The origins, genomic diversity and global spread of SARS-CoV-2

Juan Li[1,5#], Shengjie Lai[2#], George F. Gao[3,4], Weifeng Shi[1,5]*

[1]School of Public Health, Shandong First Medical University & Shandong Academy of Medical Sciences, Taian 271000, China.

[2]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK.

[3]National Institute for Viral Disease Control and Prevention, China CDC, Beijing 102206, China.

[4]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China; Center for Influenza Research and Early-warning (CASCIRE), CAS-TWAS Center of Excellence for Emerging Infectious Diseases (CEEID), Chinese Academy of Sciences, Beijing 100101, China.

[5]Key Laboratory of Etiology and Epidemiology of Emerging Infectious Diseases in the Universities of Shandong, Shandong First Medical University & Shandong Academy of Medical Sciences, Taian 271000, China.

#Contributed equally to this article.

*Corresponding author.

Preface

It has been 19 months since COVID-19 was first documented in Wuhan, China. Since this time, the world has witnessed a devastating global pandemic, with more than 209 million infections, over four million fatalities, and cases rising rapidly on a daily basis. Herein, we describe the currently available data on the origins of the causative virus, SARS-CoV-2, outline its early spread in Wuhan and transmission patterns in China and globally, and highlight how genomic surveillance has helped trace the spread and genetic variation of the virus, comprising a key element of pandemic control. We devote particular attention to characterizing and describing the international spread of the major 'variants of concern' that were first identified in SARS-CoV-2 in late 2020 and demonstrate that virus evolution has entered a new phase. More broadly, we highlight our currently limited understanding of coronavirus diversity in nature, the rapid spread of the virus and its variants in such an increasingly connected world, the reduced protection of vaccines, and the urgent need for coordinated global surveillance using genomic techniques. Overall, we provide important information for the prevention and control of both the ongoing COVID-19 pandemic and the novel diseases that will inevitably emerge in humans in future generations.

1. Introduction

On the last day of 2019, the Wuhan Municipal Health Commission reported an outbreak of pneumonia on its official website. Shortly after, scientists reported the discovery of a novel coronavirus from the respiratory system from some of these patients that was different from all known coronaviruses including severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV) and Middle East respiratory syndrome (MERS) coronavirus (MERS-CoV)[1-4]. Shortly afterwards, the World Health Organization (WHO) named the disease COVID-19 and the International Committee on Taxonomy of Viruses named the novel infectious agent SARS-CoV-2[5], the seventh coronavirus that can cause epidemics. Dramatically and unexpectedly, COVID-19 rapidly spread through the global population, generating several variants of concern and developing into a major and devastating pandemic. Herein, we summarize our

current understanding of the origins, global spread and genetic diversity of SARS-CoV-2.

2. The origins of SARS-CoV-2

2.1 SARS-CoV-2 related coronaviruses

Many early COVID-19 cases from Wuhan were associated with the Huanan seafood market[2] which, given the presence of wildlife at the market, made it an obvious candidate for the location of the initial zoonotic (i.e. cross-species transmission) event. However, none of the animals from the market (including rabbits, snakes, stray cats, badgers, and bamboo rats) tested positive for SARS-CoV-2[6], and viral genome sequences of environmental samples from the market may not occupy basal positions on the viral phylogeny (although the position of the rooting on the tree is uncertain)[7]. In addition, some early COVID-19 cases from Wuhan were not epidemiologically linked to the market[8], and some were linked to other markets[9,10]. Hence, although not fully resolved, current evidence suggests that the Huanan seafood market might be the location of an early 'superspreading' event.

From the earliest genomic comparisons it was clear that SARS-CoV-2 possessed a similar genomic organization to SARS-CoV. Both had similar three-dimensional structures in the spike protein, suggesting that these viruses might utilize the same cell surface receptor - human angiotensin-converting enzyme (hACE2)[2]: this was soon confirmed *in vitro*[4] and by structural biology[11]. However, SARS-CoV-2 differs from SARS-CoV in two fundamental ways[12]. First, there are six amino acid positions in the receptor binding domain (RBD) of the spike protein that mediate attachment of the SARS-CoV and SARS-CoV-2 spike to the hACE2 receptor[13]. However, amino acids at five of the six positions differed between SARS-CoV and SARS-CoV-2[2,12]. Intriguingly, such differences have endowed SARS-CoV-2 with higher binding efficiency to the hACE2 receptor[11], and may contribute to the higher transmissibility of SARS-CoV-2 than SARS-CoV. Second, there was a 12-nucleotide (nt) insertion at the cleavage site of the spike protein of SARS-CoV-2 that has not yet been identified in closely related betacoronaviruses, but which has a complex evolutionary history across the coronaviruses as a whole indicating that it is evolutionarily volatile[14]. This insertion encoded four amino acids, PRRA, that can be recognized by a protease

-furin- extensively expressed in different tissues and organs[15]. This insertion may decrease the overall stability of SARS-CoV-2 spike, thereby facilitating the adoption of the open conformation required for the spike-ACE2 binding[16], and SARS-CoV-2 without this furin-cleavage site had reduced replication in a human respiratory cell line and was attenuated in laboratory animals[17]. Notably, amino acid substitutions have been documented at all four positions in the PRRA motif, with a P-to-H substitution (HRRA) identified in >487,000 viral genomes as of June 2021.

SARS-CoV-2, like many other members of the genus *Betacoronavirus* (including SARS-CoV), seemingly has its evolutionary roots in those viruses that commonly infect bats[2]. Not surprisingly, shortly after the identification of SARS-CoV-2, a close relative of SARS-CoV-2 was described - RaTG13 that was identified from a bat (*Rhinolophus affinis*) sampled in Yunnan province in 2013[4]. Interestingly, it was collected from a mine cave where four workers were sent to clean bat faeces and subsequently developed severe pneumonia[18]. Although RaTG13 exhibits 96.2% sequence identity to SARS-CoV-2 at the scale of whole genome, it does not possess similar RBD or cleavage site sequences. Further analyses suggest that RaTG13, rather than SARS-CoV-2, was a recombinant, and they likely diverged over 30 years ago[19]. Therefore, the SARS-CoV-2 RBD was an ancestral trait shared with bat viruses[19].

Subsequently, a number of groups reported the identification of SARS-CoV-2 related coronaviruses in Malayan pangolins (*Manis javanica*) smuggled into Guangxi and Guangdong provinces, China[20,21]. These pangolin coronavirus genomes exhibited 85.5% to 92.4% sequence similarity to SARS-CoV-2[20]. Notably, however, these pangolin derived coronaviruses formed two sub-lineages, with the Guangdong sub-lineage clustering with RaTG13 and SARS-CoV-2 and sharing 97.4% amino acid similarity to SARS-CoV-2 in the RBD, with identical amino acids at the six critical residues of the RBD. Also of note was that the Guangdong pangolins appeared to suffer a similar disease manifestation to humans suffering from COVID-19[22]. Thus, although the role, if any, played by pangolins in the genesis of SARS-CoV-2 and the ecology of coronaviruses in general is unknown, it is clear that wildlife coronaviruses exist that possess SARS-CoV-2 like RBD and high binding efficiency to hACE2.

Furthermore, a novel bat coronavirus, RmYN02, was reported, having been collected during routine surveillance of *R. malayanus* bats in Yunnan province on June 25,

2019[23]. RmYN02 shared 97.2% sequence identity with SARS-CoV-2 in the 1ab open reading frame (ORF), the largest in coronaviruses at approximately 21,300 nt. In June 2021, we reported four SARS-CoV-2 related coronaviruses genomes from Yunnan province[24]. Of these, RpYN06, found in *R. pusillus*, exhibited 94.5% sequence identity to SARS-CoV-2. However, for the genome excluding the spike gene which has a recombination history, the similarity to SARS-CoV-2 was 97.2%, making it the closest genomic backbone to SARS-CoV-2 identified to date. The other three SARS-CoV-2 related coronaviruses were more distant from SARS-CoV-2. However, they carried a genetically distinct spike gene that could bind to the hACE2 receptor *in vitro*, though weakly.

SARS-CoV-2-like coronaviruses have also been identified in bat populations from other parts of Asia, including Japan[25], Cambodia[26], and Thailand[27]. Notably, although two betacoronaviruses (STT182 and STT200) from *R. shameli* bats sampled in 2010 from Cambodia share 92.6% nucleotide identity with SARS-CoV-2 across the genome as a whole, they share five of the six critical RBD sites observed in SARS-CoV-2 and the Guangdong pangolin coronavirus[26].

2.2 Emergence pathways of SARS-CoV-2

There are several hypotheses regarding the origin and emergence of SARS-CoV-2 that have been thoroughly clarified in the WHO-China joint report[6]. These contradictory hypotheses have raised standing debates, with the central point being two competing hypotheses: zoonotic emergence (including direct zoonotic introduction or introduction through an intermediate host) and a laboratory escape. The discovery of more and more SARS-CoV-2 related coronaviruses from wild animals provides evidence of a zoonotic origin of SARS-CoV-2[4,20,21,23-27]. Importantly, all the SARS-CoV-2 related coronaviruses mentioned above are evidently not the direct ancestor of SARS-CoV-2. Any such direct ancestral virus, which has yet to be identified, would be expected to exhibit >99% similarity to SARS-CoV-2 across the genome as a whole. However, the discovery of these viruses again highlights that more closely related viruses in bats and other wildlife species will be identified with enhanced sampling in a broader geographic region, including most parts of Southeast Asia with high diversity of *Rhinolophus* species[24]. Since it has been seldom seen that a bat coronavirus is able to efficiently transmit among humans without adaptation and

the repeated human-animal contacts[9], introduction through an intermediate host, such as raccoon dogs, is more likely than direct zoonotic introduction.

Whether SARS-CoV-2 is introduced through a laboratory incident or it has been genetically manipulated is highly debatable. After a thorough analysis of the genetic characterizations of SARS-CoV-2 from both the early and later stages, as well as its close relatives from wild animals, the global scientific community have reached the consensus that SARS-CoV-2 is unlikely to be a laboratory escape and there is no scientific evidence that SARS-CoV-2 has been genetically manipulated[9]. However, the exact spillover and emergence process of SARS-CoV-2 is still obscure, and more information from the earliest stage of the epidemic is clearly important to understand how SARS-CoV-2 reached humans.

3. Global genetic diversity of SARS-CoV-2

3.1 Genomic surveillance of SARS-CoV-2

Mutations are a natural part of the replication cycle of any RNA virus, leading to the diversification of viral lineages when coupled with inter-host transmission. This is also true of SARS-CoV-2, even though coronaviruses contain certain proofreading mechanisms that enhance the genome fidelity[28]. Genomic surveillance has generated unprecedented amount of sequencing data for a single virus (**Box 1**), and has proven an essential tool[29,30] to trace the spread of SARS-CoV-2 at various scales, from individual transmission events to the intercontinental spread of the virus. In addition, it plays a central role in monitoring the evolution of SARS-CoV-2 and identifying novel variants with enhanced transmissibility and/or pathogenicity, decreased susceptibility to therapeutic agents and evading natural or vaccine-induced immunity (**Fig. 1**). Genomic surveillance has demonstrated the effectiveness in tracking local transmission cases, recognizing importation sources and superspreading events in Australia[31,32], in informing public health decision-making in the Netherlands[33], and in adopting social distancing measures to reduce viral spread in Israel[34]. In January 2021, du Plessis and colleagues described the analysis of 50,887 SARS-CoV-2 genomes[35], quantifying the viral genetic structure of the UK epidemic at fine scale, including the size, spatio-temporal origins and persistence of lineages as well as the impact of intervention measures.

Herein, we take Guangdong province, China and the USA as examples to illustrate how genomic surveillance has facilitated our understanding of this pandemic.

Guangdong, China

Guangdong is a populous province in Southeast China, with resident population >100 million. After the SARS-CoV outbreak, believed to have originated in Guangdong[36], long-term reforms in the public health agencies have greatly improved the infrastructures and enhanced the capacity of disease control and prevention. The first case of COVID-19 in Guangdong had symptom onset on January 1 and was reported on January 19, 2020[10,37]. Like many other Chinese provinces, Guangdong experienced three phases - domestic importation, local community transmission and international importation - with the epidemic peak in early February[37]. Large-scale surveillance (~1.6 million tests by March 19, 2020 identifying 1,388 SARS-CoV-2 cases) and intervention measures were implemented from the very beginning the outbreak, and after February 22 no more than one case was being reported daily[37]. The genomic epidemiology of SARS-CoV-2 in Guangdong showed that most of the infections before March were imported from Hubei province, particularly Wuhan. Although some early cases were caused by community transmission, local transmission chains were limited both in size and duration[37]. These results suggest the efficacy of intensive testing and contact tracing even in such a densely populated urban region. Intensive surveillance also identified two SARS-CoV-2 variants with deletions in the spike gene[38]. In addition, the Guangdong CDC successfully identified the imported Alpha and Beta variants on January 2, 2021[39] and January 6, 2021[40], respectively.

The USA

The first COVID-19 case in the USA (sequence WA1) was reported on January 20, 2020, representing a traveler from Wuhan[41]. By February 15, 2020, the number of laboratory confirmed and clinically diagnosed COVID-19 cases reached 15[42]. By combining multiple sources of information, Worobey and colleagues showed that the WA1 (belonging to lineage A) case was successfully contained, and the subsequent larger outbreaks in the Washington State might have been caused by multiple independent introductions of the virus from China in late January or early February, 2020[43]. However, evidence from various studies revealed that the early viruses present between February 29 and March 18, 2020 in the New York City were imported from

Europe and other parts of the United States via multiple, independent introductions[44]. In addition, cryptic transmission and a prolonged period of unrecognized community spread has been documented in Northern California[45], Washington State[46] and New York City[47] from late January to March 2020. For example, SARS-CoV-2 sequences sampled from Connecticut during March 6-14, 2020 group with those from Washington State, highlighting the long-distance domestic transmission[48]. Genomic surveillance in Dane and Milwaukee counties in Wisconsin between March and April, 2020 provided evidence for reduced viral spread following the statewide "Safer at Home" order[49]. Combined, these genomic surveillance studies clearly depict the early transmission of SARS-CoV-2 and highlight the efficacy of intensive testing, contact tracing and decreasing public gatherings in containing SARS-CoV-2.

3.2 Mutational diversity of SARS-CoV-2

By January 2021, approximately 25,000 out of the 29,800 sites (the length of the complete SARS-CoV-2 genome) have been shown to carry mutational differences (https://bigd.big.ac.cn/ncov/), and it has been estimated that approximately two mutations are fixed in the SARS-CoV-2 genome per month[43,50,51]. Although most of these mutations represent standard replication errors, host-dependent RNA editing may also shape the short- and long-term evolution of SARS-CoV-2. Indeed, the SARS-CoV-2 genome is characterized by frequent biased C→U hypermutation that is likely due to a human APOBEC-like editing process[52,53].

Similar to other coronaviruses, the spike protein of SARS-CoV-2 contains important antigen epitopes[54,55]. As such, mutations in the spike protein will likely affect the receptor binding efficiency and potentially lead to immune escape and even weaken vaccine efficacy. The first notable mutation was A23403G that caused the D614G amino acid substitution in the spike protein. This mutation might have arisen separately as early as in late January 2020 in China and later in Europe, representing an interesting mutation of convergence evolution, and greatly increased in frequency during the European outbreak[56,57]. There is now compelling evidence that D614G has increased virus infectivity and transmissibility[56-61], and molecular epidemiological studies suggest that this mutation increased R_0 from 3.1 (614D) to 4.0 (614G)[57]. In addition, a so-called 'cluster V' (also called B.1.1.298) SARS-CoV-2 variant was identified in Danish mink that also carried mutations in the spike protein, including

Y453F, I692V, M1229I and the deletion of two amino acids (69-70) (**Fig. 2**)[14,62].

Not surprisingly, as number of COVID-19 cases continue to rise, mutational variants with a likely greater impact of fitness have also emerged, including some that might result in immune escape. Indeed, there are putative escape mutations to the ten human monoclonal antibodies (mAbs) targeting the SARS-CoV-2 RBD[63]. Of particular note are the major SARS-CoV-2 'variants of concern' (VOC) that arose in late 2020: Alpha (formerly B.1.1.7, and also called VOC-202012/01), Beta (formerly B.1.351, and also denoted 501Y.V2), Gamma (formerly P.1), and Delta (formerly B.1.617.2) first identified in the UK[64,65], South Africa[66], Brazil[67,68], and India[69], respectively (**Box 2, Fig. 1-2**).

The emergence of these variant lineages has raised concerns that the virus has entered a new phase in its evolution[70,71], characterized by ongoing immune escape in the face of rising levels of infected hosts that likely impacts vaccine efficacy[72], as well as the possibility of selection for increased transmission due to the imposition of nonpharmaceutical interventions (NPIs)[71]. The Alpha variant has been associated with increased rates of virus population growth[64,65], and has been reported to be able to escape neutralization by most mAbs targeting the NTD of the spike[73]. However, there is no widespread escape of the Alpha variant from mAbs or antibody responses generated by natural infection or vaccination[73-75], such that its spread may reflect increased transmissibility. In particular, some of the Alpha variants acquired additional mutations in the spike protein, particularly E484K, and exhibited a substantial loss of neutralizing activity by vaccine-elicited antibodies and mAbs resistance to convalescent plasma[76]. More worryingly, the Beta variant can escape neutralization by most RBD mAbs and substantially escape from neutralizing antibodies in COVID-19 convalescent plasma[73,77,78]. Similarly, the Gamma variant shows significant decreases in neutralization with post-vaccination sera[79] although, surprisingly, it is significantly less resistant to naturally acquired or vaccine-induced antibody responses than the Beta lineage[80]. In addition, neutralization of the Delta lineage is reduced when compared with ancestral circulating strains[74,75], and convalescent sera from patients infected with the Beta and Gamma variants show markedly more reduction in neutralization of the Delta lineage[74].

As well as nucleotide substitutions, the SARS-CoV-2 genome has experienced many

deletion events. For example, some viruses from Singapore and Taiwan, China carried a 382-nt deletion truncating ORF7b and covering almost the entire ORF8[81-83]. This variant showed significantly higher replicative fitness *in vitro* than the wild type[81], but seemed to be associated with a milder infection clinically[82] and has not been reported in recent months. Su and colleagues described other ORF7b/8 deletions of various lengths, including viruses from Australia (138-nt), Bangladesh (345-nt) and Spain (62-nt)[81]. Long deletion events were also found in clinical samples from Beijing, with 120-nt deletion within the ORF7a and 154-nt deletion within ORF8, respectively[84].

4. Global spread of SARS-CoV-2

4.1 Initial spread of SARS-CoV-2 in China

Generally, China experienced three distinct phases of SARS-CoV-2 transmission: (i) initial rapid spread within Wuhan, (ii) seeding from Wuhan to cause community transmission in other regions of China, and (iii) sporadic outbreaks caused by international importations after China controlled the first wave[37,84].

Early spread of SARS-CoV-2 in Wuhan

The original SARS-CoV-2 outbreak in Wuhan can itself be divided into three phases[85]: (i) rapid transmission prior to the implementation of the large-scale population "lockdown" of the city on January 23, 2020[8], with an estimated effective reproduction number (R_e) of 3.5 (95% credible interval 3.4-3.7) during this period[86]; (ii) reduction of the rate of virus transmission during the period January 23 to February 1 (via lockdown and home quarantine), producing an average R_e of 1.2 (95% CI 1.1-1.3)[86]; and (iii) the interruption of transmission through intensified stringent interventions during February 2-16, 2020 (centralized isolation and treatment of the cases) and February 17 - March 8 (community screening). Population-based serological surveys conducted during March-May 2020 revealed that the overall seropositivity rate in Wuhan was 3.2%-4.4%[87-89], indicating that many infections went undetected due to asymptomatic and mild infections and the limited laboratory diagnosis capacity during the early stages of the outbreak[86,90,91]. However, a city-wide nucleic acid screening of SARS-CoV-2 between May 14 and June 1, 2020 among nearly ten million residents of Wuhan only found ~300 asymptomatic cases after the lockdown was lifted on April 8, 2020[92], and no symptomatic local cases have been found in the

city after May 10, 2020.

Spread from Wuhan to other provinces

The coincidence of the SARS-CoV-2 emergence and the massive seasonal human migration (*Chunyun*, starting from January 10 in 2020) for the Chinese Lunar New Year holiday likely exacerbated the seeding of the virus across China[93,94]. Movement restrictions from Wuhan, the key transportation hub in central China, commenced on January 23, 2020, and reduced the peak population numbers leaving the city two days before the Lunar New Year; unfortunately, however, the disease had spread to every province in mainland China by this time[95,96]. In general, following the rapid implementation of stringent and integrated NPIs, R_e in provinces outside Hubei decreased below the epidemic threshold (1.0) from February 8, 2020[97]. Compared with Wuhan, the seropositivity rate in cities outside Wuhan was far lower. According to a national COVID-19 sero-epidemiological survey in China during March-May 2020[89], only 0.44% of the sampled population in other cities of Hubei were positive, and only two out of more than 12,000 people outside Hubei tested positive, suggesting that SARS-CoV-2 transmission was well contained across the country during the first wave[95,98,99].

Frequent international importation events

Over six thousand incoming travelers from abroad infected with SARS-CoV-2 had been reported in mainland China by June 15, 2021, although RT-PCR testing at the border control and the 14-day centralized quarantine implemented in China since March 2020 greatly reduced any transmission risk. For example, in Guangzhou, Guangdong province in southern China, 73.5% of the imported positive cases were detected at the immigration checkpoint and 19.0% during centralized quarantine in hotels[100]. Although SARS-CoV-2 is predominantly associated with respiratory transmission, since June 2020, multiple Chinese provinces have detected SARS-CoV-2 RNA or live virus from the packages of frozen products[101]. Indeed, cold-chain food or package contamination was proposed to have triggered the resurgence in Beijing in June 2020[102] as well as other sporadic outbreaks in China[101], although this warrants further investigation. It is notable that the number of confirmed cases was low in the Xinfadi outbreak, Beijing, June 2020. Similarly, all the COVID-19 outbreaks in China triggered by internationally imported travelers are

small-scale, with a few sustained cases. This was mainly due to the citywide, grid-based mass-screening protocol using reverse-transcriptase–polymerase-chain-reaction (RT-PCR) testing[103].

4.2 Intercontinental spread of SARS-CoV-2

From China to other regions

The global spread of SARS-CoV-2 shows how rapidly geographically disparate countries can be reached by an emerging pathogen (**Fig. 3a**)[104-105]. Two distinct transmission phases of international exportations of SARS-CoV-2 were identified at the early stage of the pandemic[106]. In the first phase, a high volume of international airline passengers left Wuhan for hundreds of destinations across the world during the two weeks prior to the Wuhan lockdown[91]. Cities across Asia, Europe and North America are the main destinations and have reported several imported cases at the early stage of the outbreak[105,107], and the WHO declared a Public Health Emergency of International Concern on January 30, 2020. Containment of the outbreak in China, particularly the implementation of travel restrictions since late January 2020, significantly reduced the further spread of SARS-CoV-2 beyond China[95,96,98,108,109].

From Europe to other regions

However, international travel outside of China from mid-February to late-March facilitated the second phase of international SARS-CoV-2 spread and onward transmissions[106,110], with the epicenter quickly shifted to the Middle East[111] and Europe (**Fig. 3c**). Although France was the first country to identify COVID-19 cases in Europe, Italy soon became the first major hotspot in the continent[107,108,112,113], and Spain, Belgium and the UK reported the highest numbers of deaths in Europe during the first wave[114]. The virus exported from Europe acted as a major source of global spread[44], and the WHO eventually declared a pandemic on March 11, 2020. Countries quickly placed restrictions on flights from Europe during March-April 2020, although these measures could not prevent local community transmission[72,110].

By late March 2020 cases surged in the USA, with North America becoming the global epicenter[115,116]. By the end of 2020, the total number of confirmed cases recorded in the USA surpassed 20 million, including more than 350 thousand deaths. Although the first SARS-CoV-2 case in the USA was reported in a traveler returning

from China on January 20, 2020[41], phylogenetic evidence suggests that importations from Europe mainly contributed to the wide spread of the virus across the country[106,115]. Latin America and South Asia have also been badly affected. SARS-CoV-2 was confirmed in Brazil on February 25, 2020 and a month later it was found in every state, with confirmed cases exceeding one million on June 19, 2020[117,118]. Although the first COVID-19 case was confirmed in India on January 30, 2020 and the situation was seemingly under control until the end of March 2020[119], India has reported the second highest number of COVID-19 cases since September 2020[120]. Additionally, most African countries experienced community transmission by May 31, 2020, with most imported cases returning from Europe and the USA[121], and it is believed that the disease is generally underreported across Africa due to the limited testing and health care capacity[122-125].

Secondary waves across countries

NPIs, such as travel restrictions, case isolation and contact tracing, physical distancing, face covering, hand washing, and even closures of business and schools, have been widely implemented to mitigate the transmission of SARS-CoV-2[108,126,127]. Full or partial lockdowns during specific periods has also been imposed in many countries[114]. Although the effectiveness of different interventions and their combinations have varied, these measures have played an important role in the response to the first wave of the pandemic[128,129].

Unfortunately, following the relaxation of interventions, the recovery of population movements, and the spread of novel variants with higher transmissibility, a new wave of infections has swept through many nations since October 2020 (**Fig. 3d-3e, Supplementary Table 1**)[130-132]. The first wave in the USA mainly affected the Northeast of that country[133], whereas the second wave in summer mainly hit the south and west, and almost every state has seen a spike in cases during the third wave since October 2020[134]. Brazil has experienced a major second wave since November 2020 and even had death toll second only to the USA in early 2021[135]. Similarly, Europe also suffered from the spread of novel SARS-CoV-2 variants throughout the continent after travel resumed in the summer of 2020, with the highest daily number of cases recorded in many countries between October 2020 and March 2021. Following NPIs implemented and even the second or third lockdown, combined with ongoing and

large-scale vaccination, many countries passed the second wave by the end of May 2021. This has reduced the pressure on the healthcare system and bought time to vaccinate people at the greatest risk of severe disease[136].

However, the emergence and rapid spread of various SARS-CoV-2 VOCs and VOIs that are more contagious and/or potentially evade immunity has triggered new waves in many countries (**Fig. 3b, Extended Data Fig. 1**). For example, India has experienced a major second wave from March to June 2021, mostly due to the Delta variant. As of August 10, 2021, a total of 142 countries, territories and areas across the world have reported the Delta variant[69] (**Extended Data Fig. 1d**), even in countries with mass vaccination, e.g. the UK and Israel[137]. In particular, community transmission of this variant has also been reported in many countries[69]. In mid-June 2021, the WHO declared that the Delta variant has displaced most of the other VOCs and become the dominant lineage across the world[137,138].

5. Challenges and outlook

Although of vital importance to the prevention of future emerging infectious diseases that will inevitably impact human populations, current understanding of the initial SARS-CoV-2 spillover event is limited. Although the closest relatives to SARS-CoV-2 are found in horseshoe bats, it is unclear whether the virus directly jumped from bats to humans or was passed through an intermediate animal host as was the case for previous coronavirus epidemics, although the latter seems more reasonable[6,9].

The genomic surveillance of SARS-CoV-2 is by far the largest pathogen genomic sequencing project ever undertaken, with more than 2.8 million complete genomes generated as of August 2021. This endeavor has played an essential role in the prevention and control of COVID-19 and shed light on the transmission patterns of SARS-CoV-2 at different scales, such as the time and source of the introduction events, the spatio-temporal characterizations of local spread, the role of superspreading events, and also the viral factors associated with the fitness, transmissibility, infectivity and disease severity. Of particular note is the identification of the major SARS-CoV-2 variants of concern, as well as several Variants of Interest (VOI; denoted Epsilon to Lambda)[139,140] that emerged in different countries and have

caused an increased proportion of cases both locally and globally.

The emergence of these SARS-CoV-2 variants has shaped the complex global transmission dynamics of COVID-19. More importantly, there is mounting evidence[141] that these SARS-CoV-2 variants are able to cause decreases in neutralizing titers from convalescent patients and vaccine recipients, and escape neutralization by the mAbs targeting the NTD and RBD of the spike to different degree. However, genomic surveillance would be more informative if coupled with a system for risk assessing and phenotyping these mutations. For example, the infectivity and antigenicity of 106 mutations in the SARS-CoV-2 spike was assessed using pseudotyped viruses[142]. Deep mutational scanning has also been used to assess all single amino acid variants of the SARS-CoV-2 spike protein[143,144]. In addition, more and more data on antigenic variations of the SARS-CoV-2 variants, with different sets of single amino acid mutations, to mAbs and vaccines are available. A risk assessment system that integrates pathogen surveillance and immune escape data is desirable, although confounded by the different classes of neutralizing antibodies, the NTD antibodies, vaccine strategies, and even the host heterogeneity.

That the major SARS-CoV-2 VOCs have reduced the efficacy of mAbs and vaccines has posed serious challenges in the control of the COVID-19 pandemic. First, although vaccines can protect people with SARS-CoV-2 variants against severe disease, vaccine manufacturers are exploring redesigns of their products to gain more effective protection – to eventually prevent virus transmission. Second, the suboptimal protection provided by vaccines[145] and the deployment of antibody-based treatments of limited or undemonstrated efficacy[146] has raised concerns that this would accelerate the emergence of new variants, although there is a strong argument for mass vaccination even if vaccines can only provide partial immunity[147,148]. Third, this has also raised the possibility that SARS-CoV-2 will become a recurrent seasonal infection[149,150]. Fourth, since vaccines cannot completely prevent transmission of the major variants, stringent NPIs should have to be implemented in order to reduce transmission of the virus, as unlimited, large-scale spread of the variants would likely generate more novel variants.

The genomic surveillance of SARS-CoV-2 is also facing several major challenges. First, despite this enormous endeavor, in reality only a tiny proportion (~1.3%) of

cases have been sequenced. In addition, the majority of sequences come from a small number of countries, and remarkably, as of August 2021, ~50% of genomes have generated in the UK and the USA that have led the world in this respect. In marked contrast, other countries with very major outbreaks, such as India and Brazil, have sequenced a far smaller number of cases, which may cause delays in identifying variants with novel phenotypic characteristics. Therefore, it is likely that there are additional new variants that are yet undetected given the limited genomic surveillance in a number of regions. Indeed, because the major VOCs are genetically divergent, it is possible that they had been circulating cryptically in unsampled locations, or have also emerged in chronically infected hosts that shed virus for extended periods[151,152]. Second, the complex transmission dynamics caused by different SARS-CoV-2 variants and their continuous evolution clearly necessitate increased genomic surveillance. Third, it is possible that recombination among viruses will also change the genetic structure of SARS-CoV-2, perhaps generating viruses of altered phenotype. Indeed, there have already been suggestions of recombination between the Alpha and Epsilon variants in California in early 2021[153]. Similarly, the potential recombination between SARS-CoV-2 and other mild human coronaviruses should not be neglected.

In summary, SARS-CoV-2 has led to a new understanding of coronavirus evolution and the virus has entered a new evolutionary phase characterized by the frequent emergence and spread of variants that impact immune escape and reduce the efficacy of vaccines. Of particular concern is that the limited genomic surveillance in many low-income countries may cause delays in identifying variants with novel phenotypic characteristics. To contain this and future pandemics, we urgently call for closer international cooperation, increased vaccine supply and sharing, rapid information exchange, and the establishment of both the infrastructure and trained personnel required for the effective genomic surveillance of SARS-CoV-2 and other emerging viruses.

Figures

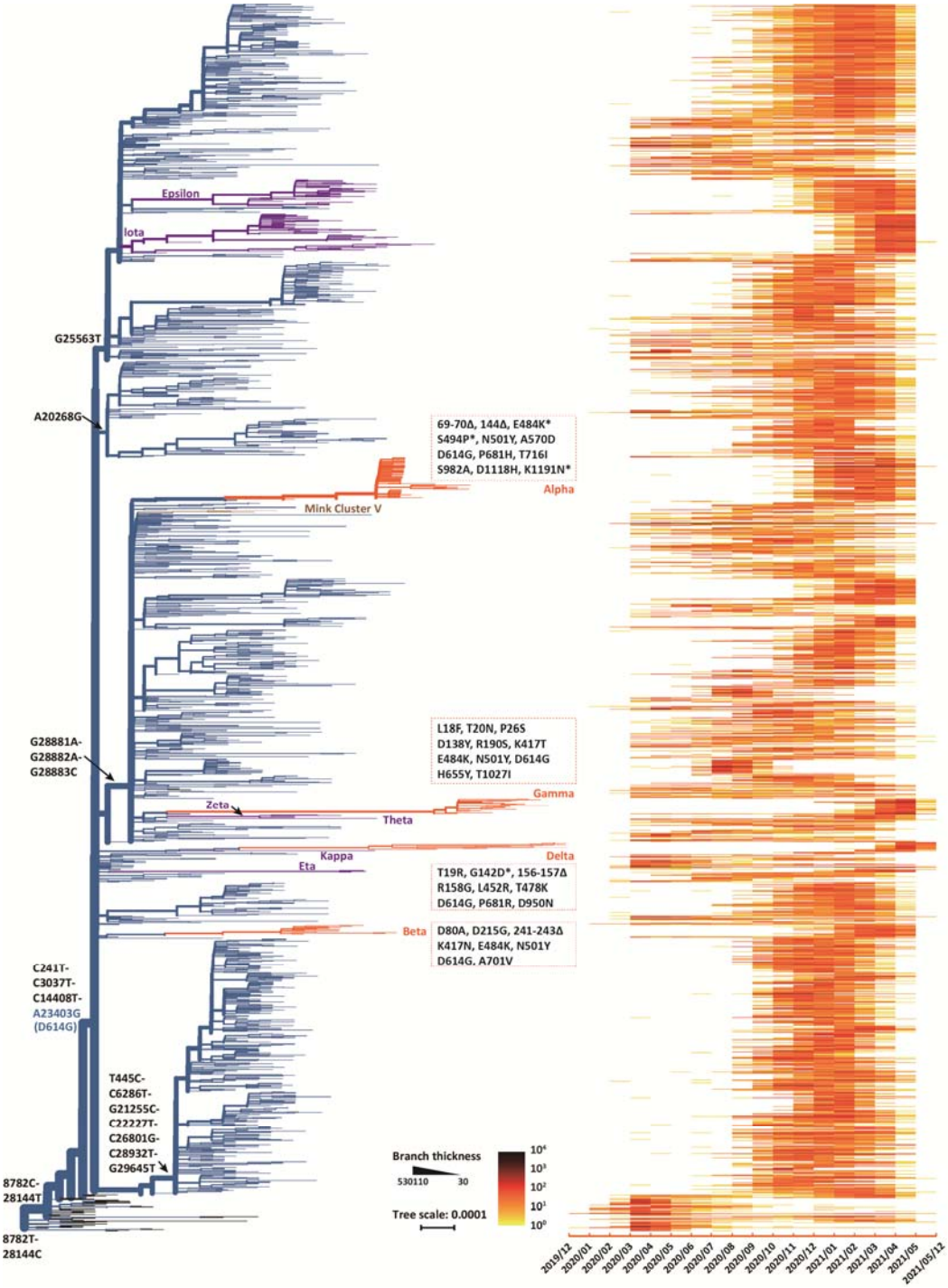


Fig. 1 | Phylogenetic tree of global SARS-CoV-2 and the temporal distribution of major sequence variants.

Phylogenetic analysis was performed using full-length genome sequences of SARS-CoV-2 collected from GISAID as of May 12, 2021. A maximum likelihood tree of 1715 representative high-quality SARS-CoV-2 sequences carrying specific accumulative mutations was estimated using RAxML[156], with 1,000 bootstrap replicates and the GTR nucleotide substitution model. The major VOCs (Alpha to Delta) are shown in orange, and the major VOIs (Epsilon to Lambda) are shown in purple. Both the thickness of each branch in the phylogenetic tree and the shading from light to dark in the heatmap indicate the number of sequences carrying specific sets of mutations. Specific nucleotide substitutions are highlighted on the major branches of the tree. The branches with the D614G mutation are colored blue.

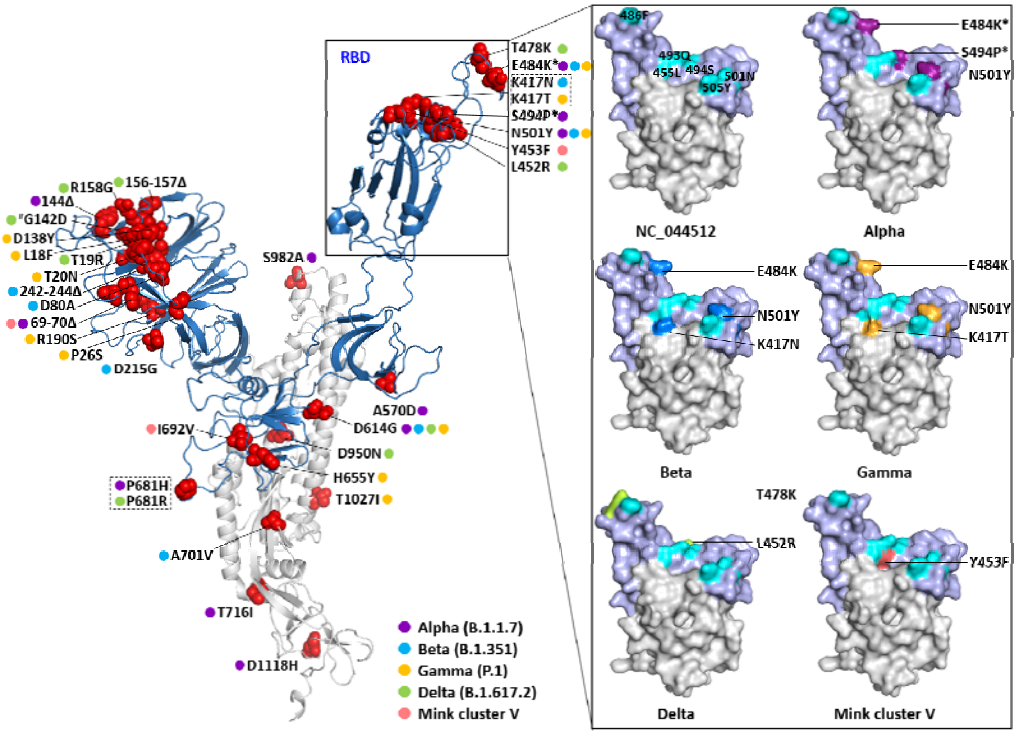


Fig. 2 | SARS-CoV-2 spike mutations in the Alpha, Beta, Gamma, Delta, and mink 'cluster V' variants.

Three-dimensional structures are modeled using the Swiss-Model program employing the spike protein of SARS-CoV-2 (PDB: 7CWU.1.G) as a template. In the left panel, the blue spheres represent the residues of NC_045512, and the red spheres represent the mutations found in the Alpha[157], Beta[64], Gamma[140], Delta[140] variants of concern,

as well as the mink 'cluster V' variants[62]. The amino acid positions of all the strains are numbered according to the template. The triangle represents a nucleotide deletion event. In the right panel, the surfaces of the six amino acid residues (L455, F486, Q493, S494, N501, and Y505) at the RBD are colored cyan. The molecular surfaces of the mutations in the Alpha (purple), Beta (blue), Gamma (yellow), Delta (orange), and mink 'cluster V' (pink) variants are highlighted. *Not all Alpha variants possess the E484K and S494P mutations. #Not all Delta variants possess the G142D mutation. It should be noted that we just use this figure to highlight the locations of the mutations in the variants based on the three-dimensional structure of one ancestral Wuhan strain (NC_045512), and this figure does not really represent the true three-dimensional structure of the variants.

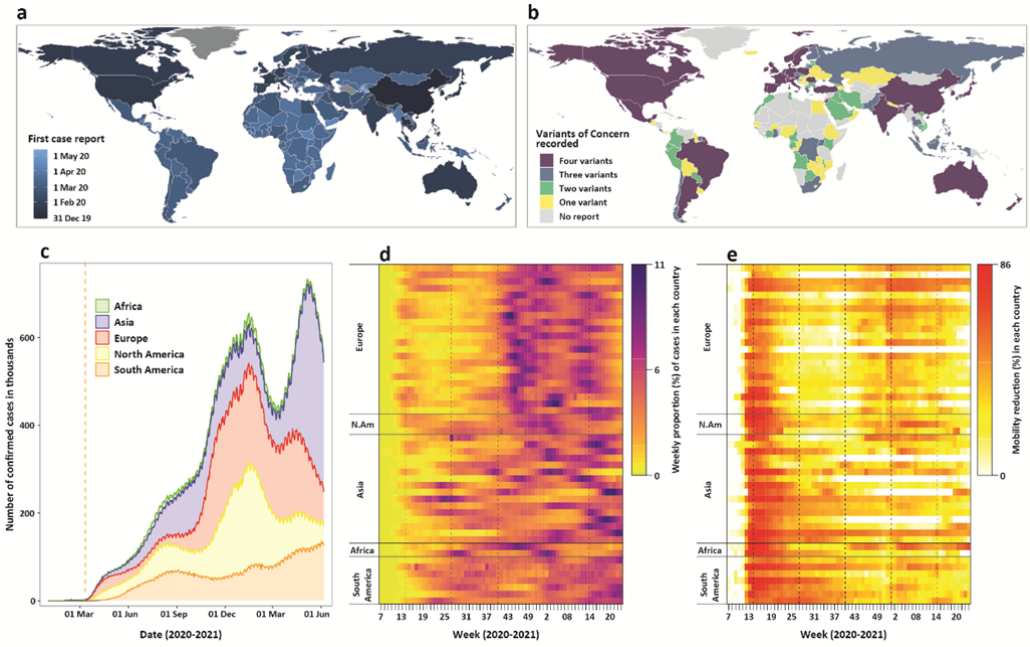


Fig. 3 | Global spread of SARS-CoV-2 and cases reported across countries.

a, The date of the first COVID-19 report in each country, territory or area. The areas without data are shown in grey. **b**, Reports of "Variants of Concern" (now denoted VOC Alpha to Delta) based on records published at the COVID-19 Weekly Epidemiological Update by the World Health Organization (https://covid19.who.int/), as of August 10, 2021. **c**, The 7-day rolling average of the number of confirmed COVID-19 cases reported by continent. The orange vertical dashed line indicates the date of COVID-19 declared as a pandemic by the WHO. **d**, The weekly proportion of case number in the top 50 ranked countries with the highest

number of COVID-19 cases and the available mobility data in panel **e**, as of August 8, 2021. The weekly proportion was calculated as the case count in a specific week and country, divided by the total number of cases reported in each country. **e**, The changes of human mobility (by August 8, 2021) in the 50 countries as presented in panel **d**, compared to the normal mobility from January 3 to February 6, 2020. Each row in panels **d** and **e** represents a country, grouped by continent and then sorted by the latitudes of capital cities from North to South (the country list is available in Supplementary Table 1). The grey dotted vertical lines in panels **d** and **e** from left to right indicate the first week of April, July, and October in 2020, and January, April, and July in 2021, respectively. The data set of case numbers was obtained from the data repository collated by the Johns Hopkins University (github.com/CSSEGISandData/COVID-19). The anonymized and aggregated data of population mobility in transit stations were obtained from the Google COVID-19 Community Mobility Reports (www.google.com/covid19/mobility/). The administrative boundary maps were obtained from the Natural Earth (www.naturalearthdata.com).

Acknowledgements

We thank the researchers who generated and shared the sequencing data from the GISAID and NCBI GenBank databases. We also acknowledge the efforts of the World Health Organization in sharing the COVID-19 Weekly Epidemiological Update, and the researchers that are part of the cov-lineages.org team (https://cov-lineages.org/) in assembling the records on new strains, and the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) in collating the COVID-19 case data (github.com/CSSEGISandData/COVID-19), and Google in producing and sharing the COVID-19 Community Mobility Reports (www.google.com/covid19/mobility/). We thank T. Hu, P. Wang, X. Yao and H. Song for helping produce the figures, and E. C. Holmes, A. Tatem and J. Floyd for commenting on the manuscript. This work was supported by Key research and development project of Shandong province (Grant No. 2020SFXGFY01 and 2020SFXGFY08), the National Key Research and Development Programme of China (Grant No. 2020YFC0840800), the National Science and

Technology Major Project (Grant No. 2018ZX10101004-002 and 2016ZX10004222-009), the National Natural Science Fund of China (81773498), the Academic Promotion Programme of Shandong First Medical University (2019QL006), and the Bill & Melinda Gates Foundation (INV-024911). W.S. was supported by the Taishan Scholars Programme of Shandong Province (ts201511056).

Competing interests

The authors declare no competing interests.

Author contributions

W.S. conceived the study. J.L. performed phylogenetic analysis and homology modelling. S.L. conducted the literature review on the global spread of SARS-CoV-2 and VOCs, and collected, analyzed and visualized the data of case number, VOC reports, and human mobility, using publicly available data resources. W.S., J.L., and S.L. wrote the first draft of the manuscript. W.S., and G.F.G. proofread the manuscript and the pre-submission inquiry.

References

1 Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).
The first paper to describe the identification of SARS-CoV-2 in Wuhan in late December 2019, providing the sequences of three full-length viral genomes.

2 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574 (2020).
Described the genomic structure and phylogenetic position of nine SARS-CoV-2 genomes sampled from eight patients from different hospitals in Wuhan in late December, 2019.

3 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).

4 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
As well as providing an early identification of SARS-CoV-2, this paper provides a description of the closest relative of SARS-CoV-2 identified to date – the bat-derived coronavirus RaTG13.

5 Gorbalenya, A. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature microbiology* **5**, 536–544 (2020).
6 Study, J. W.-C. WHO-convened Global Study of Origins of SARS-CoV-2: China Part.
7 Hill, V. & Rambaut, A. *Phylodynamic analysis of SARS-CoV-2*, https://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420 (2020).
8 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **382**, 1199-1207 (2020).
The first estimate of the epidemiological characteristics of the early COVID-19 outbreak, providing important evidence that human-to-human transmission from the middle of December 2019 in Wuhan.
9 Holmes, E. C. *et al.* The Origins of SARS-CoV-2: A Critical Review. *Cell*, doi: https://doi.org/10.1016/j.cell.2021.08.017 (2021).
10 Chan, J. F. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514-523 (2020).
11 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263 (2020).
12 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature medicine* **26**, 450-452 (2020).
This paper describes the fundamental genomic features of SARS-CoV-2, particularly the receptor binding domain and the furin cleavage site, and outlines the case for its zoonotic origin.
13 Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J Virol* **94**, e00127-00120 (2020).
14 Garry, R. F. & Gallaher, W. R. *Naturally occurring indels in multiple coronavirus spikes*, https://virological.org/t/naturally-occurring-indels-in-multiple-coronavirus-spikes/560 (2020).
15 Li, X. *et al.* A furin cleavage site was discovered in the S protein of the 2019 novel coronavirus. *Chinese Journal of Bioinformatics (In Chinese)* **18**, 103-108 (2020).
16 Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat Struct Mol Biol* **27**, 763-767 (2020).
17 Johnson, B. A. *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293-299 (2021).
18 Zhou, P. *et al.* Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **588**, E6 (2020).
19 Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* **5**, 1408-1417(2020).
20 Lam, T. T. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285 (2020).
21 Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286-289(2020).
22 Li, X. *et al.* Pathogenicity, tissue tropism and potential vertical transmission of SARSr-CoV-2 in Malayan pangolins. *bioRxiv*, doi:10.1101/2020.06.22.164442 (2020).
23 Zhou, H. *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural

Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr Biol* **30**, 2196-2203 (2020).

24 Zhou, H. *et al.* Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*, doi:10.1016/j.cell.2021.06.008 (2021).

25 Murakami, S. *et al.* Detection and Characterization of Bat Sarbecovirus Phylogenetically Related to SARS-CoV-2, Japan. *Emerg Infect Dis* **26**, 3025-3029 (2020).

26 Hul, V. *et al.* A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *bioRxiv*, doi:10.1101/2021.01.26.428212 (2021).

27 Wacharapluesadee, S. *et al.* Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun* **12**, 972 (2021).

28 Bouvet, M. *et al.* RNA 3 '-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9372-9377 (2012).

29 Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature reviews. Genetics* **19**, 9-20 (2018).

30 Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**, 10-19 (2019).

31 Rockett, R. J. *et al.* Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat Med* **26**, 1398-1404 (2020).

32 Popa, A. *et al.* Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Science translational medicine* **12**, eabe2555 (2020).

33 Oude Munnink, B. B. *et al.* Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat Med* **26**, 1405-1410 (2020).

34 Miller, D. *et al.* Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nature communications* **11**, 5518 (2020).

35 du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708-712 (2021).

Landmark paper describing the genomic surveillance of SARS-CoV-2 in the UK. Used an analysis of over 50,000 genome sequences to quantify the structure of SARS-CoV-2 lineages in the UK epidemic at fine scale.

36 Zhong, N. S. *et al.* Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet* **362**, 1353-1358 (2003).

37 Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003 e1009 (2020).

Detailed description of the genomic epidemiology of SARS-CoV-2 in Guangdong province, China. Illustrated how genomic surveillance facilitated outbreak containment.

38 Liu, Z. *et al.* Identification of Common Deletions in the Spike Protein of Severe Acute Respiratory Syndrome Coronavirus 2. *J Virol* **94**, e00790-00720 (2020).

39 Xinhua. *Imported coronavirus variant case reported in Guangdong*, http://www.xinhuanet.com/english/2021-01/03/c_139637931.htm (2021).

40 Xinhua. *Guangdong reports imported coronavirus variant case*, http://www.xinhuanet.com/english/2021-01/06/c_139646690.htm (2021).

41 Holshue, M. L. *et al.* First Case of 2019 Novel Coronavirus in the United States. *New England*

Journal of Medicine **382**, 929-936 (2020).

42 WHO. *Coronavirus disease 2019 (COVID-19) Situation Report – 26*, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200215-sitrep-26-covid-19.pdf?sfvrsn=a4cc6787_2 (2020).

43 Worobey, M. *et al*. The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564-570 (2020).

44 Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297-301 (2020).

The first paper described the early transmission dynamics of SARS-CoV-2 in New York City, highlighting viral introductions from Europe into the USA.

45 Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582-587 (2020).

46 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571-575 (2020).

Notable paper descriing the cryptic transmission of SARS-CoV-2 in the USA.

47 Maurano, M. T. *et al.* Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. *Genome research* **30**, 1781-1788 (2020).

48 Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996 e995 (2020).

Described the long-distance domestic spread of SARS-CoV-2 in the USA.

49 Moreno, G. K. *et al.* Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat Commun* **11**, 5558 (2020).

50 Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus evolution* **6**, veaa061 (2020).

51 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).

52 Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science advances* **6**, eabb5813 (2020).

53 Simmonds, P. Rampant C -> U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, e00408-00420 (2020).

54 Lu, G., Wang, Q. & Gao, G. F. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol* **23**, 468-478 (2015).

55 Wang, Q., Wong, G., Lu, G., Yan, J. & Gao, G. F. MERS-CoV spike protein: Targets for vaccines and therapeutics. *Antiviral research* **133**, 165-177 (2016).

56 Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827 (2020).

An important paper that provided the initial evidence that the D614G mutation increased the infectivity, and hence transmissibility, of SARS-CoV-2.

57 Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75 (2021).

58 Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116-121 (2020).

59 Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* **11**, 6013 (2020).
60 Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739-751 e738 (2020).
61 Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464-1468 (2020).
62 Ria Lassaunière *et al. SARS-CoV-2 spike mutations arising in Danish mink and their spread to humans*, https://files.ssi.dk/Mink-cluster-5-short-report_AFO2 (2020).
63 Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell host & microbe* **29**, 44-57.e49 (2021).
64 Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*, doi:10.1101/2020.12.30.20249034 (2021).
65 Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2020).
66 Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*, doi:10.1101/2020.12.21.20248640 (2020).
67 Faria, N. R. *et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings*, https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (2021).
68 Naveca, F. *et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein*, https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585 (2021).
69 O'Toole, Á. & Hill, V. *SARS-CoV-2 lineages: B.1.617.2 report*, https://cov-lineages.org/global_report_B.1.617.2.html (2021).
70 MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol* **19**, e3001115 (2021).
71 Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv*, doi:10.1101/2021.02.23.21252268 (2021).
72 Kupferschmidt, K. Viral evolution may herald new pandemic phase. *Science* **371**, 108-109 (2021).
73 Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130-135 (2021).
74 Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236 (2021).
75 Liu, J. *et al.* BNT162b2-elicited neutralization of B.1.617 and other SARS-CoV-2 variants. *Nature* **596**, 273-275 (2021).
76 Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136-141 (2021).

77 Nelson, G. *et al.* Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv*, doi:10.1101/2021.01.13.426558 (2021).

78 Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat Med* **27**, 622-625 (2021).

79 Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2523 (2021).

80 Dejnirattisai, W. *et al*. Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939-2954 e9 (2021).

81 Su, Y. C. F. *et al.* Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *mBio* **11**, e01610-01620 (2020).

82 Young, B. E. *et al.* Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* **396**, 603-611 (2020).

83 Gong, Y. N. *et al.* SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerg Microbes Infect* **9**, 1457-1466 (2020).

84 Du, P. *et al*. Genomic surveillance of COVID-19 cases in Beijing. *Nat Commun* **11**, 5503 (2020).

85 Pan, A. *et al.* Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. *Jama* **323**, 1915-1923 (2020).

86 Hao, X. *et al.* Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584**, 420-424 (2020).

A comprehensive study of the transmission dynamics of COVID-19 in Wuhan through time, providing important lessons learnt from the interventions in the city.

87 Xu, X. *et al.* Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. *Nat Med* **26**, 1193-1195 (2020).

88 Liu, A. *et al.* Seropositive Prevalence of Antibodies Against SARS-CoV-2 in Wuhan, China. *JAMA Netw Open* **3**, e2025717 (2020).

89 ChinaCDC. a. *Results of national COVID-19 seroepidemiological survey in China*, http://www.chinacdc.cn/yw_9324/202012/t20201228_223494.html (2020).

90 Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489-493 (2020).

91 Niehus, R., De Salazar, P. M., Taylor, A. R. & Lipsitch, M. Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *Lancet Infectious Diseases* **20**, 803-808 (2020).

92 Cao, S. Y. *et al.* Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. *Nature communications* **11**,5917 (2020).

93 Jia, J. S. *et al.* Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* **582**, 389-394 (2020).

94 Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling

study. *Lancet* **395**, 689-697 (2020).
95 Lai, S. *et al.* Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* **585**, 410-413 (2020).
Quantified the impact of various non-pharmaceutical interventions and their timings on COVID-19, providing early evidence to inform response efforts across the world.
96 Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395-400 (2020).
97 Zhang, J. *et al.* Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* **20**, 793-802 (2020).
98 Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493-497 (2020).
99 Leung, K., Wu, J. T., Liu, D. & Leung, G. M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* **395**, 1382-1393 (2020).
100 Zhang, Z.-B. *et al.* Countries of origin of imported COVID-19 cases into China and measures to prevent onward transmission. *Journal of travel medicine* **27**, taaa139 (2020).
101 Bai, L. *et al.* Controlling COVID-19 Transmission due to Contaminated Imported Frozen Food and Food Packaging. *China CDC Weekly* **3**, 30-33 (2021).
102 Pang, X. *et al.* Cold-chain food contamination as the possible origin of Covid-19 resurgence in Beijing. *National Science Review* **7**, 1861-1864 (2020).
103 Xing, Y., Wong, G. W. K., Ni, W., Hu, X. & Xing, Q. Rapid Response to an Outbreak in Qingdao, China. *N Engl J Med* **383**, e129 (2020).
104 Bogoch, I. I. *et al.* Potential for global spread of a novel coronavirus from China. *Journal of travel medicine* **27**, taaa0111 (2020).
105 Lai, S., Isaac I. Bogoch, Alexander Watts, Kamran Khan & Tatem, A. *Preliminary risk analysis of 2019 novel coronavirus spread within and beyond China*, https://www.worldpop.org/events/china (2020).
106 Yang, J. *et al.* Uncovering two phases of early intercontinental COVID-19 transmission dynamics. *Journal of travel medicine* **27**, taaa200 (2020).
107 Pullano, G. *et al.* Novel coronavirus (2019-nCoV) early-stage importation risk to Europe, January 2020. *Eurosurveillance* **25**, 2000057 (2020).
108 Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638-642 (2020).
109 Wells, C. R. *et al.* Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proc Natl Acad Sci U S A* **117**, 7504-7509 (2020).
110 Russell, T. W. *et al.* Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study. *Lancet Public Health* **6**, e12-e20 (2021).
111 Devi, S. COVID-19 resurgence in Iran. *Lancet* **395**, 1896 (2020).
112 Giordano, G. *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med* **26**, 855-860 (2020).
113 Salje, H. *et al.* Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208-211 (2020).
114 Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in

Europe. *Nature* **584**, 257-261 (2020).
115 Bushman, D. *et al.* Detection and Genetic Characterization of Community-Based SARS-CoV-2 Infections — New York City, March 2020. *MMWR. Morbidity and mortality weekly report* **69**, 918-922 (2020).
116 Rossen, L. M., Branum, A. M., Ahmad, F. B., Sutton, P. & Anderson, R. N. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity — United States, January 26–October 3, 2020. *MMWR. Morbidity and mortality weekly report* **69**, 1522-1527 (2020).
117 Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255-1260 (2020).
118 Silveira, M. F. *et al.* Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. *Nat Med* **26**, 1196-1199 (2020).
119 Acharya, R. & Porwal, A. A vulnerability index for the management of and response to the COVID-19 epidemic in India: an ecological study. *The Lancet. Global health* **8**, e1142-e1151 (2020).
120 Laxminarayan, R. *et al.* Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science* **370**, 691-697 (2020).
121 Massinga Loembe, M. *et al.* COVID-19 in Africa: the spread and response. *Nat Med* **26**, 999-1003 (2020).
122 Gilbert, M. *et al.* Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *Lancet* **395**, 871-877 (2020).
123 Pullano, G. *et al.* Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature* **590**, 134-139 (2020).
124 Rice, B. L. *et al.* Variation in SARS-CoV-2 outbreaks across sub-Saharan Africa. *Nat Med* **27**, 447-453 (2021).
125 Salyer, S. J. *et al.* The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet* **397**, 1265-1275 (2021).
126 Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the United States. *Science advances* **6**, eabd6370 (2020).
127 Brauner, J. M. *et al.* Inferring the effectiveness of government interventions against COVID-19. *Science*, eabd9338 (2020).
A chronological and global data set was constructed to compare the effectiveness of different NPIs in reducing COVID-19 transmission among countries during the first wave of the pandemic.
128 Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav* **4**, 1303-1312 (2020).
129 Dehning, J. *et al.* Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* **369**, eabd9338 (2020).
130 Baker, R. E., Yang, W. C., Vecchi, G. A., Metcalf, C. J. E. & Grenfell, B. T. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. *Science* **369**, 315-319 (2020).
131 Han, E. *et al.* Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. *Lancet* **396**, 1525-1534 (2020).
132 Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *Lancet Infectious Diseases* **20**, 1247-1254 (2020).
133 Perkins, T. A. *et al.* Estimating unobserved SARS-CoV-2 infections in the United States. *Proc*

Natl Acad Sci U S A **117**, 22597-22602 (2020).

134 CDC. a. *CDC COVID Data Tracker*, https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days (2020).

135 Health, B. M. o. *Coronavirus, Brazil*, https://covid.saude.gov.br/ (2021).

136 ECDC. *COVID-19 pandemic*, https://www.ecdc.europa.eu/en/covid-19-pandemic (2021).

137 Kupferschmidt, K. & Wadman, M. Delta variant triggers new phase in the pandemic. *Science* **372**, 1375 (2021).

138 WHO *Weekly epidemiological update on COVID-19 - 29 June 2021*, https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---29-june-2021 (2021).

139 WHO. *Tracking SARS-CoV-2 variants*, https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (2021).

140 CDC. a. *SARS-CoV-2 Variant Classifications and Definitions*, https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (2021).

141 Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature reviews. Microbiology* **19**, 409-424 (2021).

142 Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284-1294 e9 (2020).

143 Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476 e6 (2021).

144 Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44-57 e9 (2021).

145 Saad-Roy, C. M. *et al.* Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes. *Science* **372**, 363-370 (2021).

146 Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277-282 (2021).

147 Cobey, S., Larremore, D. B., Grad, Y. H. & Lipsitch, M. Concerns about SARS-CoV-2 evolution should not hold back efforts to expand vaccination. *Nature reviews. Immunology* **21**, 330-335 (2021).

148 Hanage, W. P. & Russell, C. A. Partial immunity and SARS-CoV-2 mutations. *Science* **372**, 354 (2021).

149 Murray, C. J. L. & Piot, P. The Potential Future of the COVID-19 Pandemic: Will SARS-CoV-2 Become a Recurrent Seasonal Infection? *Jama* **325**, 1249-1250 (2021).

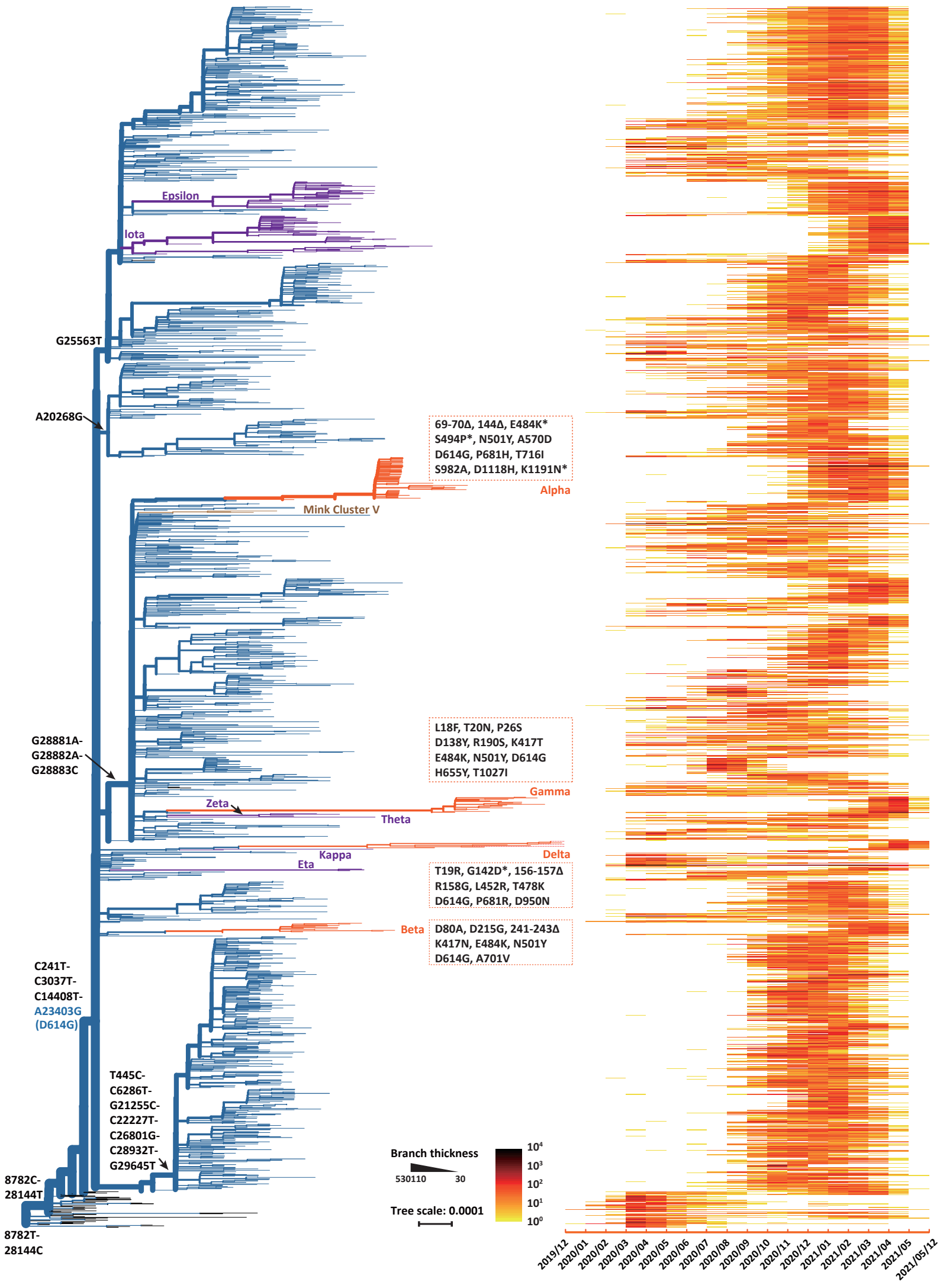
150 Phillips, N. The coronavirus is here to stay-here's what that means. *Nature* **590**, 382-384 (2021).

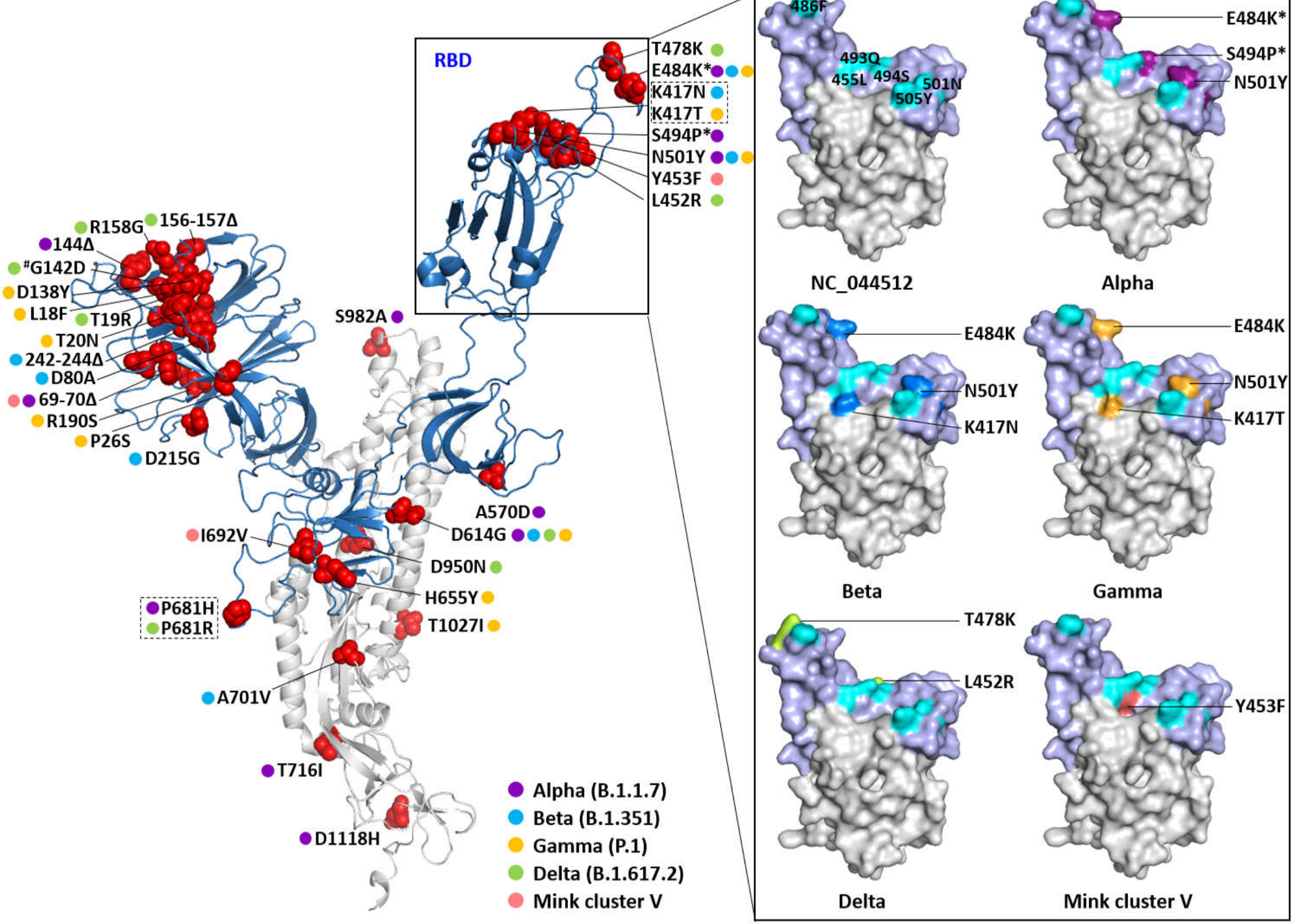
151 Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912 e9 (2020).

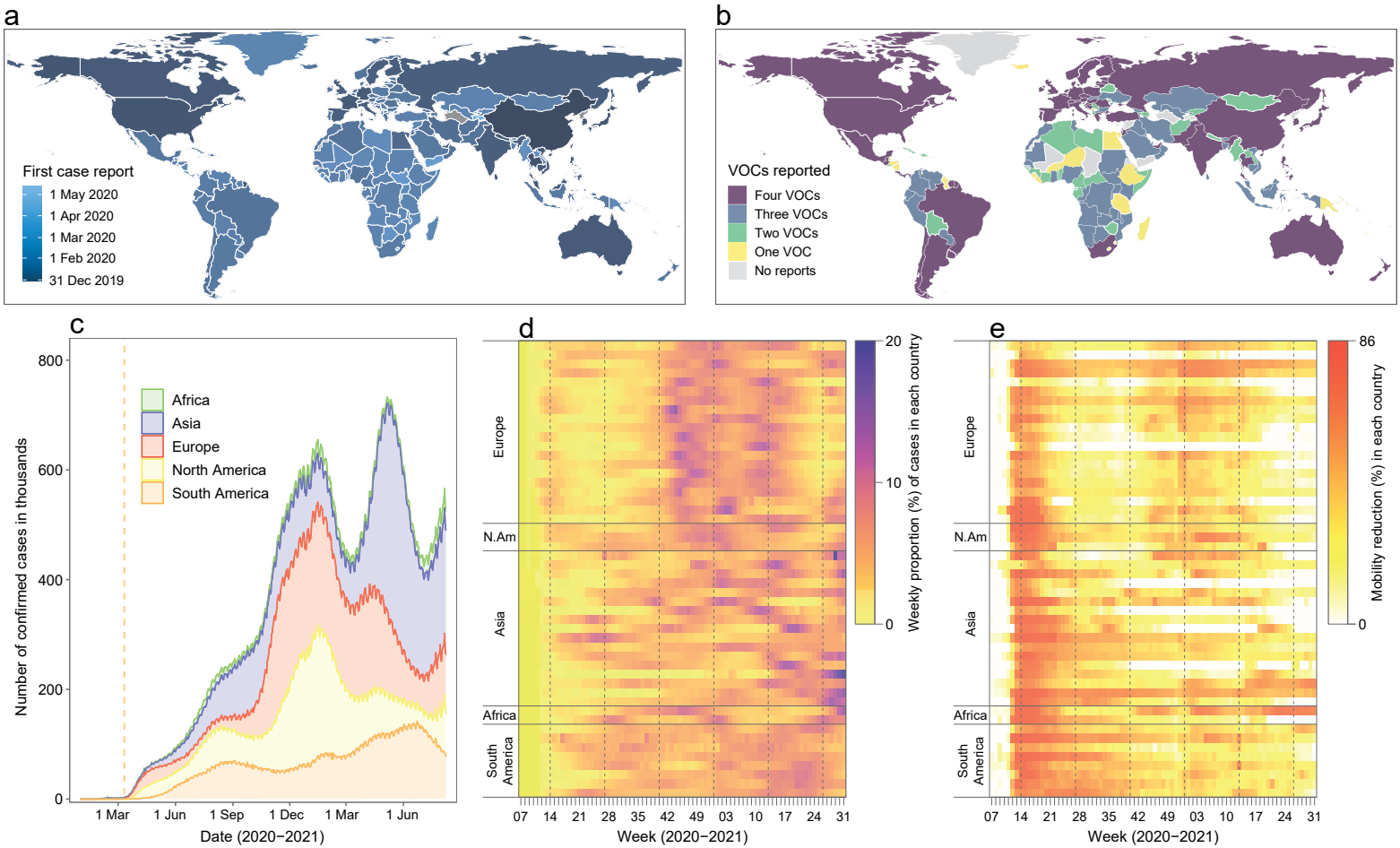
152 Aydillo, T. *et al.* Shedding of Viable SARS-CoV-2 after Immunosuppressive Therapy for Cancer. *N Engl J Med* **383**, 2586-2588 (2020).

153 Lawton, G. *Exclusive: Two variants have merged into heavily mutated coronavirus*, https://www.newscientist.com/article/2268014-exclusive-two-variants-have-merged-into-heavily-mutated-coronavirus/ (2021).

154 Song, S. *et al.* The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR. *Genomics Proteomics Bioinformatics*, doi:10.1016/j.gpb.2020.09.001 (2020).
155 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020).
156 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
157 GISAID. *UK reports new variant, termed VUI 202012/01*, https://www.gisaid.org/references/gisaid-in-the-news/uk-reports-new-variant-termed-vui-2020 1201/ (2021).
158 Davies, N. G. *et al.* Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* **593**, 270-274 (2021).
159 Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *Lancet Infect Dis*, doi:10.1016/S1473-3099(21)00170-5 (2021).
160 Graham, M. S. *et al.* Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Health* **6**, e335-e345.
161 Organization, W. H. *Weekly epidemiological update on COVID-19 - 10 August 2021*, https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---10-au gust-2021 (2021).
162 Control, E. C. f. D. P. a. *Emergence of SARS-CoV-2 B.1.617 variants in India and situation in the EU/EEA*. (2021).
163 *Delta Plus: Key things to know about new coronavirus variant* https://economictimes.indiatimes.com/news/et-explains/delta-plus-key-things-to-know-about-new-coronavirus-variant-in-india/articleshow/83739996.cms (2021).
164 Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603-1607 (2020).


Epsilon
Iota
G25563T
A20268G
69-70Δ, 144Δ, E484K*
S494P*, N501Y, A570D
D614G, P681H, T716I
S982A, D1118H, K1191N*
Alpha
Mink Cluster V
G28881A-
G28882A-
G28883C
L18F, T20N, P26S
D138Y, R190S, K417T
E484K, N501Y, D614G
H655Y, T1027I
Gamma
Zeta
Theta
Kappa
Delta
Eta
T19R, G142D*, 156-157Δ
R158G, L452R, T478K
D614G, P681R, D950N
Beta
D80A, D215G, 241-243Δ
K417N, E484K, N501Y
D614G, A701V
C241T-
C3037T-
C14408T-
A23403G
(D614G)
T445C-
C6286T-
G21255C-
C22227T-
C26801G-
C28932T-
G29645T
8782C-
28144T
8782T-
28144C
Branch thickness
530110
30
Tree scale: 0.0001
10^4
10^3
10^2
10^1
10^0
2019/12
2020/01
2020/02
2020/03
2020/04
2020/05
2020/06
2020/07
2020/08
2020/09
2020/10
2020/11
2020/12
2021/01
2021/02
2021/03
2021/04
2021/05
2021/05/12


RBD
T478K
E484K*
K417N
K417T
S494P*
N501Y
Y453F
L452R
R158G
156-157Δ
144Δ
#G142D
D138Y
L18F
T19R
T20N
242-244Δ
D80A
69-70Δ
R190S
P26S
D215G
S982A
A570D
D614G
I692V
D950N
H655Y
P681H
P681R
T1027I
A701V
T716I
D1118H
Alpha (B.1.1.7)
Beta (B.1.351)
Gamma (P.1)
Delta (B.1.617.2)
Mink cluster V
486F
493Q
455L
494S
501N
505Y
NC_044512
E484K*
S494P*
N501Y
Alpha
E484K
N501Y
K417N
Beta
E484K
N501Y
K417T
Gamma
T478K
L452R
Delta
Y453F
Mink cluster V


a
First case report
1 May 2020
1 Apr 2020
1 Mar 2020
1 Feb 2020
31 Dec 2019
b
VOCs reported
Four VOCs
Three VOCs
Two VOCs
One VOC
No reports
c
Africa
Asia
Europe
North America
South America
Number of confirmed cases in thousands
800
600
400
200
0
1 Mar
1 Jun
1 Sep
1 Dec
1 Mar
1 Jun
Date (2020−2021)
d
Europe
N.Am
Asia
Africa
South America
20
10
0
Weekly proportion (%) of cases in each country
07
14
21
28
35
42
49
03
10
17
24
31
Week (2020−2021)
e
Europe
N.Am
Asia
Africa
South America
86
0
Mobility reduction (%) in each country
07
14
21
28
35
42
49
03
10
17
24
31
Week (2020−2021)

BOX 1. Sources of SARS-CoV-2 genomic data and surveillance

The GISAID database (Global Initiative on Sharing All Influenza Data, https://www.gisaid.org/)

There have been more than 2.8 million complete SARS-CoV-2 genomes and metadata available from GISAID EpiCoV^{TM} as of August 2021. Useful tools, including Blast search, phylogenetic trees, PrimerChecker, spike glycoprotein mutation and emerging variants surveillance are provided, and related analyses are constantly updated.

The NCBI database (National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/)

More than 1.1 million SARS-CoV-2 nucleotide records and nine hundred thousand SRA runs have been deposited in the NCBI GenBank and SRA databases. The NCBI SARS-CoV-2 Resources (https://www.ncbi.nlm.nih.gov/sars-cov-2/) also provided a comprehensive access to other related data sources and numerous online analysis tools.

The CNBC/NGDC database (National Bioinformatics Center/National Genomics Data Center, https://bigd.big.ac.cn/ncov/)

This database integrates the SARS-CoV-2 genomes and related metadata from other sources, e.g. the GISAID, NCBI, GWH (Genome Warehouse, https://bigd.big.ac.cn/gwh/), NMDC (National Microbiology Data Center), and CNGB (China National GeneBank)[155]. It provides a variety of useful online analysis tools, including sequence integrity and quality assessment, spatiotemporal dynamics, Haplotype network, variant distribution, molecular mutation, and also published clinical trials.

PANGO lineages (https://cov-lineages.org/)

This is a useful nomenclature system for SARS-CoV-2 genomes. As of Aug 2021, the Pango system contains over 1500 designated lineages covering all the SARS-CoV-2 sequences from GISAID. Web-based or the open source code of applications e.g. Pangolin, Scorpio, Civet, Polecat are internally developed to cluster identify. Via the Pangolin web interface (https://pangolin.cog-uk.io/), sequences uploaded by the users can be assigned the most likely lineage based on the Pango dynamic nomenclature[156]. Information of the SARS-CoV-2 variants is also provided.

Nextstrain SARS-CoV-2 resources (https://nextstrain.org/sars-cov-2/)

Genomic epidemiological analysis of global SARS-CoV-2 is continually updated on the open source platform Nextstrain, based on the genomic data from GISAID. It provides a variety of visualization options for users. The nucleotide and amino acid diversity of the spike protein and the frequencies of the Nextstrain clades are provided and updated. In addition, Nextclade can perform clade assignment, mutation calling, and sequence quality check for the SARS-CoV-2 sequences uploaded by users.

BOX 2. Genetic characterizations of the major VOCs

The Alpha variant

The Alpha variant is defined by 17 amino acid-altering mutations (14 non-synonymous mutations and 3 deletions), including eight in the spike protein (**Fig. 1-2**). Notably, three of these mutations are of potential biological significance - N501Y, P681H and the deletion of two amino acids 69-70[65,66]. Notably, this new variant has increased infectiousness across all age groups, being 43% to 90% more transmissible than previously circulating strains[65,66]. In addition, the infection with the Alpha variant has the potential to cause substantial additional mortality, with an increased risk of death from 32% to 104%[159]. However, there are also reports of no association between this variant and increased severity[160,161]. As of August 10, 2021, 185 countries, territories or areas have identified this variant[162] (**Fig. 3b, Extended Data Fig. 1a**).

The Beta variant

The Beta variant is characterized by eight lineage-specific mutations in the spike protein, including three at important residues in the RBD (K417N, E484K and N501Y) (**Fig. 1-2**)[67]. Besides South Africa, 135 additional countries, territories or areas have also reported the identification of this variant as of August 10, 2021 (**Fig. 3b, Extended Data Fig. 1b**), with community transmission mainly found in Africa, Europe, and North America[162].

The Gamma variant

The Gamma variant contains a number of potentially important mutations, such as K417T, E484K, and N501Y in the spike protein (**Fig. 1-2**)[68,69]. The Gamma variant might be 1.7- to 2.4- fold more transmissible than previous (non-Gamma) strains in Brazil. As of August 10, 2021, identification of this variant has been reported in 81 countries, territories and areas (**Fig. 3b, Extended Data Fig. 1c**), with most of them located in America and Europe[162].

The Delta variant

The Delta variant contains several important amino acid mutations in the spike protein, including a three-amino acid-altering mutations (two deletions at 156 and 157, and one substitution of R158G) in the N-terminal domain (NTD), L452R, T478K, and P681R (**Fig. 1-2**)[163]. The Delta variant itself has been subject to ongoing evolution and a so-called “Delta plus” variant with an additional K417N mutation in the spike protein was identified in India in June 2021[138,164].

Despite their independent emergence (**Fig. 1**), the Alpha, Beta, and Gamma variants possess the N501Y mutation found in the mouse-adapted SARS-CoV-2 variant[165]. In addition, the Beta and Gamma lineages share E484K[65,66,68,69], which was also identified in the late, rather than early, Alpha variants[77].