Cell Host & Microbe

CellPress

Review

HIV-1 and SARS-CoV-2: Patterns in the evolution of two pandemic pathogens

Will Fischer,[1,2] Elena E. Giorgi,[1,2] Srirupa Chakraborty,[1,3] Kien Nguyen,[1] Tanmoy Bhattacharya,[4] James Theiler,[5] Pablo A. Goloboff,[6,7] Hyejin Yoon,[1] Werner Abfalterer,[1] Brian T. Foley,[1] Houriiyah Tegally,[8] James Emmanuel San,[8] Tulio de Oliveira,[8] Network for Genomic Surveillance in South Africa (NGS-SA), Sandrasegaram Gnanakaran,[1] and Bette Korber[1,2,*]

[1]T-6: Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA
[2]New Mexico Consortium, Los Alamos, New Mexico, 87545, USA
[3]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA
[4]T-2: Nuclear and Particle Physics, Astrophysics and Cosmology, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545 USA
[5]ISR-3: Space Data Science and Systems, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA
[6]Unidad Ejecutora Lillo, Consejo Nacional de Investigaciones Científicas y Técnicas - Fundación Miguel Lillo, S. M. de Tucumán, Miguel Lillo 251 4000, Argentina
[7]Research Associate, American Museum of Natural History, New York 10024, USA
[8]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Department of Laboratory Medicine & Medical Sciences, University of KwaZulu-Natal, Durban, South Africa
*Correspondence: btk@lanl.gov
https://doi.org/10.1016/j.chom.2021.05.012

SUMMARY

Humanity is currently facing the challenge of two devastating pandemics caused by two very different RNA viruses: HIV-1, which has been with us for decades, and SARS-CoV-2, which has swept the world in the course of a single year. The same evolutionary strategies that drive HIV-1 evolution are at play in SARS-CoV-2. Single nucleotide mutations, multi-base insertions and deletions, recombination, and variation in surface glycans all generate the variability that, guided by natural selection, enables both HIV-1's extraordinary diversity and SARS-CoV-2's slower pace of mutation accumulation. Even though SARS-CoV-2 diversity is more limited, recently emergent SARS-CoV-2 variants carry Spike mutations that have important phenotypic consequences in terms of both antibody resistance and enhanced infectivity. We review and compare how these mutational patterns manifest in these two distinct viruses to provide the variability that fuels their evolution by natural selection.

INTRODUCTION

In the past half century, two distinct, novel RNA viruses have caused global pandemics: human immunodeficiency virus type 1 (HIV-1) and severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2). Both emerged as zoonotic pathogens.

HIV-1 is most closely related to immunodeficiency viruses found in wild chimpanzees (SIVcpz; Hahn et al., 2000; Sharp and Hahn, 2011); zoonoses involving SIVcpz have occurred multiple times (Li et al., 2012). The vast majority of HIV/AIDS cases worldwide are associated with the M-group (Main), which comprises subgroups A, B, C, D, F, G, H, and J in addition to circulating recombinant forms (CRFs). In this review, we use the term HIV-1 in a restricted sense to refer to the M-group only. HIV-1 is transmitted sexually or through other body fluids such as blood and breast milk. The ancestor to the HIV-1 M group, which ultimately gave rise to the pandemic, is likely to have first entered the human population in Africa early in the 20th century (Zhu et al., 1998; Korber et al., 2000; Worobey et al., 2008). HIV-1 is a chronic infection that over many years results in immunodeficiency and eventually causes death by impairing immune defenses, allowing opportunistic infections to arise. As a consequence, acquired immune deficiency syndrome (AIDS) was not recognized as a distinct disease until 1981 (Centers for Disease Control (CDC), 1981; Hymes et al., 1981). A major challenge in eliminating HIV-1 is latency, which is not an issue in SARS-CoV-2.

By contrast, SARS-CoV-2 is most closely related to a virus isolated from bats (Zhou et al., 2020). An ancestral recombination event in the ACE-2 receptor binding region and an unusual insertion at the furin cleavage site of this virus are likely to have potentiated its transmission in humans (Li et al., 2020b; Walls et al., 2020). Its dominant form of transmission is respiratory (Meyerowitz et al., 2021). There have been three major zoonotic outbreaks of betacoronaviruses in the past two decades. The first outbreak, in 2002–2003, was of severe acute respiratory syndrome coronavirus [type 1] (SARS-CoV), which infected over 8,000 people and killed around 800 (Graham and Baric, 2010). The next was of Middle East respiratory syndrome coronavirus, MERS-CoV, which requires close contact for transmission but is a more lethal virus, with 35% mortality (Cui et al., 2019; Graham and Baric, 2010). There were 2,442 cases and 842 killed by MERS between 2012 and May 2019 (Donnelly et al., 2019). The third has been caused by SARS-CoV-2 (Gorbalenya et al., 2020). From its first detection in Wuhan, China (Zhou et al., 2020), it was just a matter of months before the global spread

of SARS-CoV-2 was formally recognized by the World Health Organization (WHO) as a pandemic (Ghebreyesus, 2020).

Over the course of 40 years since its discovery, at the close of 2019, the WHO estimated that HIV-1 has infected between 56 and 100 million people and that AIDS-related illnesses have taken between 25 and 42 million human lives. Despite great progress in treatment and in strategies to prevent HIV-1 transmission, between 32 and 45 million people were estimated to be living with HIV-1 at the close of 2019. In contrast, the WHO estimates that there have been over 110 million COVID-19 cases and 2.4 million deaths globally in the first year of the SARS-CoV-2 pandemic.

The two viruses are from different families and differ substantially in their ecology, mode of infection, genome content, and pathogenesis. HIV-1 belongs to the genus *Lentivirus* in the family *Retroviridae*. Its name derives from the Latin word *lenti*, meaning "slow"; this refers to the long chronic period of the virus, which typically lasts many years before the disease AIDS manifests (Giorgi et al., 2002). Like all retroviruses, HIV-1 replicates its genome in a multistep process: (1) reverse transcription of the viral genetic material, single-stranded RNA, to DNA; (2) insertion of the DNA copy into the genome of the infected host cell; and (3) subsequent expression of new genomic transcripts (as well as mRNAs for protein production) from the integrated DNA. When activated CD4+ T cells are infected, some revert to a resting memory state that is non-permissive for viral expression, creating a latent reservoir of infected cells that can eventually be reactivated. This latent reservoir makes it very difficult to clear the virus using antiretroviral or immune therapy, a major challenge for the HIV field. The HIV-1 reverse transcriptase RT, part of the Pol polyprotein, is an RNA-dependent DNA polymerase without proofreading activity. Compared to other viral families, all retroviruses, including HIV-1, have a high per-replication-cycle mutation rate (Mansky and Temin, 1995). Retroviruses are enveloped viruses, with a membrane, derived from the host cell, that incorporates host proteins (Burnie and Guzzo, 2019). The viral membrane is also studded with Envelope (Env) trimers that interact with host cell-surface proteins to enable cellular entry. During the course of an HIV-1 infection, the virus evolves under continuous immune pressure from the host (reviewed in Korber et al., 2017). In turn, both arms of the host immune response, T cell (Liu et al., 2013) and B cell responses (Gao et al., 2014; Bonsignori et al., 2016, 2017; Bhiman et al., 2015), adapt and respond to the changing antigenic profile of the virus. During the course of a single HIV-1 infection, the replicating viral population diversifies extensively, so that every HIV-1 infection is distinct, with each individual carrying a diverse HIV-1 quasispecies. In particular, diversification of Env proteins during chronic infection often includes mutations in key epitope regions and hypervariable regions that affect sensitivity to neutralizing antibodies (see Stephenson et al., 2020, for a recent review that addresses the role of hypervariable region evolution).

SARS-CoV-2 belongs to the *Coronaviridae*, a family of enveloped RNA viruses in which the genetic material is a single strand of positive-sense RNA that can serve directly as either an mRNA for protein production in an infected cell or as a template for genome production (in which it is copied by RNA polymerase to a negative strand, which then serves as the template for new positive strands that are packaged into viral progeny (Denison et al., 2011; Pal et al., 2020). Some members of this family of viruses cause severe respiratory or gastrointestinal diseases in mammals and birds. Coronaviruses have a proofreading mechanism that reduces the replication error rate (Romano et al., 2020), which is one factor in the relatively slow accumulation of mutations in the SARS-CoV-2 pandemic. Another critical factor is that the SARS-CoV-2 peak infectivity window is brief, occurring early in infection (He et al., 2020), so there is typically little time for *in vivo* viral evolution in any individual host prior to transmission. The combination of these two factors explains the low levels of diversity observed in the COVID-19 pandemic. Several variant lineages with relatively large numbers of mutations are emerging, and some may have accumulated these mutations in the context of prolonged infections in immunocompromised individuals (Rambaut et al., 2020b), and multiple studies documenting instances of such accumulation over time have been reported (Avanzato et al., 2020; Choi et al., 2020; Baang et al., 2021; Hensley et al., 2021; Kemp et al., 2021b). The B.1.1.7 lineage (named using the phylogenetically-based "Pango" nomenclature system; Rambaut et al., 2020a) that was first sampled in England in the fall of 2020, soon becoming highly prevalent there and spreading internationally, is an example of such a multiple-mutation lineage, though its specific origin is unknown. Although these levels of mutational distance would be trivial in the context of circulating HIV-1 (Figure 1), the lineages are quite divergent in the context of SARS-CoV-2; when they alter amino acids that display phenotypic differences in infectivity and relative antibody resistance, they may become epidemiologically significant.

Here, we discuss striking parallels and profound differences in the modes of evolution of these two viruses. We focus primarily, but not exclusively, on the docking/fusion proteins that mediate cell entry, Env (HIV-1) and Spike (SARS-CoV-2), because of the immediate importance of these proteins for immune-based therapeutics, potential small-molecule drugs, and vaccines. Although the tempo and pattern of change differ between the two viruses, HIV-1 and SARS-CoV-2 employ similar evolutionary tool kits in their adaptation for propagation through the human population. Single nucleotide mutations can alter antigenic profiles and infectivity in both viruses, but evolutionary change also occurs via insertions and deletions (indels) and recombination. While each of these mutational events can directly modify a protein, they can also alter the proteins' display of surface glycans, an important modulator of sensitivity to neutralizing antibodies. Finally, in addition to sharing their human host and a set of evolutionary mechanisms, these two pathogens have one additional critical thing in common: many in the HIV-1 research community, which for decades has been working on challenges in HIV-1 vaccine development, are now translating their experience to SARS-CoV-2 immunity and vaccines (Dumiak, 2020; Moore and Wilson, 2021). Thus, the evolutionary trajectory of HIV-1, which is very familiar to us and others in this research community, informs our perspective on the new pandemic.

In this review, we provide some basic analyses to enable a comparative framework for considering the overlapping sets of evolutionary strategies employed by HIV-1 and SARS-CoV-2. We compare the frequencies of mutations, their genomic and phylogenetic distributions, and the importance of insertions and deletions in variant lineages among thousands of HIV-1 genomes and hundreds of thousands of SARS-CoV-2 genomes sequenced since that outbreak came to worldwide attention.

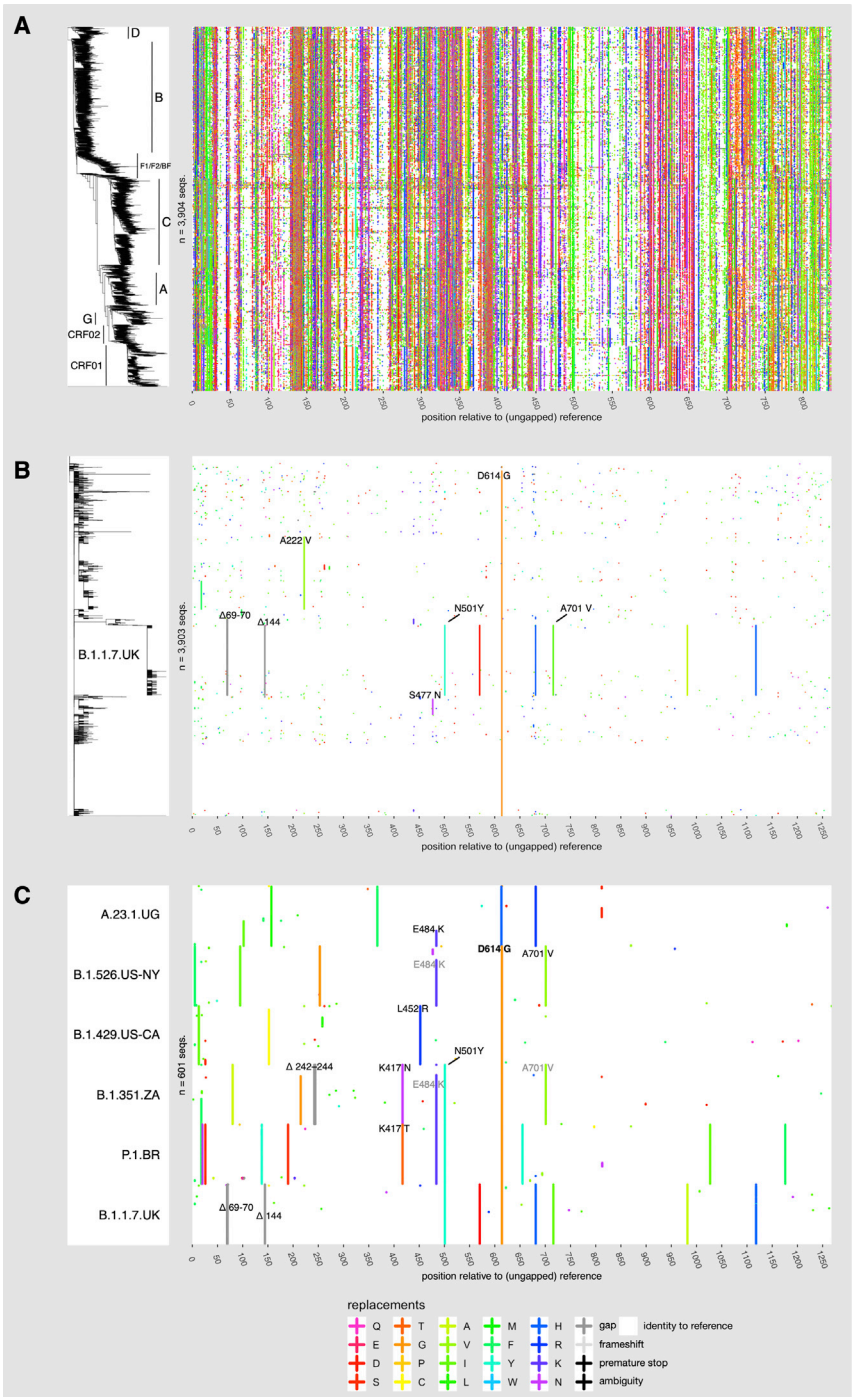


Figure 1. Variability of HIV-1 and SARS-CoV-2 docking/fusion proteins

Right panel: Variant-visualized amino-acid sequence alignments of HIV-1 Env (A) and SARS-CoV-2 Spike (B and C). The colored panels are a matrix in which each row represents a single sequence, and the columns are positions in a sequence alignment, where colored marks ("+") denote positions that vary compared to a reference sequence and reference-identical positions are white. A consensus sequence based on the most common base in each position serves as the reference for HIV-1 Env, and the outbreak strain (NC_045512) is the reference sequence for the SARS-CoV-2 Spike. Amino acid colors are based on Taylor (1997). Sequences are ordered top to bottom according to the phylogenetic tree in the left panel. Each phylogenetic tree is derived from a whole-genome nucleotide alignment: an approximation to the maximum-likelihood tree, generated with RAxML-NG (Kozlov et al., 2019) for HIV-1 Env (A) and a parsimony tree generated with TNT (Goloboff and Catalano, 2016) for SARS-CoV-2 Spike (B). As a consequence, continuous vertical stripes indicate lineage-specific mutations that are shared by related sequences (see text). Variant amino acids for the 100 most recent sequences (as of 2020-03-07) of 6 SARS-CoV-2 lineages with multiple Spike mutations, including 5 VOC/VOIs, are shown in (C). Mutations of particular interest that are discussed in the text are labeled in (B) and (C). The SARS-CoV-2 sequence data in this figure used data from the GISAID 2021-02-25 release date, "near-complete" alignment as described in Korber et al., (2020b); alignment statistics at https://cov.lanl.gov/.

But first, a comparison of the data we used and its sources follows. All SARS-CoV-2 sequence data used here were provided through GISAID, The Global Initiative for Sharing All Influenza Data, a global repository of both Influenza and SARS-CoV-2 viral sequence data (http://gisaid.org). The sequences were aligned and processed at cov.lanl.gov (https://cov.lanl.gov/content/index). The SARS-CoV-2 sequence data in GISAID were contributed by laboratories from throughout the world. Unless otherwise noted, most of the summaries and analyses provided here are based on the GISAID March 1, 2021 release, including 519,035 Spike protein sequences. At the time of this writing, May 23, 2021, just 18 months since SARS-CoV-2 was first detected in humans, 1,675,199 sequences are available through GISAID, reflecting an astonishing global effort.

We also evaluate the three-dimensional structures of the Spike and Env docking/fusion proteins in terms of immunological masking of critical epitopes by glycan shielding. We discuss the evidence for and the importance of recombination in both viruses and provide examples of evidence for recombination in SARS-CoV-2 data from South Africa. Finally, we summarize and evaluate the development of distinct epidemiologically relevant lineages as affected by virus ecology and consider the potential impact of viral diversity on vaccines going forward.

The HIV-1 sequence data used here originated at GenBank (https://www.ncbi.nlm.nih.gov/genbank/), but are taken directly from a curated alignment of 7,590 global HIV Env sequences publicly available through the HIV database (https://www.hiv.lanl.gov/), and restricted to include only one sequence per sampled individual. Unlike SARS-CoV-2, published HIV-1 sequences are very often sampled in great depth and longitudinally, and they

sometimes include many hundreds of sequences from a single individual, so restricting to one sequence per person provides a way to better reflect diversity at the population level.

NUCLEOTIDE AND AMINO-ACID DIVERSITY IN SARS-CoV-2 COMPARED TO HIV-1

The paucity of mutations in SARS-CoV-2 compared to HIV-1 is immediately evident (see Figure 1 for Spike and Env amino acids and Figure S1 for full-genome base mutations). As touched on above, this is due in part to the comparatively high fidelity of the coronavirus RNA polymerase (Denison et al., 2011): its 3′-to-5′ exonuclease activity removes mis-incorporated nucleotides, providing effective proofreading of the nascent transcript (Bouvet et al., 2012; reviewed by Robson et al., 2020). HIV-1's RNA polymerase has a high intrinsic error rate and no such proofreading mechanism (Mansky and Temin, 1995). As noted above, the relative lack of SARS-CoV-2 diversity is also due in part to the timing of the transmission window, which has a narrow peak, early in infection. Consequently, very little selection pressure is imposed by the adaptive immune system in response to a first infection. In contrast, within-host evolution of HIV-1 is extensive: it begins early in infection and continues through years of chronic infection, and selection for immune resistance drives significant diversification (Phillips et al., 1991; Arendrup et al., 1992; Fischer et al., 2010; Boutwell et al., 2010; Liu et al., 2013; Bonsignori et al., 2016, 2017; Bhiman et al., 2015; Bar et al., 2012). Natural selection can act decisively upon rare but favorable mutations; this was first noted for SARS-CoV-2 when the Spike mutation D614G, which confers greater infectivity, rapidly became the globally dominant form of the virus during the spring of 2020 (Korber et al., 2020b). In the case of influenza, antigenic drift results from the accumulation of mutations during a flu season arising as a consequence of the interplay between selection for immunological resistance and viral fitness (Wu et al., 2020); this drives the need to develop new influenza vaccines every few seasons. Many new SARS-CoV-2 variants that carried mutations conferring resistance to neutralizing antibodies and sera, as well as enhanced infectivity, began to emerge in late 2020 (Deng et al., 2021, West et al., 2021, Rambaut et al., 2020b, Bugembe et al., 2021). This suggests that, even as the virus continues to adapt to its new human host, antigenic drift may be well underway.

The major HIV-1 clades were well-established within 30–50 years of the origin of the HIV-1 M group (Korber et al., 2000; Worobey et al., 2008) and have a star-like phylogenetic distribution. As the HIV-1 pandemic continues, within-clade Env proteins show ever greater levels of diversity. As a consequence, over time, sera from infected people will become less potent even against viruses of the same clade as the infecting virus (Hraber et al., 2014; Rademeyer et al., 2016).

The mutational patterns observed in the two different viruses are distinctive. We inferred approximate maximum likelihood trees (Kozlov et al., 2019) from HIV-1 M-group and SARS-CoV-2 nearly-whole-genome datasets, using the same number of taxa (3,903) to facilitate direct comparisons. The parameterization of this model offers a compact summary of the mutational landscape of these viruses (Table S1). In terms of base frequencies, HIV-1 is A-rich (the genome is 45% adenine), while SARS-CoV-2, though A-T rich, is biased toward T (37% thymine in the coding strand). HIV-1 is subject to mutagenesis via APOBEC proteins (part of the host antiviral response, which acts on the pre-integration complementary DNA copy), but this is countered in part by the HIV-1 Vif protein (Malim, 2009). SARS-CoV-2 may likewise be subject to editing by host deaminases, including APOBEC3C, as part of the host innate response to viral infection (Di Giorgio et al., 2020; Wei et al., 2020).

Since the ancestral outbreak sequence for the SARS-CoV-2 pandemic is known, we can use it as a reference sequence to illustrate observed mutations in an alignment of SARS-CoV-2 sequence data; organizing the aligned sequences by the phylogenetic tree highlights the mutational patterns that distinguish the emerging lineages (Figure 1B). To provide a visual comparison with HIV-1, we created an HIV consensus sequence using the Los Alamos HIV-1 database curated reference set, simply taking the most common base found in each position in the alignment to use as a central reference point. This choice of a reference sequence was made to minimize the number of differences highlighted in our HIV figures (Figures 1A and S1). For illustration, we condense the alignments into a single figure that displays the 35 million bases in the 3,903-sequence curated full-genome HIV-1 alignment (containing only one sequence per infected individual) and >100 million bases in a randomly sampled 3,903-sequence near-full-genome SARS-CoV-2 alignment (Figure S1). The mutational patterns in the Env and Spike proteins, most relevant to vaccine design, are shown in Figures 1A and 1B, respectively; the high density of Env mutations reflects the formidable challenge of creating a HIV vaccine that can elicit cross-reactive immune responses.

Other conspicuous features of the data are the different degrees of ''bushiness'' in the phylogenetic trees (Figure 1) and the large numbers of lineage-specific mutations in HIV-1 Env; these patterns of enriched mutations are specific to, and help to define, major clades and circulating forms (Figures 1 and S1). The HIV-1 tree itself has a star-like structure that is consistent with a rapidly expanding infection in a homogeneously susceptible population. Contrastingly, in Spike, the emerging clades and variants of interest (VOIs) of SARS-CoV-2 are associated with a small number of amino-acid changes across the protein (or base mutations across the ~30,000 bases of genome). The long vertical lines in Figures 1B and S1 represent mutations that are shared among phylogenetically clustered sequences. Some clade-defining prominent mutations in Spike are apparent (Figure 1B). These include D614G, (Korber et al., 2020b), which is embedded in a 4-mutation haplotype that defines the G clade (Figure S1); A222V, which became common in the UK and Europe in the summer of 2020 (Hodcroft et al., 2020; Bartolini et al., 2020); and S477N, which dominated the Australian sampling in the summer of 2020. Both A222V and S477N became relatively less common in late 2020 and early 2021 as the lineages with these mutations were replaced (Shen et al., 2021) by VOIs or variants of concern (VOCs)—see the CDC SARS-CoV-2 Variant Classifications and Definitions website, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html.

In the first quarter of 2021, different VOIs/VOCs have been increasing in prevalence in some geographically distinct local populations at a rapid pace (Deng et al., 2021; West et al.,

2021; Rambaut et al., 2020b; Bugembe et al., 2021), but only the B.1.1.7 form was sampled at a high enough frequency globally to be visually apparent in the subset of sampled viruses included in Figure 1B. We therefore highlight the changes in the Spike protein (Figure 1C) and in the full-length genome (Figure S1C) that characterize the baseline forms in six of the variants that are increasing in frequency in local populations and are spreading globally. Two things are evident in these figures: first, that each VOI/VOC is itself a lineage—continuing to evolve, sampling additional mutations over time—and also that multiple VOIs/VOCs share particular mutations. To wit, the mutation N501Y, noticed first in the B.1.1.7 lineage and considered important due to its location in the Receptor Binding Motif (Rambaut et al., 2020b), is also found in the distinct lineages B.1.351 (South Africa; also called 501Y.V2; Wibmer et al., 2021) and P.1 (Brazil). N501Y enhances infectivity with a modest impact on neutralization (Leung et al., 2021; Rathnasinghe et al., 2021). L452R, found in the B.1.427 and B.1.429 lineages from California and more recently in the B.1.617-related lineages from India, can enhance infectivity and impart resistance to many RBD-targeting antibodies and sera (Deng et al., 2021, McCallum et al., 2021b). E484K is found in B.1.351, P.1, and in a sublineage of A.23.1; it confers neutralizing antibody resistance (Wibmer et al., 2021). Two distinct mutations from amino acid K417, K417T and K417N, appear in P.1 and B.1.351, respectively; mutations in K417 also contribute to neutralizing antibody resistance (Wibmer et al., 2021). B.1.351 shares a further mutation, A701V, with the B.1.526 variant first reported from New York (Figure 1C); A701V is an example of a shared mutation that is as yet unexplored for phenotypic consequences. The observation of any single mutation in multiple expanding lineages suggests convergent evolution, i.e., that fitness effects of that mutation help drive lineage expansion. As noted above, several of these mutations have been shown experimentally to be advantageous to the virus. The observation of multiple such mutations in particular lineages suggests that these fitness effects can be cumulative.

INSERTIONS AND DELETIONS: AN ADAPTIVE MECHANISM FOR HIV-1 AND SARS-COV-2

Insertions and deletions (indels) are a critical adaptive mechanism for both HIV-1 and SARS-CoV-2, but they manifest differently. Indels originate via non-homologous recombination and can happen anywhere in the HIV genome. However, viable indels that do not introduce frameshifts are most commonly found in hypervariable regions (Wood et al., 2009). HIV-1 insertions generally manifest as direct, short-repeat duplications (Wood et al., 2009). Four of the variable loops of the HIV-1 Env protein, (V1, V2, V4, and V5, but not V3) contain hypervariable sections that have an extraordinary capacity to change by insertion and deletion (Bricault et al., 2019); the variability in these regions is dramatic and plays an important role in neutralizing antibody resistance. The extraordinary length variation in these four hypervariable loops (Figure 2A) is accompanied by changes in net charge and in the number of N-linked glycosylation sites (Tian et al., 2016). For example, the V1 loop in Env can accommodate lengths that range between 5 and 66 amino acids (median 30; Figure 2A); some hypervariable V1 loops include no N-linked glycosylation sites, others up to 11 (median 4), and they have a net charge ranging from −6 to 8 (median −1). Much of the observed population-level variation in these loops can be recapitulated in a single individual during the course of infection (Stephenson et al., 2020). Hypervariable region indel evolution begins early in HIV-1 infections and contributes to viral escape from the earliest antibodies as they begin to impose selective pressure during antibody/viral co-evolution (Bar et al., 2012; Gao et al., 2014; Bonsignori et al., 2017; Roark et al., 2021; Korber et al., 2017).

Because indels are the primary evolutionary driver within the hypervariable regions of Env, it is inappropriate to assume sequence homology and base-substitution drive evolution in these regions. Thus, alignment dependent strategies for identifying positive selection associated with the acquisition of immunological resistance can be misleading. Generally, therefore, we explore the impact of hypervariable regions on neutralizing antibody sensitivity by using three attributes of the variable loops that are independent of alignment: loop length, net charge, and number of glycosylation sites. Particular characteristics of certain loops are associated with resistance to particular classes of broadly neutralizing antibodies (Bricault et al., 2019). The loop lengths, their charge, and variable glycosylation patterns all affect loop conformation, directly modulating access to critical epitopes. A remarkable aspect of Env hypervariable region evolution is that the location of hypervariable indels observed in a human host during early infection with HIV-1 will often be precisely recapitulated when that same Env is incorporated into a SHIV construct and used to infect Rhesus macaques (Roark et al., 2021).

SARS-CoV-2 is also accumulating indels that can have critical phenotypic consequences, but thus far only a few specific deletions have become prominent among pandemic variants. As in HIV-1 Env, these deletions often occur in, or proximal to, structurally flexible loop regions. The Spike ΔH69/V70 deletion (Δ69/70) is the most common globally and is found in many lineages and distinctive Spike contexts (Figures 2B and 2C). It first came to prominence in association with mink farm outbreaks in Denmark (European Centre for Disease Prevention and Control, 2020; Lassaunière et al., 2020; van Dorp et al., 2020) paired with either the N439K or Y454F mutations in the RDB (Shen et al., 2021; Kemp et al., 2021a). One study suggested that Δ69/70 has minimal impact on the neutralization potency of serum from convalescents or vaccinees (Shen et al., 2021), although another found that this deletion could affect antibody binding and/or neutralization (McCarthy et al., 2021); a third study reported that Δ69/70 can enhance infectivity *in vitro* (Kemp et al., 2021a). These forms of the virus were a significant presence in the European epidemic in the summer and early fall of 2020 (Shen et al., 2021). Their prevalence, however, like that of the G clade they descended from, began to decline soon after the more transmissible B.1.1.7 variant (also a G clade descendant) was first sampled in the United Kingdom in late September 2020 (Volz et al., 2021; Rambaut et al., 2020b; Davies et al., 2020). B.1.1.7 carries *both* the Δ69/70 deletion and a common deletion in the NTD supersite, ΔY144 (Δ144), which can confer resistance to multiple neutralizing antibodies that target this region (McCallum et al., 2021). Like Δ69/70, Δ144 also recurs in multiple lineages (Figures 2B, 2C, and S2A), although much of the growing global

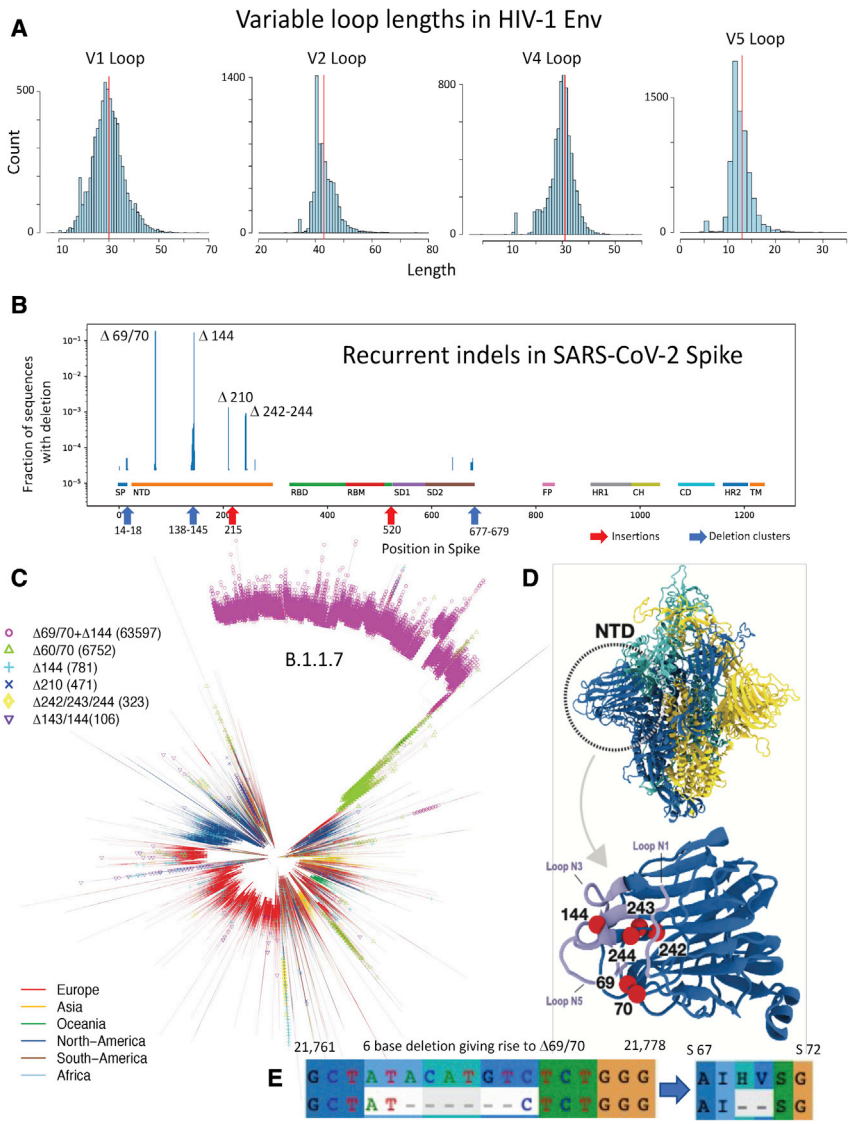


Figure 2. Distributions of Env loop lengths in HIV-1 and indel lengths and positions in SARS-CoV-2

(A) The distribution of hypervariable loop lengths (for loops V1, V2, V4, and V5) from the global Env sequence alignment from the HIV-1 database (1 sequence per individual). Hypervariable region lengths are calculated via the HIV-1 Los Alamos Database (https://cov.lanl.gov/) "Variable Region Characteristics" web interface; net charge and number of potential N-linked glycosylation sites can also be calculated using this tool.

(B) Frequencies of Spike indels from the GISAID 1-March-2021 release. A log scale is used, as most indels are quite rare except Δ69/70 and Δ144, which are common because they are present in the highly sampled B.1.1.7 lineage. Both, however, are also frequently sampled in other contexts: Δ69/70 was found an additional 10,168 times and Δ144 an additional 1,513 times. Focused regions of rare but recurring indels are highlighted here, and details are provided in Figures S1 and S2. The different regions in Spike are highlighted and include: the signal peptide (SP), the N-terminal domain (NTD), the receptor binding domain (RBD) and motif (RBM), subdomain 1 and 2 (SD1 and SD2), the fusion peptide (FP), heptad repeat 1 and 2 (HR1 and HR2), the central helix (CH), and the connecting domain (CD) and the transmembrane region (TM).

(C) A parsimony tree based on the cov.lanl.gov 17-March-2021 full-genome alignment (inferred with TNT 1.5; Goloboff and Catalano, 2016), showing the recurrence of the most common indel patterns in multiple lineages in the phylogeny. Branches are colored by the geographic region of the viral sample to illustrate that these mutations are geographically as well as phylogenetically dispersed.

(D) Structural representation of the SARS-CoV-2 Spike trimer, with three protomers shown in light blue, yellow, and cyan. Dashed circle indicates the NTD domain of one protomer. In the (lower) close-up view of NTD, the positions of the most common deletions—69/70, 144, and 242–244—are depicted as red beads. Residues shown in light blue are loops N1 (14-26), N3 (141-156), and N5 (246-260) that define the supersite for NTD-binding neutralizing antibodies (Cerutti et al., 2021). Since deletion sites are near or in the supersite, those deletions can alter the shape, hydrophobicity, and/or surface charge distribution of the supersite. These factors may perturb the binding of antibodies to NTD. The Spike structure shown here was modeled by Mansbach et al. (2021) based on the cryo-EM reconstruction from Walls et al. (2020) (PDB ID: 6VXX). Modeling was required because numerous regions were not resolved in the 6VXX structure, including loops N1, N3, and N5. Molecular visualization was prepared using VMD (Humphrey et al., 1996).

(E) The position of the 6-nucleotide, 3-codon deletion at SARS-CoV-2 genome positions 21,766–21,771 that causes most instances of the Spike Δ69/70 2-amino-acid deletion. Note that the third position of the original isoleucine codon "ATA" (I68) is replaced by the "C" that was originally the third base of the "GTC" codon encoding V70. The 6-base, out-of-frame nucleotide deletion translates to a 2-amino-acid in-frame deletion.

presence of both Δ69/70 and Δ144 can be attributed to their presence in the rapidly expanding B.1.1.7 lineage (Figure 1). Both deletions likely contribute to B.1.1.7's relative fitness and increasing prevalence in some regions.

There are also small, spatially localized clusters of distinct indels found in Spike that are rare but likely to be viable and transmitted, as they often are sampled multiple times (Figures 2B and S2). The most interesting of these clusters is in the region between Spike positions 137–148 (Figure S2A). While this Spike variable region is much less variable than the hypervariable regions of HIV-1, it shares some features with them: (1) the region overlaps with an exposed loop on Spike, the N3 loop (Chi et al., 2020); (2) there are many distinctive patterns of local deletions observed in this region—along with the very frequently observed Δ144 deletion, a spectrum of 24 other distinct deletion patterns are found among 341 different Spike sequences (Figure S2); and finally, (3) it is embedded in the NTD supersite, and so, like Δ144 (McCallum et al., 2021; Cerutti et al., 2021), the other deletions in this region are also likely to impact antibody resistance. There are also rare deletions that are near to or span the furin cleavage site (Figure S2A, positions 671–693); a deletion of the furin cleavage site augmented viral growth in culture but produced virus that was attenuated *in vivo* (Johnson et al., 2021).

A third deletion, ΔL242/A243/L244 (Δ242–244), is found in the B.1.351 lineage that has come to dominate the South African epidemic (Wibmer et al., 2021). This variant has a formidable

neutralization resistance profile, and Δ242–244 has been proposed to alter the loop structure and contact region for NTD-targeting neutralizing antibodies (Wibmer et al., 2021). The positions 242–244 are not themselves in a loop, but represent three hydrophobic residues at the end of a strand in a β-hairpin motif; deleting them would likely alter the N5 loop structure that connects the strands, a contact region for NTD-targeting neutralizing antibodies. Interestingly, Δ242–244 is found not only in the B.1.351 backbone but also in a small number of B.1.1.7 lineage sequences as well as in a few sequences with no additional Spike mutations (e.g., 3 from South Africa in Dec. 2020 and one from China in Feb. 2020; in all, 4 cases in the context of an ancestral form of Spike at position 614, D614). The ancestral D614 is currently (March 2021) very rarely sampled. D614G confers a fitness advantage in terms of transmissibility, and global samples had almost entirely shifted to the mutated form by early summer of 2020 (Korber et al., 2020b; Hou et al., 2020). Still, the D614G mutation may come at a cost for the virus, as some have found the ancestral D614 form to be more resistant to neutralization by sera. A 4- to 6-fold increase in vaccine sera sensitivity was observed for D614G in one study (Weissman et al., 2021), while in another, an average 1.7-fold difference was observed among sera from hamsters infected with D614 virus against the D614G variant (Plante et al., 2021). However, not all studies find such a difference. For example, Hou and colleagues did not see a significant difference between the two forms in terms of neutralization sensitivity to human convalescent sera (Hou et al., 2020). As the virus is increasingly confronted with convalescent and vaccine sera over the course of 2021, the greater neutralization sensitivity of the D614G form (if this is indeed the case) may come to outweigh its increased transmissibility as a selective force at the population level, and D614 may begin to re-emerge. Of note in this regard, the ancestral D614 is part of the Spike signature of the VOI in the A23.1 lineage that recently emerged from Uganda. D614 has also recently resurfaced in combination with Δ69/70 and with Δ144 (Figure 2C). A small number of interesting D614 sequences sampled in Wales and England carry both Δ69/70 and Δ144 (as of March 1, 2021, Figure 2) as well as a third distinctive 2-amino-acid deletion, Δ243–244. This could lead to concerted conformational changes in the NTD supersite region due to the spatial proximity of Δ243/244 and Δ69/70.

Spike deletions Δ69/70, Δ144, and Δ242-244 recur in multiple lineages, indicating that their indel boundary specificity may be biochemically favored by local RNA structure. For example, the Spike Δ69/70 2-amino-acid deletion is generally encoded, in the B.1.1.7 variant and in many other contexts, by a precise 6-base deletion that overlaps 3 codons encoding Spike amino acids 68–70 (Figure 2E). As noted above, such repeated precise mutations were also found in SHIV studies with specific HIV-1 Envs even in different infected hosts (Roark et al., 2021). In addition to spontaneous indel recurrence, recombination may contribute to indel movement through the population, enabling selection and increasing the frequencies of distinct variants, and variants-of-variants, that carry Δ60/70, Δ144, and Δ242–244 in Spike backbones (Giorgi et al., 2021; Varabyou et al., 2020).

More extensive indel patterns have been increasingly observed in recent regionally emerging variants through the spring of 2021. Some examples include a complex variant, increasingly sampled in Chile and spreading internationally, that carries a seven-amino-acid deletion at Spike Δ246–252 (previously Pango lineage B.1.1.1; now called C.37); a variant that is increasingly sampled in the Philippines that carries two deletions, Spike Δ141–143 and Δ243–244 (lineage P.3); a variant increasingly frequently sampled globally, first sampled in India (lineage B.1.617.2, a CDC-listed VOI), that carries a two-amino-acid deletion in a distinctive region, Spike Δ156–157; and a still very rare but particularly interesting variant in terms of indels that was first sampled in Russia (lineage AT.1) with a large nine-amino-acid deletion, Spike Δ136–144, and a four-amino-acid insertion at Spike 679, GIAL, very near the furin cleavage site.

LARGE-SCALE COMPARISONS OF ENV AND SPIKE STRUCTURES

The SARS-CoV-2 Spike protein and HIV-1 Envelope (Env) protein are Class I viral fusion glycoproteins (White et al., 2008) that are trimeric in both pre- and post-fusion states. Env is the smaller of the two. In its native state, it forms a heterodimeric trimer comprised of the gp120 and gp41 subunits. It has approximately 650 residues per protomer in the extracellular domain (ECD; Kwon et al., 2015) and a solvent-accessible surface area (SASA) of approximately 830 nm^2 (Figure 3). The Spike trimer has ~1,200 residues per protomer (Walls et al., 2020) in the ECD and a SASA of 1,250 nm^2 in the closed conformation. Both proteins undergo large-scale structural rearrangements. The Spike protein transitions from a closed to an open conformation with the upward movement of a single receptor binding domain (RBD; see Figure 3C; Wrobel et al., 2020); this increases the SASA to 1,300 nm^2. HIV-1 Env also undergoes substantial conformational change during the pre-fusion process of docking to a target cell. Upon Env binding to its primary receptor (CD4), the variable V1 and V2 loops move away from the Env apex, exposing the CCR5 co-receptor binding site. CCR5 binding triggers a conformational transition that enables the gp41 fusion machinery to access the target cell membrane and initiate fusion (Wang et al., 2016). Thus, both viral surface proteins exhibit significant conformational plasticity that protects the receptor-binding interface until the critical moment, enabling the preservation of high-affinity binding to host receptors in the face of immune pressure.

STRUCTURAL COMPARISONS OF THE GLYCAN SHIELDS OF SPIKE AND ENV

Both Spike and Env proteins are highly glycosylated, primarily with N-linked glycans, although both also include less well-characterized O-glycans (Shajahan et al., 2020; Silver et al., 2020). Glycans are extremely dynamic and are much more flexible than the underlying protein, so any single glycan can sample a large volume in space. Multiple glycans in combination become, in effect, a physical shield that blocks antibody access to the antigenic surface of the protein for both HIV-1 Env (Wei et al., 2003; Berndsen et al., 2020) and coronavirus Spikes (Walls et al., 2016; Watanabe et al., 2019). The SARS-CoV-2 Spike protein typically has 22 N-glycans per protomer, although N-linked glycosylation sites occasionally vary, giving a total of 63–69

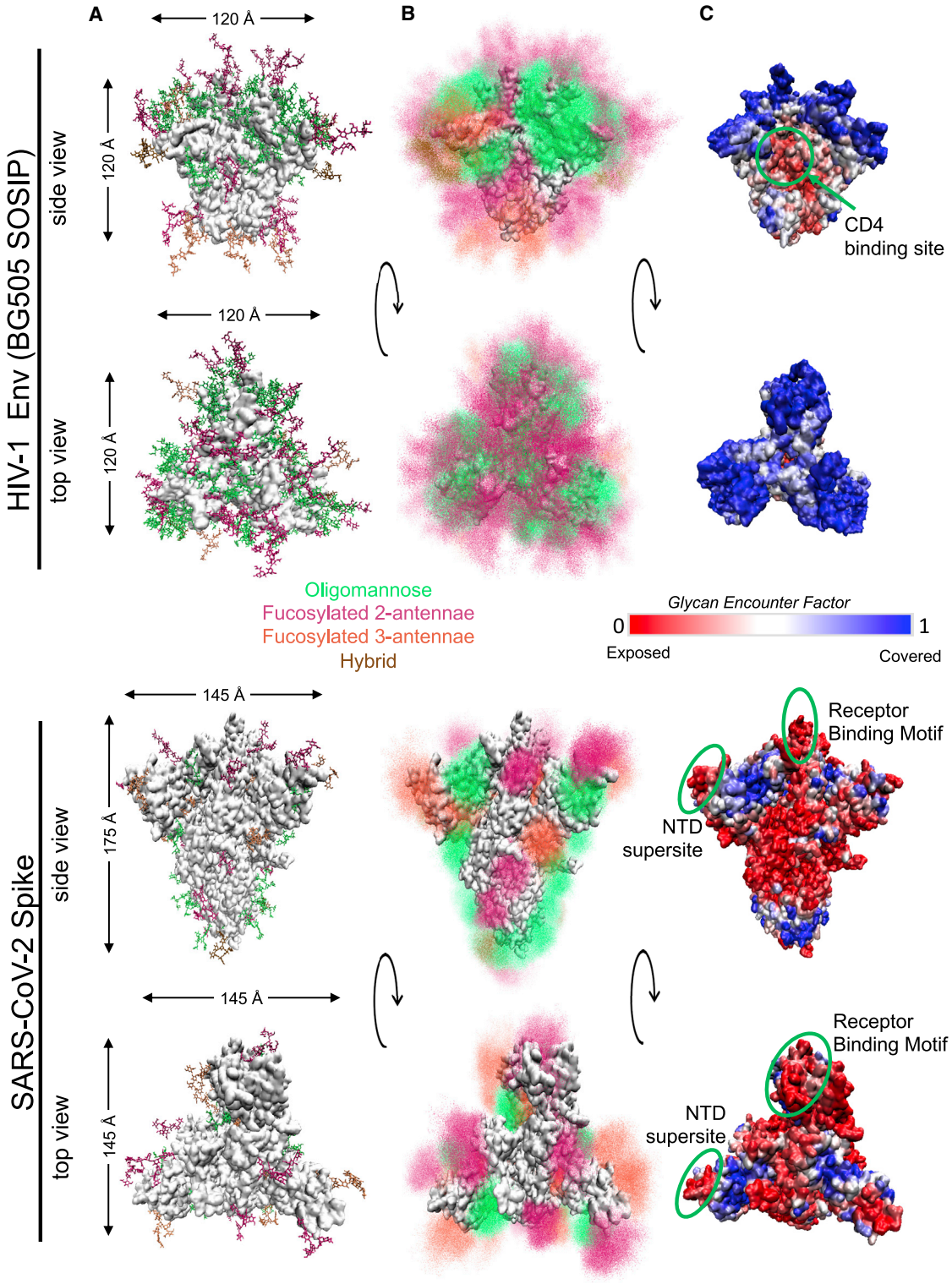


Figure 3. Structural comparison of HIV-1 Env and SARS-CoV-2 Spike glycoproteins

(A) Single conformation of Env (top) and Spike in 1-RBD-up state (bottom) from side and top view of trimers. Glycans are shown as stick representations and are colored by class (oligomannose, fucosylated 2-antennae, fucosylated 3-antennae and hybrid; see key given). Protein surface is shown in white. Protein sizes are to scale, with the maximum dimensions of the underlying protein represented by arrows.

(B) Ensemble picture of the dynamic glycan shield, including 500 different conformations of each glycan represented as point densities based on fraction of occupancy. Glycans are colored as in part (A).

(legend continued on next page)

glycans per trimer. Other coronavirus species have more glycans, e.g., 87 per trimer for the HCoV-HKU1 (a betacoronavirus, like SARS-CoV-2; Watanabe et al.2020b), and 102 for HCoV-NL63 (an alphacoronavirus; Walls et al., 2016), but there is little variation in glycan count within coronavirus species. In contrast, the number of potential N-linked glycosylation sites in HIV-1 Env varies throughout the course of every infection, with a typical range of between 23–32 per protomer (69-96 for the trimer). Site-specific mass spectrometry studies (Behrens et al., 2016; Watanabe et al., 2020a) indicate that, in both proteins, each glycosylation site is occupied by heterogeneous mixtures of glycoforms: high oligomannose, complex (with or without fucosylated cores and negatively charged sialic acid tips), and hybrids of the two. Relative glycoform frequencies depend on expression cell lines and their glycosylation enzyme repertoire (Goh and Ng, 2018). Both the chemical composition of individual glycans and of the amino acids in physical contact with them can affect the orientation of glycans and inter-glycan interactions (Chakraborty et al., 2020). All these factors affect the overall topology and the immunological protection conferred by the glycan shield.

For HIV-1, glycan shield evolution is an important immune evasion strategy. Glycans on the HIV-1 Env are concentrated on long, flexible loops, increasing their dynamic range and spatial coverage (Figure 3B). The glycan encounter factor (GEF) gives the probability of a probe's encountering a sugar's heavy atoms as it approaches the Env surface. This provides a metric to quantify the glycan coverage over the surface of the protein, illustrating how well the glycan shield can protect against approaching antibodies (Figure 3C; Chakraborty et al., 2020). Due to the dense and dynamic glycan coverage, the Env protein has high GEF across almost the entire antigenic surface, although the CD4 binding site where the Env interacts with the host receptor (Figure 3C) remains relatively exposed. In HIV-1 Env, certain glycan sites can sometimes shift by a few amino acids with important immunological consequences. For example, a shift of a glycan from position 332 to 334 can result in significant antibody resistance for broadly neutralizing antibodies (bNAbs) targeting V3-glycans (Bricault et al., 2019). HIV-1 bNAbs with breadth and potency favor long heavy-chain third complementarity-determining regions (HCDR3s) (Haynes et al., 2019), which enable these antibodies to reach through the glycan shield to target epitopes at the protein surface (Dashti et al., 2019). Since the glycan shield topology varies between HIV-1 viruses, glycan holes with low GEF regions may vary likewise, greatly affecting Env variant sensitivity to different antibody responses.

Due to the smaller number of glycans on a larger surface area, the density of glycosylation is much lower on the SARS-CoV-2 Spike than it is on HIV-1 Env (Figure 3). Since the Spike protein surface is less effectively shielded by glycan coverage relative to the Env surface, two regions that harbor critical neutralizing antibody epitopes, the RBD and the N-terminal Domain (NTD) supersite region, are relatively exposed. The RBD forms the functional interface between the Spike protein and the host ACE2 receptor and contains several mutations seen in the recent variants. The recurrent N501Y mutation alters specific interactions with ACE2 and may lead to increased binding affinity as well as enhanced infectivity (Leung et al., 2021; Rathnasinghe et al., 2021). When a single RBD is rotated up, effecting the change from "all-down" to "one-up" conformation that enables ACE-2 receptor binding (Wrapp et al., 2020), all-atom molecular dynamics simulations show that there is an accompanying transition in the glycan shield (Figure 3C): in the "one-up" conformation, the glycan coverage at the apex of the trimer in disappears (Mansbach et al., 2021). Of note, when the RBD is in the "one-up" conformation (Figure 3C) the amino acid at site 501 is exposed to solvent with no glycan coverage, not even by neighboring RBD glycans N331 and N343.

An N-terminal domain (NTD) "supersite" (Figure 3C) is recognized by several specific neutralizing antibodies. The supersite comprises Spike residues 14–20 (the N1 loop at the NTD terminus), 140–158 (the N3 loop, β-hairpin), and 245–264 (the N5 loop) (McCallum et al., 2021); it remains largely exposed with very low GEF. Mutations at or within contact distance of the supersite region can disrupt the structural motif or change the loop lengths, altering antibody recognition and binding. Intriguingly, there are also four potential glycosylation sites (asparagines N17, N74, N122, and N149) where glycans could potentially interact with antibodies bound to the NTD (Cerutti et al., 2021). The occasional loss of glycosylation motifs at these sites (386, 542, 28, and 280 times, respectively, out of a dataset of 519,035 Spike sequences sampled March 1, 2021) could affect antibody interactions. Of note, one of the multiple Spike mutations in the Brazilian P.1 variant (Toovey et al., 2021) introduces a novel N-linked glycosylation site, T20N. Thus, in SARS-CoV2, variation in glycosylation sites is just beginning to emerge. It remains to be seen whether these mutations are immunologically relevant or whether they will become increasingly epidemiologically relevant.

RECOMBINATION

Recombination is an important evolutionary mechanism for RNA viruses, including retroviruses like HIV-1 and coronaviruses like SARS-CoV-2. Natural recombination can occur when two distinctive viral variants co-infect the same host, and has the potential to accelerate evolution by bringing together advantageous mutations that arose independently. It can be challenging to detect recombination in situations of low diversity like SARS-CoV-2, and recombination can arise *in vitro* or as methodological artifacts of sequence assembly, thus a clear understanding of the role of recombination in viral evolution can be complicated by both false negatives and false positives. In this section we briefly describe the major role of recombination in HIV-1 in terms of global diversity and within-host evolution, as well as the role of recombination in SARS-CoV-2's origins. We also provide an illustration of likely recombination events among locally co-circulating SARS-CoV-2 variants in S. Africa.

(C) Glycan Encounter Factor represented as a color map on the surface of the Env (top) and Spike (bottom) proteins. Blue indicates high glycan shielding and red indicates regions of relatively high shield vulnerability. The CD4 binding site of Env and the receptor binding motif and NTD supersite of the Spike protein are marked by green circles.

Recombination in HIV-1

Recombination has played a critical role in HIV-1 evolution (Zhang et al., 2010). HIV-1 nomenclature (Robertson et al., 2000) recognizes both major clades, specified A–K, and over 100 circulating recombinant forms (CRFs) (characterized and listed at the Los Alamos HIV database, http://www.hiv.lanl.gov). Such inter-subtype recombination events are readily detected by sequence analysis; within-subtype recombination is more challenging to resolve, but can still be identified (Kiwelu et al., 2013; Nikolaitchik et al., 2015). Recombination is also a very important evolutionary mechanism over the course of a natural HIV-1 infection within a single individual (Shriner et al., 2004, Song et al., 2018). A bioinformatic tool developed to track within-subject recombination in the low-diversity setting of HIV early infection (RAPR, Song et al., 2018) can also be usefully applied in the low-diversity setting of SARS-CoV-2 in the COVID-19 pandemic.

Recombination in SARS-CoV-2

Coronavirus infections are frequent and widespread across different animal reservoirs, where distinct viruses may coexist in the same hosts and often recombine (Denison et al., 2011; Su et al., 2016). At high multiplicities of infection, more than 25% of viral progeny may be recombinant (Baric et al., 1990). Recombination is an important element of coronavirus evolution, can be observed even between different coronavirus families, and has been implicated in the origin of SARS-1, MERS, and SARS-CoV-2 (Sabir et al., 2016; Li et al., 2020a; Lam et al., 2020; Hon et al., 2008). In low-diversity settings, such as the first year of the SARS-CoV-2 pandemic, many standard bioinformatic strategies for detecting recombination will be insufficiently sensitive (this includes, e.g., strategies developed to detect recombination between major HIV-1 clades). Nevertheless, several studies have found occurrences of recombination among SARS-CoV-2 pandemic variants (De Maio et al., 2020; Korber et al., 2020a; Varabyou et al., 2020). Varabyou and colleagues found evidence of recombination in SARS-CoV-2 based on major clades and their defining mutations. By using this method to screen the full GISAID database (https://gisaid.org/), they found hundreds of instances of likely recombinants, some of which persisted in the population. They could demonstrate that at least some of these recombinants were not the result of sequencing from mixed infections, and that some were parts of transmitted lineages (Varabyou et al., 2020).

Using the RAPR tool, which was designed specifically for low-diversity settings (Song et al., 2018), we find strong evidence of recombination among geographically regional sets of SARS-CoV-2 sequences. RAPR uses the full set of variable positions in its analysis, not just major clade defining positions, which may enhance sensitivity in some cases, but it is computationally intensive (as it compares all possible sequence triplets), which limits its use to fairly small datasets. Here we present, as an illustration, four examples of likely recombination from sequences recently sampled in South Africa (Figure 4). Three of these recombination events involve B.1.351 variants (Wibmer et al., 2021) recombining with viruses outside of that lineage. A careful review of the original sequencing data from each of these four examples found no indication that the recombinants were from individuals with mixed infection, and the relevant base calls were well supported. While recurring mutations are a plausible explanation of the frequent observation of identical SNPs in different lineages (e.g., E484K and N501Y, Figure 1, and the recurrent deletions discussed above), the observed patterns of a series of clustered mutations presented here are more parsimoniously explained by recombination events that grafted sets of mutations from co-circulating forms onto the backbone of existing variants in regions with high rates of infections.

TRANSITIONS IN GLOBAL DIVERSITY

The level of sequence diversity differs conspicuously between SARS-CoV-2 and HIV-1 (Figures 1 and S1). Outside of closely related transmission chains, within-subtype Hamming distances between Env proteins typically range from 10% to 25% with between-subtype differences of 20% to 40%. Within each single host, HIV evolves at a rate approaching 1% per year (Krakoff et al., 2019). The baseline form of the most divergent lineage of SARS-CoV-2 characterized to date, P.1, carries only 12 changes in the 1,273 amino acids of the Spike protein, less than 1% (Toovey et al., 2021). But a comparison of the global transitions in variant frequencies over time merits consideration, as for both viruses, variants and global diversity will shape the future success of vaccines. A general feature of the HIV-1 epidemic is the gradual increase in diversity within clades in local geographic populations; these changes are accompanied by greater levels of resistance to sera derived from natural infection (Hraber et al., 2014; Rademeyer et al., 2016).

To illustrate the degree of large-scale change in the HIV-1 pandemic over time, we compare the global subtype and CRF distribution between two 6-year windows, 2000–2005 and 2015–2020 (Figure 5A). A striking feature of this analysis is the relatively consistent frequency of sampling of different subtypes in different major geographic regions (Bbosa et al., 2019). This consistency, which is likely a consequence of the much lower transmission rate of sexually transmitted, as opposed to respiratory, pathogens, may enable regional deployment of subtype-specific immunological strategies for prevention if they can be successfully developed. The C clade continues to dominate in Southern Africa and India. The Western Hemisphere remains a predominantly B-clade epidemic, with the C clade just beginning to make inroads in North America. The A6 sub-lineage of the A clade continues to be the most common form in Russia and the former Soviet Union. CRF02, of an AG recombinant origin, continues to dominate in West and Central Africa, but co-exists with a very diverse viral population.

In other regions of the world, subtype distributions of sampled sequences are gradually changing. The United Kingdom and Europe have gotten more diverse over time (UK Collaborative Group on HIV Drug Resistance, 2014). In Uganda, there has been a shift from D to A clade (Figure 5A), and recombinants are increasing in prevalence (Grant et al., 2020); such an increase in the number of recombinant forms is common in regions with complex HIV-1 epidemics (Hemelaar et al., 2019). Since standard pseudovirus panels typically include only major clades and CRF01 and CRF02 (Bricault et al., 2019), an increased prevalence of recombinants complicates the extrapolation of laboratory studies of antibody neutralization breadth to real-world global diversity. Previously under-sampled regions can reveal unexpected patterns of

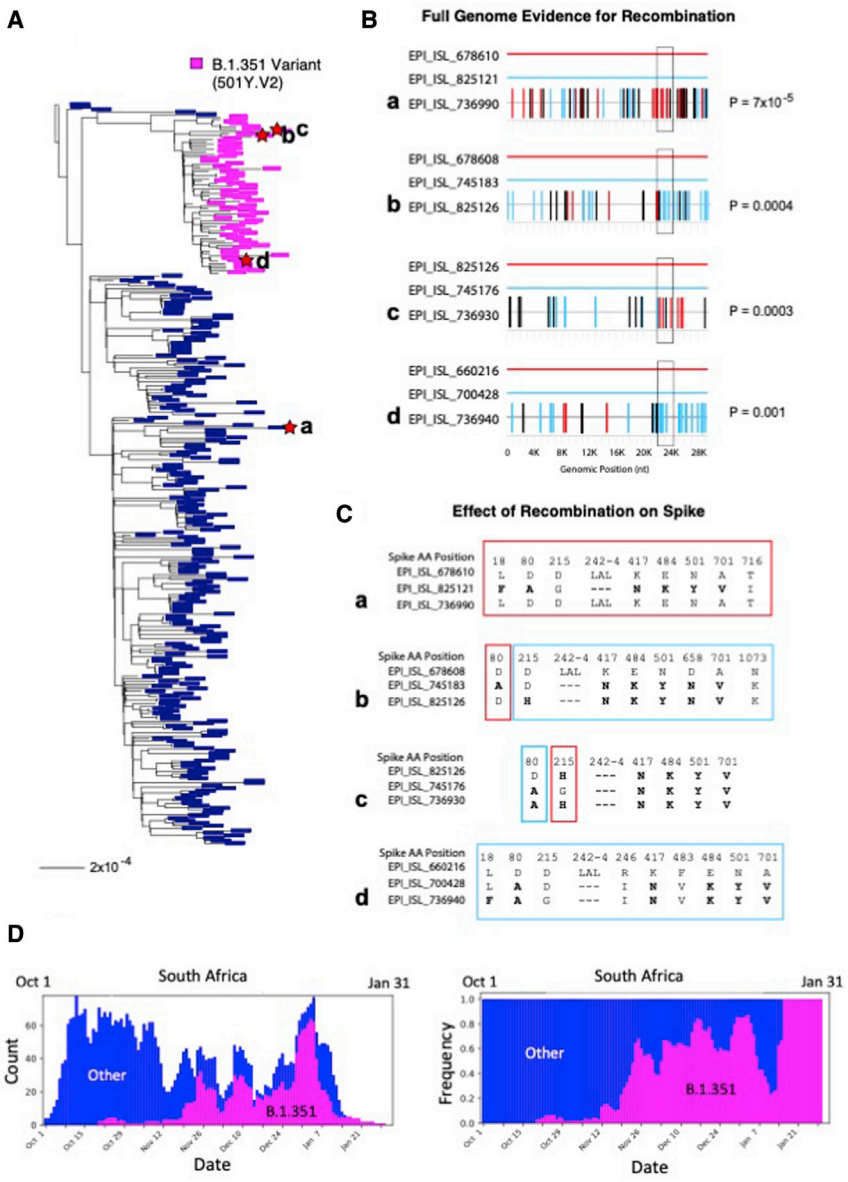


Figure 4. Phylogenetic tree and recombinant triplets from South Africa

(A) Phylogenetic tree of 298 SARS-CoV-2 sequences sampled in South Africa from 10/01/2020 to 01/31/2021. Sequences bearing the set of Spike mutations L18F, D80A, D215G, Δ242-244, K417N, E848K, N501Y, D614G, and A701V, characteristic of the most common form of Spike in the B.1.351 lineage, are labeled in magenta; all other regional variants are labeled in blue. Lowercase letters a though d mark the 4 recombinants shown in the right panels, and red stars indicate the recombinant leaves on the tree.

(B) Each graph represents a recombinant triplet. The full genome of each parental strain is shown as a solid line, one in red and one in light blue, and the recombinant is shown below with mutations marked in either light blue or red, according to the parental strain they match, or black if they match neither parent. The Spike gene is demarcated with a black box. Recombination p values, calculated via the Runs Test statistic (obtained using the tool RAPR; Song et al., 2018), are shown to the left of each graph. The top graph (recombinant a) shows the strongest recombination signal detected in the full alignment ($p = 7 \times 10^{-5}$); however, while the parental strain in light blue is a B.1.351 variant, the recombinant is not. The other three recombinants (b through d) are all B.1.351 variants.

(C) Each graph shows the Spike positions and corresponding amino acid at which the triplets shown in (B) differ. Color-coded boxes are either blue or red depending on which parental strain the recombinant matches at those position(s). Mutations typical of the B.1.351 variant are highlighted in bold. Note that, given sampling limitations, the sequences identified by RAPR are not expected to be the precise parents and child giving rise to the recombinant; rather, each member of the triplet represents a lineage to which the true parents and child belong.

(D) The time period used to identify examples of likely recombination between co-circulating strains was selected to be 10/1/2020 through 1/31/21 because during this period B.1.351 came to dominate the South African epidemic and the B.1.351 variant was co-circulating with other natural variants, providing an opportunity for detectable natural recombination to arise. Weekly average counts of sampling of B.1.351 (magenta) relative to other variants (dark blue) during this study period are shown on the left; the same data is plotted as sampling frequencies on the right. B.1.351 was initially rare, but came to dominate the South African epidemic during this 3-month time frame.

change. For example, a CRF02/B recombinant form was found to be commonly circulating in Nigeria, an unexpected finding since B clade variants have rarely been sampled in Africa (Billings et al., 2019). One of the more unfortunate recent trends evident in Figure 5 is that CRF01 viruses, which once had a more limited distribution focused in Southeast Asia, are now more commonly sampled in China and Australia. Due to glycan shifts, as discussed above, CRF01 viruses are almost completely insensitive to V3-glycan-targeting bNAbs, a key focus of current vaccine design efforts.

In contrast to the slow transitions in HIV subtypes and diversification over decades, SARS-CoV-2 variants that carry an advantageous set of mutations can move very globally swiftly and effect near-total turnover of local populations on the timescale of a few months (Figures 5B and 5C). If this occurs repeatedly without lineage extinction, a star-like phylogeny results; in contrast, if less-fit lineages are repeatedly driven to very low levels, a ladder-like tree is expected, as with influenza. A visual signature of this phenomenon is present in Figures 1 and S1, in which rapidly expanding lineages have reduced background mutations. Rapid lineage expansion in SARS-CoV-2 was first observed as the G clade rapidly replaced the ancestral virus that had initially seeded the global pandemic in the spring of 2020 (Korber et al., 2020a). By the autumn of 2020, the viruses carrying the ancestral D614 form of the virus were very rarely seen (Figure 5B). Similarly, the B.1.1.7 virus, first detected in the UK in late September 2020, within a matter of months had replaced the G and GV clade variants as the regionally prevalent form (Figure 5C) and had begun its global spread. Unlike the recent scenario in the UK, however, other forms of the virus with different combinations of advantageous mutations are concurrently circulating in other geographic regions, so multiple high-fitness variants are co-circulating in the same geographic regions. Thus, as the B.1.1.7 virus has spread

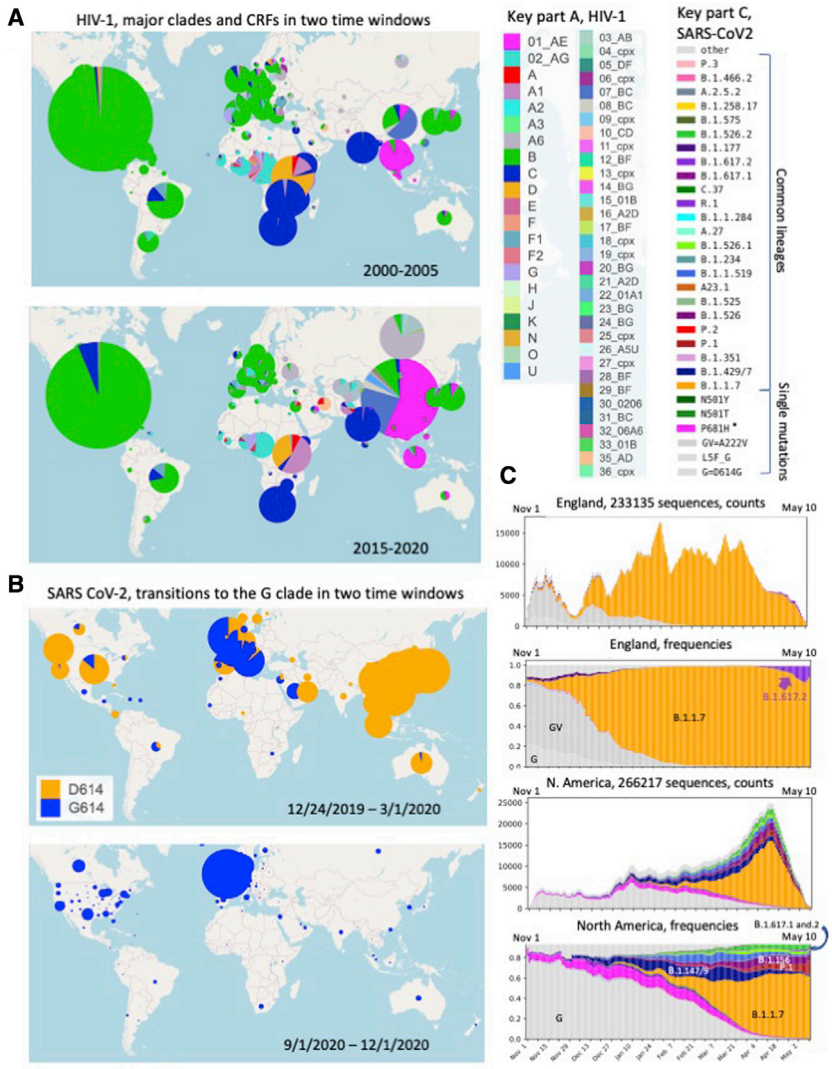


Figure 5. A comparison of transition patterns in major clades

(A) Major HIV-1 clades and CRF sampling frequencies in two 6-year windows: 2000–2005 and 2015–2020. The circle area reflects the relative number of sequences available from a given region within each map.

(B) Frequency of sampling of the SARS-CoV-2 G clade (carrying D614G) and its descendants (shown in blue) versus the frequency of sampling of the ancestral form of the virus that carried D614 (shown in orange) in two time-windows, roughly the first 10 weeks of the pandemic (through March 1, 2020), and the last 3 months of 2020.

(C) The top two graphs show the frequency of sampling of different variant forms in the United Kingdom between November 1, 2020 and May 10, 2021. In the fall, the G clade (light gray), and the GV clade (the G clade with an additional A222V mutation (darker gray) were co-circulating, with a gradual relative increase in the GV clade relative to G clade over the summer and fall. B.1.1.7 (orange) was first sampled in September, and rapidly increased in prevalence in the UK, comparable to the global transitions we found when the G clade became globally dominant (Korber et al., 2020b). In the spring of 2021, B.1.627.2, initially sampled in India, had begun to rise significantly in frequency in the UK. In this evolutionary pattern, one form gave way successively to another: G to GV to B.1.1.7. Currently B.1.617.2 has begun to be increasingly sampled; over the next few months we will learn if B.1.617.2 continues in this upward trajectory in the UK and elsewhere. The same data are plotted two ways: weekly average tallies of each form, to give a sense of sampling, and weekly average frequencies. Below, the same data is plotted for North America. The G clade is dominant in the fall. G clade forms which carried additional mutations near the furin cleavage site (magenta and purple) became increasingly frequently sampled, but then gave way to variants with more complex forms of Spike, which often still carried a positive charge near the furin cleavage site. When the B.1.1.7 variant began to be sampled in early December, there are already distinct forms with an established presence and relative fitness advantages co-circulating, and VOI/VOCs first sampled from California, Brazil, and New York all had a significant presence. Still, B.1.1.7 has been increasingly sampled throughout North America, although P.1 and B.1.526 are also continuing to maintain or increase in frequency in some regions states in the USA. As of early May, 2021, B.1.617.2 is still rare but present and increasing in frequency in North America.

globally, it has been introduced into communities with complex epidemics, and although the increase in prevalence of B.1.1.7 has been observed in most regions of the world with recent sampling, some variants in local regions may be persisting. Current North American diversity provides an example of this (Figure 5C); all VOIs currently being tracked by GISAID have a presence in North America (https://www.gisaid.org/hcov19-variants/), and while B.1.147 and B.1.149 are diminishing, others, including P.1 and B.1.562, are persisting. One of several CDC-designated B.1.617-related VOIs that were originally detected in India, B.1.617.2, has begun to be increasingly sampled globally, is rapidly increasing in prevalence in England, and also has a presence in North America (Figure 5C). So although B.1.1.7 is the currently the dominant form in much of the northern hemisphere, it remains to be seen if B.1.1.7, B.1.617.2, yet another VOC/VOI, a recombinant, or further mutated descendants of a currently circulating VOC/VOI will emerge from this complex milieu to become the globally dominant form over the coming year.

DISCUSSION

As we have highlighted throughout this review, there are some similarities between SARS-CoV-2 and HIV-1, but there are key differences as well. Both are enveloped RNA viruses, both are animal viruses that crossed into humans, and both gave rise to pandemics. HIV-1 is primarily sexually transmitted and took many decades to acquire a global presence, whereas SARS-CoV-2 is a respiratory infection that became a global pandemic within months of its initial detection. Both viruses evolve using insertions, deletions, and recombination in addition to base substitution. Both viruses evolve under immune pressure and have circulating variants with mutations in key epitope regions that confer relative resistance to neutralizing antibodies; indels are important for the evolution of antibody resistance in both HIV-1 and SARS-CoV-2. Both viruses have heavily glycosylated receptor-binding surface proteins that enable entry into host cells, but the glycan shield conferred by HIV-1 is far denser. A fundamental

biological difference is that HIV-1 is a retrovirus and its genetic material can be harbored in latently infected cells, making it very difficult to achieve virological cure. As a chronic infection, HIV-1 continues to evolve under immune pressure in every infected individual. In contrast, SARS-CoV-2 infections are typically soon cleared, although rare chronic cases of COVID-19 may be contributing to the more extensively mutated variants of interest and concern. Vaccines for HIV-1 are very challenging, in part because of the difficulty of inducing antibodies that can penetrate the glycan shield and in part because of the tremendous diversity of the virus. Vaccines for SARS-CoV-2 were enabled by the relative accessibility of the key epitopes for neutralizing antibodies in the RBD and the NTD supersite and by the limited variability of these epitope regions in the initial phases of the pandemic. As further variation in these regions continues to emerge, it will be critical to document the impact of arising mutations and to remain agile in our response to COVID-19.

In the spring of 2020, SARS-CoV-2 was advancing through an immunologically naive population and swiftly spreading in the context of its new human host. In such a scenario, it was reasonable to suppose that enhanced infectivity and transmissibility would confer a primary selective advantage, and this indeed proved to be the case. There was a repeated and very rapid shift in prevalence to G-clade viruses (which carried the D614G mutation) essentially whenever that variant entered a new geographic region, even if that region had an ongoing, well-established, ancestral Wuhan-variant epidemic (Korber et al., 2020b). G-clade viruses were found to be more infectious in pseudotype assays, were associated with higher viral loads in the upper respiratory tract (Korber et al., 2020b), and were shown to be more infectious in laboratory animals (Hou et al., 2020).

What changed over the course of 2020 is that the virus began to encounter and propagate through populations with varying levels of immunity from prior exposure. Under these conditions, immunological resistance has greater potential to be favored as a force for positive selection; widespread vaccination will further increase such selective pressure during 2021. The G-clade viruses may be somewhat more susceptible to serum neutralization (Weissman et al., 2021; Plante et al., 2021), and selection for antibody resistance may in the future counter-select for viruses with the ancestral D614 form. Some early evidence for this is that the A.23.1 viral lineage, which was increasing in prevalence in Central Africa (Bugembe et al., 2021), carries the ancestral D614 form. Many new VOIs and VOCs have begun to emerge that simultaneously carry both multiple neutralizing antibody resistance mutations and enhanced-infectivity mutations, and many of the mutations that phenotypically benefit the virus are being resampled concurrently in different lineages (Figures 1, 2, and 4). This suggests that the virus may be exploring and re-exploring a favored mutational landscape within the context of currently circulating forms. Thus, understanding and defining recurrent mutation events may serve to guide us as we prepare for the possibility of second-generation vaccine designs to contend with growing viral diversity. By the summer of 2021, the virus will be moving through mostly-vaccinated populations in some countries and through populations increasingly enriched for recovered individuals, and so the evolutionary pressures driving selection may once again be altered.

The future course of the SARS-CoV-2 pandemic may well be set by the events surrounding large-scale vaccine rollout in the first and second quarter of 2021. Although complete viral eradication is unlikely, there are different possible modes of viral persistence with different implications for future vaccine control. In influenza, selection for resistance and seasonal bottlenecking give rise to a ladder-like tree topology: there is extensive diversity over many years but relatively little in any one season. In HIV-1, long-term persistent infection and co-evolution with immune responses produce a "bushy" tree with the simultaneous and temporally extended epidemiological presence of extremely diverse viral lineages. Forcing case counts to low levels (as in the influenza seasonal bottleneck) reduces the variation available for viral evolution. If this were achieved in the current pandemic, SARS-CoV-2 might be driven to an influenza-like evolutionary trajectory. The developing phylogeny would then have a more ladder-like topology, with population-level immunity influencing shifts in dominant variants over time. In this case, a strategy with periodically updated vaccines specifically targeting the currently circulating variants may suffice for continuing vaccine protection. If, on the other hand, a wide range of variants continues to circulate, diversify, and recombine, the eventual result could be simultaneous and continuous circulation of various phylogenetically and immunologically distinctive variants, and possibly emerging recombinants between them. This situation—more reminiscent of HIV-1 than of influenza—would present a different kind of challenge for vaccines, and might require vaccines to be designed to induce broad responses specifically addressing frequently sampled antibody-resistance mutations that arise in multiple lineages.

Several highly efficacious COVID-19 vaccines were deployed within a year of the emergence of SARS-CoV-2 (Richman, 2021; Tumban, 2020), whereas 40 years of research has failed to produce any comparable vaccine for HIV-1. To begin to understand this discrepancy, one only has to look at Figures 1 and 3. In HIV-1, the key vaccine-elicited antibody epitopes are diverse at the sequence level, and they are well-protected by a dense, highly variable, and dynamic glycan shield. These challenges require innovative and complex strategies to ultimately enable the design of an effective HIV-1 vaccine that can achieve broadly neutralizing antibody induction. In contrast, the SARS-CoV-2 key epitope regions have been comparatively slow in accumulating small numbers of mutations (Figure 1), and key neutralizing epitope regions are exposed and hence vulnerable to antibodies (Figure 3). Spike vaccine antibody responses are very potent and target multiple epitopes, and to date they have generally been resilient and able to offer protection against variants with modest numbers of mutations. The SARS-CoV-2 vaccines may also elicit cross-reactive T cell responses; known SARS-CoV-2 T cell epitopes are highly conserved (Tarke et al., 2021; Redd et al., 2021). The impact of such responses is still being determined. Rhesus macaques (RMs) are naturally resistant to severe disease, and one CD4+ or CD8+ T cell depletion experiment in RMS only slightly prolonged recovery from infection and did not impact re-infection (Hasenkrug et al., 2021), while another found that CD8+ T cell depletion of convalescent macaques partially abrogated protective

immunity against rechallenge (McMahan et al., 2021). In the more diverse HIV-1, many T cell epitopes are also highly variable, and relatively few vaccine-elicited responses to full HIV-1 proteins are likely to cross-react with circulating variants (Korber and Fischer, 2020). As variants shift in prevalence and the SARS-CoV-2 pandemic takes on new forms, our capacity to track and test these variants and to use past mutational patterns to anticipate the future may enable us to keep up with our viral foe's evolutionary twists and turns.

CONSORTIA

The members of the Network for Genomic Surveillance in South Africa (NGS-SA) are Eduan Wilkinson, Nokukhanya Msomi, Arash Iranzadeh, Vagner Fonseca, Deelan Doolabh, Emmanuel James San, Koleka Mlisana, Anne von Gottberg, Sibongile Walaza, Mushal Allam, Arshad Ismail, Thabo Mohale, Allison J Glass, Susan Engelbrecht, Gert Van Zyl, Wolfgang Preiser, Francesco Petruccione, Alex Sigal, Diana Hardie, Gert Marais, Marvin Hsiao, Stephen Korsman, Mary-Ann Davies, Lynn Tyers, Innocent Mudau, Denis York, Caroline Maslo, Dominique Goedhals, Shareef Abrahams, Oluwakemi Laguda-Akingba, Arghavan Alisoltani-Dehkordi, Adam Godzik, Constantinos Kurt Wibmer, Bryan Trevor Sewell, Jose Lourenco, Sergei L Kosakovsky Pond, Steven Weaver, Marta Giovanetti, Luiz Carlos Junior Alcantara, Darren Martin, Jinal N Bhiman, and Carolyn Williamson.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2021.05.012.

ACKNOWLEDGMENTS

This work was supported by Los Alamos National Laboratory (LANL) LDRD project 20200554ECR and by LANL Technology Innovation funds, as well as through the NIH NIAID, DHHS Interagency Agreement R-00441015-0/AAI12007. We thank the staff at GISAID for kindly supporting our efforts at cov.lanl.gov and the many groups throughout the world that provide the global SARS-CoV-2 viral sequence data. We acknowledge specifcally the following laboratory groups whose work we highlighted in the text: the MRC/UVRI & LSHTM Uganda Research Unit of the Uganda Virus Research Institute, the Rwanda National Reference Laboratory, the Laboratorio de Ecologia de Doencas Transmissiveis na Amazonia (Instituto Leonidas e Maria Deane), CDL Laboratorio Santos e Vidal LTDA, the Laboratorio de Referencia Nacional de Virus Respiratorio (Instituto Nacional de Salud Peru), Fulgent Genetics, KwaZulu-Natal – National Health Laboratory Service (Inkosi Albert Luthuli Central Hospital, Durban), and the New York City Pandemic Response Laboratory, as well as the many other dedicated sequencing groups whose efforts enable the global tracking of emergent variants. Special thanks to Duncan McBranch and Joseph "Pat" Fitch for their leadership of the COVID pandemic response at Los Alamos National Laboratory.

DECLARATION OF INTERESTS

B.K., W.F., J.T., T.B., and S.G. have provisional patents and patents relating to vaccine design to address viral diversity as applied to HIV-1 and/or SARS-CoV-2.

REFERENCES

Arendrup, M., Nielsen, C., Hansen, J.E., Pedersen, C., Mathiesen, L., and Nielsen, J.O. (1992). Autologous HIV-1 neutralizing antibodies: emergence of neutralization-resistant escape virus and subsequent development of escape virus neutralizing antibodies. J. Acquir. Immune Defic. Syndr. (1988) *5*, 303–307.

Avanzato, V.A., Matson, M.J., Seifert, S.N., Pryce, R., Williamson, B.N., Anzick, S.L., Barbian, K., Judson, S.D., Fischer, E.R., Martens, C., et al. (2020). Case study: Prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell *183*, 1901–1912.e9.

Baang, J.H., Smith, C., Mirabelli, C., Valesano, A.L., Manthei, D.M., Bachman, M.A., Wobus, C.E., Adams, M., Washer, L., Martin, E.T., and Lauring, A.S. (2021). Prolonged severe acute respiratory syndrome coronavirus 2 replication in an immunocompromised patient. J. Infect. Dis. *223*, 23–27.

Bar, K.J., Tsao, C.Y., Iyer, S.S., Decker, J.M., Yang, Y., Bonsignori, M., Chen, X., Hwang, K.K., Montefiori, D.C., Liao, H.X., et al. (2012). Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. PLoS Pathog. *8*, e1002721.

Baric, R.S., Fu, K., Schaad, M.C., and Stohlman, S.A. (1990). Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. Virology *177*, 646–656.

Bartolini, B., Rueca, M., Gruber, C.E.M., Messina, F., Giombini, E., Ippolito, G., Capobianchi, M.R., and Di Caro, A. (2020). The newly introduced SARS-CoV-2 variant A222V is rapidly spreading in Lazio region, Italy. medRxiv. https://doi.org/10.1101/2020.11.28.20237016.

Bbosa, N., Kaleebu, P., and Ssemwanga, D. (2019). HIV subtype diversity worldwide. Curr. Opin. HIV AIDS *14*, 153–160.

Behrens, A.J., Vasiljevic, S., Pritchard, L.K., Harvey, D.J., Andev, R.S., Krumm, S.A., Struwe, W.B., Cupo, A., Kumar, A., Zitzmann, N., et al. (2016). Composition and antigenic effects of individual glycan sites of a trimeric HIV-1 Envelope glycoprotein. Cell Rep. *14*, 2695–2706.

Berndsen, Z.T., Chakraborty, S., Wang, X., Cottrell, C.A., Torres, J.L., Diedrich, J.K., López, C.A., Yates, J.R., 3rd, van Gils, M.J., Paulson, J.C., et al. (2020). Visualization of the HIV-1 Env glycan shield across scales. Proc. Natl. Acad. Sci. USA *117*, 28014–28025.

Bhiman, J.N., Anthony, C., Doria-Rose, N.A., Karimanzira, O., Schramm, C.A., Khoza, T., Kitchin, D., Botha, G., Gorman, J., Garrett, N.J., et al. (2015). Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. Nat. Med. *21*, 1332–1336.

Billings, E., Kijak, G.H., Sanders-Buell, E., Ndembi, N., O'Sullivan, A.M., Adebajo, S., Kokogho, A., Milazzo, M., Lombardi, K., Baral, S., et al.; MHRP Viral Sequencing Core and the TRUST/RV368 Study Group (2019). New subtype B containing HIV-1 circulating recombinant of sub-Saharan Africa origin in Nigerian men who have sex with men. J. Acquir. Immune Defic. Syndr. *81*, 578–584.

Bonsignori, M., Kreider, E.F., Fera, D., Meyerhoff, R.R., Bradley, T., Wiehe, K., Alam, S.M., Aussedat, B., Walkowicz, W.E., Hwang, K.K., et al. (2017). Staged induction of HIV-1 glycan-dependent broadly neutralizing antibodies. Sci. Transl. Med. *9*, eaai7514.

Bonsignori, M., Zhou, T., Sheng, Z., Chen, L., Gao, F., Joyce, M.G., Ozorowski, G., Chuang, G.Y., Schramm, C.A., Wiehe, K., et al.; NISC Comparative Sequencing Program (2016). Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. Cell *165*, 449–463.

Boutwell, C.L., Rolland, M.M., Herbeck, J.T., Mullins, J.I., and Allen, T.M. (2010). Viral evolution and escape during acute HIV-1 infection. J. Infect. Dis. *202* (*Suppl 2*), S309–S314.

Bouvet, M., Imbert, I., Subissi, L., Gluais, L., Canard, B., and Decroly, E. (2012). RNA 3¢-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proc. Natl. Acad. Sci. USA *109*, 9372–9377.

Bricault, C.A., Yusim, K., Seaman, M.S., Yoon, H., Theiler, J., Giorgi, E.E., Wagh, K., Theiler, M., Hraber, P., Macke, J.P., et al. (2019). HIV-1 neutralizing antibody signatures and application to epitope-targeted vaccine design. Cell Host Microbe *26*, 296.

Bugembe, D.L., Phan, M.V.T., Ssewanyana, I., Semanda, P., Nansumba, H., Dhaala, B., Nabadda, S., O'Toole, Á.N., Rambaut, A., Kaleebu, P., and Cotten, M. (2021). A SARS-CoV-2 lineage A variant (A.23.1) with altered spike has emerged and is dominating the current Uganda epidemic. medRxiv, 2021.02.08.21251393.

Burnie, J., and Guzzo, C. (2019). The Incorporation of Host Proteins into the External HIV-1 Envelope. Viruses *11*, 85.

Centers for Disease Control (CDC) (1981). Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men–New York City and California. MMWR Morb. Mortal. Wkly. Rep. *30*, 305–308.

Cerutti, G., Guo, Y., Zhou, T., Gorman, J., Lee, M., Rapp, M., Reddem, E.R., Yu, J., Bahna, F., Bimela, J., Huang, Y., Katsamba, P.S., Liu, L., Nair, M.S., Rawi, R., Olia, A.S., Wang, P., Chuang, G.Y., Ho, D.D., Sheng, Z., Kwong, P.D., and Shapiro, L. (2021). Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. bioRxiv, 2021.01.10.426120.

Chakraborty, S., Berndsen, Z.T., Hengartner, N.W., Korber, B.T., Ward, A.B., and Gnanakaran, S. (2020). Quantification of the resilience and vulnerability of HIV-1 native glycan shield at atomistic detail. iScience *23*, 101836.

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y., et al. (2020). A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. Science *369*, 650–655.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Kuo, H.H., Boucau, J., Bowman, K., et al. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. N. Engl. J. Med. *383*, 2291–2293.

Cui, J., Li, F., and Shi, Z.L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. *17*, 181–192.

Dashti, A., DeVico, A.L., Lewis, G.K., and Sajadi, M.M. (2019). Broadly neutralizing antibodies against HIV: Back to blood. Trends Mol. Med. *25*, 228–240.

Davies, N.G., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J., Pearson, C.A.B., Russell, T.W., Tully, D.C., Abbott, S., Gimma, A., Waites, W., Wong, K.L.M., van Zandvoort, K., Eggo, R.M., Funk, S., Jit, M., Atkins, K.E., and Edmunds, W.J. (2020). Estimated transmissibility and severity of novel SARS-CoV-2 variant of concern 202012/01 in England. medRxiv, 2020.12.24.20248822.

De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., and Goldman, N. (2020). Issues with SARS-CoV-2 sequencing data. *https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473*.

Deng, X., Garcia-Knight, M.A., Khalid, M.M., Servellita, V., Wang, C., Morris, M.K., Sotomayor-González, A., Glasner, D.R., Reyes, K.R., Gliwa, A.S., et al. (2021). Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. medRxiv, 2021.03.07.21252647.

Denison, M.R., Graham, R.L., Donaldson, E.F., Eckerle, L.D., and Baric, R.S. (2011). Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol. *8*, 270–279.

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

Donnelly, C.A., Malik, M.R., Elkholy, A., Cauchemez, S., and Van Kerkhove, M.D. (2019). Worldwide Reduction in MERS Cases and Deaths since 2016. Emerg. Infect. Dis. *25*, 1758–1760.

van Dorp, L., Tan, C.C., Lam, S.D., Richard, D., Owen, C., Berchtold, D., Orengo, C., and Balloux, F. (2020). Recurrent mutations in SARS-CoV-2 genomes isolated from mink point to rapid host-adaptation. bioRxiv. https://doi.org/10.1101/2020.11.16.384743.

Dumiak, M. (2020). The race is on. IAVI Rep. *24*.

European Centre for Disease Prevention and Control (2020). Detection of new SARS-CoV-2 variants related to mink. Technical Report (European Centre for Disease Prevention and Control).

Fischer, W., Ganusov, V.V., Giorgi, E.E., Hraber, P.T., Keele, B.F., Leitner, T., Han, C.S., Gleasner, C.D., Green, L., Lo, C.C., et al. (2010). Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS One *5*, e12303.

Gao, F., Bonsignori, M., Liao, H.X., Kumar, A., Xia, S.M., Lu, X., Cai, F., Hwang, K.K., Song, H., Zhou, T., et al. (2014). Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. Cell *158*, 481–491.

Ghebreyesus, T.A. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020 (World Health Organization (WHO)), (press release). 11 March 2020. *https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020*.

Giorgi, E.E., Bhattacharya, T., Fischer, W., Yoon, H., Abfalterer, W., and Korber, B. (2021). Recombination and low-diversity confound homoplasy-based methods to detect the effect of SARS-CoV-2 mutations on viral transmissibility. bioRxiv, 2021.01.29.428535.

Giorgi, J.V., Lyles, R.H., Matud, J.L., Yamashita, T.E., Mellors, J.W., Hultin, L.E., Jamieson, B.D., Margolick, J.B., Rinaldo, C.R., Jr., Phair, J.P., and Detels, R.; Multicenter AIDS Cohort Study (2002). Predictive value of immunologic and virologic markers after long or short duration of HIV-1 infection. J. Acquir. Immune Defic. Syndr. *29*, 346–355.

Goh, J.B., and Ng, S.K. (2018). Impact of host cell line choice on glycan profile. Crit. Rev. Biotechnol. *38*, 851–867.

Goloboff, P.A., and Catalano, S.A. (2016). TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics *32*, 221–238.

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.; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

Graham, R.L., and Baric, R.S. (2010). Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J. Virol. *84*, 3134–3146.

Grant, H.E., Hodcroft, E.B., Ssemwanga, D., Kitayimbwa, J.M., Yebra, G., Esquivel Gomez, L.R., Frampton, D., Gall, A., Kellam, P., de Oliveira, T., Bbosa, N., Nsubuga, R.N., Kibengo, F., Kwan, T.H., Lycett, S., Kao, R., Robertson, D.L., Ratmann, O., Fraser, C., Pillay, D., Kaleebu, P., and Leigh Brown, A.J. (2020). Pervasive and non-random recombination in near full-length HIV genomes from Uganda. Virus Evol. *6*, veaa004.

Hahn, B.H., Shaw, G.M., De Cock, K.M., and Sharp, P.M. (2000). AIDS as a zoonosis: scientific and public health implications. Science *287*, 607–614.

Hasenkrug, K.J., Feldmann, F., Myers, L., Santiago, M.L., Guo, K., Barrett, B.S., Mickens, K.L., Carmody, A., Okumura, A., Rao, D., et al. (2021). Recovery from acute SARS-CoV-2 infection and development of anamnestic immune responses in T cell-depleted rhesus macaques. bioRxiv, 2021.2004.2002.438262.

Haynes, B.F., Burton, D.R., and Mascola, J.R. (2019). Multiple roles for HIV broadly neutralizing antibodies. Sci. Transl. Med. *11*, eaaz2686.

He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. *26*, 672–675.

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., Williams, B., Gouws-Williams, E., and Ghys, P.D.; WHO–UNAIDS Network for HIV Isolation Characterisation (2019). Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. Lancet Infect. Dis. *19*, 143–155.

Hensley, M.K., Bain, W.G., Jacobs, J., Nambulli, S., Parikh, U., Cillo, A., Staines, B., Heaps, A., Sobolewski, M.D., Rennick, L.J., et al. (2021). Intractable COVID-19 and prolonged SARS-CoV-2 replication in a chimeric antigen receptor-modified T-Cell therapy recipient: A case study. Clin. Infect. Dis. ciab072.

Hodcroft, E.B., Zuber, M., Nadeau, S., Crawford, K.H.D., Bloom, J.D., Veesler, D., Vaughan, T.G., Comas, I., Candelas, F.G., Stadler, T., and Neher, R.A. (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv, 2020.10.25.20219063.

Hon, C.C., Lam, T.Y., Shi, Z.L., Drummond, A.J., Yip, C.W., Zeng, F., Lam, P.Y., and Leung, F.C.C. (2008). Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. J. Virol. *82*, 1819–1826.

Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science *370*, 1464–1468.

Hraber, P., Korber, B.T., Lapedes, A.S., Bailer, R.T., Seaman, M.S., Gao, H., Greene, K.M., McCutchan, F., Williamson, C., Kim, J.H., et al. (2014). Impact

of clade, geography, and age of the epidemic on HIV-1 neutralization by antibodies. J. Virol. *88*, 12623–12643.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38, 27–8.

Hymes, K.B., Cheung, T., Greene, J.B., Prose, N.S., Marcus, A., Ballard, H., William, D.C., and Laubenstein, L.J. (1981). Kaposi's sarcoma in homosexual men-a report of eight cases. Lancet *2*, 598–600.

Johnson, B.A., Xie, X., Bailey, A.L., Kalveram, B., Lokugamage, K.G., Muruato, A., Zou, J., Zhang, X., Juelich, T., Smith, J.K., et al. (2021). Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. Nature *591*, 293–299.

Kemp, S., Meng, B., Ferriera, I., Datir, R., Harvey, W., Collier, D., Lytras, S., Papa, G., Carabelli, A., Kenyon, J., et al. (2021a). Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/V70. bioRxiv, 2020.12.14.422555.

Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A.T.M., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., et al.; CITIID-NIHR BioResource COVID-19 Collaboration; COVID-19 Genomics UK (COG-UK) Consortium (2021b). SARS-CoV-2 evolution during treatment of chronic infection. Nature *592*, 277–282.

Kiwelu, I.E., Novitsky, V., Margolin, L., Baca, J., Manongi, R., Sam, N., Shao, J., McLane, M.F., Kapiga, S.H., and Essex, M. (2013). Frequent intra-subtype recombination among HIV-1 circulating in Tanzania. PLoS One *8*, e71131.

Korber, B., and Fischer, W. (2020). T cell-based strategies for HIV-1 vaccines. Hum. Vaccin. Immunother. *16*, 713–722.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., Parker, M.D., et al. (2020a). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv, 2020.04.29.069054.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al.; Sheffield COVID-19 Genomics Group (2020b). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell *182*, 812–827.e19.

Korber, B., Hraber, P., Wagh, K., and Hahn, B.H. (2017). Polyvalent vaccine approaches to combat HIV-1 diversity. Immunol. Rev. *275*, 230–244.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. Science *288*, 1789–1796.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics *35*, 4453–4455.

Krakoff, E., Gagne, R.B., VandeWoude, S., and Carver, S. (2019). Variation in intra-individual lentiviral evolution rates: a systematic review of human, nonhuman primate, and felid species. J. Virol. *93*, e00538-19.

Kwon, Y.D., Pancera, M., Acharya, P., Georgiev, I.S., Crooks, E.T., Gorman, J., Joyce, M.G., Guttman, M., Ma, X., Narpala, S., et al. (2015). Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env. Nat. Struct. Mol. Biol. *22*, 522–531.

Lam, T.T.Y., Jia, N., Zhang, Y.W., Shum, M.H.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature *583*, 282–285.

Lassaunière, R., Fonager, J., Rasmussen, M., Frische, A., Strandh, C.P., Rasmussen, T.B., Bøtner, A., and Fomsgaard, A. (2020). Working paper on SARS-CoV-2 spike mutations arising in Danish mink, their spread to humans and neutralization data, Technical Report (Statens Serum Institut.).

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., et al. (2020a). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell *182*, 1284–1294.e9.

Li, X., Giorgi, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020b). Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science Advances *6*, eabb9153.

Li, Y., Ndjango, J.B., Learn, G.H., Ramirez, M.A., Keele, B.F., Bibollet-Ruche, F., Liu, W., Easlick, J.L., Decker, J.M., Rudicell, R.S., et al. (2012). Eastern chimpanzees, but not bonobos, represent a simian immunodeficiency virus reservoir. J. Virol. *86*, 10776–10791.

Liu, M.K.P., Hawkins, N., Ritchie, A.J., Ganusov, V.V., Whale, V., Brackenridge, S., Li, H., Pavlicek, J.W., Cai, F., Rose-Abrahams, M., et al.; CHAVI Core B (2013). Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. J. Clin. Invest. *123*, 380–393.

Leung, K., Shum, M.H., Leung, G.M., Lam, T.T., and Wu, J.T. (2021). Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Euro Surveill. *26*, 2002106.

Malim, M.H. (2009). APOBEC proteins and intrinsic resistance to HIV-1 infection. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 675–687.

Mansbach, R.A., Chakraborty, S., Nguyen, K., Montefiori, D.C., Korber, B., and Gnanakaran, S. (2021). The SARS-CoV-2 Spike variant D614G favors an open conformational state. Sci. Adv. *7*, eabf3671.

Mansky, L.M., and Temin, H.M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. *69*, 5087–5094.

McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello, M., Chen, A., Liu, Z., Zatta, F., et al. (2021). N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. Cell *184*, 2332–2347.e16.

McCallum, M., Bassi, J., Marco, A., Chen, A., Walls, A.C., Iulio, J.D., Tortorici, M.A., Navarro, M.J., Silacci-Fregni, C., Saliba, C., et al. (2021b). SARS-CoV-2 immune evasion by variant B.1.427/B.1.429. bioRxiv. 2021.03.31.437925. https://doi.org/10.1101/2021.03.31.437925.

McCarthy, K.R., Rennick, L.J., Nambulli, S., Robinson-McCarthy, L.R., Bain, W.G., Haidar, G., and Duprex, W.P. (2021). Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science *371*, 1139–1142.

McMahan, K., Yu, J., Mercado, N.B., Loos, C., Tostanoski, L.H., Chandrashekar, A., Liu, J., Peter, L., Atyeo, C., Zhu, A., et al. (2021). Correlates of protection against SARS-CoV-2 in rhesus macaques. Nature *590*, 630–634.

Meyerowitz, E.A., Richterman, A., Gandhi, R.T., and Sax, P.E. (2021). Transmission of SARS-CoV-2: A review of viral, host, and environmental factors. Ann. Intern. Med. *174*, 69–79.

Moore, J.P., and Wilson, I.A. (2021). Decades of basic research paved the way for today's 'warp speed' Covid-19 vaccines (STAT News). https://www.statnews.com/2021/01/05/basic-research-paved-way-for-warp-speed-covid-19-vaccines/.

Nikolaitchik, O., Keele, B., Gorelick, R., Alvord, W.G., Mazurov, D., Pathak, V.K., and Hu, W.S. (2015). High recombination potential of subtype A HIV-1. Virology *484*, 334–340.

Pal, M., Berhanu, G., Desalegn, C., and Kandi, V. (2020). Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. Cureus *12*, e7423.

Phillips, R.E., Rowland-Jones, S., Nixon, D.F., Gotch, F.M., Edwards, J.P., Ogunlesi, A.O., Elvin, J.G., Rothbard, J.A., Bangham, C.R., Rizza, C.R., et al. (1991). Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature *354*, 453–459.

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. Nature *592*, 116–121.

Rademeyer, C., Korber, B., Seaman, M.S., Giorgi, E.E., Thebus, R., Robles, A., Sheward, D.J., Wagh, K., Garrity, J., Carey, B.R., et al. (2016). Features of recently transmitted HIV-1 clade C viruses that impact antibody recognition: Implications for active and passive immunization. PLoS Pathog. *12*, e1005742.

Rambaut, A., Loman, N.J., Pybus, O.G., Barclay, W., Barrett, J., Carabelli, A.M., Connor, T., Peacock, T., and Robertson, D.L.E.,V. (2020b). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020a). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407.

Rathnasinghe, R., Jangra, S., Cupic, A., Martínez-Romero, C., Mulder, L.C.F., Kehrer, T., Yildiz, S., Choi, A., Mena, I., De Vrieze, J., et al. (2021). The N501Y mutation in SARS-CoV-2 spike leads to morbidity in obese and aged mice and is neutralized by convalescent and post-vaccination human sera. medRxiv. https://doi.org/10.1101/2021.01.19.21249592.

Redd, A.D., Nardin, A., Kared, H., Bloch, E.M., Pekosz, A., Laeyendecker, O., Abel, B., Fehlings, M., Quinn, T.C., and Tobian, A.A.R. (2021). CD8+ T cell responses in COVID-19 convalescent individuals target conserved epitopes from multiple prominent SARS-CoV-2 circulating variants. medRxiv, 2021.02.11.21251585.

Richman, D.D. (2021). COVID-19 vaccines: implementation, limitations and opportunities. Glob Health Med *3*, 1–5.

Roark, R.S., Li, H., Williams, W.B., Chug, H., Mason, R.D., Gorman, J., Wang, S., Lee, F.H., Rando, J., Bonsignori, M., et al. (2021). Recapitulation of HIV-1 Env-antibody coevolution in macaques leading to neutralization breadth. Science *371*, eabd2638.

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., et al. (2000). HIV-1 nomenclature proposal. Science *288*, 55–56.

Robson, F., Khan, K.S., Le, T.K., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P., and Ng, W.L. (2020). Coronavirus RNA proofreading: Molecular basis and therapeutic targeting. Mol. Cell *79*, 710–727.

Romano, M., Ruggiero, A., Squeglia, F., Maga, G., and Berisio, R. (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. Cells *9*, 1267.

Sabir, J.S.M., Lam, T.T.Y., Ahmed, M.M.M., Li, L., Shen, Y., Abo-Aba, S.E.M., Qureshi, M.I., Abu-Zeid, M., Zhang, Y., Khiyami, M.A., et al. (2016). Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science *351*, 81–84.

Shajahan, A., Supekar, N.T., Gleinich, A.S., and Azadi, P. (2020). Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology *30*, 981–988.

Sharp, P.M., and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. Cold Spring Harb. Perspect. Med. *1*, a006841.

Shen, X., Tang, H., McDanal, C., Wagh, K., Fischer, W., Theiler, J., Yoon, H., Li, D., Haynes, B.F., Sanders, K.O., et al. (2021). SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral spike vaccines. Cell Host Microbe *29*, 529–539.e3.

Shriner, D., Rodrigo, A.G., Nickle, D.C., and Mullins, J.I. (2004). Pervasive genomic recombination of HIV-1 in vivo. Genetics *167*, 1573–1583.

Silver, Z.A., Antonopoulos, A., Haslam, S.M., Dell, A., Dickinson, G.M., Seaman, M.S., and Desrosiers, R.C. (2020). Discovery of O-linked carbohydrate on HIV-1 Envelope and its role in shielding against one category of broadly neutralizing antibodies. Cell Rep. *30*, 1862–1869.e4.

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.

Stephenson, K.E., Wagh, K., Korber, B., and Barouch, D.H. (2020). Vaccines and broadly neutralizing antibodies for HIV-1 prevention. Annu. Rev. Immunol. *38*, 673–703.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., and Gao, G.F. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. *24*, 490–502.

Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., da Silva Antunes, R., Moore, E., Rubiro, P., et al. (2021). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Rep. Med. *2*, 100204.

Taylor, W.R. (1997). Residual colours: a proposal for aminochromography. Protein Eng. *10*, 743–746.

Tian, J., López, C.A., Derdeyn, C.A., Jones, M.S., Pinter, A., Korber, B., and Gnanakaran, S. (2016). Effect of glycosylation on an immunodominant region in the V1V2 variable domain of the HIV-1 Envelope gp120 protein. PLoS Comput. Biol. *12*, e1005094.

Toovey, O.T.R., Harvey, K.N., Bird, P.W., and Tang, J.W.W. (2021). Introduction of Brazilian SARS-CoV-2 484K.V2 related variants into the UK. J. Infect. *82*, e23–e24.

Tumban, E. (2020). Lead SARS-CoV-2 candidate vaccines: Expectations from phase III trials and recommendations post-vaccine approval. Viruses *13*, 54.

UK Collaborative Group on HIV Drug Resistance (2014). The increasing genetic diversity of HIV-1 in the UK, 2002-2010. AIDS *28*, 773–780.

Varabyou, A., Pockrandt, C., Salzberg, S.L., and Pertea, M. (2020). Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. bioRxiv, 2020.09.21.300913v2.

Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, A., et al. (2021). Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv, 2020.12.30.20249034.

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.e6.

Walls, A.C., Tortorici, M.A., Frenz, B., Snijder, J., Li, W., Rey, F.A., DiMaio, F., Bosch, B.J., and Veesler, D. (2016). Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat. Struct. Mol. Biol. *23*, 899–905.

Wang, H., Cohen, A.A., Galimidi, R.P., Gristick, H.B., Jensen, G.J., and Bjorkman, P.J. (2016). Cryo-EM structure of a CD4-bound open HIV-1 envelope trimer reveals structural rearrangements of the gp120 V1V2 loop. Proc. Natl. Acad. Sci. USA *113*, E7151–E7158.

Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., and Crispin, M. (2020a). Site-specific glycan analysis of the SARS-CoV-2 spike. Science *369*, 330–333.

Watanabe, Y., Berndsen, Z.T., Raghwani, J., Seabright, G.E., Allen, J.D., Pybus, O.G., McLellan, J.S., Wilson, I.A., Bowden, T.A., Ward, A.B., and Crispin, M. (2020b). Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nat. Commun. *11*, 2688.

Watanabe, Y., Bowden, T.A., Wilson, I.A., and Crispin, M. (2019). Exploitation of glycosylation in enveloped virus pathobiology. Biochim. Biophys. Acta, Gen. Subj. *1863*, 1480–1497.

Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes, J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., et al. (2003). Antibody neutralization and escape by HIV-1. Nature *422*, 307–312.

Wei, Y., Silke, J.R., Aris, P., and Xia, X. (2020). Coronavirus genomes carry the signatures of their habitats. PLoS One *15*, e0244025.

Weissman, D., Alameh, M.G., de Silva, T., Collini, P., Hornsby, H., Brown, R., LaBranche, C.C., Edwards, R.J., Sutherland, L., Santra, S., et al. (2021). D614G Spike mutation increases SARS-CoV-2 susceptibility to neutralization. Cell Host Microbe *29*, 23–31.e4.

West, A.P., Barnes, C.O., Yang, Z., and Bjorkman, P.J. (2021). SARS-CoV-2 lineage B.1.526 emerging in the New York region detected by software utility created to query the spike mutational landscape. bioRxiv. https://doi.org/10.1101/2021.02.14.431043.

White, J.M., Delos, S.E., Brecher, M., and Schornberg, K. (2008). Structures and mechanisms of viral membrane fusion proteins: multiple variations on a common theme. Crit. Rev. Biochem. Mol. Biol. *43*, 189–219.

Wibmer, C.K., Ayres, F., Hermanus, T., Madzivhandila, M., Kgagudi, P., Oosthuysen, B., Lambson, B.E., de Oliveira, T., Vermeulen, M., van der Berg, K., et al. (2021). SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nat. Med. *27*, 622–625.

Wood, N., Bhattacharya, T., Keele, B.F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B.F., McMichael, A., et al. (2009). HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. *5*, e1000414.

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.J., Kabongo, J.M.M., Kalengayi, R.M., Van Marck, E., et al. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature *455*, 661–664.

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*, 1260–1263.

Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., and Gamblin, S.J. (2020). SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat. Struct. Mol. Biol. *27*, 763–767.

Wu, N.C., Otwinowski, J., Thompson, A.J., Nycholat, C.M., Nourmohammad, A., and Wilson, I.A. (2020). Major antigenic site B of human influenza H3N2 viruses has an evolving local fitness landscape. Nat. Commun. *11*, 1233.

Zhang, M., Foley, B., Schultz, A.-K., Macke, J.P., Bulla, I., Stanke, M., Morgenstern, B., Korber, B., and Leitner, T. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology *7*, 25.

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.

Zhu, T., Korber, B.T., Nahmias, A.J., Hooper, E., Sharp, P.M., and Ho, D.D. (1998). An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature *391*, 594–597.

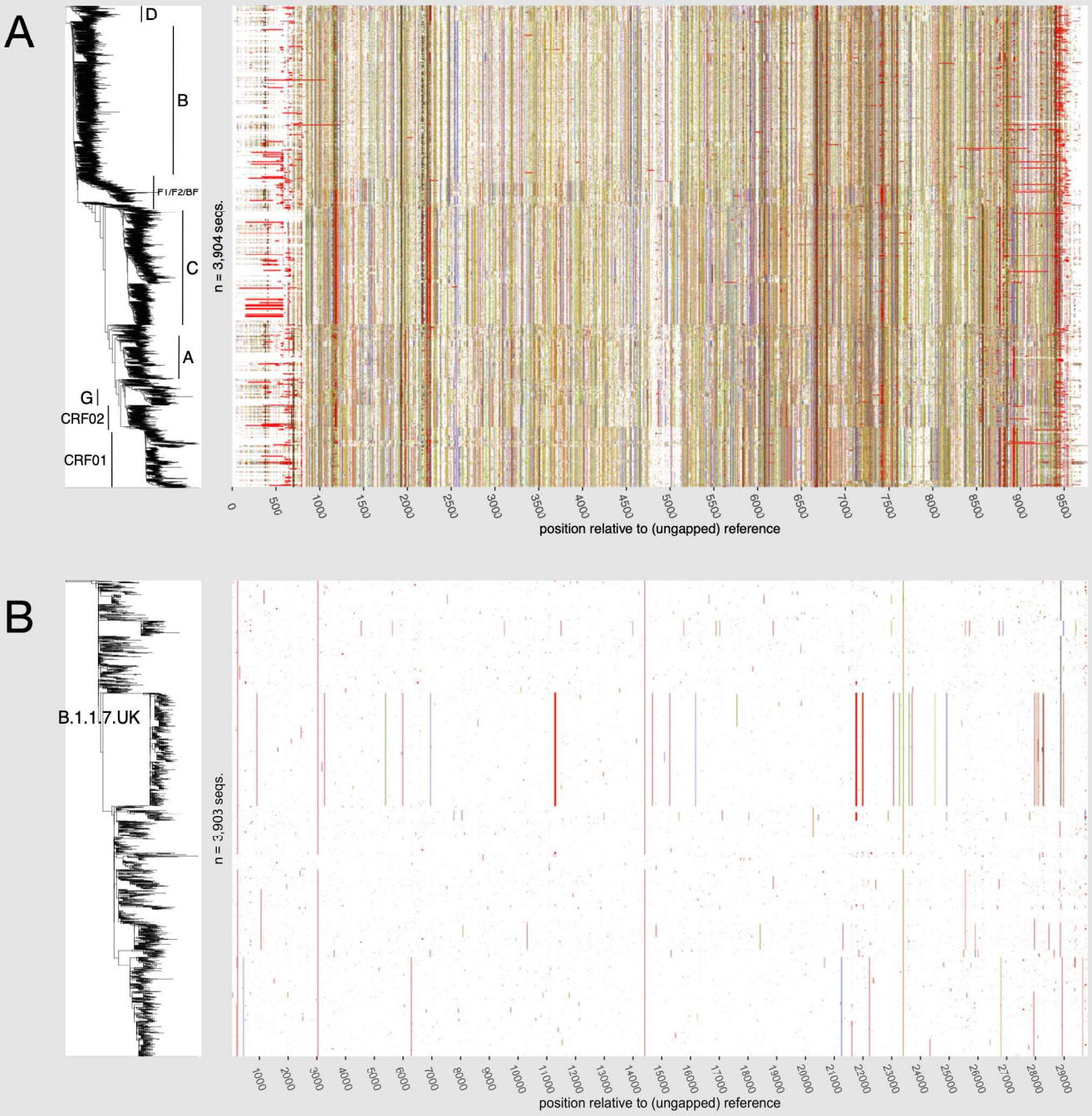
Cell Host & Microbe, Volume *29*

Supplemental information

HIV-1 and SARS-CoV-2: Patterns

in the evolution of two pandemic pathogens

Will Fischer, Elena E. Giorgi, Srirupa Chakraborty, Kien Nguyen, Tanmoy Bhattacharya, James Theiler, Pablo A. Goloboff, Hyejin Yoon, Werner Abfalterer, Brian T. Foley, Houriiyah Tegally, James Emmanuel San, Tulio de Oliveira, Network for Genomic Surveillance in South Africa (NGS-SA), Sandrasegaram Gnanakaran, and Bette Korber


A
D
B
F1/F2/BF
C
A
G
CRF02
CRF01
n = 3,904 secs.
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
position relative to (ungapped) reference
B
B.1.1.7.UK
n = 3,903 seqs.
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
27000
28000
29000
position relative to (ungapped) reference
C
A.23.1.UG
B.1.429.US-CA
B.1.1.7.UK
B.1.1.248.BR
B.1.526.US-NY
B.1.352.ZA
n = 599 seqs.
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
27000
28000
29000
position relative to (ungapped) reference
replacements
A to -
C to -
G to -
T to -
- to A
C to A
G to A
T to A
- to C
A to C
G to C
T to C
- to G
A to G
C to G
T to G
- to T
A to T
C to T
G to T

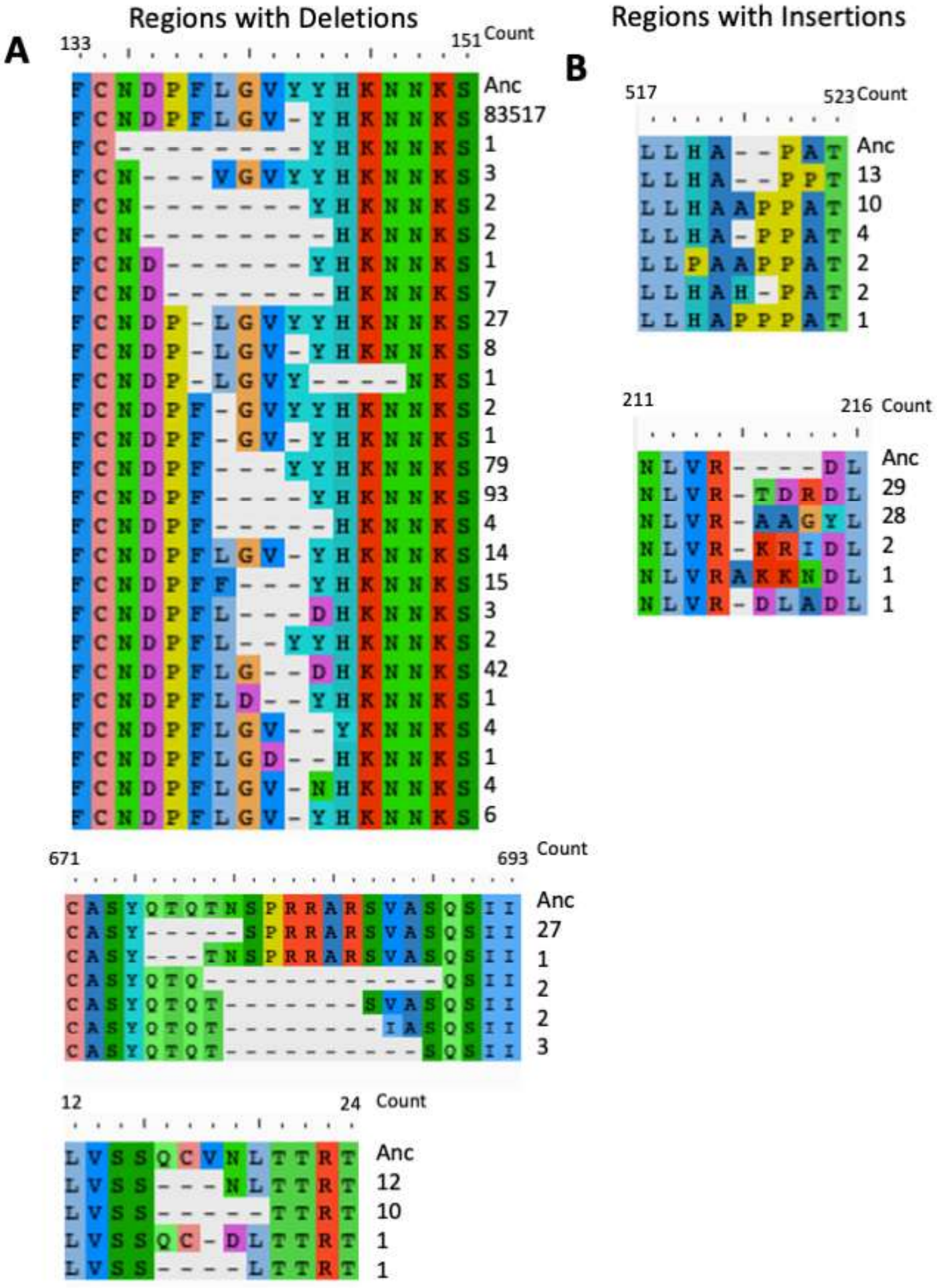

A
Regions with Deletions
133
151
Count
Anc
83517
1
3
2
2
1
7
27
8
1
2
1
79
93
4
14
15
3
2
42
1
4
1
4
6
671
693
Count
Anc
27
1
2
2
3
12
24
Count
Anc
12
10
1
1
B
Regions with Insertions
517
523
Count
Anc
13
10
4
2
2
1
211
216
Count
Anc
29
28
2
1
1

Figure legends

Figure S1. Variability of HIV-1 and SARS-CoV-2 genomes at the nucleotide level. This figure is associated with Figure 1 in the main text.

Right panel: Variant-visualized whole-genome sequence alignments of HIV-1 (a), and SARS-CoV-2 (b). The colored panels are a matrix where each row represents a single sequence, and the columns are positions in a sequence alignment, where colored marks ("+") denote positions that vary compared to a reference sequence. Reference-identical positions are shown as white. A "plurality" consensus sequence, with the most common alignment base at each position, serves as the reference for HIV-1; the outbreak strain (NC 045512) is the reference sequence for SARS-CoV-2. Nucleotide changes are colored based on the mutant base. Sequences are ordered top-to-bottom according to the phylogenetic tree in the left panel; consequently, continuous vertical stripes indicate lineage-specific mutations that are shared by related sequences (see text). The trees in the left panel are in each case derived from a whole-genome nucleotide alignment: an approximation to the maximum-likelihood tree, generated with RAxML-NG (Kozlov et al., 2019) for HIV-1 Env, and for SARS-CoV-2 Spike, a neighbor-joining tree inferred with PAUP (Swofford, 2003) using Log-Determinant (logdet) distances with uninformative sites removed (see Swofford et al., 1996, pp. 459-462), saved with parsimony branch lengths.

Figure S2. Rare but repeated indel patterns in SARS-CoV-2. This figure is associated with Figure 2B in the main text.

Five short regions of the Spike protein are shown, with their positions in Spike relative to the Ancestral Wuhan reference sequence (NC_045512) indicated above each the alignment. All indels are aligned to the reference sequence, indicated as "Anc" for Ancestor in the figure. Regions with deletions are shown on the left (A), and with insertions on the right (B). Most of these indels are rare (the counts are from the GISAID sample of 487,073 Spike sequences available in the cov.lanl.gov alignment on 2/25/2021) but they all come in several forms and are found repeated in several geographic regions, and recur, so are likely to be viable. (A, upper)

Unique repeated deletions patterns in the region between positions 133-151. There were many Spike sequence patterns of deletions in this region, the only common one being the loss of Y144, Δ144.This deletion was found 82,017 times in the context of the B.1.1.7 lineage, and over 1,500 times in the context of other variants and lineages. There was a total of 313 additional sequences with other deletion patterns in this region. To illustrate this, we included a representative example of each distinct deletion pattern between positions 133-151 in Spike, and include a count of how many times the form was repeated in the data set. (A, middle) Repeated deletion patterns in the region between positions 671-693. This small set of deletion patterns includes Spike sequences in which the furin cleavage site motif, RRxR, is deleted (Walls et al., 2020). These may represent sequences from cultures or viruses that are attenuated in vivo (Johnson et al., 2021). (A, lower) Repeated deletion patterns near the Spike signal peptide cleavage site between positions 12 and 24. . (B, Upper) The alignment between positions 517-523. Additional prolines are occasionally added to Spike near 520/521, which are located at the end of the RBD. In the top case (shown in the row beneath Anc), we found a repeated A522P change, no insertion, but a distinctive way to add two prolines in this local domain, so we included it here. In the other cases, either a single P, or a two-amino-acid duplication, AP, were introduced. The insertions in this region were all sampled from one source in Houston, Texas. (B. Lower) Distinctive repeated insertions found between Spike positions 211-216. These insertions are carried in lineages that are rare but increasing in frequency. In the May 12 sampling of GISAID, the insertion 214 TDR, which is carried in the lineage B.1.214.2, was found in 12 countries, and had been sampled 602 times. The insertion 215 AGG, which is carried in the lineage A.2.5.2, was found in 7 countries and sampled 122 times. Note that additional indel patterns have been recently observed in these regions and others that are discussed in the text, including several distinctive deletions patterns near the 242-244 deletion commonly found in the B.1.351 variant, the deletion at 156-157 found in B.1.617.2, and a four amino acid insertion near the furin cleavage site.

Table S1: This table is associated with Figure 1 in the main text.

Estimated mutational parameters based on maximum likelihood trees generated using RAxML-NG (Kozlov et al., 2019), with a general time-reversible model with rate categories estimated using an 8-category Gamma rate distribution (GTR+G8).

	HIV-1	SARS-CoV-2
Base frequencies:		
A	0.4456	0.2900
C	0.1741	0.1757
G	0.1974	0.1626
T	0.1828	0.3716
Exchangeabilities (R):		
AC	1.0546	0.2127
AG	3.0375	0.6778
AT	0.5612	0.1490
CG	0.7638	0.3050
CT	4.2699	2.1128
GT	1.0000	1.0000
Gamma shape	0.4804	0.2989

Network for Genomic Surveillance in South Africa (NGS-SA) author list and affiliations

Eduan Wilkinson, Nokukhanya Msomi[2], Arash Iranzadeh[4], Vagner Fonseca[1], Deelan Doolabh[5], Emmanuel James San[1], Koleka Mlisana[7,8], Anne von Gottberg[9,10], Sibongile Walaza[9,11], Mushal Allam[9], Arshad Ismail[9], Thabo Mohale[9], Allison J Glass[10,12], Susan Engelbrecht[13], Gert Van Zyl[13], Wolfgang Preiser[13], Francesco Petruccione[14,15], Alex Sigal[16,17,18], Diana Hardie[19], Gert Marais[19], Marvin Hsiao[19], Stephen Korsman[19], Mary-Ann Davies[20,21], Lynn Tyers[5], Innocent Mudau[5], Denis York[22], Caroline Maslo[23], Dominique Goedhals[24], Shareef Abrahams[25], Oluwakemi Laguda-Akingba[25,26], Arghavan Alisoltani-Dehkordi[27,28], Adam Godzik[28], Constantinos Kurt Wibmer[9], Bryan Trevor Sewell[29], José Lourenço[30], Sergei L Kosakovsky Pond[31], Steven Weaver[31], Marta Giovanetti[32], Luiz Carlos Junior Alcantara[32],Darren Martin[4,5], Jinal N Bhiman[9,10], Carolyn Williamson[5,8,19]

Affiliations:

[1] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Department of Laboratory Medicine & Medical Sciences, University of KwaZulu-Natal, Durban, South Africa

[2] Discipline of Virology, University of KwaZulu-Natal, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service, Durban, South Africa

[4] Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, 7925, South Africa

[5] Division of Medical Virology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[7] National Health Laboratory Service, Johannesburg, South Africa

[8] Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

[9] National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa

[10] School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[11] School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[12] Department of Molecular Pathology, Lancet Laboratories, Johannesburg, South Africa

[13] Division of Medical Virology at NHLS Tygerberg Hospital and Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

[14] Centre for Quantum Technology, University of KwaZulu-Natal, Durban, South Africa [15] National Institute for Theoretical Physics (NITheP), KwaZulu-Natal, South Africa

[16] Africa Health Research Institute, Durban, South Africa

[17] School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal, Durban, South Africa

[18] Max Planck Institute for Infection Biology, Berlin, Germany

[19] Division of Medical Virology at NHLS Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa

[20] Centre for Infectious Disease Epidemiology and Research, University of Cape Town, Cape Town, South Africa

[21] Western Cape Government: Health, Cape Town, South Africa

[22] Molecular Diagnostics Services, Durban, South Africa

[23] Department of Quality Leadership, Netcare Hospitals, Johannesburg, South Africa

[24] Division of Virology at NHLS Universitas Academic Laboratories, University of The Free State, Bloemfontein, South Africa

[25] National Health Laboratory Service, Port Elizabeth, South Africa

[26] Department of Laboratory Medicine and Pathology, Faculty of Health Sciences, Walter Sisulu University, Mthatha, South Africa

[27] Division of Medical Virology, Department of Pathology, University of Cape Town, Cape Town, South Africa

[28] Division of Biomedical Sciences, University of California Riverside School of Medicine, Riverside, California, USA

[29] Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Rondebosch, South Africa

[30] Department of Zoology, University of Oxford, Oxford, United Kingdom

[31] Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA

[32] Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil