr>
<tr>
<td>Aliview</td>
<td>Anders Larsson</td>
<td><a href="https://ormbunkar.se/aliview/">https://ormbunkar.se/aliview/</a></td>
</tr>
<tr>
<td>ARTIC network protocol (accessed the 19th of April)</td>
<td>ARTIC network</td>
<td><a href="https://artic.network/ncov">https://artic.network/ncov</a> 2019</td>
</tr>
<tr>
<td>Python Matplotlib A 2D Graphics Environment v 3.2.2</td>
<td>Hunter, 2007</td>
<td><a href="https://matplotlib.org">https://matplotlib.org</a></td>
</tr>
<tr>
<td>The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC</td>
<td>Schrödinger</td>
<td><a href="https://pymol.org/2/">https://pymol.org/2/</a></td>
</tr>
<tr>
<td>Los Alamos National Lab HIV Database: Analyze Align and Entropy</td>
<td>Los Alamos National Lab</td>
<td><a href="http://cov.lanl.gov/content/index">http://cov.lanl.gov/content/index</a></td>
</tr>
<tr>
<td>Los Alamos National Lab SARS CoV 2 Analysis Pipeline: SARS CoV 2 map, Relative Frequency Change by Geographical Region, Rainbow Tree</td>
<td>This study</td>
<td><a href="http://cov.lanl.gov/content/index">http://cov.lanl.gov/content/index</a></td>
</tr>
<tr>
<td>PAUP</td>
<td>David Swofford</td>
<td><a href="https://paup.phylsolutions.com">https://paup.phylsolutions.com</a></td>
</tr>
<tr>
<td>GraphPad Prism 8</td>
<td>GraphPad Software, Inc</td>
<td><a href="https://www.graphpad.com">https://www.graphpad.com</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCES AVAILABILITY

Lead Contact
Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Bette Korber (btk@lanl.gov).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
All sequence data used here are available from The Global Initiative for Sharing All Influenza Data (GISAID), at https://www.gisaid.org/. The user agreement for GISAID does not permit redistribution of sequences. Other data have been deposited to Mendeley Data: https://doi.org/10.17632/hn3h9gdrj.1.

Web-based tools to recreate much of the analyses provided in this paper but based on contemporary GISAID data downloads are available at https://cov.lanl.gov/.

Code to create the alignments as described in Figure S1 and to perform the Isotonic regression analysis in Figure 3 will be available through https://cov.lanl.gov, at also GitHub, once permission from our funders is obtained.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects
999 individuals presenting with active COVID-19 disease were sampled for SARS CoV-2 sequencing at Sheffield Teaching Hospitals NHS Foundation Trust, UK using samples collected for routine clinical diagnostic use. This work was performed under approval by the Public Health England Research Ethics and Governance Group for the COVID-19 Genomics UK consortium (R&D NR0195). SARS-CoV-2 sequences were generated using samples taken for routine clinical diagnostic use from 999 individuals presenting with active COVID-19 disease: 593 female, 399 male, 6 no gender specified; ages 15-103 (median 55) years.

METHOD DETAILS

Detection and Sequencing of Sars-Cov-2 Isolates from Clinical Samples
Samples for PCR detection of SARS-CoV-2 (Figure 5A) were all obtained from either throat or combined nose/throat swabs. Nucleic acid was extracted from 200μl of sample on MagnaPure96 extraction platform (Roche Diagnostics Ltd, Burgess Hill, UK). SARS-CoV-2 RNA was detected using primers and probes targeting the E gene and the RdRp genes for routine clinical diagnostic purposes, with thermocycling and fluorescence detection on ABI Thermal Cycler (Applied Biosystems, Foster City, United States) using previously described primer and probe sets (Corman et al., 2020). Nucleic acid from positive cases underwent long-read whole genome sequencing (Oxford Nanopore Technologies (ONT), Oxford, UK) using the ARTIC network protocol (accessed the 19th of April; https://artic.network/ncov-2019) Following base calling, data were demultiplexed using ONT Guppy using a high accuracy model. Reads were filtered based on quality and length (400 to 700bp), then mapped to the Wuhan reference genome and primer sites trimmed. Reads were then downsampled to 200x coverage in each direction. Variants were called using nanopolish (https://github.com/jts/nanopolish) and used to determine changes from the reference. Consensus sequences were constructed using reference and variants called.
Pseudotyped Virus Infectivity

**VSV System**

Plasmids for full-length SARS-CoV-2 Spike were generated from synthetic codon-optimized DNA (Wuhan-Hu-1 isolate, GenBank: MN908947.3) through sub-cloning into the pHCMV3 expression vector, with a stop codon included prior to the HA tag. The D614G variant was generated by site-directed mutagenesis. Positive clones were fully sequenced to ensure that no additional mutations were introduced.

Lentiviruses for stable cell line production were generated by seeding 293T cells at a density of $1 \times 10^6$ cells/well in a 6-well dish. Once the cells reached confluency, they were transfected with 2ug pCaggs-VSV-G, 2ug of lentiviral packaging vector pSPAX2, and 2ug of lentiviral expression plasmid pCW62 encoding ACE2-V5 and the puromycin resistance gene (pCW62-ACE2.V5-PuroR) or TMPRSS2-FLAG and the blasticidin resistance gene (pCW62-TMPRSS2.FLAG-BlastR) using TransIT transfection reagent according to manufacturer’s instructions. 24 hours post-transfection, media was replaced with fresh DMEM containing 10% FBS and 20mM HEPES. 48 hours post-transfection, supernatants were collected and filtered using a 0.45um syringe filter (VWR Catalog #28200-026).

293T-ACE2 cells were generated by seeding 293T cells at a density of $1 \times 10^6$ cells/well in a 6-well dish. At confluency, cells were transduced with 100uL of ACE2.V5-PuroR lentivirus. 48 hours post-transduction, cells were placed under 5ug/ml puromycin. 293T-AEC2+TMPRSS2 cells were generated by seeding 293T-AEC2 cells at a density of $1 \times 10^5$ cells/well in a 6-well dish. At confluency, cells were transduced with 100uL of TMPRSS2.FLAG-BlastR lentivirus. 48 hours post-transduction, cells were placed under 10ug/ml blasticidin selection.

Recombinant SARS-CoV-2-pseudotyped VSV-ΔG-GFP were generated by transfecting 293T cells with pHCMV3 expressing the indicated version of codon-optimized SARS-CoV-2 Spike using TransIT according to the manufacturer’s instructions. At 24 hr post-transfection, the medium was removed, and cells were infected with rVSV-G-pseudotyped ΔΔG-GFP parent virus (VSV-GΔΔG-GFP) at MOI = 2 for 2 hours with rocking. The virus was then removed, and the cells were washed twice with OPTI-MEM containing 2% FBS (OPTI-2) before fresh OPTI-2 was added. Supernatants containing rVSV-SARS-2 were removed 24 hours post-infection and clarified by centrifugation.

Viral titrations were performed by seeding cells in 96-well plates at a density sufficient to produce a monolayer at the time of infection. Then, 10-fold serial dilutions of pseudovirus were made and added to cells in triplicate wells. Infection was allowed to proceed for 12-16 hr at 37°C. The cells were then fixed with 4% PFA, washed two times with 1xPBS and stained with Hoescht (1ug/ml in PBS). After two additional washes with PBS, pseudovirus titers were quantified as the number of fluorescent forming units (ffu/mL) using a CellInsight CX5 imager (ThermoScientific) and automated enumeration of cells expressing GFP.

**Lentiviral System**

Additional assessments of corresponding D614 and G614 Spike pseudotyped viruses were performed by using lentiviral vectors and infection in 293T/ACE2 MF and TZM-bl/ACE2 MF cells (both cell lines kindly provided by Drs. Mike Farzan and Huilui Mu at Scripps). Cells were maintained in DMEM containing 10% FBS, 1% Pen Strep and 3 ug/ml puromycin. An expression plasmid encoding codon-optimized full-length spike of the Wuhan-1 strain (VRC7480), was provided by Drs. Barney Graham and Kizzmekia Corbett at the Vaccine Research Center, National Institutes of Health (USA). The D614G amino acid change was introduced into the lentiviral backbone pCMVΔR8.2 and firefly luciferase reporter gene (pHR' CMV Luc) (Naldini et al., 1996). Pseudovirions for TZM-bl/ACE2 infection were produced by co-transfection with a lentiviral backbone (pCMV ΔR8.2) and firefly luciferase reporter gene (pHR' CMV Luc) (Naldini et al., 1996). Pseudovirions for TZM-bl/ACE2 infection were produced by co-transfection with the Env-deficient lentiviral backbone pSG3ΔEnv (kindly provided by Drs Beatrice Hahn and Feng Gao). Culture supernatants from transfections were clarified of cells by low-speed centrifugation and filtration (0.45 μm filter) and used immediately for infection in 96-well culture plates. 293T/ACE2 MF cells were preseeded at 5,000 cells per well in 96-well black/white culture plates (Perkin-Elmer Catalog # 6005060) one day prior to infection. Sixteen wells were inoculated with 50 ul of a 1:10-dilution of each pseudovirus and incubated for three days. Luminescence was measured using the Promega Luciferase Assay System (Catalog # E1501). For infection of TZM-bl/ACE2 MF cells, 10,000 freshly trypsinized cells were added to 16 wells of a 96-well clear culture plate (Fisher Scientific) and inoculated with undiluted pseudovirus. Luminescence was measured after 2 days in a solid black plate using the Britelite Plus Reporter Gene Assay System (Perkin-Elmer). Luminescence in both assays was measured using a PerkinElmer Life Sciences, Model Victor2 luminometer. HIV-1 p24 content (produced by the backbone vectors) was quantified using the Alliance p24 ELISA Kit (PerkinElmer Health Sciences, Cat# NEK050B001KT). Reported relative luminescence units (RLUs) were adjusted for p24 content.

**Neutralization Assay**

Pre-titrated amounts of rVSV-SARS-CoV-2 (D614 or G614 variant) were incubated with serially diluted human sera at 37°C for 1 hr before addition to confluent Vero monolayers in 96-well plates. Infection proceeded for 12-16 hr at 37°C in 5% CO2 before cells were fixed in 4% paraformaldehyde and stained with 1ug/mL Hoescht. Cells were imaged using a CellInsight CX5 imager and infection was quantitated by automated enumeration of total cells and those expressing GFP. Infection was normalized to the percent cells infected with rVSV-SARS-CoV-2 incubated with normal human sera. Data are presented as the relative neutralization for each concentration of sera.
Data Pipeline

Background and General Approach

The Global Initiative for Sharing All Influenza Data (GISAID) (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) has been coordinating SARS-CoV-2 genome sequence submissions and making data available for download since early in the pandemic. At time of this writing, hundreds of sequences were being added every day. These sequences result from extraordinary efforts by a wide variety of institutions and individuals: while an invaluable resource, but are mixed in quality. The complete sequence download includes a large number of partial sequences, with variable coverage, and extensive ‘N’ runs in many sequences. To assemble a high-quality dataset for mutational analysis, we constructed a data pipeline using some off-the-shelf bioinformatic tools and a small amount of custom code.

From the SARS-CoV-2 sequences available from GISAID, we derived a “clean” codon-aligned dataset comprising near-complete viral genomes, without large insertions or deletions (“indels”) or runs of undetermined or ambiguous bases. For convenience in mutation assessment, we generated a codon-based nucleotide multiple sequence alignment, and extracted translations of each reading frame, from which we generated lists of mutations. The cleaning process was in general a process of deletion, with alignment of retained sequences; the following criteria were used to exclude sequences:

1. Fragmented matching (> 20 nt gap in match to reference)
2. Gaps at 5’ or 3’ end (> 3 nt)
3. High numbers of mismatched nucleotides (> 20), ‘N’ or other ambiguous IUPAC codes.
4. Regions with concentrated ambiguity calls: > 10 in any 50 nt window

Any sequence matching any of the above criteria was excluded in its entirety.

Sequence Mapping and Alignment

Sequences were mapped to a reference (bases 266:29674 of GenBank entry NC 045512; i.e., the first base of the ORF1ab start codon to the last base of the ORF10 stop codon) using “nucmer” from the MUMmer package (version 3.23; Kurtz et al., 2004). The nucmer output “delta” file was parsed directly using custom Perl code to partition sequences into the various exclusion categories (Sequence Mapping Table) and to construct a multiple sequence alignment (MSA). The MSA was refined using code derived from the Los Alamos HIV database “Gene Cutter” tool code base. At this stage, alignment columns comprising an insertion of a single “N” in a single sequence (generating a frameshift) were deleted, and gaps were shifted to conform with codon boundaries.

Using the initial “good-sequence” alignment, a low-effort parsimony tree was constructed. Initially, trees were built using PAUP* (Swofford, 2003) with a single replicate heuristic search using stepwise random sequence addition; subsequently, a parsimony ratchet was added; currently, oblong (Goloboff, 2014) is used. Sequences in the alignment were sorted vertically to correspond to the (ladderized) tree, and reference-sequence reading frames were added. See Figure S1 for a pipeline schematic.

Data partitioning and phylogenetic trees

Alignments were made and trees inferred for three distinct data partitions, the longer the alignments, the fewer sequences the sequences (Figure S1.) The full genome tree was used for Figure 7. Trees were inferred by either of two methods: 1. neighbor-joining using a p-distance criterion, (Swofford, 2003) or 2. parsimony heuristic search using a version of the parsimony ratchet (Goloboff, 2014), the general conclusions in Figure 7 were substantiated in both; the parsimony tree is shown.

Global Maps

The Covid-19 pie chart map is generated by overlaying Leaflet (a JavaScript library for interactive maps) pie charts on maps provided by OpenStreetMap. The interface is presented using rocker/shiny, a Docker for Shiny Server.

QUANTIFICATION AND STATISTICAL ANALYSIS

Systematic Regional Analysis of D614/G614 Frequencies

To observe a significant change in the frequency of two SARS-CoV-2 variants in a geographic region, three minimal requirements must be met. Both variants must have been introduced into an area and be co-circulating, data must be sampled for a long enough period to observe a change in frequency, and there must be enough data to be powered adequately to detect a difference.

We use the bioinformatic approaches described above to extract from GISAID all the politically defined geographic regions within the data that met these criteria, to track changes in frequency in a systematic way using all available data. The political/geographical regions we use are strictly hierarchically segmented based on the naming conventions used in GISAID. GISAID data is labeled such that the geographic source is noted first as a continent or Oceania; we call this Level 1. Level 2 is the country of origin of a sample. Level 3 are subcountries and states, and although occasionally level 3 includes a major city in a small country. For this purpose, England, Scotland, and Wales are considered sub-countries of the United Kingdom, and assigned level 3; the sampling in the UK has been the most extensive globally to date. Level 4, is the county or city of origin. The levels are strictly hierarchical, and within a given level, the geographical regions do not overlap. In some cases (e.g., Nepal Kathmandu and Nepal, Greece Athens and Greece, Ita-
ly Veneto Verona and Italy Veneto, or Iceland Reykjavik and Iceland) the sampling in a sub-level exactly matches the sampling in the corresponding upper level, in which case the sub-level is not presented. Levels 3 and 4 are not always available, and the day of sampling is also not always available.

The statistical strategies we use are then applied separately in each country, region or city, and we do not assume that outbreaks in each political subdivision are independent and identically distributed. Instead, our model assumption is that the individuals we test within a region are independent. This assumption may fail if there are sampling biases in a region that change over a given period of time. The G614 form is part of the G clade haplotype that is introduced by travelers, as we discuss in the text, and it is rare for it to arise independently. Our null hypothesis is that the observed shifts in frequency are random nondirectional drift. We have taken two statistical approaches to test this.

**Fisher’s exact comparison**

For this comparison, we used a two-sided Fisher’s exact test to compare the G614 and D614 counts in the pre-onset and the post-delay periods, as described in the text, and provides a p value against the null hypothesis that the fraction of D614 and G614 sequences did not change. To be included in the analysis, 15 sequences were required pre-onset, with a mixture of D614 and G614 present such that the rarer form was present at least 3 times; we also required a minimum of 15 sequences be sampled at least 2 weeks later, to create a post-delay set. Only regions for which p < 0.05 are considered, based on a two sided-test. We then use a binomial test to evaluate the null hypothesis that in regions where we saw significant change in sampling frequency over time, the shift was as likely to be an increase or a decrease in G614 across geographic regions. This analysis is presented in Figure 1B.

**Isotonic Regression**

Isotonic regression forms the basis of a one-sided test of the hypothesis for positive selection based on fitting the indicator that the typed strain is G as a logistic regression in which the logarithm of the odds ratio is a non-decreasing function of time. We use the residual deviance of the fitted model as our test statistics. To be included in this analysis, a region was required to have at least 5 sequences each of D614 and G614, and a minimum of 14 sampling days of data available. While we have a composite null hypothesis (the log-odds ratio is non-increasing), assuming that the log-odds ratio remains constant over time leads to tests that have largest power. While the classical chi-square approximation does not hold, we can sample from the constant log-odds ratio by permuting the vector of variant labels, and refitting the isotonic logistic regression. We performed 400 randomizations of the data in each region. Hence the lowest p value we can obtain is 0.0025. The reverse hypothesis, namely than the fraction of G variant decreases with time is also tested by fitting a non-increasing function of time. The isotonic logistic regression was done using R and the cgam package. We applied the binomial test across regions with a significant change in sampling frequency over time, the shift was as likely to be an increase or a decrease in G614 across geographic regions. This analysis is presented in Figure 3 and Data S1.

**Clinical Data and Modeling**

**Baseline Comparisons of Clinical Parameters**

Univariate analysis showed no associations between the age of individuals and their D614 (median 54.8, IR 39.4-77) or G614 status (median 54.6 (38.7-72.8) (Wilcoxon rank sum p = 0.37), nor with D614 and G614 and sex (Fisher’s exact p = 0.32). Comparing hospitalization and age, the median (IR) are: for all hospitalized, (IP+OCU), 74 years (59-83); for all OP, 44 (32-54), Wilcoxon p < 2.2e-16. 67% of males were hospitalized, versus 33% of females (Fisher’s exact p = p value < 2.2e-16).

**Modeling PCR Ct**

Two PCR Ct methods were used as a surrogate for estimating in vivo viral load in the upper respiratory tract, switching methods in mid-April due a shortage of kits. The first method involved nucleic acid extraction; the second method, heat treatment (Fomsgaard and Rosenstierne, 2020).

To assess the impact of available clinical parameters on viral load as measured by PCR Cts, we used a linear model, predicting Ct from PCR method, Sex, Age and D614G variant. This revealed that only the PCR method and the D614G variant were statistically significant. A negative coefficient for the G variant indicated that patients infected by the latter have, on average, a higher viral load, but that that viral load is not impacted by neither age nor sex.

The results from the smaller model are:

Coefficients:

<table>
<thead>
<tr>
<th>Estimate Std. Error t value Pr(&gt;</th>
<th>t</th>
<th>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept) 24.301 0.3166 76.757 &lt; 2e-16 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G614 0.7763 0.3718 2.088 0.037 *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method 2.3.1979 0.3658 8.743 &lt; 2e-16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results comparing D614G status for the two methods were also evaluated independently, and the first method showed a significant association between lower Ct values and presence of G614 (Wilcoxon p = 0.033), but the second method, with many fewer samples, did not reach significance.

**Predicting Hospitalization**

The simple Fisher’s exact test analysis in Figure 5 indicates that the D614G status is not predictive of hospitalization, even though it is predictive of viral load. We can make a first analysis to predict hospitalization from viral load, gender, age and D614G status:
Coefficients:

|        | Estimate | Std. Error | z value | Pr(>|z|) |
|--------|----------|------------|---------|---------|
| (Intercept) | 7.548823 | 0.624270   | 12.092  | < 2e-16 *** |
| G614     | 0.112038 | 0.214107   | 0.523   | 0.600779 |
| Male     | 1.490789 | 0.181695   | 8.205   | 2.31e-16 *** |
| Age      | 0.089444 | 0.005664   | 15.791  | < 2e-16 *** |
| CT       | 0.069376 | 0.018243   | 3.803   | 0.000143 *** |
| Method 2 | 0.358397 | 0.218856   | 1.638   | 0.101506 |

As somewhat expected, the D614G status is not statistically significant, even though viral load is, but the coefficient goes in the opposite direction than we would have intuited: a lower viral load is predictive of a higher probability of hospitalization. Sex (Male) and Age both increase the probability of hospitalization.

**Predicting Hospitalization, revisited**

Although the above analysis indicates that aa614G does not predict hospitalization directly, it does predict viral load and viral load predicts hospitalization; so there is a concern that aa614G might affect hospitalization, but that this effect is “masked” by the viral load. To explore this hypothesis, we “unmask” the aa614G by using the residuals from the regression of Ct on extraction method and D614G status to get a second predictive model for hospitalization:

Coefficients:

|        | Estimate | Std. Error | z value | Pr(>|z|) |
|--------|----------|------------|---------|---------|
| (Intercept) | 5.889991 | 0.393950   | 14.951  | < 2e-16 *** |
| G614     | 0.029858 | 0.209349   | 0.143   | 0.886587 |
| Corrected Ct | 0.069276 | 0.018225   | 3.801   | 0.000144 *** |
| Male     | 1.490690 | 0.181584   | 8.209   | 2.22e-16 *** |
| Age      | 0.089714 | 0.005661   | 15.849  | < 2e-16 *** |

In these regression analyses, the estimated coefficients for age, sex and viral load (corrected or not for method and strain) remain mostly unchanged, and strain still does not have an effect.

All other comparisons were not significant. All coding was done using R. Results of these analysis are presented in the main text and in Figure 5.

**Modeling pseudotype virus infectivity**

We used a log-normal generalized linear model (GLM) to test whether the G614 variant grew to higher titers than the wild-type D614 virus in Vero, 293T-ACE2 and 293T-ACE2-TMPRSS2 cell lines. The full experiment was repeated twice, each time in triplicate, and the 2 experimental repeats were considered random effects. Viral variant and cell line were considered as fixed effects. On average, across all cell lines, G614 grows to about a 3-fold (2.95) higher titer than D614 (p = 9x10^-11). A significant interaction was found between viral variant and cell line (p = 0.002), indicating that the relative increase of G614 compared to D614 was significantly different across cell lines (p = 0.002).

Results of these analyses are presented in Figure 6A.

**Sequence quality control**

We discovered a sequencing processing error that gave rise to what appeared at first to be a mutation of interest at position 943 (24389 A > C and 24390 C > G) in Spike that was evident in sequences from Belgium. It was frequent enough to be a site of interest, and was tracked. We contacted the group in Belgium, the source of the data, who were already aware of the issue, concurred with our interpretation, and they had been in touch with GISAID with a request to remove the problematic sequences.

We identified the issue with this site as part of another study using a method to detect systematic sequencing errors (Freeman et al., 2020); we are interrogating the quality of available sequencing data and these positions were highlighted as suspect. We interrogated these positions in the raw sequencing data from Sheffield, and although these two variants are not present in the final consensus sequence from any of the Sheffield isolates, the raw, un-trimmed bam files show their presence in only one of the amplicons covering the site (Figure S7A and S7B). We noticed that in fact this position is to the left of the 5’ primer of amplicon 81 in what we believe to be an adaptor sequence. Comparison of the Wuhan reference and the adaptor sequence reveals similarity around this position:

Nanopore adaptor sequence:

CAGCACCTT

The Wuhan reference sequence:

CAGCAAGTT
In our validation set, we see a C present at around 50% of called bases at both these positions in raw data but this region is trimmed by the ARTIC pipeline and is therefore not used to call variants and contribute to the final consensus sequence. Although it is evident in amplicon 81, in this region, there is no evidence for these variants in the data from amplicon 80, which also covers these positions. We include a figure (Figure S7) to explain our finding.

In summary this is an error that has arisen due to a combination of improper trimming of adaptor and primer regions from raw sequencing reads before downstream analysis, and the coincidental homology between the nanopore adaptor sequence and the Wuhan reference genome in this region. This is included here as a cautionary note; resolving rare biological mutations and sequencing error will be an important balance going forward in terms of interpretation of rare mutations (De Maio et al., 2020). A recurrent amino acid change like L5F (Figure 7) could potentially result from a recurrent sequencing or sequence processing error (De Maio et al., 2020), or alternatively, it may be of particular interest if it is naturally recurring homoplasy.

**ADDITIONAL RESOURCES**

Current data updates, analytical results, and webtools: https://cov.lanl.gov
Figure S1. Schematic of the SARS-CoV-2 Outbreak Sequence Processing Pipeline, Related to Figure 1

The intent of these procedures is to generate, for each of several regions, a set of contiguous codon aligned sequences, complete in that region, without extensive uncalled bases, large gaps, or regions that are unalignable or highly divergent, in reasonable running time for n > 30,000 /C24 30kb viral genomes. This allows daily processing of GISAID data to enable us to track mutations. This process provides the foundational data to enable the generation of Figures 1, 2, 3, and 7.

A. Processing procedures:
1. Download all SARS-CoV-2 sequences from GISAID.org (34,607 as of 2020 05 29). The downloaded sequences are stored in compressed form (via bzip2: https://sourceware.org/git/bzip2.git).
2. Align sequences to the SARS-CoV-2 reference sequence (NC 045512), trim to desired endpoints, and filter for coverage and quality. These steps are incorporated in a single Perl script, ‘align to ref.pl’, briefly summarized here: sequences are compressed for identity, then mapped against the given reference sequence using ‘nucmer’ from the ‘MUMmer’ package (Kurtz et al., 2004). The nucmer ‘delta’ file contains locations of matching regions and is parsed and used to, first, partition the sequences into ‘good’ and ‘bad’ subsets, and then to generate alignments from the ‘good’ sequences.
B. Categories of sequences included and excluded from our automated alignments. A series of criteria is used successively to exclude sequences with large internal gaps, excessive five and three prime gaps, large numbers of mismatches or ambiguities (> 30) overall, or regions with a high concentration of mismatches or ambiguities (> 10 in any 100 nt subsequence): the counts of these categories of ‘bad’ sequence are shown, for different regional genome alignments.

Sequences are trimmed to the extent of the reference (with minimum allowed gaps at 5' and 3' ends), following which the pairwise alignments are generated from the matching regions, and a multiple sequence alignment is constructed from the pairwise alignments.

3. De duplicate1. To reduce computational demands, sequences are compressed by identity following trimming to the desired region, by computing a hash value for each sequence (currently the SHA 1 message digest, 160 bits encoded as a 40 character hex string). To prevent the loss of parsimony informative characters when they occur in identical strings, however, multiple sequences are reduced to a minimum of two occurrences.

Notes:
1. Multiply occurring identical sequences are reduced to 2 occurrences so that parsimony-informative sites do not become unique.
2. “Coding regions” subset includes sequences passing error filtering, bounded from the orf1ab start codon to the ORF10 stop codon (NC 045512 genome positions 286-29874).

FDA-CBER-2022-908-0013967
4. Codon align. Gaps are introduced into the entire compressed alignment so that the alignment column containing the last base of each codon has a number divisible by three; this simplifies processing of translations. Code for this procedure is derived from the GeneCutter tool from the LANL HIV database (https://www.hiv.lanl.gov/content/sequence/GENECUTTER/cutter.html).

5. Partition (full/spike only). For subalignments that encompass the spike protein and substantial additional sequence, the spike region is extracted separately, to allow matched comparisons.

6. Build parsimony trees. A brief parsimony search (parsimony ratchet, with 5 replicates) is performed with ‘oblong’ (Goloboff, 2014). This is intended as an efficient clustering procedure rather than an explicit attempt to achieve an accurate phylogenetic reconstruction, but it appears to yield reasonable results in this situation of a very large number of sequences with a very small number of changes, where more complex models may be subject to overfitting. When multiple most parsimonious trees are found, only the shortest of these (under a p distance criterion) is retained. Distance scoring is performed with PAUP* (Swofford, 2003).

7. Re duplicate (expand, i.e., uncompress). The original sequence names and occurrence counts are restored to FASTA format files and the appropriate leaf taxa added to the parsimony trees.

8. Sort alignment by tree. Sequences in the FASTA files are sorted by the expanded tree, allowing patterns of mutation to be discerned by inspection.

9. Mutations of interest can be readily tracked on the trees to resolve whether they are identified in predominantly in single clades or distributed throughout the tree and likely to be recurring (e.g., Figure 7 (sites of interest with low frequency amino acid substitutions) and Figure S6 (Site 614)).
Figure S2. The Increasing Frequency of the D614G Variant over Time in North America, Related to Figure 1
Maps of the relative frequencies of D614 and G614 in North America in two different time windows. B. Weekly running counts of G614 illustrating the timing of its spread in North America. This figure complements Figures 2 and S3, and Figure 1 has details about how to read these figures. When a particular stay at home order date was known for a state or county it is shown as a pink line, followed by a light pink block indicating the maximum two week incubation time. Different counties in California had different stay at home order dates (Mar. 16 19) so are not highlighted, but more detail can be seen regarding California in Figure S4. The decline in D614 frequency often continues well after the stay at home orders were initiated, and sometimes beyond the 14 day maximum incubation period, when serial reintroduction of the G614 would be unlikely. On the right, Washington State is shown, with details from two heavily sampled counties, Snohomish and King. Both counties had well established ongoing D614 epidemics when G614 variants were introduced, undoubtedly by travelers. Washington state’s stay at home order was initiated March 24. At this time there were 1170 confirmed cases in King County, and 614 confirmed cases in Snohomish County. (Confirmed COVID19 case count data from: COVID 19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University). Testing was limited, and so this is lower bound on actual cases. Of the sequences sampled by March 24, 95% from King County (153/161) and 100% from Snohomish County (33/33) were the original D614 form (Part B, details at https://cov.lanl.gov/). By mid April, D614 was rarely sampled. Whatever the geographic origin of the G614 variants that entered these counties, and whether one or if multiple G614 variants were introduced, the rapid expansion of G614 variants occurred in the framework of well established local D614 variant epidemics. Santa Clara county is one of the two exceptions to the pattern of D614 decline in Figure 1B: details are provided in Figures S4A and S4C.
Figure S3. The Increasing Frequency of the D614G Variant over Time in Australia and Asia, Related to Figures 1 and 2

This figure complements Figure 2 and Figure S2, and Figure 1 has details about how to read these figures. The plot representing national sampling in Australia is on the left, with two regional subsets of the data on the right. In each case a local epidemic started with the D614 variant, and despite being well established, the G614 variant soon dominates the sampling. Only limited recent sampling from Asia is currently available in GISAID; to include more samples on the map the 10 day period between March 11–20, is shown rather than the period between March 21–30; even the limited sampling mid March supports the repeated pattern of a shift to G614. The Asian epidemic was overwhelmingly D614 through February, and despite this, G614 repeatedly becomes prominent in sampling by mid March.
A. Details regarding Santa Clara county, the only exceptional pattern at the county/city level in Figure 1B. Many samples from the Santa Clara County Department of Public Health (DPH) were obtained from March into May, and D614 has steadily dominated the local epidemic among those samples. The subset of Santa Clara county samples specifically labeled “Stanford,” however, were sampled over a few weeks mid March through early April, and have a mixture of both the G614 and D614 forms. These distinct patterns suggest relatively little mixing between the two local epidemics. Why Santa Clara county DPH samples should maintain the original form is unknown, but one possibility is that they may represent a relatively isolated community that had limited exposure to the G614 form, and G614 may not have had the opportunity to become established in this community though this may be changing, see Part C. The local stay at home orders were

B. Iceland and Austria had more targeted testing of travelers. In Iceland, the G614 variant was present in March – early April, but then lost its dominance. In Austria, the G614 variant was more widespread, with a peak in early April.

C. Sampling California, June 19, 2020. The patterns in California show a similar trend to those in Santa Clara county, with a mix of G614 and D614 forms observed in the early months of the pandemic and a decrease in the later months.
initiated relatively early, March 16, 2020. B. Details regarding Iceland, the only country with an exceptional pattern from Figure 1B. All Icelandic samples are from Reykjavik, and only G614 variants were initially observed there, with a modest but stable introduction of the original form D614 in mid March. This atypical pattern might be explained by local sampling. The Islanders conducted a detailed study of their early epidemic (Gudbjartsson et al., 2020), and all early March samples were collected from high risk travelers from Europe and people in contact with people who were ill; the majority of the traveler samples from early March were from people coming in from Italy and Austria, and G614 dominated both regions. On March 13, they began to sequence samples from local population screening, and on March 15, more travelers from the UK and USA with mixed G614/D614 infections began to be sampled in the high risk group, and those events were coincident with the appearance of D614. C. Updated data regarding California from the June 19, 2020 GISAID sampling. Most of the analysis in this paper was undertaken using the May 29, 2020 GISAID download, but as California was an interesting outlier, and more recent sampling conducted while the paper was under review was informative, we have included some additional plots from California data that were available at the time of our final response to review, on June 19th. Informative examples from well sampled local regions are shown. Stay at home order dates are shown as a pink line, followed by a light pink block indicating the maximum two week incubation time. N indicates the number of available sequences. Overall California, and specifically, San Diego and San Joaquin, show a clear shift from D614 to G614. The transition for San Joaquin was well after the stay at home orders and incubation period had passed. San Francisco shows a trend toward G614. Santa Clara DPH, which was essentially all D614 in our May 29th GISAID download, had 7 G614 forms sampled in late May that were evident in our June 19th GISAID download. Ventura is an example of a setting that was essentially all G614 when it began to be sampled significantly in early April, so a transition cannot be tracked; i.e., we cannot differentiate in such cases whether the local epidemic originated as a G614 epidemic, or whether it went through a transition from D614 to G614 prior to sampling. The figures in Parts A, B, and C can be recreated with more current data at http://cov.lanl.govcontent/index.
Weekly running counts of the earliest forms of the virus carrying G614. B. The GISAID G clade is based on a 4 base haplotype that distinguishes it from the original Wuhan form. Part B shows the tallies of all of the variants among the 4 base mutations that are the foundation of the G clade haplotype, versus the bases found in the original Wuhan reference strain, and highlights of some of the earliest identified sequences bearing these mutations. We first consider just the 3 mutations that are in coding regions, to enable using an alignment that contains a larger set of sequences. One is in the RdRp protein (nucleotide C14408T resulting in a P323L amino acid change), one in Spike (nucleotide A23403G resulting in the D614G amino acid change) and one is silent (C3037T). The other mutation is in the 5' UTR (C241T), and tallies based on all 4 positions are done separately, as they are based on an alignment with fewer sequences. The earliest examples of a partial haplotype, TTCG, are found in Germany and China (Parts A and B). This form was present in Shanghai but never expanded (Part A). A cluster of infections that all carried this form were identified in Germany, but they did not expand and were subsequently replaced with the original Wuhan form, only to be replaced again by sequences that carry the full 4 base haplotype variant TTTG. The first example in our alignments (see Figure S1) found to carry the full 4 base haplotype was sampled in Italy on Feb 20. The first cases in Italy were of the original Wuhan form CCCA form, but by the end of February TTTG was the only form sampled in Italy, and it is the TTTG form that has come the dominate the pandemic. Of note, the TTCG form did not expand, and it lacked the RdRp P323L change, raising the possibility that the P323L change may contribute to a selective advantage of the haplotype. TTTG and CCCA are almost always linked in SARS CoV 2 genomes, > 99.9% of the time (Part B). The number of cases where the clade haplotype is disrupted, and the form of the disruption, are all noted in part B (not including ambiguous base calls). Some cases of a disrupted haplotype may be due to recombination events and not de novo mutations. Given the fact that these two forms are co circulating in many communities throughout the spring of 2020, and the fact that disruptions in the 4 base pattern are rare, suggests that recombination is overall relatively rare among pandemic sequences.
Figure S6. Distribution of A23403G (D614G) Mutation and Other Mutations on an Approximate Phylogenetic Tree Using Parsimony, Related to Figure 7

This tree the same as the tree shown in Figure 6A, but highlights complementary information: the G614 substitution, and patterns of bases that underly the clades. It is based on the “FULL” alignment of 17,760 sequences, from the June 2 alignment, described in the pipeline in Figure S1, from the beginning of (orf1ab) to the last stop codon (ORF10, NC 045512 bases 266 29,674). The outer element is a radial presentation of a full coding region parsimony tree; branches are colored by the global region of origin for each virus isolate. The inner element is a radial bar chart showing the identity of common mutations (any of the top 20 single nucleotide mutations from the June 2 alignment), so that sectors of the tree containing a particular mutation at high frequency are subtended by an inner colored arc; mutations not in the top 20 are presented together in gray. The tree is rooted on a reference sequence derived from the original Wuhan isolates (GenBank accession number NC 045512), at the 3 o’clock position. Branch ends representing sequence isolates bearing the D614G change are decorated with a gray square; sectors of the tree containing that mutation are subtended by a dark blue arc in the inner element; other mutations are denoted by different colors. As an example, in this tree, the region from approximately 12:30 to 3 o’clock represents GISAID’s “GR” clade, defined both by mutations we are tracking in this paper that carry the G614 variant (the GISAID G clade, defined by mutations A23403G, C14408T, C3037T, and a mutation in the 5' UTR (C241T, not shown here), and an additional 3 position polymorphism: G28881A + G28882A + G28883C. These base substitutions are contiguous and result two amino acid changes, including N204R, hence GISAID’s “GR clade” name. Close examination of this triplet in sequences from the Sheffield dataset suggests the mutations are not a sequencing artifact. The outer phylogenetic tree was computed using oblong (see STAR Methods), and plotted with the APE package in R. The inner element is a bar chart plotted with polar coordinates using the gglot2 package in R. The frequency of the GR clade appears to be increased in the UK and Europe as a subset of the regional G clade expansion, given that both carry G614D.
A. IGV plots showing bam files from nanopore sequencing data of amplicons produced by the Artic network protocol. Raw data from amplicon 81 contains a portion of adaptor sequence which is homologous to the reference genome, apart from the C variants which lead to a S943P mutation call. This region is therefore included in variant calling if location based trimming is not carried out. Subsequent panels show that this region is soft clipped when trimming adapters and primers and is therefore not available for variant calling.

B. Base frequencies at position 24389 in 23 samples from the Sheffield data show that C is present in half of the reads in the raw data, but is absent from trimmed and primer trimmed data. This figure is also associated with the STAR Methods.