bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2

Korber B[1], Fischer WM[1], Gnanakaran S[1], Yoon H[1], Theiler J[1], Abfalterer W[1], Foley B[1], Giorgi EE[1], Bhattacharya T[1], Parker MD[3], Partridge DG[4], Evans CM[4], de Silva TI[4,5], on behalf of the Sheffield COVID-19 Genomics Group[#], LaBranche CC[2], and Montefiori DC[2]

[1]T6: Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545 USA
[2]Duke Human Vaccine Institute & Department of Surgery, Durham, North Carolina, 27710 USA
[3]Sheffield Biomedical Research Centre & Sheffield Bioinformatics Core, University of Sheffield, Sheffield, UK.
[4]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.
[5]Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK.

[#]Members of Sheffield COVID-19 Genomics Group: Adrienne Angyal, Rebecca L. Brown, Laura Carrilero, Luke R. Green, Danielle C. Groves, Katie J Johnson, Alexander J Keeley, Benjamin B Lindsey, Paul J Parsons, Mohammad Raza, Sarah Rowland-Jones, Rachel M. Tucker, Dennis Wang, Matthew D. Wyles

Corresponding Author and Lead Contact: Bette Korber, btk@lanl.gov

Declaration of Interests: The authors declare no competing interests.

Summary
We have developed an analysis pipeline to facilitate real-time mutation tracking in SARS-CoV-2, focusing initially on the Spike (S) protein because it mediates infection of human cells and is the target of most vaccine strategies and antibody-based therapeutics. To date we have identified fourteen mutations in Spike that are accumulating. Mutations are considered in a broader phylogenetic context, geographically, and over time, to provide an early warning system to reveal mutations that may confer selective advantages in transmission or resistance to interventions. Each one is evaluated for evidence of positive selection, and the implications of the mutation are explored through structural modeling. The mutation Spike D614G is of urgent concern; it began spreading in Europe in early February, and when introduced to new regions it rapidly becomes the dominant form. Also, we present evidence of recombination between locally circulating strains, indicative of multiple strain infections. These finding have important implications for SARS-CoV-2 transmission, pathogenesis and immune interventions.

KEYWORDS: COVID-19, pandemic, diversity, evolution, Spike, positive selection, recombination, reinfection, antibodies, transmission rates

Introduction:

The past two decades have seen three major highly pathogenic zoonotic outbreaks of betacoronaviruses (Cui et al., 2019; de Wit et al., 2016; Fehr et al., 2017; Lu et al., 2020; Wu et al., 2020a). The first was Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) in

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2002, which infected over 8,000 people and killed 800 (Graham and Baric, 2010). This was followed in 2012 by Middle East Respiratory Syndrome, MERS-CoV in 2012, a difficult to transmit but highly lethal virus, with 2,294 cases as of 2019, and 35% mortality (Cui et al., 2019; Graham and Baric, 2010). The third, SARS-CoV-2 is the cause of the severe respiratory disease COVID-19 (Gorbalenya et al., 2020). It was first reported in China in late December 2019 (Zhou et al., 2020), and triggered an epidemic (Wu et al., 2020b) that rapidly spread globally to become a pandemic of devastating impact, unparalleled in our lifetimes; today's World Health Organization (WHO) situation report reads: over 2.5 million confirmed cases of COVID-19, and over 175,000 deaths (WHO Situation Report Number 94, April 23); tomorrow's report will bring us new and markedly higher tallies of suffering, as the WHO continues to track the remarkable pace of expansion of this disease.

Three related factors combine to make this disease so dangerous: human beings have no direct immunological experience with this virus, leaving us vulnerable to infection and disease; it is highly transmissible; and it has a high mortality rate. Estimates for the basic reproductive ratio, R_0, vary widely, but commonly range between 2.2-3.9 (Lv et al., 2020). Estimates of mortality, deaths per confirmed cases, also vary widely, and range between 0.5-15% (Zhou et al., 2020) (Mortality Analyses, John Hopkins University of Medicine). Differences in mortality estimates will reflect regional access to testing, as milder cases are confirmed more often when more testing is deployed, as well as regional differences in clinical care, and population differences in associated risk factors such as age. These basic numbers, R_0 and mortality, are critical for public health response planning, but are difficult to resolve with confidence or to generalize across populations, given limited diagnostic testing and variations in the strategies of estimation.

Although the observed diversity among pandemic SARS-CoV-2 sequences is low, its rapid global spread provides the virus with ample opportunity for natural selection to act upon rare but favorable mutations. This is analogous to the case of influenza, where mutations slowly accumulate in the hemagglutinin protein during a flu season, and there is a complex interplay between amino acids that can confer immune resistance to the virus, and the fitness landscape of the particular variant in which they arise (Wu et al., 2020c). Antigenic drift in influenza, the accumulation of mutations by the virus during an influenza season, provides the baseline variation needed to enable selection for antibody resistance across populations, and this drift is the primary reason we need to develop new influenza vaccines every few seasons. The seasonality of influenza is likely to be dictated in part by weather patterns (Chattopadhyay et al., 2018); longer seasonal epidemics allow selection pressure to continue over a more extended period, enhancing opportunities for the development of virus with novel antigenic surfaces that resist pre-existing immunity (Boni et al., 2006). SARS-CoV-2 is new to us; we do not yet know if it will wane seasonally as the weather warms and humidity increases, but our lack of pre-existing immunity and its high transmissibility relative to influenza are among the reasons it may not. If the pandemic fails to wane, this could exacerbate the potential for antigenic drift and the accumulation of immunologically relevant mutations in the population during the year or more it will take to deliver the first vaccine. Such a scenario is plausible, and by attending to this risk now, we may be able avert missing important evolutionary transitions in the virus that if ignored could ultimately limit the effectiveness of the first vaccines to clinical use.

There is clearly an urgent need to develop an effective vaccine against SARS-CoV-2, as well as antibody-based therapeutics (Kumar et al., 2020). Over 62 vaccine approaches are currently being explored, and a wide variety of candidate SARS-CoV-2 vaccines are in development (Landscape of COVID-19 Variants, WHO). Most of these vaccine approaches target the trimeric Spike protein (S) with the goal of eliciting protective neutralizing antibodies. Spike mediates

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

binding and entry into host cells and is a major target of neutralizing antibodies (Chen et al., 2020; Yuan et al., 2020). Each Spike monomer consists of an N-terminal S1 domain and a membrane-proximal S2 domain that mediate receptor binding and membrane fusion, respectively (Hoffmann et al., 2020; Walls et al., 2020; Wrapp et al., 2020). Notably, current immunogens and testing reagents are generally based on the Spike protein sequence from the index strain from Wuhan (Wang et al., 2020). SARS-CoV-2 is closely related to SARS-CoV; the two viruses share ~79% sequence identity (Lu et al., 2020) and both use angiotensin converting enzyme-2 (ACE2) as their cellular receptor (Hoffmann et al., 2020; Li et al., 2005; Wrapp et al., 2020), however the SARS-CoV-2 S-protein has a10-20-fold higher affinity for ACE2 than the corresponding S-protein of SARS-CoV (Wrapp et al., 2020). It remains to be seen to what extent lessons learned from SARS-CoV are helpful in formulating hypotheses about SARS-CoV-2, but SARS-CoV studies suggest that the nature of the antibody responses to the Spike protein are complex. In SARS-CoV infection, neutralizing Abs are generally thought to be protective; however, rapid and high neutralizing Ab responses that decline early are associated with greater disease severity and a higher risk of death (Ho et al., 2005; Liu et al., 2006; Temperton et al., 2005; Zhang et al., 2006). Furthermore, some antibodies against Spike mediate antibody dependent enhancement (ADE) of SARS-CoV (Jaume et al., 2011; Kam et al., 2007; Wan et al., 2020; Wang et al., 2014; Yilla et al., 2005; Yip et al., 2016; Yip et al., 2014). Because of the short duration of the outbreak, there were no efficacy trials of SARS-CoV vaccines and therefore we lack critical information that would help guide SARS-CoV-2 vaccine development.

Given Spike's vital importance both in terms of viral infectivity and as an antibody target, we felt an urgent need for an "early warning" pipeline to evaluate Spike pandemic evolution. Our primary intent is to identify dynamically changing patterns of mutation indicative of positive selection for Spike variants. Also, because recombination is an important aspect of coronavirus evolution (Graham and Baric, 2010; Li et al., 2020; Rehman et al., 2020), we also set out to determine if whether recombination is playing a role in SARS-CoV-2 pandemic evolution. Here we describe a three-stage data pipeline (analysis of daily GISAID data, structural modeling of sites of interest, and experimental evaluation) and the identification of several sites of positive selection, including one (D614G) that may have originated either in China or Europe, but begin to spread rapidly first in Europe, and then in other parts of the world, and which is now the dominant pandemic form in many countries.

Results

Overview

Over the past two months, the HIV database team at Los Alamos National Laboratory has turned to developing an analysis pipeline to track in real time the evolution of the SARS-CoV-2 Spike (S) protein in the COVID-19 pandemic, using the Global Initiative for Sharing All Influenza Data GISAID SARS-CoV-2 sequence database as our baseline (Sup. Item 1 is the GISAID acknowledgments table, listing all the groups who contribute sequences to this global effort) (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). GISAID is the primary COVID-19 sequence database resource, and our intent is to complement what they provide with visualizations and summary data specifically intended to support the immunology and vaccine communities, and to alert the broader community to changes in frequency of mutations that might signal positive selection and a change in either viral phenotype or antigenicity. New global sequences arrive at GISAID at a furious pace; currently, hundreds of new SARS-CoV-2 sequences are added each day (Sup. Fig 1). By automating a series of key analysis steps with

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

daily inputs of GISAID data, critical analysis is provided in real time. The figures for this paper were based on an April 13, 2020 download of the GISAID data to enable the preparation of this manuscript. Our analysis pipeline (www.cov.lanl.gov) enables readers to reproduce the key figures (*e.g.* Figs. 1-3, and an updated table of sites of interest, Item S2) based on contemporary data downloaded daily from GIASID. The pipeline was developed in collaboration with the neutralizing antibody evaluation team at Duke University, who are in the process of creating a neutralizing antibody pseudovirus testing facility to experimentally resolve the virological and immunological implications of mutations of interest in Spike. While developing the computational tools to enable the analysis pipeline, the Los Alamos group has concurrently provided reports, at roughly two-week intervals, to colleagues at Duke as well as to others in the community who are working in the early design and testing phases of Spike-targeting vaccines and immunotherapeutics. This paper is essentially a formalized version of the fourth such report, based on the April 13th GISAID data, and marks the transition of our analysis pipeline to a public resource and webpage (www.cov.lanl.gov).

The analysis pipeline begins by downloading the GISAID data daily, then discarding partial and problematic sequences using stringent inclusion criteria, and then trimming jagged ends of the sequences back to the beginning of the first open reading frame (orf) and the end of the last orf. We use this to create two basic codon-aligned alignments (Kurtz et al., 2004): a more comprehensive Spike alignment (6,346 sequences as of April 13th), to monitor mutations that are beginning to accrue in Spike while maximizing the sample size; and a full genome alignment (4,535 sequences as of April 13th), to enable tracking the Spike mutations in a phylogenetic context informed by the evolutionary trajectories of the full genome (Fig. 1). Daily updates of analyses based on these alignments are provided through the pipeline.

Mutations among the pandemic SARS-CoV-2 sequences are sparse, limiting the applicability of traditional sequence-based methods for detecting positive selection. An alternative analysis framework for identifying positive selection can be taken, however, based on GISAID data, which provides a rich database of thousands of sequences linked to geographic information and sampling dates. This enables the tracking of sites for early indications of positive selection by identifying shifts in mutational frequencies over time. Early indicators include: (i) an increasing frequency of sequences that exhibit a particular mutation over time in a local region; (ii) frequent recurrent identification of a particular mutation in different geographic regions, and in different regions of the phylogenetic tree; (iii) the use of different codons to encode the same recurrent amino acid; and (iv) tight clusters of mutations in linear or structural sequence space. Because Spike mutations are rare, we set a low threshold for a site to be deemed “of interest” for further tracking. Thus, when a mutation is found in 0.3% of the sequences, we begin to track it by exploring its evolutionary trajectory and modeling its structural implications, *e.g*., its potential impact on antibody binding sites, trimer stability, and glycosylation patterns. Accumulation of mutations in known antibody epitopes are a special focus, including sites in or near the Spike Receptor Binding Domain (RBD), as well as sites that are near an immunodominant enhancing antibody epitope from the first SARS epidemic that centers on the mutation D614G (Wang et al., 2016); thus, a lower tracking threshold of 0.1% has been adopted for sites in Spike that are within 4 Å of these epitopes in the Spike 3D structure. If, given this information, a site merits further characterization, the experimental pipeline is triggered, and reagents are ordered and put into the queue for experimental evaluation of its impact on infectivity, antigenicity, neutralization sensitivity, and capacity to bind to the ACE2 receptor.

A summary of the sites of interest (as of April 13th) included 14 positions, as well as one local cluster of mutations (Table 1); a detailed report for each site is provided in supplemental data Item S2. Some of the sites are diminishing in frequency, and may have simply reflected a local

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sampling artifact from an earlier data set; they are still included in Item S2 because they reached the 0.3% threshold in a past sample of GISAID. Still others are increasing in frequency or persisting (Sup. Item 2), or have other interesting features suggesting they merit continued monitoring, such as recurrence in many regions of a phylogenetic tree (Fig. 1), or being located in a part of Spike that is of structural interest. We discuss two Spike mutations of particular interest, D614G and S943P, at some length, and then briefly summarize the other sites of interest. D614G is increasing in frequency at an alarming rate, indicating a fitness advantage relative to the original Wuhan strain that enables more rapid spread. S943P is located in the fusion core region, and is of particular interest as it is spreading via recombination.

The D614G mutation

Increasing frequency and global distribution. The mutation D614G (a G-to-A base change at position 23,403 in the Wuhan reference strain) was the only site of interest identified in our first Spike mutation report in early March (Fig. 1A); it was found 7 times in 183 sequences that were available at the time. Four of these seven first D614G strains were sampled in Europe, and one each in Mexico, Brazil, and in Wuhan. In 5/7 cases, D614G was accompanied by 2 other mutations: a silent C-to-T mutation in the nsp3 gene at position 3,037, and a C-to-T mutation at position 14,409 which results in a RNA-dependent RNA polymerase (RdRp) amino acid change (RdRp P323L).The combination of these three mutations forms the basis for the clade that soon emerged in Europe (Fig. 1). By the time of our second report in mid-March, D614G was being tracked at the GISAID due to its high frequency, and referred to as the “G” clade; it was present in 29% of the global samples, but was still found almost exclusively in Europe. It was recently reported by Pachetti *et al.* to be found in Europe and absent from regions globally, presumably because they capture GISAID data in roughly the time frame as our second report (Pachetti et al., 2020). The data available for study mid-March, given an approximate 2-week lag time between sampling and reporting, were consistent with the possibility of a founder effect in Europe resulting in spread across the continent, coupled with an increase in European sampling in the database. However, an early April sampling of the data from GISAID showed that G614’s frequency was increasing at an alarming pace throughout March, and it was clearly showing an ever-broadening geographic spread (Fig. 1A).

To differentiate between founder effects and a selective advantage driving the increasing frequency of the G clade in the GISAID data, we applied the suite of tools that we had been developing for the SARS-CoV-2 analysis pipeline (Fig. 2, Fig. 3, Fig. S2 and Fig. S3). A clear and consistent pattern was observed in almost every place where adequate sampling was available. In most countries and states where the COVID-19 epidemic was initiated and where sequences were sampled prior to March 1, the D614 form was the dominant local form early in the epidemic (orange in Figs. 2 and 3). Wherever G614 entered a population, a rapid rise in its frequency followed, and in many cases G614 became the dominant local form in a matter of only a few weeks (Fig. 3 and S3).

In Europe, where the G614 first began its expansion, the D614 and G614 forms were co-circulating early in the epidemic, with D614 more common in most sampled countries, the exceptions being Italy and Switzerland (Fig. 2A). Through March, G614 became increasingly common throughout Europe, and by April it dominated contemporary sampling (Fig. 2 and 3). In North America, infections were initiated and established across the continent by the original D614 form, but in early March, the G614 was introduced into both Canada and the USA, and by the end of March it had become the dominant form in both nations. Washington state, the state with the greatest number of available GISAID SARS-CoV-2 sequences from the USA, exemplifies this pattern (Figure 3 and S3), and a similar shift over time is evident in many other

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

states with samples available throughout March (detailed data by state provided in Item S2). Sequences from New York were poorly sampled until some days into March (Fig. 3), and the G614 form was the predominant form; the G614 form was coming into prominence elsewhere in the USA by that time. Thus it is not clear whether the local pandemic was seeded by European contacts, as suggested in (Brufsky, 2020), if it was seeded by contacts from within the USA where it had already achieved high prevalence, or by a combination of both routes. Australia follows the same transition pattern, from D614 to G614 dominance, as the USA and Canada. Iceland is the single exception to the pattern; there sequencing was extensive and the epidemic seems to have started with a G614 form, but there was a transient increase in the D614 form, which then persisted at constant low level (Fig. 3 and S3). Asian samples were completely dominated by the original Wuhan D614 form through mid-March, but by mid-March in Asian countries outside of China, the G614 form was clearly established and expanding (Figs 2 and 3). The status of the D614G mutation in China remains unclear, as very few Chinese sequences in GISAID were sampled after March 1. South America and Africa remain sparsely sampled, but are shown for completeness in Figs. 2 and 3; for details see Item S2. Taken together, the data shows that G614 confers a selective advantage that is repeatedly reflected by dramatic shifts to G614 forms in regional epidemics over a period of several weeks.

The earliest D614G mutation in Europe was identified in Germany (EPI_ISL_406862, sampled 1/28/2020), and it was accompanied by the C-to-T mutation at 3,037, but not by the mutation at 14,409. Of potential interest for understanding the origin of the G614 clade, the G614 form of the virus was also found 4 times in China among early samples (Item S2). A Wuhan sequence (EPI_ISL_412982) sampled on 2/7/2020 had the D614G mutation, but it did not have either of the two accompanying mutations, and this single D614G mutation may have arisen independently (Fig. 1). The other three cases were potentially related to the German sequence. One sequence sampled on 1/24/2020 was from Zhejiang (EPI_ISL_422425), which borders Shanghai, and it had all three mutations associated with the G clade that expanded in Europe. The other two samples were both from Shanghai, and sampled on 1/28/2020 and 2/6/2020 (EPI_ISL_416327, EPI_ISL_416334); like the German sequence they lacked the mutation at 14,409. Given that these early Chinese sequences were also highly related to the German sequence throughout their genomes, it is possible that the G614 may have originated either in China or in Europe as it was present in both places in late January. There are no recent GISAID sequences from Shanghai or Zhejiang at this time, so we do not know if G614 was preferentially transmitted there, but the D614 form was prevalent in Shanghai in January and early February.

D614G and potential mechanisms for enhanced fitness. There are two distinct conceptual frameworks that may explain why the D614G mutation is associated with increased transmission. The first is based on structure. D614 is located on the surface of the spike protein protomer, where it can form contacts with the neighboring protomer. Examination of the cryo-EM structure (Wrapp et al., 2020) indicates that the sidechain of D614 potentially can form a hydrogen bond with T859 of the neighboring protomer as shown in Fig. 4C; the strength of Asp-Thr hydrogen bonding has been well documented (Kandori et al., 2001). This protomer-protomer hydrogen bonding may be of critical importance as it can bring together a residue from the S1 unit of one protomer to the S2 unit of the other protomer Fig. 4D. These two sites in the spike protein bracket both the dibasic furin- and S2-cleavage sites; thus, it is possible that the *D614G* mutation diminishes the interaction between the S1 and S2 units, facilitating the shedding of S1 from viral-membrane-bound S2. An alternative structure-based hypothesis is that this mutation may impact RBD-ACE2 binding (albeit indirectly since this site is not proximal to the binding interface). The RBD needs to be in the “up” position to engage with the ACE2 receptor, and it is possible that this site could allosterically alter the transitions of “up” and “down”. In the only known SARS-CoV-2 Spike structure with all RBDs in “down” position, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distances between *D614* and T859 remain the same between all protomers (Walls et al., 2020). However, in the structures with one RBD "up," the distances between these residues are altered (Walls et al., 2020; Wrapp et al., 2020). The distance between the "up" protomer and the "down" neighboring protomer (clockwise) is slightly longer than the rest of the protomer-protomer distances (Fig. S4D). These slight changes are, however, most likely within the conformational fluctuations of a dynamic spike trimer. More detailed experimental and modeling studies are needed to elucidate the effect of this mutation on RBD transitions.

The second way the D614G mutation might impact transmission is immunological. D614 is embedded in an immunodominant linear epitope in the original SARS-CoV Spike, $S_{597–603}$. This peptide had a very high level of serological reactivity (64%), and induced long term B-cell memory responses in convalescent-phase sera from individuals infected during the original SARS-CoV 2002 epidemic (Wang et al., 2016). Antibodies against this peptide mediate antibody-dependent enhancement (ADE) of SARS-CoV infection by an epitope-sequence-dependent mechanism, both *in vitro*, and *in vivo* in rhesus macaques (Wang et al., 2016). The minimal linear core epitope for the ADE mediating antibodies in SARS-CoV was *LYQDVNC* (SARS-CoV $S_{597-603}$), and this peptide is immediately proximal to a peptide that is targeted by potentially beneficial neutralizing antibodies (SARS-CoV $S_{604-625}$). The ADE target peptide spans the SARS-CoV-2 D614 site, and is identical to the equivalent region in SARS-CoV, $S_{611-617}$. Wang et al. noted proximity of this epitope to the RBD, and speculated that antibody binding may mediate a conformational change in Spike that increases RBD-ACE2 interaction resulting in the enhancing effect. This mechanism is notably different from the more common Fc receptor-mediated mechanism of SARS-CoV ADE reported by others, and which occurs in both the presence and absence of ACE2 (Jaume et al., 2011; Kam et al., 2007; Wan et al., 2020; Wang et al., 2014; Yilla et al., 2005; Yip et al., 2016; Yip et al., 2014). Thus, based on currently available information, there are several ways the D614G mutation may impact Spike's infectivity: it may improve receptor binding, fusion activation, or ADE antibody elicitation. Another mechanism for the shift to the G614 form at later times points might simply be through antigenic drift mediating antibody escape. If the D614G mutation in SARS-CoV-2 was impacting neutralizing antibody sensitivity as well as, or instead of, the ADE activity observed in the SARS-CoV study, D614G could also be mediating escape that makes individuals susceptible to a second infection.

D614G and clinical outcome. We were concerned that if the D614G mutation can increase transmissibility, it might also impact severity of disease. Because clinical outcome data are not available in GISAID, we focused on a single geographic region, Sheffield, England, where a large data set existed and was made available for an initial exploration of this question. SARS-CoV-2 sequences were generated from 453 individuals presenting with COVID-19 disease at the Sheffield Teaching Hospitals NHS Foundation Trust. Sheffield followed the pattern observed through much of world, starting out with D614, and shifting to predominantly G614 by the end of March. The Sheffield data included age, gender, date of sampling, cycle threshold (CT) for positive signal in E gene-based RT-PCR (used here as a surrogate for relative levels of viremia) (Corman et al., 2020), and clinical status: outpatient (OP), inpatient (IP, requiring hospitalization), or admittance into the intensive care unit (ICU). Because the numbers admitted into in the ICU were small, and because that information was not readily available for all subjects, we grouped the IP and ICU together for most of the statistical analysis. As anticipated, there was a significant relationship between hospitalization and age (Fig. S5) (Wilcoxon $p < 2.2e-16$, median age 74, interquartile range (IR) 58-83 for the hospitalized patients, versus a median of age 43 (with IR 31-53) for the outpatients (Dowd et al., 2020; Promislow, 2020). Also, males were hospitalized more often than females (Fisher's $p = 1.95e-13$) (Conti and Younes, 2020; Promislow, 2020). Furthermore, fewer PCR cycles were required for detection of virus

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

among individuals who admitted to the hospital, indicating higher viral loads (Fig. S5) (Wilcoxon p = 4.8e-06, median 25, IR 21-28, versus median 19, IR 21-25) (Fig. S5) (Liu et al., 2020). *There was, however, no significant correlation found between D614G status and hospitalization status*; although the G614 mutation was slightly enriched among the ICU subjects, this was not statistically significant (Fig. 5C).

We were concerned sampling issues might have introduced a bias, particularly as the D614 viruses were more heavily sampled early, and G614 at later times, and clinical practice might have changed over the course of the time period. Age, however, was evenly distributed between the G614 and D614 hospitalized groups (Fig 5B), and the relative number of hospitalizations stayed constant throughout the study period in both the D614 and G614 groups (Fig 5A). We also performed a multivariate Generalized Linear Model (GLM) analysis, (Bates et al., 2015) with outpatient vs. hospitalized status as the outcome, and age, gender, D614, and PCR CT as potential predictors. PCR CT was a significant predictor of disease severity, after adjusting for age and gender, but D614G status was not. Again, age was the most significant predictor ($p<2x10^{-16}$), followed by gender ($p=2 x10^{-7}$) and then PCR CT ($p=1.9x10^{-6}$). D614G status did not significantly contribute to modeling hospitalization as an outcome, but there was a marginally significant interaction with PCR CT (p=0.04).

While D614G did not predict hospitalization, there was a significant shift in cycle threshold to fewer PCR cycles being required for detection among the group that carried G614 relative to D614 (Fig. 5D). This indicated that patients carrying the G614 mutation had higher viral loads (Wilcoxon p = 0.003, median 23.1, IR 19.5-26.6, versus median 24.9, IR 21.2-27.6). This comparison is limited by uncertainty regarding the time from infection at which sample was taken, and by the fact that PCR is an indirect measure of viral load; still it is notable that despite these limitations, a significant difference was observed.

The S943P mutation

Recombination and the evolution of the S943P mutation. This site had an interesting evolutionary trajectory. It is found only in Belgium, and although the mutation is local in terms of geographic spread, it did not emerge as a single lineage from a local founder, but instead recurs in many distinctive lineages in the phylogeny that are co-circulating in Belgium, a pattern suggestive of recombination (Fig. 1B). Knowing that recombination plays an important role in coronavirus evolution generally (Graham and Baric, 2010; Li et al., 2020; Rehman et al., 2020), the possibility that recombination might be also be contributing significantly to evolution in the current pandemic seemed plausible, but difficult to detect using standard methods given how little variation there is among pandemic strains. Recombination requires simultaneous infection of the same host with different viruses, and the two parental strains have to be distinctive enough to manifest in a detectable way in the recombined sequence. Both criteria were met in Belgium, as the samples were relatively geographically clustered, providing the opportunity for coinfection, and the set was sufficiently diverse to enable detection. Thus, we decided to explore further the possibility that recombination was driving the spread of the S943P in Belgium.

In our April 13th GISAID Spike protein data set, the S943P mutation was found 23 times among the 284 SARS-CoV-2 Spike sequences from Belgium, but never among the rest of globally sampled 6,063 Spike sequences from outside of Belgium. The chances that this could be a random distribution are vanishingly small (Fisher's p < 2.2e-16), leaving three possibilities that could explain the cluster in Belgium. First, one might a priori assume that the observed concentration of sequences with the S943P mutation are the result of a local lineage circulating

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in Belgium resulting from a founder effect, but the tree shown in Fig. 1B clearly demonstrates that this is not the case. While a "local lineage" hypothesis could explain *other* sites of interest, like A831V (Fig. 1B), it was not consistent with S943P's dispersion in the phylogenetic tree. Second, one might imagine that S943P to be a recurrent mutation that was under strong positive selection. But if so, why was that selection only operative within Belgium? Furthermore, the S943P mutation comes with a double base mutation in its codon, one of which is a transversion: AGT (S) -> CCT (P), thus not one but two base mutations would be required to have arisen together and repeatedly been co-selected. This is extremely unlikely given that *de novo* mutations are exceedingly rare among SARS-CoV2 pandemic strains. For example, among the 284 Belgian sequences under consideration, of length 29,427, there were 8,357,229 unambiguous base calls. Of these, only 1198 carried mutants relative to the consensus sequence in Belgium — only one base in 10,000 was different.

The third possibility is that the evolution of the S943P mutation was a result of recombination of distinct viruses within a coinfected host, and this is well supported by the data. The diversity of the full set of data from Belgium is depicted in a highlighter plot in Fig. S6; patterns suggestive of recombination are circled. To facilitate visualization of the evolutionary relationships, Fig. 6 shows just 98 positions out of the 29,427 sites in the full alignment that were phylogenetically informative, in that they had 2 or more shared substitutions within the dataset. This facilitates visual tracking of the evolutionary relationships between the viruses. In this compact alignment, there are 16 unique variants that carry the P943 paired base mutations in the CCT codon, and one that carries S943T, due to a single base change (AGT -> ACT). Nine of these distinctive P943-carrying variants perfectly match another sequence variant within the set from Belgium in all phylogenetically informative positions except for the P943 codon mutations, indicating that the P943 region of the virus was being inserted into multiple distinctive backbones (Fig. 6). The other 7 sequences carrying the P943 CCT codon were also very similar (1-2 base difference across the full genome) to at least one other sequence in the Belgium set, again supporting the idea that the P943 mutation was being carried into additional distinctive backbones circulating in Belgium via recombination (Fig 6). We used a phylogenetic strategy to estimate the probability that this pattern could have been due to spontaneous mutation (Paradis and Schliep, 2018; Schliep, 2010). In the subtrees corresponding to sequences from Belgium, the pair of bases in the AGT -> CCT codon (once ACT) changed very often: 9 and 10 times respectively. Based on the mutation rates calculated excluding all Belgian sequences, these sites should have mutated no more than 0.18 times on average. The probability of these sites changing 9 and 10 times by chance is about 4×10^{-13} and 7×10^{-15}, respectively. Thus, this rapid change in the phylogeny could not be explained as *de novo* point mutations, but it is consistent with recombination. Of note, the two base changes in this codon that encodes position 943 in spike were the *only* two sites found to be significantly different in the Belgian subtree distribution than expected based on their rate of mutation, after multiple testing correction for the number of sites in the alignment (i.e., $p<0.05$ with Bonferroni correction). This suggests that they were being strongly selected for after the recombination events, although the frequency of the P943 mutation is not increasing among samples from Belgium over time (Fig. S7). Other patterns suggestive of recombination among circulating forms in Belgium were also evident, and are indicated in Fig 6.

To determine if recombination events could be identified in other geographically regional data sets using more established strategies, we applied a computational method called RAPR (Song et al., 2018) that we had originally developed to explore the evolutionary role of within-patient HIV recombination in acute HIV infection—another situation of low viral diversity. RAPR enables the comparison of all triplets (sets of three sequences) in an alignment, and applies a run-length statistic to evaluate the possibility of recombination. We began this analysis with the set of SARS-CoV-2 sequences from Washington State, as it was particularly well sampled set and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from a geographically local population where co-infection might occur. Using RAPR, we found several recombination candidates, and the two most significant examples of these are shown in (Fig. S8); given the very low rate of mutation among the pandemic sequences, the patterns of shared sites shown in Fig. S8 are highly unlikely to have arisen by spontaneous mutation. Furthermore, RAPR analysis indicated that one of the recombinant forms in Washington founded a lineage that continued to spread. To identify other geographic regions where recombination might be occurring among the sequences in the global data set, we next focused on sequences sampled from areas with co-circulation of both haplotypes of the D614G mutation, as defined by 3 mutational events (Fig. 1): the a C-to-T mutation at position 3,037, a C-to-T mutation at position 14,409, and the G-to-A base change at position 23,403 that gives rise to the D614G amino acid change in spike. The original Wuhan D614 form carried the bases 'C-C-G' in these positions, and the G614 form carries the bases 'T-T-A'. Given that these positions were well spaced in the genome, and the two forms were co-circulating in many communities over the month of March, we reasoned that deviations from either the 'C-C-G' or the 'T-T-A' patterns would be indicative of recombination. Such deviations were rare, found only 34 times (0.8%) in our full genome alignment of 4,535 sequences. 24/34 (70%) were from Belgium, indicating that SARS-CoV-2 samples from Belgium are particularly enriched for recombination events. (This raises the possibility that the S943P mutation, found only in Belgium to date, and an apparent target of recombination and selection (Fig. 6), may be enabling co-infection through conferring immunological resistance; this hypothesis is speculative, but something that can be experimental tested.) Three haplotype deviations were found in the Netherlands (8.8% of the 34), the other seven from scattered locations. Using RAPR to identify likely recombination sets in Belgium and the Netherlands, we were indeed able to identify recombination candidates based on RAPR output, and the most significant examples from these countries are also shown in Fig. S8. Thus, RAPR analysis further supports the findings in Fig. 6, that recombination is playing a role in the evolution of pandemic strains.

The S943P mutation and cluster of mutations at the fusion core of the HR1 region. We are tracking local clusters of mutations as well as single mutations in the pipeline by estimating the average per-position entropy of a sliding window of 10 amino acids (10-mers) scanning the Spike alignment (Item S2). The highest entropy 10-mers, not surprisingly, span position 614. The next highest entropy region spans a cluster of mutations that includes S943 (Fig. S9). S937-S943 occurs in the fusion core of the HR1 region (Xia et al., 2020), specifically where the HR1 helix is broken in the trimeric pre-fusion spike (**Fig. 4 (E**)). This cluster is rich in Ser and Thr. These residues have a high propensity to form hydrogen bonds. Upon spatially localizing in a cluster, they have a tendency to form hydrogen bonding among themselves. When such a cluster occurs in helices, it can enhance the association of helices. For example, SXXSSXXT-like motifs have been shown to drive the association of trans-membrane helices (Dawson et al., 2002) and could be even more relevant given the association with amphipathic helices. Given that, this cluster in the HR1 region of S2 unit could impact any of the following three conformational adjustments as the S2 unit transitions from pre-fusion to post-fusion (Fig. S4): (i) promoting the elongation of HR1 helices, (ii) maintaining amphipathicity when the a single HR1 helix is extended, and (iii) enhancing the association of the HR1 trimer. Even though such a cluster of Ser-Thr is not seen in HIV gp41, it is observed near the MPER region in γ-retroviral glycoproteins and has been suggested as a possible conserved mechanism to drive oligomerization (Salamango and Johnson, 2015). Finally, the mutational sites A831 and D839 occur in the region of fusion peptide. Neither position is structurally resolved in the cryo-EM reconstruction though site A831 is part of the canonical fusion peptide.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Additional sites of interest in Spike

Sites L5F and L8V. These are both signal peptide mutations, and it is difficult anticipate how they might impact the virus. Variation in the signal peptide of other viruses, for example HIV Env, can impact posttranslational modifications in the endoplasmic reticulum, including folding, expression levels, and glycosylation (Asmal et al., 2011; Upadhyay et al., 2018). The L5F mutation is intriguing because of its recurrence in many lineages throughout the SARS-CoV-2 phylogenetic tree, and in many different countries throughout the world (Fig 1). Once established, it is often regionally transmitted and contributes to multiple small local clusters (Fig. 1, Item S2) – *e.g.* a cluster of 5 infections in Iceland with identical in sequences all carrying 5F, and several comparable clusters in different states in the USA). This combination suggests that it may be a favorable mutation that tends to persist when it arises. Despite its recurrences, it has maintained but not increased in frequency, and continues to be found at roughly 0.6% of the global sampling through April.

L8V is potentially interesting because of a very different evolutionary trajectory. It is mostly found in a single lineage in Hong Kong, with one apparent recurrence in Canada. Because Hong Kong has not been recently sampled in GISAID, L8V's overall global frequency appears to decline in April; however, L8V may in fact be increasing in frequency in Hong Kong over time and thus merits continuing scrutiny (Fig. S7)

Sites V367F, G476S, and V483A. There are three mutational sites, V367F, G476S, and V483A, that are found within the RBD domain (Fig. 4 A,B). Of these three, only G476S occurs directly at the binding interface of RBD and pepidase domain of ACE2. Fig. 4B shows that G476 is at the end of the binding interface that is predominantly driven by polar interactions (Yan et al., 2020). The closest ACE2 residue to G476 is Q24, which is near the end of α-1, the helix of ACE2 most engaged in RBD recognition. The ACE2 residue Q24 makes distinct electrostatic interactions with the backbone of A475, the neighboring residue to G476 (Fig. S4 B). The G476S mutation may contribute to the rigidification of the local loop region in RBD near the binding interface, or bring this flexible loop even closer to the interface. The mutation at site V483 does not directly contact ACE2 although it is on the same face of the RBD that forms the binding interface with the ACE2. Site V367 is at the opposite end from where RBD binds to ACE2 receptor; it is on the same face as the epitope of CR3022, a neutralizing antibody that was isolated from a convalescent SARS-CoV patient when at least two of the RBD regions of the Spike trimer are "up", (Fig. S4 C) though no direct contacts between V367 and CR3022 are observed (Yuan et al., 2020).

Sites V367F and G476S were identified as mutations of interest in early smaller datasets, and appear to be diminishing in overall global frequency in later samples, but V367F merits continued scrutiny due to its potential for interactions with ACE2. The mutation V483A has predominately appeared in Washington State (Item S2). It is not increasing over time among samples from Washington, but it maintains a steady, albeit very low, presence (Fig. S7).

Sites H49Y, Y145H/del, Q239K. Each of these mutations are located in S1 N-terminal domain (NTD), a domain not well characterized functionally. The sites were identified as sites of interest in early smaller datasets, but appear to be diminishing in overall frequency in later samples. They recur in different countries, although K239 is predominantly found in the Netherlands.

Sites A831V and D839Y/N/E. A831V and D839 are both being maintained at ~0.4% in the global sample through April. A831V is found only in Iceland and is in a single lineage that is

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stable in frequency over time (Fig. S6), whereas D839 has been sampled in many countries in Europe. Both are located in S2, and as noted above, near the potential fusion peptide.

Site P1263L. This mutation is not included in the SARS-CoV-2 structure, but is near the end of the cytoplasmic tail of the Spike protein. The mutation is found mostly in the UK, in both England and Wales, but also in Australia, and is emerging as a single related lineage (Fig. 1). It is maintaining its frequency both globally as well as locally in the UK.

Discussion

When we embarked on our SARS-CoV-2 analysis pipeline, our motivation was to identify mutations that might be of potential concern in the SARS-CoV-2 Spike protein as an early warning system for consideration as vaccine studies progress; we did not anticipate such dramatic results so early in the pandemic. In a setting of very low genetic diversity, traditional means of identification of positive selection have limited statistical power, but the incredibly rich GISAID data set provides an opportunity to look more deeply into the evolutionary relationships among the SARS-CoV-2 sequences in the context of time and geography. This approach revealed that viruses bearing the mutation Spike D614G are replacing the original Wuhan form of the virus rapidly and repeatedly across the globe (Fig. 2-3). We do not know what is driving this selective sweep, nor for that matter if it is indeed due the modified Spike and not one of the other two accompanying mutations that share the GISAID “G-clade” haplotype. The Spike D614G change, however, is consistent with several hypotheses regarding a fitness advantage that can be explored experimentally. D614 is embedded in an immunodominant antibody epitope, recognized by antibodies isolated from recovered individuals who were infected with the original SARS-CoV; this epitope is also targeted by vaccination in primate models (Wang et al., 2016). Thus, this mutation might be conferring resistance to *protective* D614-directed antibody responses in infected people, making them more susceptible to reinfection with the newer G614 form of the virus. Alternatively, the advantage might be related to the fact that D614 is embedded in an immunodominant ADE epitope of SARS-CoV (Wang et al., 2016), and perhaps the G614 form can facilitate ADE. Finally, the D614G mutation is predicted to destabilize inter-protomer S1-S2 subunit interactions in the trimer, and this may have direct consequences for the infectivity of the virus (Fig. 4). Increased infectivity would be consistent with rapid spread, and also the association of higher viral load with G614 that we observed in the clinical data from Sheffield, England (Fig. 5).

Each of the ways we anticipated we might find evidence of positive selection in Spike are being manifested among subset of the sites that are accruing mutations. While the D616G mutation is the only one that is dramatically increasing in frequency globally (Fig. 2-3), the L8V mutation may be on the rise in the local epidemic in Hong Kong (Fig. S7). To date, mutations are extremely rare in the Spike RBD, but the mutation G476S is directly in an ACE2 contact residue. The mutation L5F occurs in many geographic regions in many distinct clades, suggesting it repeatedly arose independently, and was selected to the extent that was frequent enough to be resampled. Finally, the mutation S943P seems to have been transferred by recombination into diverse viral backbones that are co-circulating in Belgium (Fig. 6); we also found strong evidence of recombination in other regional sample sets (Fig. S8). Recombination among pandemic SARS-CoV-2 strains is not surprising, given that it is also found among more distant coronaviruses with higher diversity levels (Graham and Baric, 2010; Li et al., 2020; Rehman et al., 2020). Still, it has important implications. First, recombination cannot be detected without simultaneous coinfection of distinct viruses in one host. It is not clear if such co-infections might

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be happening prior to the adaptive immune response, or in series with reinfection occurring after the initial infection stimulated a response. Recombination may be more common in communities with less rigorous shelter-in-place and social distancing practices, in hospital wards with less stringent patient isolation because all patients are assumed to already be infected or in geographic, or in regions where antigenic drift has already begun to enable serial infection with more resistant forms of the viruses. Also, recombination provides an opportunity for the virus to bring together, into a single recombinant virus, multiple mutations that independently confer distinct fitness advantages but that were carried separately in the two parental strains.

Tracking mutations in Spike has been our primary focus to date because of the urgency with which vaccine and antibody therapy strategies are being developed; the interventions under development now cannot afford to miss their contemporary targets when they are eventually deployed. To this end, we built a data-analysis pipeline to explore the potential impact of mutations on SARS-CoV-2 sequences. The analysis is performed as the data becomes available through GISAID. Experimentalists can make use of the most current data available to best inform vaccine constructs, reagent tests, and experimental design. While the GISAID data used for the figures in this paper was frozen at April 13, 2020, many of the key figures included here are rebuilt each day based on the newly available GISAID data. While our initial focus is on Spike, the tools we have developed can be extended to other proteins and mutations in subsequent versions of the pipeline. Meanwhile understanding both how the D614G mutation is overtaking the pandemic and how recombination is impacting the evolution of the virus will be important for informing choices about how best to respond in order to control epidemic spread and resurgence.

Table

Spike Mutation	Spike Location Possible Impact	04/13/20 Count	Geographic Sampling	Phylogenetic Pattern .
D614G	SARS-CoV epitope Interprotomer stabilization	3577	Global	One main lineage & recurrent emergence
L5F	Signal Peptide	37	13 countries	Recurrent emergence
L8V/W	Signal Peptide	18	Hong Kong	
H49Y	S1 NTD domain	12	China	
Y145H/del	S1 NTD domain	10	6 countries	
Q239K	S1 NTD domain	8	Europe	
V367F	Up/Down conformations	12	Europe/Hong Kong	
G476S	Directly in the RBD	8	Washington, USA	
V483A	Up/Down conformations	21	Washington, USA	
V615I/F	In SARS-CoV ADE epitope	13	Wales	
A831V	Potential fusion peptide in S2	28	Iceland	One lineage
D839Y/N/E	S2 subunit	27	Europe	
S943P	Fusion core of HR1	23	Belgium	Local recombination
P1263L	Cytoplasmic Tail	49	UK/Iceland/Australia	One lineage
Mutational cluster 937-943:	Fusion core of HR1	65	Different forms in 9 countries	

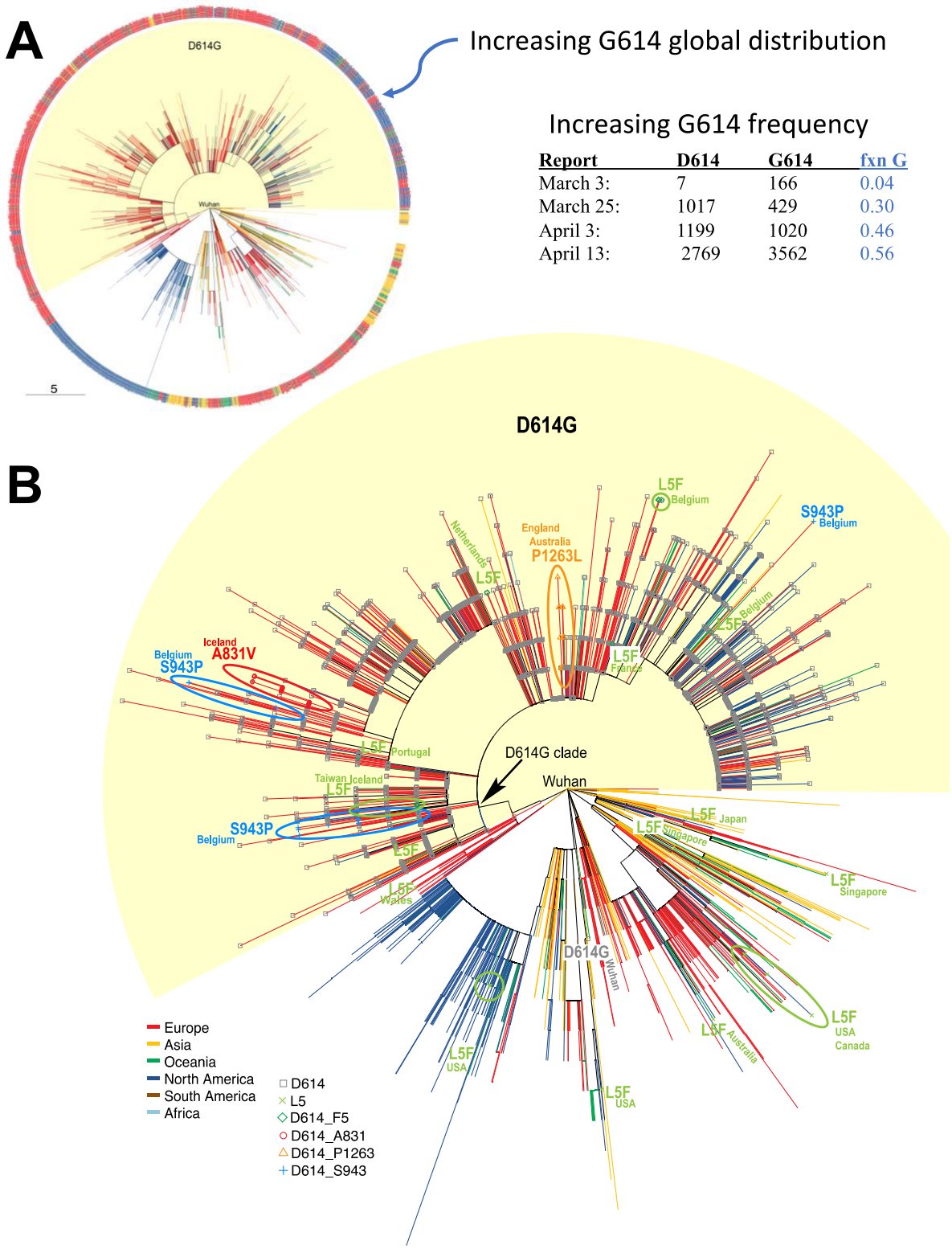
Table 1. Table summarizing the mutations we are following in Spike. With the exception of D614G, all other mutations in Spike remain rare; we will nonetheless monitor them for potential immunological impact and/or for increased frequencies regionally or globally as the pandemic progresses. The NTD is the N-terminal domain, S2 is a membrane fusion subunit, and HR1 the first heptad repeat region

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Lv et al., 2020). Up/Down conformations refer to a change in state in which the up conformation exposes the RBD (Kirchdoerfer et al., 2016; Kirchdoerfer et al., 2018). The SARS-CoV epitope was identified from the first SARS epidemic, and is the immunodominant linear antibody epitope observed in natural infection and in animal models. For the geographic sampling, we only list one country or region if it dominates a sample. For details see the complete listing of sequences with given mutation in Item S1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures



Report	D614	G614	fxn G
March 3:	7	166	0.04
March 25:	1017	429	0.30
April 3:	1199	1020	0.46
April 13:	2769	3562	0.56

Fig. 1. Phylogenetic Trees based on 4,535 trimmed full genome SARS C0V02 alignments from GISAID. A) A basic neighbor joining tree, centered on the Wuhan reference strain, with the GISAID

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

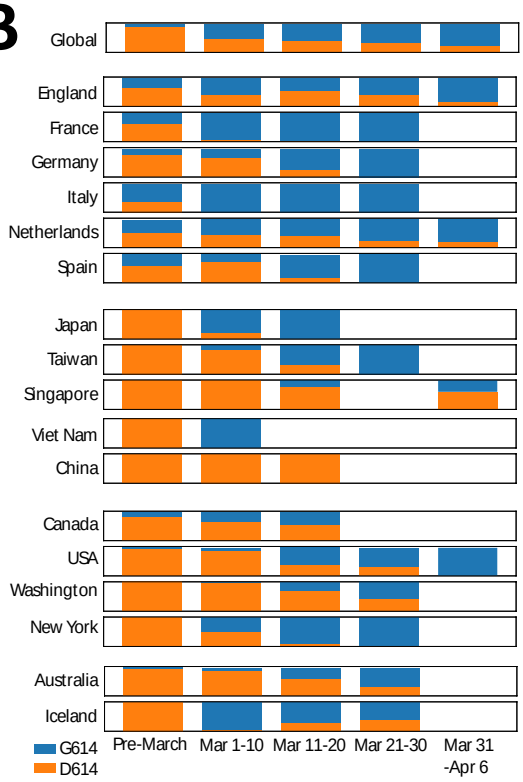
G clade (named for the D614G mutation, though a total of 3 base changes define the clade) are highlighted in yellow. The regions of the world where sequences were sampled are indicated by colors. By early April, G614 was more common than the original D614 form isolated from Wuhan, and rather than being restricted to Europe (red) it had begun to spread globally. **B)** The same tree expanded to show interesting patterns of Spike mutations that we are tracking against the backdrop of the phylogenetic tree based on the full genome. Note three distinct patterns: mutations that predominantly appear to be part of a single lineage (P1263L, orange in the UK and Australia, and also A831V, red, in Iceland); a mutation that is found in very different regions both geographically and in the phylogeny, indicating the same mutation seems to be independently arising and sampled (L5F green, rare but found in scattered locations worldwide); and a mutation in sequences from the same geographic location, but arising in very distinct lineages in the phylogeny (S943P), blue, found only in Belgium. A chart showing how GISAID sequence submissions increase daily is provided in Fig. S1. The tree shown here can be recreated with contemporary data downloaded from GISAID at www.cov.lanl.gov. The tree shown here was created using PAUP (Swofford, 2003); the trees generated for the website pipeline updates are based on parsimony (Goloboff, 2014).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

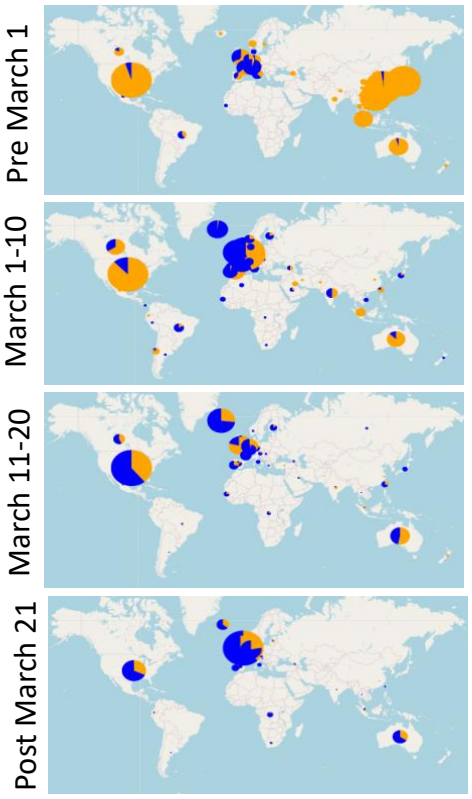
A

Region	N	pre-March			March 1-10			March 11-20			March 21-30			March 31 - April 6		
		D614	G614	fxnG	D614	G614	fxnG	D614	G614	fxnG	D614	G614	fxnG	D614	G614	fxnG
Global	6244	640	71	0.099	473	555	0.540	935	1463	0.609	639	1292	0.669	37	139	0.790
Belgium	299	1	1	0.500	9	96	0.914	5	38	0.884	13	94	0.879	6	37	0.860
England	1099	15	8	0.348	43	63	0.594	114	102	0.472	271	421	0.608	8	51	0.864
France	198	12	7	0.368	3	57	0.950	1	63	0.984	0	55	1.000	na	na	na
Germany	50	7	2	0.222	2	1	0.333	8	29	0.784	0	1	1.000	na	na	na
Italy	37	3	5	0.625	0	10	1.000	0	10	1.000	0	9	1.000	na	na	na
Netherlands	527	3	3	0.5	81	98	0.547	56	81	0.591	37	124	0.771	9	37	0.804
Spain	87	3	2	0.4	41	14	0.255	3	16	0.842	0	8	1.000	na	na	na
Switzerland	30	1	19	0.95	0	10	1.000	na	na	na	na	na	na	na	na	na
Wales		0	1	1	0	1	1.000	154	41	0.210	235	245	0.510	na	na	na
Japan	95	76	0	0	1	4	0.800	0	14	1.000	na	na	na	na	na	na
Taiwan	40	10	0	0	4	1	0.200	7	16	0.696	0	2	1	na	na	na
Singapore	44	23	0	0	12	0	0.000	3	1	0.250	na	na	na	3	2	0.400
Hong Kong	60	60	0	0	na	na	na	na	na	na	na	na	na	na	na	na
South Korea	12	12	0	0	na	na	na	na	na	na	na	na	na	na	na	na
Viet Nam	8	5	0	0	0	3	1.000	na	na	na	na	na	na	na	na	na
China	265	257	4	0.015	1	0	0.000	3	0	0.000	na	na	na	na	na	na
Canada	114	5	1	0.167	27	14	0.341	48	37	0.435	na	na	na	na	na	na
USA	1330	98	4	0.039	166	23	0.122	309	485	0.611	74	166	0.692	0	7	1.000
USA-WA	627	46	0	0	146	10	0.064	214	111	0.342	41	59	0.59	na	na	na
USA-NY	389	1	0	0	1	1	0.5	24	296	0.925	2	64	0.97	na	na	na
USA-MN	55	na	na	na	4	1	0.2	15	35	0.7	na	na	na	na	na	na
USA-CA	25	16	0	0	5	0	0	1	3	0.75	na	na	na	na	na	na
Austrailia	354	23	1	0.042	19	3	0.136	103	75	0.421	37	82	0.689	na	na	na
Iceland	499	1	0	0	1	51	0.981	99	270	0.732	25	45	0.643	na	na	na
Senegal	21	0	1	1.000	0	4	1.000	4	12	0.75	na	na	na	na	na	na
DRC	31	na	na	na	0	1	1.000	3	7	0.7	1	16	0.941	na	na	na
Brazil	21	2	3	0.6	2	10	0.833	2	2	0.5	na	na	na	na	na	na
Ecuador	4	na	na	na	1	0	0	na	na	na	0	3	1.000	na	na	na
Kuwait	7	na	na	na	4	0	0	1	2	0.667	na	na	na	na	na	na

B



C



http://cov-int.lanl.gov/apps/covid-19/map/

Fig. 2. The proportion of sequences carrying the D614G mutation is increasing in every region that was well sampled in the GISAID database through the month of March. A) A table

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

showing the tallies of each form, D614 and G614 in different countries and regions, starting with samples collected prior to March 1, then following in 10 day intervals. **B)** Bar charts illustrating the relative frequencies of the original Wuhan form (D614, orange), and the form that first emerged in Europe (G614, blue) based on the numbers in part (A). A variation of this figure showing actual tallies rather than frequencies, so the height of the bars represent the sample size, is provided as Fig. S2. **C)** A global mapping of the two forms illustrated by pie charts over the same periods. The size of the circle represents sampling. An interactive version of this map of the April 13th data, allowing one to change scale and drill down in to specific regions of the world is available at https://cov.lanl.gov/apps/covid-19/map, and daily updates of this map based on contemporary data from GISAID are provided at www.cov.lanl.gov.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

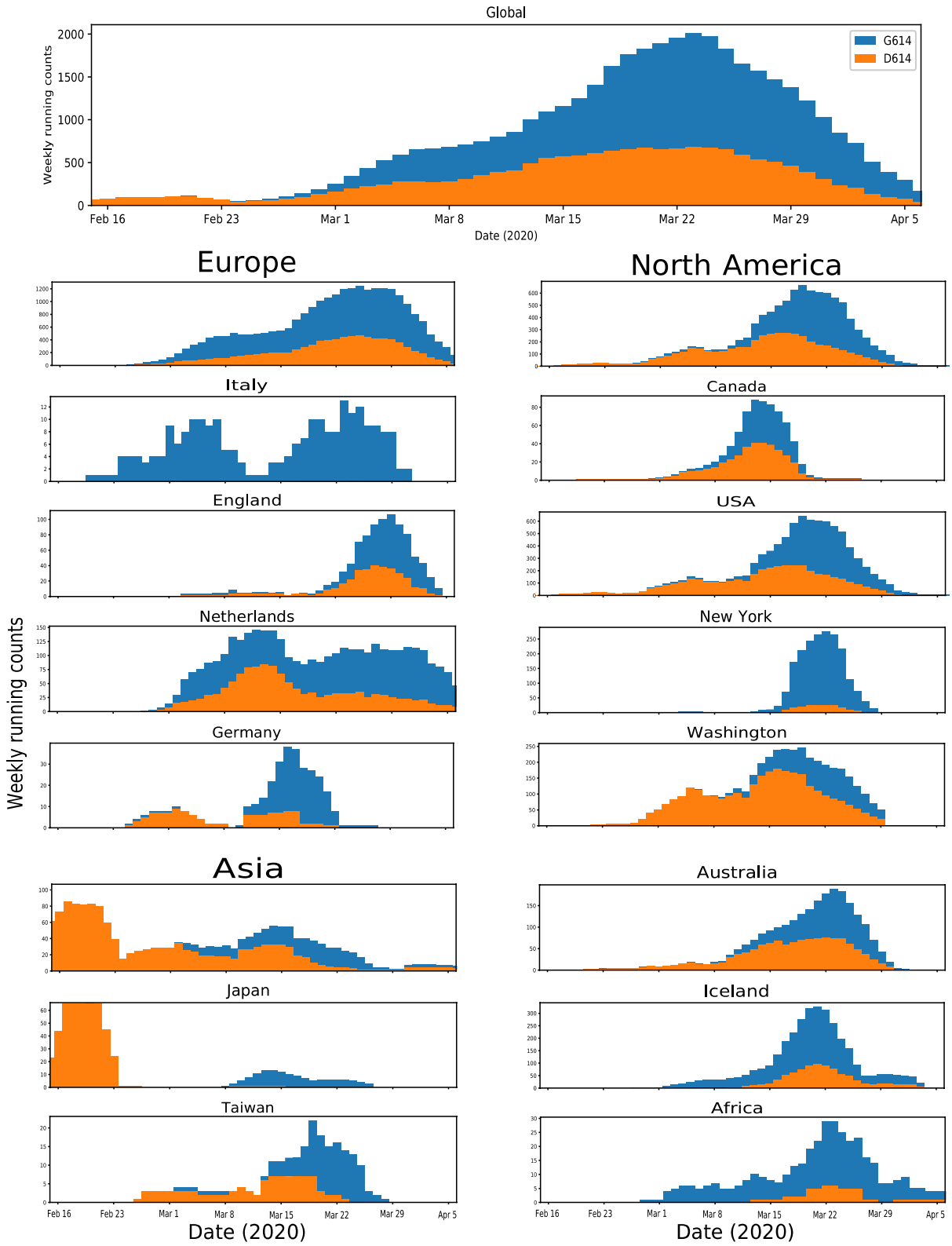


Fig. 3. Running weekly average counts showing the relative amount of D614 (orange) and G614, (blue) in different regions of the world. In almost every case soon after G614 enters a region, it begins to dominate the sample. Fig. S3 shows the same data, illustrated as a daily

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cumulative plot. Plots were generated with Python Matplotlib (Hunter, 2007). The plots shown here and in Fig. S3 can be recreated with contemporary data from GISAID at www.cov.lanl.gov.

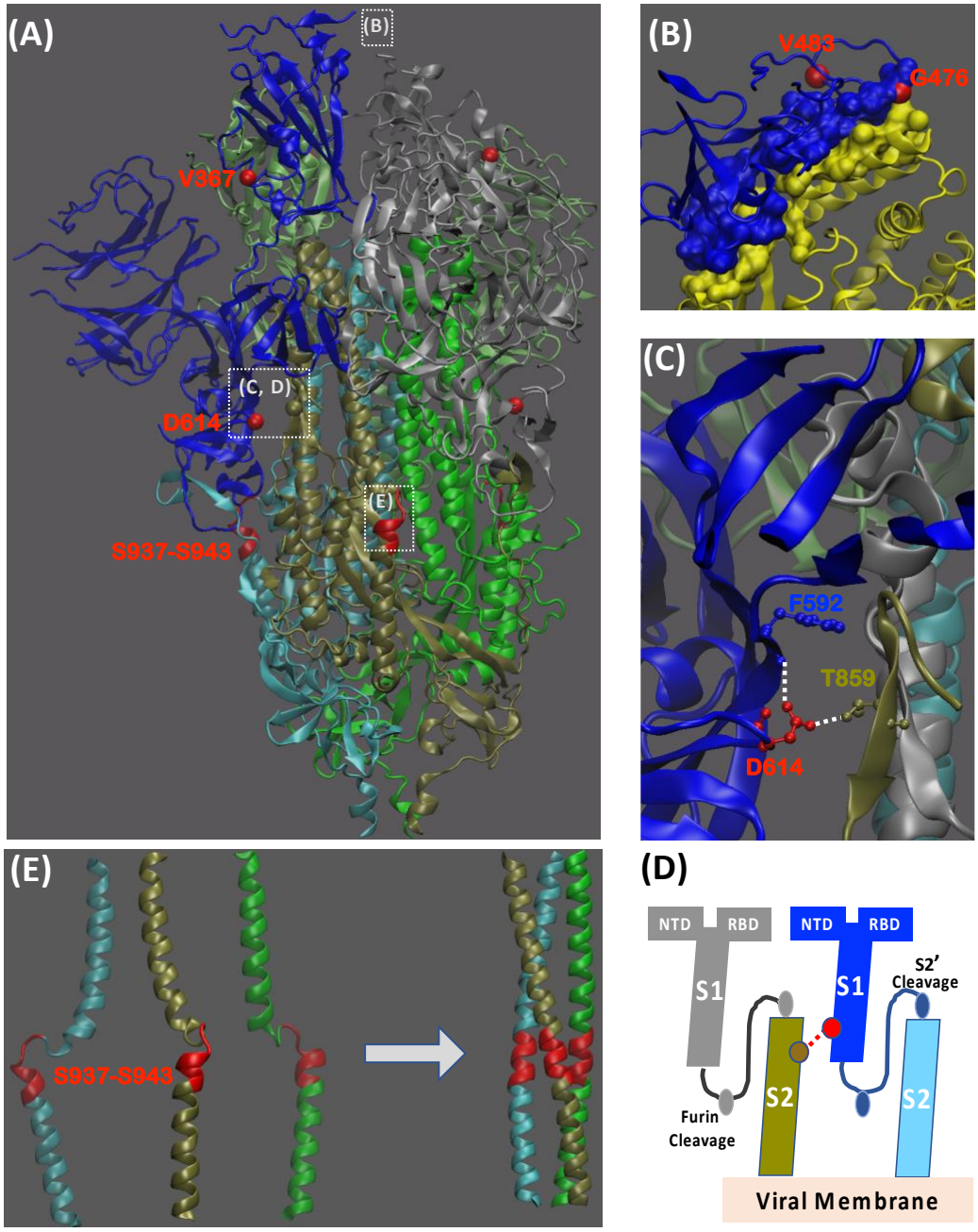


Fig. 4. Structural mapping key mutational sites in the Spike protein. A) Mutational sites span S1 and S2 structural units of the Spike protein (PDB:6VSB). Different colors are used to distinguish the protomers and sub-units. S1 and S2 sub-units are defined based on the furin cleavage site (protomer #1: S1-blue, S2-cyan, protomer #2: S1-grey, S2-tan, protomer #3: S1-light green, S2-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dark green). The RBD of protomer #1 is in "UP" position for engagement with ACE2 receptor. Red color is used to indicate individual mutational sites (ball). The dashed squares with labels indicate forthcoming detailed investigations in subsequent images. **B)** Mutational sites near the RBD (blue)-ACE2 (yellow) binding interface. The interfacial region is shown as a surface (PDB: 6M17). **C)** The proximity of D614 to T859 from the neighboring protomer. The white dashed lines indicate the possibility for forming hydrogen bonds. **D)** A cartoon is used to capture how the potential protomer-protomer interactions shown in (C) brings together D614 from S1 unit of one protomer to the T859 from S2 unit of the neighboring protomer. **E)** Cluster of mutations, *S937-S943*, in the HR1 region of the Spike protein. These residues occur in a region that undergoes conformational transition during fusion. The left and right images show the pre-fusion (PDB:6VSB) and post-fusion (PDB: 6LXT) conformations of this HR1 region. Structural implications of different sites are noted in Fig. S4. Structural evaluations and rendering of three dimensional images were carried out using Visual Molecular Dynamics (VMD) (Humphrey et al., 1996).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

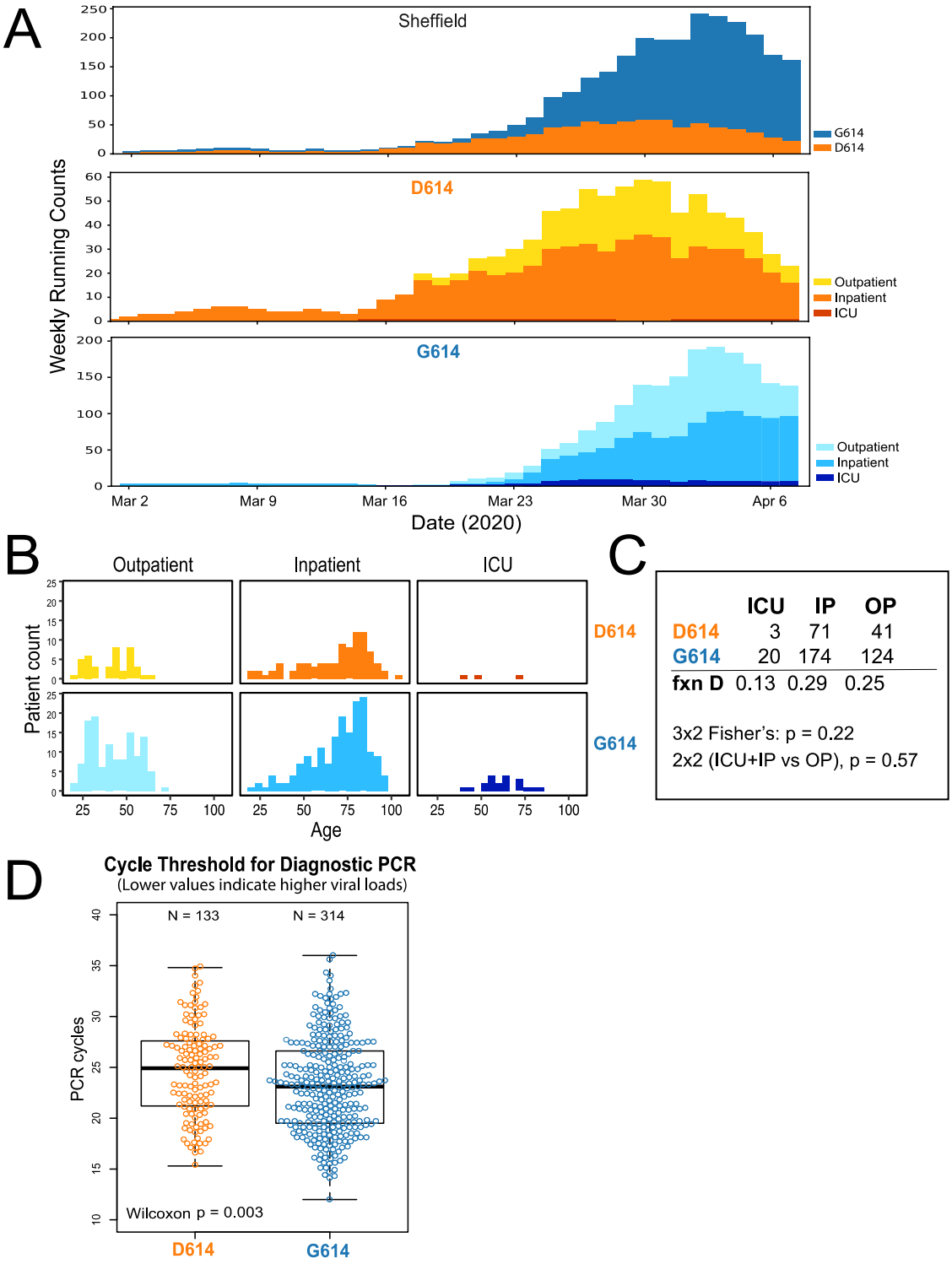


Fig. 5. A comparison of the two forms of S D614G to clinical outcomes in 447 subjects with COVID-19, over the time span of the epidemic in Sheffield, England. A) Tracking the epidemic forms in Sheffield. The top panel shows that the course of the epidemic in Sheffield is the same as

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the course we documented throughout the globe (see Fig. 3), with G614 overtaking D614 as the dominant form of SARS-CoV-2. We were concerned that the rate of hospitalization might have varied over this time period, which could have biased the sample as G614 tended to be sampled later. The D614 and G614 panels show that the rate of hospitalized individuals from whom sequences were obtained, averaged per week, remained relatively constant across this time period for both groups. **B)** Age distribution between clinical status groups. We were also concerned that the age distribution of people visiting the hospital might have differed between the groups, as age is highly associated with high risk, but part **(B)** shows that this distribution is very similar, and there was no statistical difference between the groups overall (Wilcoxon $p = 0.77$, D614 had a median age of 60 (IR 44-80), while G614 had a median age of 59 (IR 43-77). **C)** D614G status was not statistically associated with hospitalization status. **D)** G614 was associated with fewer rounds of PCR required for detection, suggesting that people with a G614 virus had higher viral loads. Other associations with hospital status are shown in Fig. S5.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

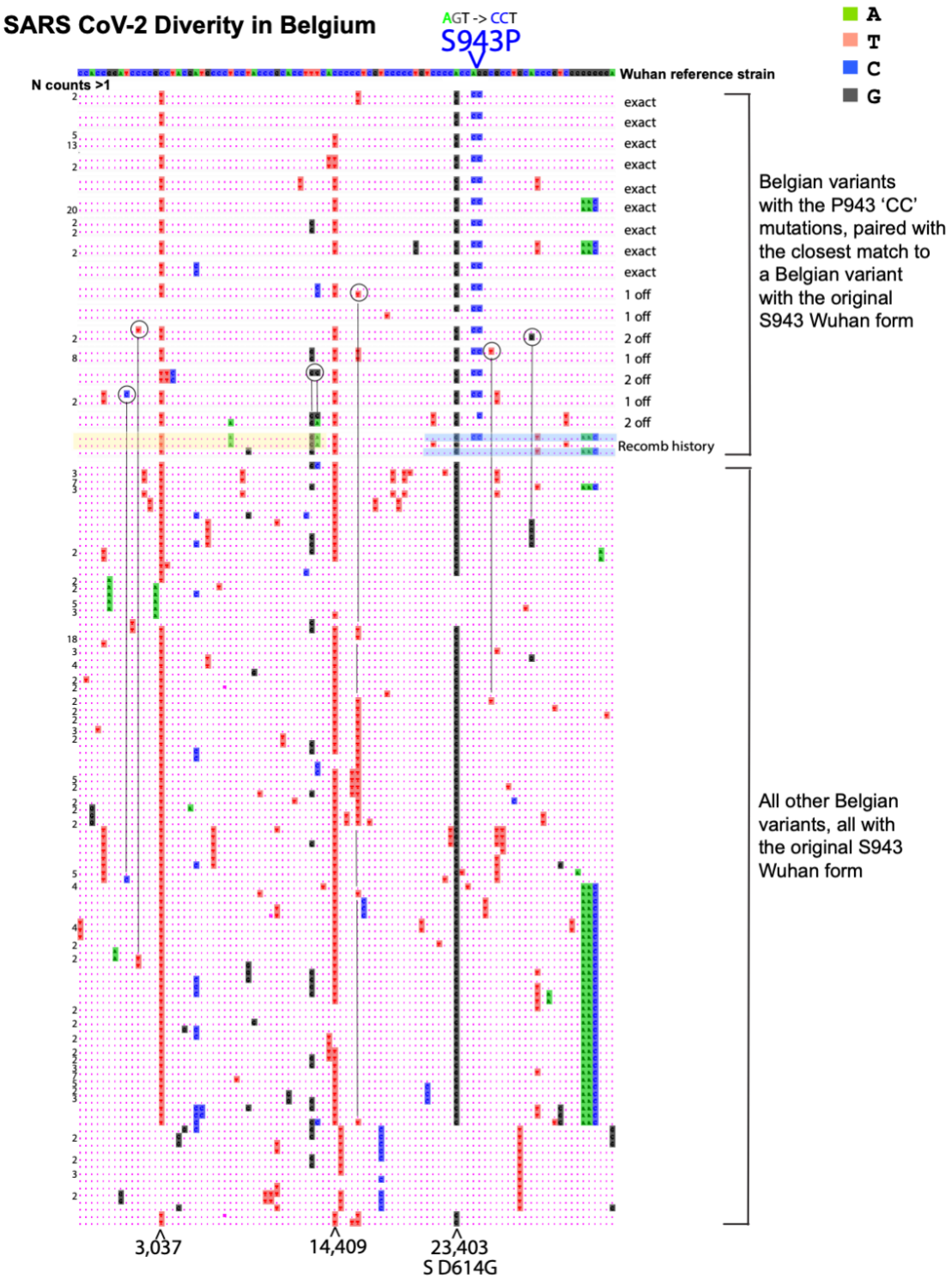


Fig. 6. The evolutionary trajectory of the S943P mutation in Belgium supports recombination and selection. Phylogenetically informative sites (including only positions that varied relative to the consensus in Belgium in 2 or more sequences) were extracted from the full genome sequence

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alignment to enable visualization of the relevant mutations; and duplicates among this set were reduced to a single representative sequence with the count noted on the left if it was found more than one time. The baseline display was generated using AliView (Larsson, 2014), and the alignment is shown relative to the Wuhan reference strain, with pink dots indicating base identities. To emphasize the introduction of the two base changes to C in the S943P mutation codon (AGT -> CCT) into distinct backbone sequences circulating in Belgium, each unique sequence carrying the S943P amino is aligned adjacent to the nearest sequence in the Belgium set that did *not* have the S943P codon with its double mutation. Nine times the match was identical outside of S943, and the other sequences that carried the P943 codon were very similar to other sequences in the Belgium set, with only one or two base differences. Of note, each of these bases were also carried in other sequences found in the Belgian set; these patterns are indicated by a circled distinctive base and a vertical line to its match among other Belgian sequences. Given the rareness of these mutations overall (1/10,000), this kind of pattern suggests other SARS-CoV-2 recombination events were commonly occurring in the Belgian population. Finally there was also evidence of more complex recombination history among the backbone sequences, and an example of this is highlighted in the sequence triplet labeled “Recomb history”. In this example, the S943 is not identical to a particular strain in Belgium; rather its 5’ half is similar to the strain marked in yellow, and its 3’ half to the strain marked in blue.

ACKNOWLEDGEMENTS

We thank Sir Andrew McMichael, Professor Sarah Rowland-Jones, and Dr. Xiao-Ning Xu for their invaluable help with bringing together the clinical and theoretical biology teams necessary for this study. Sequencing of SARS-CoV-2 samples was undertaken by the Sheffield COVID-19 Genomics Group as part of the COG-UK CONSORTIUM. COG-UK and supported by funding from 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. TIdS is supported by a Wellcome Trust Intermediate Clinical Fellowship (110058/Z/15/Z). Analyses strategies presented in this article were developed with the support of the Laboratory Directed Research and Development program of Los Alamos National Laboratory. Recombination analysis was conducted under project number 20200554ECR. The sequence data pipeline design and analysis of the structural immunological implications of Spike mutations was conducted under project number (20200706ER). The sequence data pipeline implementation was funded through the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Interagency Agreement No. AAI12007-001-00000. We gratefully acknowledge the team at GISAID for creating the remarkable COVID-19 outbreak global database and resources, and the many authors from the oringiating and submitting laboratories of sequence data on which this analysis is based; see Item S1 is the acknowledgment table from GISAID at the time of our April 13th data download, listing the many people responsible for generating the sequence data. Bless them.

STAR METHODS

RESOURCE 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Sequence data are available from The Global Initiative for Sharing All Influenza Data (GISAID), at https://gisaid.org. The user agreement for GISAID does not permit redistribution of sequences, but lists of the sequences used in our analyses, high-resolution figures, and code will be made available at www.cov.lanl.gov.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

SARS-CoV-2 sequences were generated using samples taken for routine clinical diagnostic use from 454 individuals presenting with active COVID-19 disease: 243 female, 209 male, 2 no gender specified; ages 18-103 (median 59.6) years.

METHOD DETAILS

Detection and Sequencing of SARS-CoV-2 isolates from clinical samples

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 basecalling, 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.

Data Pipeline

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 writing, dozens to hundreds of sequences were being added every day. These sequences result from extraordinary efforts by a wide variety of institutions and individuals: they are an invaluable resource, but are somewhat 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 off-the-shelf bioinformatic tools and a small amount of custom code.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

General approach

From theSARS-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 frame-shift) were deleted, and gaps were shifted to conform with codon boundaries.

Using the initial "good-sequence" alignment, a low-effort parsimony tree was constructed using PAUP: a single replicate heuristic search using stepwise random sequence addition. Sequences in the alignment were sorted vertically to correspond to the (ladderized) tree, and reference-sequence reading frames were added.

Data partitioning and phylogenetic trees

Estimated phylogenies were inferred for three distinct data partitions: the full sequence set (near-complete genomes), the spike open reading frame, and the full set with the spike open reading frame (orf) excluded. The full genome tree was used for Fig. 1. The tree with the spike orf excluded was intended to allow independent assessment of the phylogenetic distribution of changes within the spike protein, by preventing convergent or homoplastic mutations driven by phenotypic selection upon the spike protein from overwhelming phylogenetic signal from the rest of the genome. We confirmed that the phylogenetic observation discussed in Fig. 1 were supported in the Spike-excluded tree, but do not include it here. It is available for cross-checking phylogenetic based inferences at www.cov.lanl.gov.

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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Clinical data and disease severity

To assess possible associations of clinical and sequence variables with disease severity, we used a Generalized Linear Model (GLM) using outpatient vs. hospitalized status as outcome and age, gender, D614 and PCR cycle threshold for E gene amplification (E_gene_CT) as potential predictors. Outpatient vs. hospitalized status, gender, and D614 were all categorized as binomial factors, while age and E_gene_CT were considered as continuous variables. We started with the largest model that included all variables and then used ANOVA to down-select the best predicting model. All coding was done using R and the lme4 package (Bates 2015).

Mutation rate analysis of Belgian sequences

The phylogenetic tree was broken up into subtrees using the R ape package (Paradis and Schliep 2018), and those subtrees containing only Belgian or only non-Belgian tips were selected, and their total branch lengths calculated. For each subtree the R package phangorn (Schliep 2011) was used to calculate the minimum number of changes required at each site. The maximum rate, in mutations per unit branch length, compatible with the non-Belgian data was calculated as its 95% upper confidence level assuming a Poisson distribution of mutations with the Poisson parameter proportional to the branch length. The p-value of the Belgian data was estimated from a Poisson distribution with parameter given by the rate for non-Belgian data multiplied by the total branch length of the Belgian trees. Only two sites were found to be significant after Bonferroni corrected by the number of sites in the alignment.

Identification of recombination candidates using RAPR.

To identify the candidate recombination parent and child sequences shown in Fig. S8, we first did all triplet comparisons of all sequences from local region using the RAPR, and the raw p-values based on a run-length statistic all the comparisons were rank-ordered to identify the triplet candidates with strongest evidence for recombination to be used as a basis for further exploration. Thus these p-values are uncorrected for multiple testing and not formally compared against the alternative hypotheses of stepwise convergent mutation, as we traditionally do in RAPR analysis (Song et al., 2018). However, given the very low overall mutation rate among SARS-CoV-2 pandemic, it is extremely unlikely that the mutational patterns seen in the recombinant sequences were a result of stepwise convergence.

ADDITIONAL RESOURCES

Archived data for the current manuscript, and current data updates, analytical results, and webtools: https://COV.lanl.gov.

KEY RESOURCES TABLE

The R Foundation for Statistical Computing, http://www.R-project.org

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R packages (https://cran.r-project.org/ except as noted)

- The R Foundation for Statistical Computing, http://www.R-project.org

R packages (https://cran.r-project.org/ except as noted)

- 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)
- data.table (version 1.12.8): https://github.com/Rdatatable/data.table

Web tools, software, and protocols

ADD COVID tools

- GISAID: https://www.gisaid.org
- Los Alamos SARS-CoV-2 mutation analysis pipeline: https://COV.lanl.gov
- RAPR: Recombination Analysis PRogram. https://www.hiv.lanl.gov/content/sequence/RAP2017/rap.html
- VMD: Visual Molecular Dynamics: https://www.ks.uiuc.edu/Research/vmd/
- Highlighter: A tool to highlight matches, mismatches, and specific mutations in aligned protein or nucleotide sequences. https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html
- Rainbow Tree: https://www.hiv.lanl.gov/content/sequence/RAINBOWTREE/rainbowtree.html
- Aliview, a sequence alignment viewer and editor: https://ormbunkar.se/aliview/
- Add COVID tools
- Sequence resconstruction: ARTIC network protocol (accessed the 19th of April, *https://artic.network/ncov-2019*.)
- Base call resolution: Nanopolish (https://github.com/jts/nanopolish)
- Matplotlib: A 2D Graphics Environment (Hunter, 2007)
- PAUP https://paup.phylosolutions.com

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Asmal, M., Hellmann, I., Liu, W., Keele, B.F., Perelson, A.S., Bhattacharya, T., Gnanakaran, S., Daniels, M., Haynes, B.F., Korber, B.T., *et al.* (2011). A signature in HIV-1 envelope leader peptide associated with transition from acute to chronic infection impacts envelope processing and infectivity. PLoS One *6*, e23673-e23673.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software; Vol 1, Issue 1 (2015).

Boni, M.F., Gog, J.R., Andreasen, V., and Feldman, M.W. (2006). Epidemic dynamics and antigenic evolution in a single season of influenza A. Proc Biol Sci *273*, 1307-1316.

Brufsky, A. (2020). Distinct Viral Clades of SARS-CoV-2: Implications for Modeling of Viral Spread. Journal of Medical Virology *n/a*.

Chattopadhyay, I., Kiciman, E., Elliott, J.W., Shaman, J.L., and Rzhetsky, A. (2018). Conjunction of factors triggering waves of seasonal influenza. eLife *7*, e30756.

Chen, W.H., Hotez, P.J., and Bottazzi, M.E. (2020). Potential for developing a SARS-CoV receptor-binding domain (RBD) recombinant protein as a heterologous human vaccine against coronavirus infectious disease (COVID)-19. Hum Vaccin Immunother, 1-4.

Conti, P., and Younes, A. (2020). Coronavirus COV-19/SARS-CoV-2 affects women less than men: clinical response to viral infection. J Biol Regul Homeost Agents *34*.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., *et al.* (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance *25*, 2000045.

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology *17*, 181-192.

Dawson, J.P., Weinger, J.S., and Engelman, D.M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. J Mol Biol *316*, 799-805.

de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V.J. (2016). SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology *14*, 523-534.

Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M.C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences, 202004911.

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges *1*, 33-46.

Fehr, A.R., Channappanavar, R., and Perlman, S. (2017). Middle East Respiratory Syndrome: Emergence of a Pathogenic Human Coronavirus. Annual Review of Medicine *68*, 387-399.

Goloboff, P.A. (2014). Oblong, a program to analyse phylogenomic data sets with millions of characters, requiring negligible amounts of RAM. Cladistics *30*, 273-281.

Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., *et al.* (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology *5*, 536-544.

Graham, R.L., and Baric, R.S. (2010). Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. Journal of Virology *84*, 3134-3146.

Ho, M.-S., Chen, W.-J., Chen, H.-Y., Lin, S.-F., Wang, M.-C., Di, J., Lu, Y.-T., Liu, C.-L., Chang, S.-C., Chao, C.-L., *et al.* (2005). Neutralizing antibody response and SARS severity. Emerg Infect Dis *11*, 1730-1737.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., *et al.* (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell *181*, 271-280.e278.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J Mol Graph *14*, 33-38, 27-38.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering *9*, 90-95.

Jaume, M., Yip, M.S., Cheung, C.Y., Leung, H.L., Li, P.H., Kien, F., Dutry, I., Callendret, B., Escriou, N., Altmeyer, R., *et al.* (2011). Anti-severe acute respiratory syndrome coronavirus spike antibodies trigger infection of human immune cells via a pH- and cysteine protease-independent FcγR pathway. Journal of virology *85*, 10582-10597.

Kam, Y.W., Kien, F., Roberts, A., Cheung, Y.C., Lamirande, E.W., Vogel, L., Chu, S.L., Tse, J., Guarner, J., Zaki, S.R., *et al.* (2007). Antibodies against trimeric S glycoprotein protect hamsters against SARS-CoV challenge despite their capacity to mediate FcgammaRII-dependent entry into B cells in vitro. Vaccine *25*, 729-740.

Kandori, H., Yamazaki, Y., Shichida, Y., Raap, J., Lugtenburg, J., Belenky, M., and Herzfeld, J. (2001). Tight Asp-85--Thr-89 association during the pump switch of bacteriorhodopsin. Proc Natl Acad Sci U S A *98*, 1571-1576.

Kirchdoerfer, R.N., Cottrell, C.A., Wang, N., Pallesen, J., Yassine, H.M., Turner, H.L., Corbett, K.S., Graham, B.S., McLellan, J.S., and Ward, A.B. (2016). Pre-fusion structure of a human coronavirus spike protein. Nature *531*, 118-121.

Kirchdoerfer, R.N., Wang, N., Pallesen, J., Wrapp, D., Turner, H.L., Cottrell, C.A., Corbett, K.S., Graham, B.S., McLellan, J.S., and Ward, A.B. (2018). Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep *8*, 15701-15701.

Kumar, G.V., Jeyanthi, V., and Ramakrishnan, S. (2020). A short review on antibody therapy for COVID-19. New Microbes New Infect, 100682.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5*, R12-R12.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics *30*, 3276-3278.

Li, F., Li, W., Farzan, M., and Harrison, S.C. (2005). Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. Science *309*, 1864.

Li, X., Giorgi, E.E., Marichann, M.H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Korber, B., and Gao, F. (2020). Extensive recombination and strong purifying selection - origin of SARS-CoV-2. Submitted.

Liu, W., Fontanet, A., Zhang, P.-H., Zhan, L., Xin, Z.-T., Baril, L., Tang, F., Lv, H., and Cao, W.-C. (2006). Two-year prospective study of the humoral immune response of patients with severe acute respiratory syndrome. J Infect Dis *193*, 792-795.

Liu, Y., Yang, Y., Zhang, C., Huang, F., Wang, F., Yuan, J., Wang, Z., Li, J., Li, J., Feng, C., *et al.* (2020). Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury. Sci China Life Sci *63*, 364-374.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., *et al.* (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet *395*, 565-574.

Lv, M., Luo, X., Estill, J., Liu, Y., Ren, M., Wang, J., Wang, Q., Zhao, S., Wang, X., Yang, S., *et al.* (2020). Coronavirus disease (COVID-19): a scoping review. Eurosurveillance *25*, 2000125.

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., *et al.* (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine *18*, 179.

Paradis, E., and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526-528.

Promislow, D.E.L. (2020). A geroscience perspective on COVID-19 mortality. The Journals of Gerontology: Series A.

Rehman, S., Sharique, L., Ihsan, A., and Liu, Q. (2020). Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. Pathogens *9*, 240.

Salamango, D.J., and Johnson, M.C. (2015). Characterizing the Murine Leukemia Virus Envelope Glycoprotein Membrane-Spanning Domain for Its Roles in Interface Alignment and Fusogenicity. J Virol *89*, 12492-12500.

Schliep, K.P. (2010). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592-593.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance *22*, 30494.

Song, H., Giorgi, E.E., Ganusov, V.V., Cai, F., Athreya, G., Yoon, H., Carja, O., Hora, B., Hraber, P., Romero-Severson, E., *et al.* (2018). Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. Nat Commun *9*, 1928.

Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. In PAUP* Phylogenetic Analysis Using Parsimony (Sunderland, Massachusetts.: Sinauer Associates).

Temperton, N.J., Chan, P.K., Simmons, G., Zambon, M.C., Tedder, R.S., Takeuchi, Y., and Weiss, R.A. (2005). Longitudinally profiling neutralizing antibody response to SARS coronavirus with pseudotypes. Emerg Infect Dis *11*, 411-416.

Upadhyay, C., Feyznezhad, R., Yang, W., Zhang, H., Zolla-Pazner, S., and Hioe, C.E. (2018). Alterations of HIV-1 envelope phenotype and antibody-mediated neutralization by signal peptide mutations. PLoS Pathog *14*, e1006812-e1006812.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281-292.e286.

Wan, Y., Shang, J., Sun, S., Tai, W., Chen, J., Geng, Q., He, L., Chen, Y., Wu, J., Shi, Z., *et al.* (2020). Molecular Mechanism for Antibody-Dependent Enhancement of Coronavirus Entry. Journal of Virology *94*, e02015-02019.

Wang, C., Li, W., Drabek, D., Okba, N.M.A., van Haperen, R., Osterhaus, A.D.M.E., van Kuppeveld, F.J.M., Haagmans, B.L., Grosveld, F., and Bosch, B.-J. (2020). A human monoclonal antibody blocking SARS-CoV-2 infection. bioRxiv, 2020.2003.2011.987958.

Wang, Q., Zhang, L., Kuwahara, K., Li, L., Liu, Z., Li, T., Zhu, H., Liu, J., Xu, Y., Xie, J., *et al.* (2016). Immunodominant SARS Coronavirus Epitopes in Humans Elicited both Enhancing and Neutralizing Effects on Infection in Non-human Primates. ACS Infectious Diseases *2*, 361-376.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069054; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wang, S.-F., Tseng, S.-P., Yen, C.-H., Yang, J.-Y., Tsao, C.-H., Shen, C.-W., Chen, K.-H., Liu, F.-T., Liu, W.-T., Chen, Y.-M.A., *et al.* (2014). Antibody-dependent SARS coronavirus infection is mediated by antibodies against spike proteins. Biochem Biophys Res Commun *451*, 208-214.
Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.-L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260.
Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., *et al.* (2020a). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe *27*, 325-328.
Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., *et al.* (2020b). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265-269.
Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S., *et al.* (2020). Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res *30*, 343-355.
Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science (New York, NY) *367*, 1444-1448.
Yilla, M., Harcourt, B.H., Hickman, C.J., McGrew, M., Tamin, A., Goldsmith, C.S., Bellini, W.J., and Anderson, L.J. (2005). SARS-coronavirus replication in human peripheral monocytes/macrophages. Virus Res *107*, 93-101.
Yip, M.S., Leung, H.L., Li, P.H., Cheung, C.Y., Dutry, I., Li, D., Daeron, M., Bruzzone, R., Peiris, J.S., and Jaume, M. (2016). Antibody-dependent enhancement of SARS coronavirus infection and its role in the pathogenesis of SARS. Hong Kong Med J *22*, 25-31.
Yip, M.S., Leung, N.H.L., Cheung, C.Y., Li, P.H., Lee, H.H.Y., Daëron, M., Peiris, J.S.M., Bruzzone, R., and Jaume, M. (2014). Antibody-dependent infection of human macrophages by severe acute respiratory syndrome coronavirus. Virol J *11*, 82-82.
Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., and Wilson, I.A. (2020). A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. Science, eabb7269.
Zhang, L., Zhang, F., Yu, W., He, T., Yu, J., Yi, C.E., Ba, L., Li, W., Farzan, M., Chen, Z., *et al.* (2006). Antibody responses against SARS coronavirus are correlated with disease outcome of infected individuals. Journal of medical virology *78*, 1-8.
Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., *et al.* (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270-273.