A phylogeny-based metric for estimating changes in transmissibility from recurrent mutations in SARS-CoV-2

Damien Richard[1,2], Liam P Shaw[3], Rob Lanfear[4], Mislav Acman[1], Christopher J Owen[1], Cedric CS Tan[1], Lucy van Dorp[1*], François Balloux[1*]

[1]UCL Genetics Institute, University College London, London WC1E 6BT, UK
[2]Division of Infection and Immunity, University College London, London, WC1E 6BT, UK
[3]Department of Zoology, University of Oxford, Oxford OX1 3DZ, UK
[4]Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia.
*Contributed equally

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019 and spread globally to cause the COVID-19 pandemic. Despite the constant accumulation of genetic variation in the SARS-CoV-2 population, there was little evidence for the emergence of significantly more transmissible lineages in the first half of 2020. Around November 2020, several more contagious and possibly more virulent 'Variants of Concern' (VoCs) were detected near-simultaneously in various regions of the world. These VoCs share some mutations and deletions that haven arisen recurrently in distinct genetic backgrounds. Here, we build on our previous work modelling the association of mutations to SARS-CoV-2 transmissibility and characterise the contribution of individual recurrent mutations and deletions to estimated viral transmissibility. We estimate enhanced transmissibility associated to mutations characteristic of VoCs and identify a tendency for cytidine to thymidine (C→T) substitutions to be associated to a reduction in estimated transmissibility. We then assess how patterns of estimated transmissibility in all SARS-CoV-2 clades have varied over the course of the COVID-19 pandemic by summing transmissibility estimates for all individual mutations carried by any sequenced genome analysed. Such an approach recovers 501Y.v1 (B.1.1.7) as the most transmissible clade currently in circulation. By assessing transmissibility over the time of sampling, we observe a tendency for estimated transmissibility within clades to slightly decrease in most clades. Although subtle, this pattern is consistent with the expectation of a decay in transmissibility in mainly non-recombining lineages caused by the accumulation of weakly deleterious mutations. SARS-CoV-2 remains a highly transmissible pathogen, though such a trend could conceivably play a role in the turnover of different global viral clades observed over the pandemic so far.

Caveats:

- This work is not about the severity of disease. We do not analyse the severity of disease. We do not present any evidence that SARS-CoV-2 has decreased in severity.
- Lineage replacement dynamics are affected by many factors. The trend we recover for a decrease in inferred transmissibility of a clade over time is a small effect. We caution

against over-interpretation. This result would not affect the management of the SARS-CoV-2 pandemic: for example, we make no claims about any impact on the efficacy of particular non-pharmaceutical interventions (NPIs).

- Our phylogeny-based method to infer changes in estimated transmissibility due to recurrent mutations and deletions makes a number of simplifying assumptions. These may not all be valid. The consistent trend for the slight decrease we report might be due to an as-yet-unidentified systematic bias.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the viral agent of the COVID-19 pandemic, has been acquiring genomic diversity since its emergence in humans in late 2019[1–3]. The unprecedented scale of generation and sharing of SARS-CoV-2 genomes has allowed tracking of the accumulation of mutations in close to real time; supporting epidemiological tracing of transmission, mutation surveillance and phylogenetic analyses[4–8]. Such analyses have highlighted a high rate of recurrent substitutions and deletions (homoplasies) in SARS-CoV-2 alignments[1,9–12] and more recently allowed for the rapid flagging of 'Variants of Interest' (VoIs) and 'Variants of Concern' (VoCs) associated with higher transmissibility, immune evasion or higher virulence.

Mutations accumulate in SARS-CoV-2 at a rate of approximately two mutations per lineage per month[1,13]. Mutations arise following stochastic errors during replication or can be induced by host anti-viral editing proteins leaving characteristic mutational biases[9,14–17]. Mutations may also be exchanged between lineages through recombination events, combining genetic material from different viruses simultaneously infecting the same host into a new lineage[18]. While recombination is often presumed to be widespread in coronaviruses, at least within species, evidence in SARS-CoV-2 has remained scarce until recently, likely due to the limited power to detect genetic recombination provided by current levels of genetic diversity[19] (https://observablehq.com/@spond/linkage-disequilibirum-in-sars-cov-2). Recent studies however provide compelling evidence that some recombinants are circulating at low incidence levels[20–22]. In addition to point mutations, the possible evolutionary importance of genomic deletions has come to the fore, in particular in the context of antigenicity[8,10].

All SARS-CoV-2 lineages currently in circulation remain highly genetically similar to one another. Over the first ten months of the pandemic (December 2019 – September 2020) there was fairly limited evidence for marked variation in transmissibility or virulence between different SARS-CoV-2 lineages[9]. Patterns of emerging genomic diversity were well explained by neutral evolutionary processes, with the exception of the D614G mutation which has been associated with higher transmission[23,24]. However, towards the end of 2020, several lineages of SARS-CoV-2 attracted attention following their rapid increase in frequency in regions where they were first observed and the constellations of mutations they harbour. Those include 501Y.v1 (also known as PANGO lineage B.1.1.7 or NextStrain clade 20I/501Y.v1) first detected in the UK[25], 501Y.v2 (a.k.a. PANGO lineage B.1.351; Nextstrain clade 20H/501Y.v2) first detected in South Africa[26], and 501Y.v3 (a.k.a. PANGO lineage P.1; Nextstrain clade 20J/501Y.v3) first detected in cases linked to Brazil[27]. The Centers for Disease Control and Prevention (CDC) additionally list as variants of concern two PANGO lineages within clade 20C: B.1.427 and B.1.429, both first identified in California[28,29].

Notable features of the 501Y.v1, 501Y.v2 and 501Y.v3 VoCs (henceforth 501Y VoCs) are that they carry multiple mutations within the spike protein, a region crucial for human receptor binding and a dominant target for neutralizing antibodies during infection[30]. These include, besides the 501Y mutation observed in all of the aforementioned VoCs (hence the names), 484K which independently occurred in 501Y.v2, 501Y.v3, PANGO lineage P.2 identified in Brazil[31] and in some cases 501Y.v1 and changes at codon 417 of the spike protein (417N in 501Y.v2 and 417T in 501Y.v3). Such point mutations may be coupled with deletions in the spike N-terminal domain (NTD) which have been observed repeatedly, including during chronic infections[8,32–34], with H69 V70 (Δ69/70) and Y144 (Δ144) found together in 501Y.v1 and Δ243-244 found in 501Y.v2. All three 501Y VoCs carry the same deletion in NSP6 Δ106-108, which has recently been detected in tandem with shared N-terminal domain (NTD) deletions and receptor-binding domain (RBD) mutations (E484K) in recently described VoIs B.1.525 and B.1.526.1/B.1.526.2. The 501Y.v1 lineage further carries a truncated ORF8 (Q27stop). Interestingly, while many of the mutations present in VoCs, including 501Y, were observed far earlier in the course of the pandemic[7], they seemingly conferred no easily detectable adaptive advantage until more recently, pointing to a possible shift in the SARS-CoV-2 landscape of selective pressures[9,12].

Each of the three 501Y VoCs have measurable phenotypic effects compared to the Wuhan-Hu-1 reference genome (the original 'wild-type'), including enhanced receptor binding in each case[35–37], increased transmissibility (501Y.v1, 501Y.v2, 501Y.v3)[27,38] and some ability to evade past immunity from natural infection and/or vaccination for 501Y.v2 and 501Y.v3[39–43]. Real-world effects on mortality require carefully controlling for multiple factors, but infection with 501Y.v1 has been associated with higher hospitalisation rates by several studies, even if mortality in hospitalised patients seems unaffected[44–46]. Higher virulence of 501Y.v3 (P.1) has also been suggested[47].

All 501Y VoCs exhibit an excess of non-synonymous mutations (i.e. dN/dS >1) consistent with adaptive evolution[25–27]. It has been suggested, based on both the large number and nature of its mutations, that 501Y.v1 spread into community transmission following a period of rapid evolution in a chronically infected patient[12,25]. An alternative hypothesis is that the combination of mutations observed in 501Y.v1 may have been generated through recombination between circulating SARS-CoV-2 lineages; possibly also linked to a chronic infection scenario. Irrespective of how VoCs have emerged, the occurrence of a number of common recurrent mutations and deletions strongly suggests the action of convergent evolution, likely driven by the phenotypic advantage of increased human ACE2 receptor binding affinity and/or some ability to bypass prior immunity[12].

We previously developed a phylogenetic index to identify all recurrent mutations in global SARS-CoV-2 phylogenies and tested their association to variation in estimated transmissibility[9]. When applied to genome assemblies shared over the first ~seven months of the pandemic (up to 30/7/2020), our method did not identify any recurrent mutation that had a statistically significant association with increased estimated transmissibility. The multiple emergences of more transmissible VoCs in late 2020, which also involve recurrent deletions, motivated us to extend our phylogenetic scoring framework. Here, we modify our phylogeny-based method to include notable recurrent deletions and analyse a far larger dataset of over half a million globally distributed SARS-CoV-2 genome assemblies released via the GISAID Audacity platform[5,6]. We characterise the distribution and number of emergences of all mutations and deletions over the phylogeny (as of 15/04/2021), before filtering for a well-supported set of recurrent mutations to test for their association with estimated transmissibility. We express the relative contribution to estimated transmissibility provided by carriage of any individual recurrent mutation or selected deletion as Coefficients of Exponential Growth Alteration (CEGAs). Finally, we show that a simple genetic model, combining the transmissibility effect of each individual CEGA into a multilocus per-isolate score, can largely recover expected relationships in relative estimated transmissibility of major SARS-CoV-2 clades from previous work. This suggests that we can use this phylogeny-based metric to evaluate relative changes in the estimated transmissibility of viral clades over time. While our approach uncovers a global increase in estimated transmissibility of SARS-CoV-2 following the emergence of 501Y VoCs, we also find evidence that the estimated transmissibility of major clades tends to subtly decrease over time since their detection. Such an observation could be explained by the accumulation of weakly deleterious mutations.

Results

SARS-CoV-2 global genetic diversity

We consider the global genetic diversity of SARS-CoV-2 assemblies generated from samples collected between the 30th of December 2019 and the 9th of April 2021 (Supplementary Figure S1). These encompass 6,571 locations spanning six continental regions and comprise 224,824, 4,754, and 1,808 from each of the 501Y VoCs; 501Y.v1, 501Y.v2 and 501Y.v3 respectively. In contrast to the diffuse geographic structure observed early in the pandemic, when different SARS-CoV-2 lineages were essentially randomly distributed globally due to multiple introduction events in various countries[1,48], stronger patterns of geographic structure have now emerged (Supplementary Figure S1a). However, SARS-CoV-2 genetic diversity remains low at this stage and the majority of currently circulating lineages are represented by intermediate basal isolates starting from the root. Notable exceptions are a few dominant lineages located on longer branches away from the root, possibly suggestive of a burst of adaptive mutations[25]. The large number of assemblies included in this dataset means that (i) all of the 29,903 nucleotide positions of the Wuhan-Hu-1 reference genome carry single nucleotide variants (SNVs) in at least one sample, (ii) a large fraction of SNVs independently emerged multiple times, and (iii) multiple alternate alleles are often present at a single nucleotide position, which poses a challenge for present-day bioinformatics approaches.

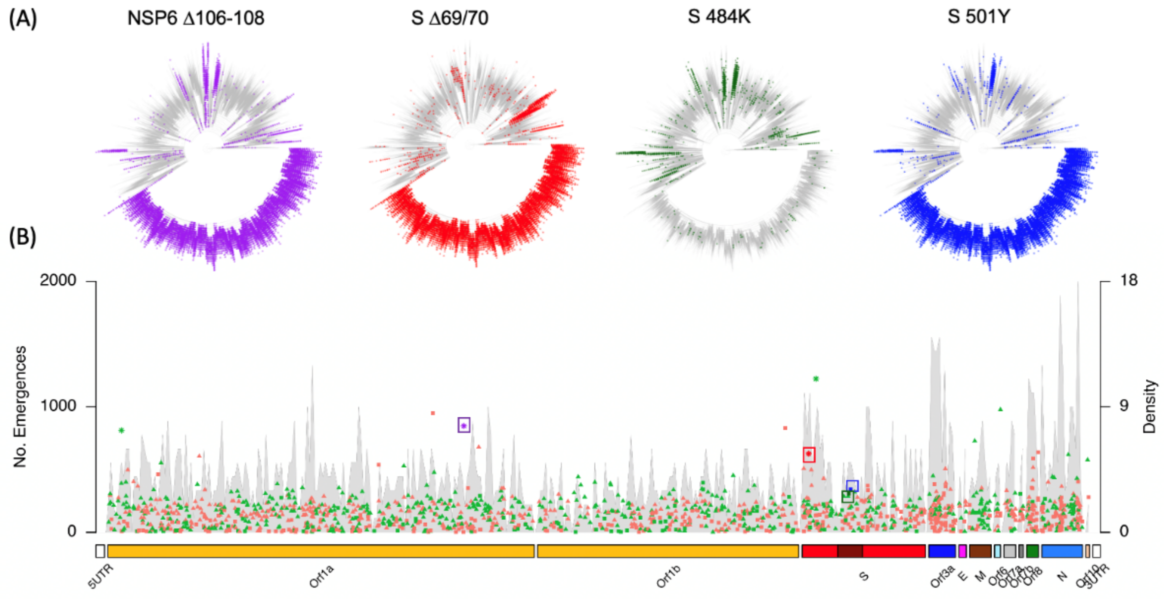

Figure 1 (A) Placement of unfiltered homoplasies associated to some 501Y VoCs in the global phylogeny; NSP6 Δ106-108 (nucleotide positions 11288-11296; 848 emergences since 12/08/2020), Spike (S) Δ69/70 (nucleotide positions 21765-21770; 627 emergences since 17/04/2020), Spike 484K (nucleotide position 23,012; 309 emergences since 15/04/2020) and Spike 501Y (nucleotide position 23,063; 344 emergences since 04/05/2020). (B) Bottom panel provides the number of emergences of all recurrent mutations detected along the SARS-CoV-2 genome, with sites depicted in the phylogenies highlighted by coloured boxes, together with the density of homoplasies over an 80-nucleotide sliding window (y-axis at right). Synonymous changes are shown in green, non-synonymous changes are shown in orange. Those sites which correspond to C→T or T→C changes are shown with a triangle. Studied deletions are shown with a '*'. All other mutations are shown with a square.

Recurrent substitutions and deletions in SARS-CoV-2

Across 550,743 SARS-CoV-2 genomes we detect 1,554 SNVs at a frequency of at least 0.1% (Figure 1, Figure S1). Most (>99%) of these can be considered homoplastic, appearing to have emerged independently more than once given the global SARS-CoV-2 phylogeny. Phylogenetic uncertainty is expected to lead to an overestimation of the number of homoplasies. This can occur because most SARS-CoV-2 genomes contain relatively little phylogenetic information, meaning that clades of closely related genomes can be interspersed due to phylogenetic noise induced by errors in genome sequencing or assembly. This interspersion is known to elevate estimates of homoplasy[49], so herein we focus on mutations that have arisen a very large number of times in the phylogeny in genetically distinct SARS-CoV-2 clades.

Consistent with previous observations[9,16], a large fraction of the 1,552 homoplasies we detect (60%) derive from C→T changes likely caused by host anti-viral RNA editing machinery[14,15,17] (Supplementary Table S2, Supplementary Figure S2). We detect a significant genome-wide excess of synonymous recurrent mutations compared to non-synonymous changes (Wilcoxon p<1x10^{-6}), a trend largely driven by patterns over the largest open reading frame: ORF1ab (Supplementary Figure S3). Beyond synonymous and non-synonymous recurrent substitutions, we also consider five recurrent in-frame deletions (see Methods), with those at highest frequency: spike H69 V70 (Δ69/70, 42% of assemblies and 627 emergences), Y144 (Δ144, 41% of assemblies and 1,224 emergences), and NSP6 Δ106-108 (43% of assemblies and 848 emergences). While homoplasies were fairly evenly distributed along the genome (see Methods, Figure 1, Figure S4, Supplementary Table S2) some sites, having excluding known sequencing artefacts[11], correspond to a large number of emergences. For instance, 22 mutations and four deletions (those three listed above and NSP1 Δ141-143) were estimated to have appeared >498 times (97.5 percentile). Amongst these highly recurrent mutations, a large fraction corresponded to C→T or T→C changes (16/22). Consistently, across all 1,552 homoplasies, C→T and T→C transitions tended to have a significantly higher number of estimated emergences (on average 179) compared to all other observed transitions (on average 159 emergences, Wilcoxon p<2e-15) (Supplementary Figure S2).

Mutations commonly associated to 501Y VoCs have appeared in multiple genetic backgrounds, already well before the first reports of VoCs towards the end of 2020 (Figure 1). For instance, spike deletion Δ69/70, present in 501Y.v1, was first detected in our dataset on 17/04/2020. Since then we estimate it has emerged independently 627 times in the phylogeny, including in combination with other mutations in the spike protein such as N439K and Y453F[10,50]. We estimate 309 emergences of E484K (nucleotide position 23,012, earliest observation 15/04/2020) while N501Y itself has emerged an estimated 344 times since 04/05/2020. While phylogenetic misplacement may inflate these raw estimates (see Methods and above), they nonetheless highlight the long duration of circulation of major mutations of interest in VoCs and their propensity to independently arise on the background of many subtly different SARS-CoV-2 lineages.

Effect of mutations and deletions on SARS-CoV-2 estimated transmissibility

By focusing our observations on mutations and deletions which have evolved independently multiple times, often in different locations and epidemiological settings, we largely exclude consideration of those sites which may have risen to high frequency solely due to founder effects[1,9,51]. To test across observations for association between individual recurrent mutations or deletions and variation in SARS-CoV-2 estimated transmissibility, we apply a phylogenetic scoring approach. The model we develop assumes that a transmission advantage conferred by a mutation/deletion translates into an excess of descendants in the phylogeny, given random sequencing of isolates. We can quantify this effect by considering the relative number of descendants that the lineage carrying the mutation/deletion gives rise to relative to its sister lineage in the phylogeny which lacks the mutation/deletion[9] (Figure S5). To do so, we study pairs of SARS-CoV-2 lineages in the phylogeny which descend from nodes corresponding to the hypothetical ancestor which acquired a particular mutation/deletion. We assess these ratios of descendants across independent emergences of the mutation/deletion in the phylogeny (homoplastic replicate). To express these ratios in terms of selection differentials (*s*), we then normalise by the estimated number of viral generations over which the lineages have been observed in circulation (see Methods). We refer to this normalised index as the Coefficients of Exponential Growth Alteration (CEGA). CEGA represents an estimate of the change in transmissibility due to a focal mutation. Hence an average CEGA of zero suggests that a mutation has no effect on transmissibility. A CEGA greater than zero suggests that the focal mutation increases the estimated transmissibility, and vice-versa for a CEGA less than zero.

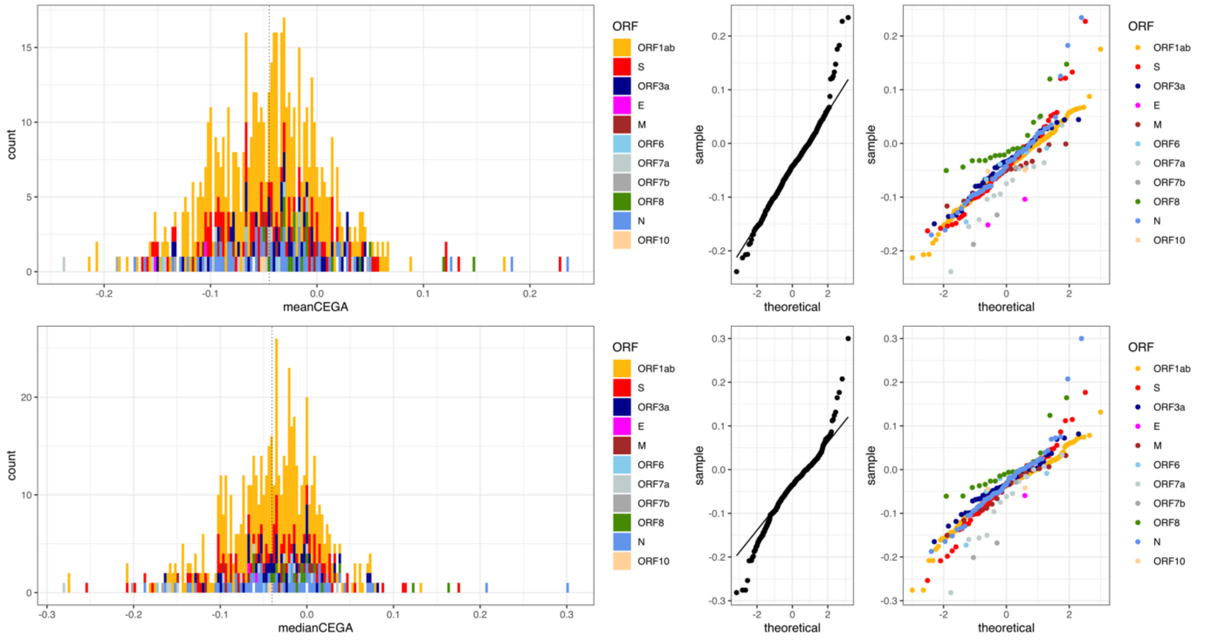

Figure 2: Distribution of mean (top) and median (bottom) CEGA score for 625 homoplastic mutations which passed our scoring filters. In both cases the mean and median values over all estimates fell below 0 (-0.045 for mean CEGA; -0.042 for median CEGA). Right hand panels provide the qqplots for included sites genome-wide and coloured by genomic feature.

As noted, recurrent mutations are to be expected within SARS-CoV-2 phylogenies given the well documented role of immune-mediated hyper-mutation in introducing genomic diversity

[9,14–17] and may also be observed due to a degree of phylogenetic misplacement[52]. We therefore implement a series of filtering steps so that we only assess, using the CEGA metric, phylogenetically well-supported recurrent mutations and deletions. These steps include: (i) only taking forward for analysis mutations for which we detect at least five emergences (replicates) with (ii) sufficient number of descending offspring displaying each allele and (iii) excluding erroneous positions associated to sequencing artefacts (see Methods). In this way we exclude from consideration low frequency and singleton emergences which are more likely to have arisen due to sequencing errors or phylogenetic misplacement[1,11,16].

Of the 1,552 identified homoplasies, 625 (40%) passed all selection criteria for downstream analysis. These 625 comprise 621 substitutions and four deletions (Supplementary Table S3). On average each of the 625 recurrent mutations was represented by 8.0 ± 4.5 (mean ± SD) independent emergences on the phylogeny that passed our filters, though with considerable variance. The number of well supported emergences by nucleotide position is provided in Supplementary Table S3. Of these 625 sites, 110 displayed a positive CEGA score (i.e., associated to on average higher estimated transmissibility). Values ranged from -0.24 to 0.23 across sites with a slightly negative mean (-0.045) and median (-0.041). Consistently, we observe a significant deviation from a normal distribution (Shapiro-Wilk *p*-value of <0.005 in each case) to a right-tailed distribution (mean skewness 0.04) suggesting that the majority of recurrent mutations/deletions in SARS-CoV-2 have a slightly deleterious effect on estimated transmissibility. We obtain highly similar distributions for the mean and median CEGA score per site (Figure 2, Supplementary Figure S6-S8) suggesting averaging CEGA scores of individual mutations/deletions can robustly capture their associations to estimated transmissibility in multiple genetic backgrounds (Figure 2, Figure 3A).

We observed that the mean CEGA score tends to be slightly higher at non-synonymous compared to synonymous sites (i.e., associated with higher estimated transmissibility) (Wilcoxon *p*=0.038), consistent with higher transmissibility tending to require amino acid altering mutations. CEGA scores associated with C→T transitions tended to fall below zero on average (p<2.2e-16), suggesting an excess of such sites leads to a slightly deleterious effect on estimated transmissibility. However, we otherwise recorded no strong trends associating any particular type of mutation to higher or lower CEGA scores (Supplementary Figures S7-S8). Considering those sites leading to the most extreme CEGA scores - 32 mutations within the upper 5% of mean CEGA scores - 11 correspond to synonymous changes, 18 to non-synonymous changes and three to deletions (Δ69/70 and Δ144 in the spike and NSP6 Δ106-108). Conversely, of the 32 mutations associated with the lowest 5% of CEGA estimates, 17 corresponded to synonymous changes and 15 to non-synonymous changes, 13 of which are C→T transitions (Figure 3B). One of the most negative scores observed is a non-synonymous change in the SARS-CoV-2 spike protein A67V, highlighting the potential for changes in the spike protein to contribute to both higher and lower estimated transmissibility. Consistently, there was no overall tendency for recurrent mutations in the spike protein to give rise to significantly higher or lower CEGA scores than those obtained in other structural genes (Supplementary Figure S9). Of our curated set of five deletions, three: spike Δ144; NSP6 Δ106-

108 and spike Δ69/70 were associated to positive CEGA values, and NSP1 Δ141-143 to a negative value (Supplementary Table S3). CEGA scores associated to our preselected deletions were marginally higher than those obtained for SNVs (*p*=0.025, Supplementary Figure S10).

The N501Y mutation and the NSP6 Δ106-108 deletion, which are both present in all 501Y VoCs have mean CEGA scores of 0.054 and 0.087, placing these in the 5% top highest scoring sites tested. Interestingly, the E484K mutation present in the spike protein of 501Y.v2 and 501Y.v3 has an associated mean CEGA score of -0.066 suggesting it slightly reduces estimated transmissibility on average. Among the VoCs, 501Y.v1 stands out as it carries 14/32 of the highest 5% of CEGA scores (Figure 3B), including two of the mutations characterising the D614G haplotype, namely the spike D614G and synonymous nucleotide position 3037 (ORF1ab F106F) both found in the majority of presently circulating lineages. We also estimate three high scoring mutations on each one of the three nucleotides encoding for nucleocapsid D3 amino-acid (28280:GAT->CTA)[53]. This triple mutation, leading to D3L, is predominately found in 501Y.v1, and has been implicated in greater subgenomic RNA expression in 501Y.v1[54] as a possible contributor to the higher transmissibility of this VoC.

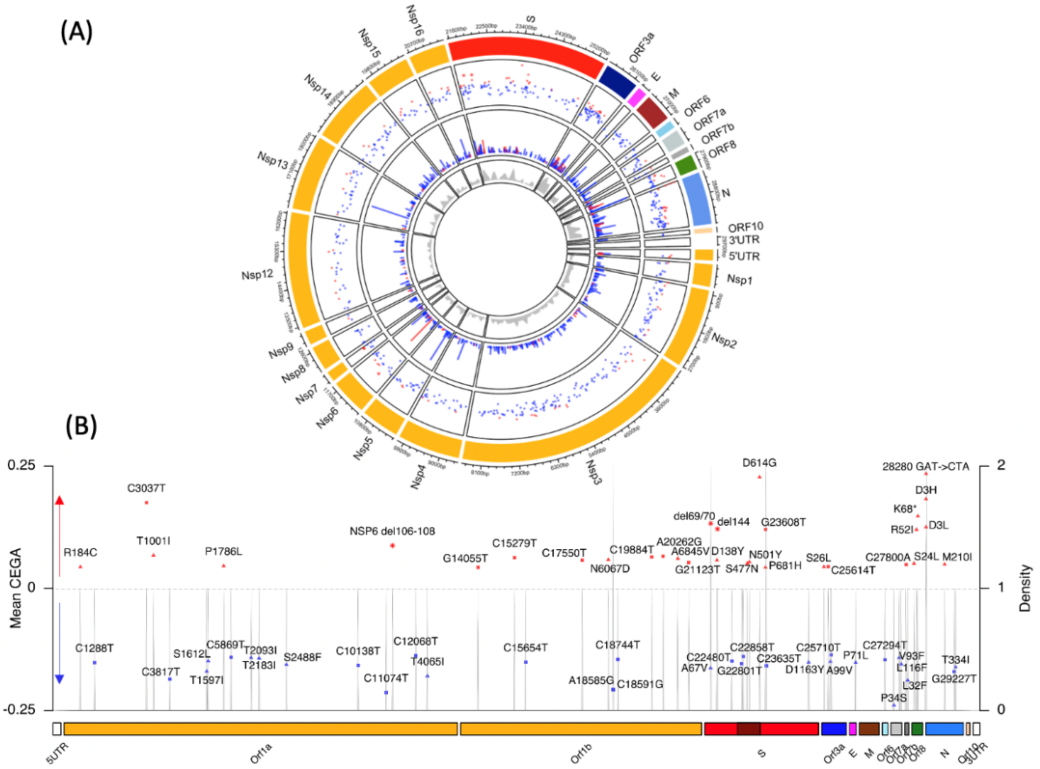

Figure 3: (A) Circular representation of the genomic structure of SARS-CoV-2 with associated mean CEGA scores. From outer to inner circles: Gene names; mean CEGA score (red: positive; blue: negative) with those estimated at deletions denoted *; Number of independent emergences used for CEGA computation; density of sites tested using a window of 20 nucleotides. (B) Sites along the genome with mean CEGA scores plotted for those falling in the upper and lower 5% of estimates. Non-synonymous sites are highlighted with a triangle with associated amino acid change, synonymous sites are depicted with a square with associated nucleotide change, deletions are shown with a *. As in (A) positive scores are depicted red with negative scores depicted blue. Figure produced using the R package circlize 0.4.12 [55] and karyoploteR 1.16.0 [56].

Three out of 11 tested recurrent mutations located in the Receptor Binding Domain (RBD) of the spike protein (nucleotide positions 22,559-23,143 [57]) displayed positive CEGA values, in increasing order of value: A520S (CEGA=0.001), S477N (CEGA=0.051) and N501Y (CEGA=0.054). While we did not recover any significant correlation between the estimates for seven of these RBD recurrent mutations and ACE2 receptor binding affinity nor ACE2 expression measured following a deep mutational scan analysis[35], our highest scoring site within the RBD is also the site predicted to have highest binding affinity (Supplementary Figure S11).

To assess the robustness of our CEGA scores to varying degrees of phylogenetic uncertainty we recomputed the CEGA values on a sub-sampled phylogeny of 130,000 assemblies simulating varying degrees of misplacement (0-10% of tips). While, as anticipated, the number of inferred homoplasies increases with the degree of misplacement (Supplementary Figure S12), the distribution of CEGA scores tends to become narrower with values estimated under the misplacement analysis becoming increasingly small (<|0.02|). This suggests our CEGA scores tend to be robust for sites with a score |>0.02|, and that while phylogenetic uncertainty will increase the number of sites with associated CEGA scores, it is not expected to produce artefactual CEGA scores associated with low or high estimated transmissibility. In addition, the 501Y.v1 (B.1.1.7) clade represents the dominant clade in SARS-CoV-2 datasets, due to its rapid increase in frequency, especially in the UK aided by the COG-UK consortium which sequenced a large fraction of SARS-CoV-2 genomes in late 2020 – early 2021. It might also be overrepresented in the dataset because of targeted sequencing of samples producing a PCR S Gene Target Failure (SGTF), caused by the Δ 69/70 spike deletion found in this clade (Figure 1A). To assess the impact of the overrepresentation of 501Y.v1, we recomputed the CEGA scores on a dataset of 370,804 assemblies, excluding all genomes annotated as 501Y.v1. Over the 308 sites for which CEGA scores could be computed both with and without 501Y.v1 isolates, correlation was high (r^2=0.582; Supplementary Figure S13). Though, removal of all 501Y.v1 isolates leads to a drop in the number of sites for which CEGA scores can be computed.

Evaluation of the transmissibility of SARS-CoV-2 clades through time

We next sought to ask whether it is possible to combine individual CEGA scores to estimate the relative transmissibility of a SARS-CoV-2 isolate from its genome. There are many ways to potentially combine CEGA scores into a single genome-wide estimate. The best way to do so likely depends on the underlying (and unknown) genetic architecture of transmissibility. Regardless, it is clear that if a method of combining CEGA scores were successful, it should at least allow one to recover the relative estimated transmissibility of any genome in the phylogeny, because this is precisely the data that was used in the estimation of individual CEGA scores. Such a combined score, or ‘genome-wide transmissibility coefficient’, offers a quantitative estimate of changes in estimated transmissibility, under the assumption of no epistatic interactions between any of the mutations and deletions considered.

To estimate a multi-locus CEGA score we considered all recurrent mutations and deletions which passed our filters, then combined them in a multiplicative fashion into a ‘poly-CEGA’ score for each genome (see Methods). Initial inspection of the poly-CEGA scores over a representative sample of the dataset highlights that the transmissibility of SARS-CoV-2 has altered in a stepwise manner over the COVID-19 pandemic so far (Figure 4), with an initial jump following the acquisition of 614G and its associated haplotype increasing the CEGA transmissibility coefficient by ~0.5 and a second increase in tandem with the emergence of 501Y VoCs by ~1.2. It is clear that the recent emergence of 501Y.v1 has resulted in a marked shift in the poly-CEGA scores of SARS-CoV-2 genomes towards increased estimated transmissibility, with 501Y.v1 now making up a very large fraction of global sequence datasets (41% of the samples in our phylogeny). As expected, the highest positive CEGA scores we estimate are enriched for those present within 501Y.v1 (Figure 3B).

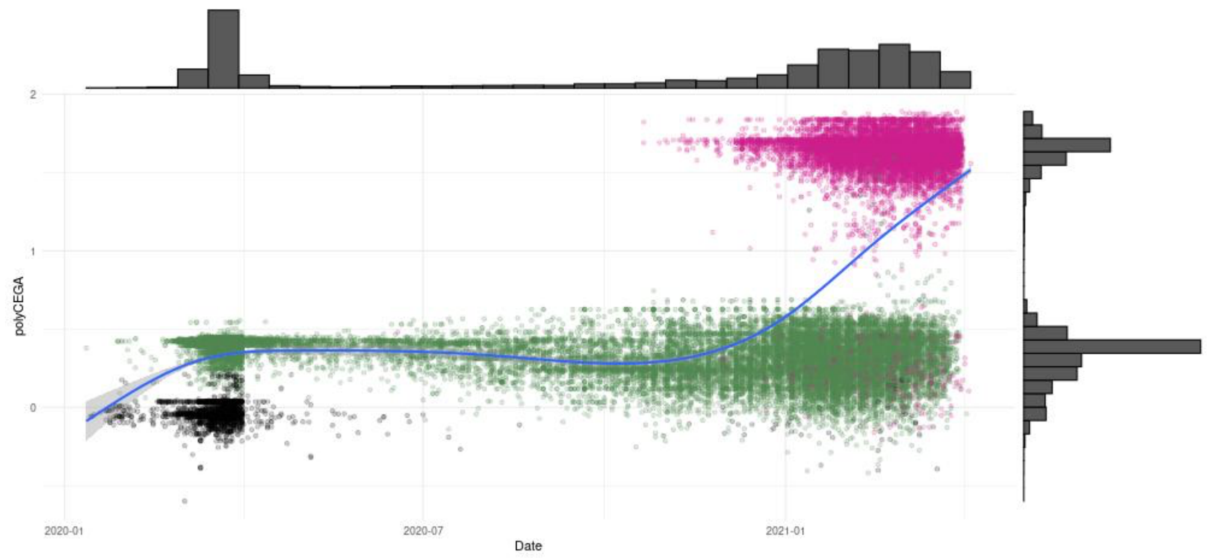

Figure 4. Poly-CEGA values (y-axis) of all available assemblies from December 2019 to April 2020 (x-axis) and a random subset from April 2020 onwards. Points coloured in black carry the ancestral spike mutation 614D with those carrying 614G coloured in dark green. Those points coloured pink label genomes carrying both 614G and 501Y.

To assess whether the poly-CEGA framework recovers relative differences in estimated transmissibility between SARS-CoV-2 clades, we computed poly-CEGA estimates for genomes assigned to 12 different clades defined by NextStrain[4]. This nomenclature defines clades which reached a global frequency of >20% for two or more months, while also including all current 501Y VoCs (https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming, accessed 13 April 2021). Calculating poly-CEGA scores over all major clades, we found the majority (9 out of 12) have an average poly-CEGA score between 0.181 and 0.442 including two of the three 501Y VoCs: 501Y.v2 and 501Y.v3. The two earliest SARS-CoV-2 clades 19A and 19B exhibited the lowest poly-CEGA scores of -0.013 and -0.085, the latter perhaps reflecting the carriage of the 614D ancestral haplotype. By far the highest estimates were for the 501Y.v1 clade (mean poly-CEGA=1.64), which exhibited poly-CEGA scores significantly different from all other SARS-CoV-2 clades (Wilcoxon *p*<1x10-6) (Figure 4, Supplementary Figures S14-S25). Interestingly we obtain consistently high estimates for clade 20G, relative to others excluding 501Y.v1, largely comprised of genomes assigned to PANGO lineage B.1.2, circulating widely in the USA at low prevalence (https://cov-lineages.org/lineages/lineage_B.1.2.html). We also observe that the third highest estimated clade transmissibility is obtained for the VoC 501Y.v3 (P.1), which has been estimated to be more transmissible, though not to the same extent as 501Y.v1[27].

These results suggest that the poly-CEGA scores can uncover relative variation in transmissibility among major lineages without the need to invoke strong non-independence between individual mutations and deletions. We note, however, that our method requires observations (emergences) of mutations present in a clade of interest to fully capture relative patterns of transmissibility. For instance, we re-computed the poly-CEGA estimates for 501Y.v1 based on the analysis fully excluding representatives of 501Y.v1 and were unable to recover the marked relative transmissibility advantage of 501Y.v1 compared to other clades (Supplementary Figure S26). This may largely be because many of the mutations associated to high CEGA scores present predominately in 501Y.v1 (Figure 3b) do not pass filters for computing individual CEGA scores in the reduced dataset, most notably the nucleocapsid triple mutation 28280:GAT>CTA. However, we note that the poly-CEGA of 501Y.v1, treated in this manner as an unseen clade, is significantly greater than NextStrain clade 20E which was dominant in the UK prior to the emergence of 501Y.v1 (Supplementary Figure S26, 20E versus B.1.1.7; mean poly-CEGA 0.13 vs. 0.15, Wilcoxon p< 2e-16).

Having explored trends in estimated transmissibility over the global dataset (Figure 4) we next asked whether there are notable trends in poly-CEGA scores within major phylogenetic clades. Within defined SARS-CoV-2 clades, application of a simple linear model (poly-CEGA estimates against sampling time) reveals that in all cases aside from 19A and 20F (for which the relationship is not significantly different from a slope of zero), we recover a tendency for the average poly-CEGA score to slightly decrease with time (Supplementary Table S4). The relative size of this effect compared to the 'initial' increased transmissibility of a lineage is statistically significant but small (Figure 5C). Considering the ratio of positive to negative individual CEGA scores, in 9/12 clades we identify a tendency towards accumulation of mutation which are associated to reduced estimated transmissibility; more negative scores (Figures S14-S25). Such an observation is consistent with the accumulation of weakly deleterious mutations over time.

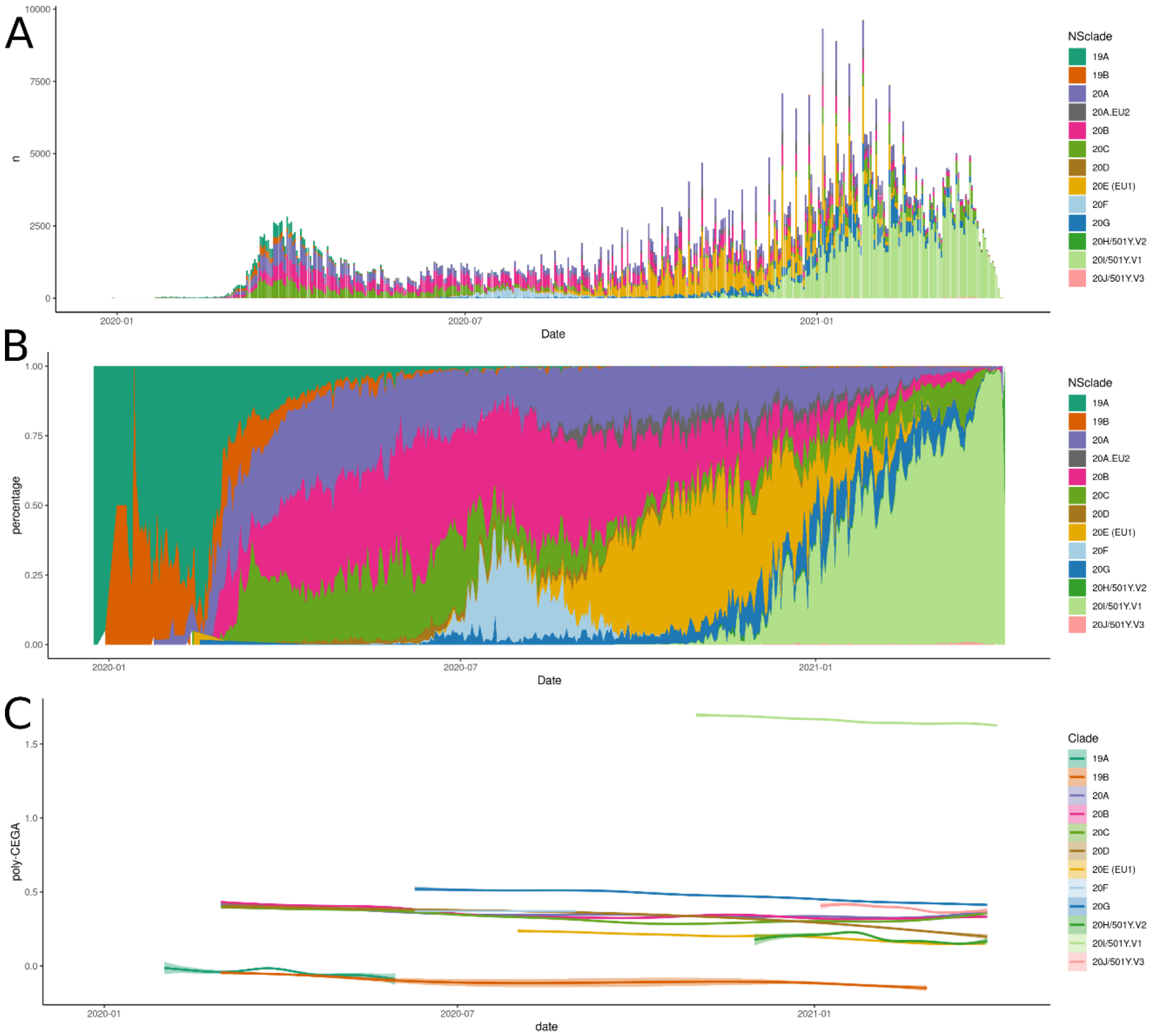

Figure 5. SARS-CoV-2 clades, defined by NextStrain, for genome assemblies available over the course of the pandemic dating from late December 2019 through to April 2021. (A) Histogram provides the daily number of sequenced SARS-CoV-2 samples from the GISAID database coloured by NextStrain clade, as given by the legend at right. (B) Daily frequency of each NextStrain clade estimated as the proportion of uploads to the GISAID genome database. (C) Smoothed curves providing the temporal evolution of the poly-CEGA score – y axis (multiplicative assessment of the transmissibility of a genome from the mutations and deletions it carries) obtained for all genomes of each NextStrain clade for which more than 50 genome assemblies have been shared per month. The underlying values, per genome, are visualised in Supplementary Figures S14-S25. All three panels use the same colour code corresponding to the NextStrain clade assignment.

Discussion/Conclusion

During the early stages of the COVID-19 pandemic the evolution of SARS-CoV-2 seemed largely neutral. Apart from the rapid emergence and diffusion of the D614G haplotype[24], little evidence for the emergence of viral lineages differing in their transmissibility or virulence was found[9,12]. This period of relative evolutionary stasis came to an end with the almost-concurrent detection of SARS-CoV-2 Variants of Concern (VoCs) towards the end of 2020. The three 501Y VoCs, each of which emerged independently in a different region, harbour strikingly similar patterns of mutations and deletions. Such recurrently emerging changes are strong candidates for convergent evolution towards altered transmissibility.

In this work, we have estimated the number of emergences of more than 1,500 recurrent mutations and deletions detected in over half a million SARS-CoV-2 genomes sampled between December 2019 through to April 2021. Focusing on a well-characterised subset of these recurrent mutations and deletions, we assessed their association to the estimated transmissibility of SARS-CoV-2. We applied a phylogeny-based scoring metric which measures the relative number of progeny descending from nodes harbouring a specific, newly acquired change. The Coefficients of Exponential Growth Alteration (CEGA) scoring index captures key components of SARS-CoV-2 evolution, including the dominance of C→T changes and the weakly deleterious overall effect of many mutations, in part driven by synonymous C→T changes. While the effect of mutation carriage on estimated transmissibility appears modest for the majority of our tested sites, we identified a subset of recurrent changes giving rise to highly positive scores. Of note, amongst the mutations most highly associated to increased transmissibility, we identify D614G in the spike protein with accompanying synonymous substitution at nucleotide position 3037, N501Y in the spike protein and NSP6 Δ106-108, changes found in all VoCs identified at this stage; together with D3L in the nucleocapsid, a mutation found predominately in 501Y.v1.

Our phylogeny-based metric provides estimates of the associations between all high frequency recurrent mutations and estimated transmissibility. While several of the mutations for which we estimated the highest scores have known phenotypic effects, this was not ubiquitously the case, suggesting the value of our approach to identify sites to take forward for further functional characterisations and to allow rapid assessment of combinations of mutations as they may appear in different backgrounds. Indeed, we found that when estimating the transmissibility of any SARS-CoV-2 genome based on its genomic makeup, poly-CEGA scores performed well in recovering key aspects of the SARS-CoV-2 phylogeny, such as the acquisition of D614G and associated changes (241, 3037, 14408) and the global increase in estimated transmissibility following the emergence of 501Y.v1 (Figure 4).

However, we can only speculate on the possible functional mechanisms underlying enhanced transmission. One hypothesis relates to the relevance of receptor binding affinity[35]. For example, 501Y has a high ACE2 binding affinity and consistently yields amongst the highest estimated transmissibility advantage in our analysis. A greater ability to bind to the receptor may result in SARS-CoV-2 harbouring 501Y requiring a lower effective dose of virions to initiate a successful infection[58]. An alternative, non-mutually exclusive mechanism for greater infective potential is the ability to escape neutralising antibodies primed by prior natural infection or vaccination, a situation which likely holds true for spike mutations at E484K and K417N/K417T[59]. A transmission increase conveyed by the ability to bypass host immunisation will depend on the rate of vaccination and prior infection of the host population. Thus, contrary to mutations providing an intrinsic transmission advantage, those allowing to bypass host immunisation may only be selected in host populations with significant immunity. Such variable selective pressure may contribute to the intriguing patterns we observe. For example, while the E484K mutation has emerged many times, our results do not indicate that it contributes to an on average increase in estimated transmissibility. Similarly, while K417N/T have also emerged multiple times (Supplementary Table S1), an insufficient number of those emergences passed our filtering criteria to enable calculation of a robust CEGA score.

Our results also highlight the importance of looking beyond the RBD of the spike protein for putative adaptive and recurrent changes. Indeed, we observe the greatest density of recurrent mutations within ORF3a, the nucleocapsid protein, and the NTD of the spike protein; the latter being a highly diverse genomic region across Sarbecoviruses[60] and a hotspot of antigenic evolution in human endemic coronaviruses[61,62]. Of particular note is the CEGA estimate for NSP6 Δ106-108 falling within the highest 5% of scores amongst all tested sites, since this deletion is found in all three 501Y VoCs and some VoIs. While the exact function of this deletion remains to be determined, NSP6 has been shown to play an important role in innate immune response, suppressing IFN-I response[63]. Consistent with this finding, it has been proposed that evasion of innate immunity may be a significant driving force underlying SARS-CoV-2 evolution[64].

A further intriguing outcome of extensive genomic surveillance efforts of SARS-CoV-2 is the marked patterns of lineage dynamics. Many clades are now essentially extinct, with others having been through fluctuations or rising/falling rapidly in frequency (Figure 5). Our poly-CEGA estimates across SARS-CoV-2 clades recovered a slight but significant decrease in estimated transmissibility over time for 10 of 12 clades analysed, largely corresponding to the accumulation of mutations associated to lower CEGA scores. Such an observation is consistent with the expectation of a decay in transmissibility for non-recombining lineages caused by the accumulation of deleterious mutations (often referred to as 'Muller's ratchet')[65,66]. Given we tend to observe significantly lower transmissibility estimates at C→T sites, one plausible contributing factor is the accumulation of mutations over time due to host-editing activity, with deamination of cytidines being a hallmark of the APOBEC family of proteins[67]. However, such a small and gradual decay in transmissibility of existing lineages is only part of the story. Lineage replacement dynamics are clearly driven by a combination of other phenomena, including the 'sudden' emergence of more transmissible lineages (such as 501Y.v1) or the impact of interventions on certain geographically restricted lineages. For instance, the 20F clade, concentrated in Australia during July and August 2020 (https://cov-lineages.org/lineages/lineage_D.2.html), exhibited no such decline in estimated transmissibility in our analysis (Supplementary Table S4, Supplementary Figure S21, Figure

5), possibly because this lineage was rapidly curtailed due to strict non-pharmaceutical interventions.

Our analyses take advantage of the unprecedented size of the GISAID dataset[5,6] thanks to the efforts of large numbers of contributing laboratories generating and openly sharing data, and aided by release of global phylogenies via the GISAID Audacity platform. While the data should be fairly representative of the extant diversity of SARS-CoV-2 in circulation globally, the dataset is still affected by significant sampling heterogeneity. This arises primarily through variable sequencing efforts by different countries, and to a lesser extent from targeted sequencing of VoIs after their identification. However, we reason that our approach should be largely robust to such sampling heterogeneities. Indeed, we rely on multiple replicates provided by the independent emergences of mutations and deletions in different genetic backgrounds. The fact that our results remain largely unaffected whether we compute our transmissibility scores using the mean or the median of the CEGA values obtained from all testable independent emergences of each mutation or deletion suggest our method produces well-behaved scores. In addition, our estimates of individual CEGA scores when excluding 501Y.v1 from the analysis remain highly correlated, suggesting the dominance of this clade in recent collections does not significantly impact our ability to estimate CEGA scores for overlapping mutations found in a variety of genetic backgrounds.

Recurrent mutations may also be detected erroneously in poorly resolved phylogenies. Uncertainty in the topology of SARS-CoV-2 phylogeny is difficult to assess and quantify because the limited available genetic diversity implies that most of the nodes of the phylogeny are supported by a sole SNV. Despite these characteristics, deep nodes of the phylogeny appear stable across trees produced by many different methods[11]. The use of parsimony to place new sequences into existing trees[52], which is used by the GISAID Audacity platform, is estimated to place 85% of SARS-CoV-2 samples correctly with the initial placement, and the Audacity pipeline further optimises these placements using pseudo-Maximum Likelihood implemented in FastTree2[68]. In order to bias the CEGA score of any given mutation, errors in the topology would have to cause strains bearing the mutation in question to be systematically misplaced by groups of at least ten (one of our filtering criteria) in lineages having a sister lineage not bearing the mutation. Given these parameters, and the results from our simulations (Supplementary Figure S12), our analysis should be robust to most phylogenetic uncertainty, because such uncertainty would tend to mix samples from related lineages and cause them to fail our filtering criteria. Low levels of recombination may also result in the detection of recurrent mutations, possibly providing a valid replicate for assessing the transmissibility effect of the mutation. Such cases would still provide an assessment of the effect of that mutation in the recombinant lineage background, with our method largely agnostic to the mechanisms underlying how a recurrent mutation arises.

Transmissibility is a complex picture and here we have analysed one aspect only: the contribution from recurrent mutations and deletions. Although we expect our scores to be relatively robust to many sources of bias, it is important to stress that the transmissibility estimates for individual mutations and deletions, as well as for whole genomes, represent relative rather than absolute transmissibility estimates (i.e. 'additional' changes in transmission). The values are also affected by choices during the normalisation procedure, including the value selected for generation time. They cannot therefore be simply compared to transmissibility estimates deriving from other approaches. A strong assumption of our method is that replicate observations at different nodes of the phylogeny provide a consistent assessment of the transmissibility differential putatively conferred by a mutation. This may not be the case. Our current approach is bound to conflate intrinsic differences in transmissibility (provided e.g. by variation in viral load) with the potential of viral lineages to bypass host immunisation. However, higher transmissibility provided by immune escape is dependent on the immunisation levels of host populations, making the transmissibility effect of an escape mutation variable in time and space. This might in part explain the lack of detection of increased transmissibility estimates for the 501Y.v2 (B.1.351) and 501Y.v3 (P.1) clades based on our current approach (Figure 5). In addition, the consistent but slight decrease in poly-CEGA scores over time could conceivably arise from a systematic bias; although we have not been able to identify one, we are actively developing the method and welcome comments.

In summary, we herein make use of an extensive genomic dataset of SARS-CoV-2 to assess the contributions of mutations and deletions to the estimated transmissibility of SARS-CoV-2 through time. The per-generation scoring metric we developed highlights a transmissibility increase associated to mutations and deletions of the spike protein that were previously known or suspected to affect transmissibility, but also sheds light on the potential role of mutations in other proteins. More fundamentally, generation of genome-wide estimates of transmissibility based on the individual contributions of mutations and deletions allows recovery of marked selective shifts in the transmissibility of SARS-CoV-2, with 501Y.v1 the most transmissible clade to date. In addition, we identify a tendency for phylogenetic clades to slightly decline in transmissibility through time after their emergence through the accumulation of mutations estimated to be weakly deleterious with respect to transmissibility. Such a trend may, in part, contribute to the patterns of lineage dynamics observed over the COVID-19 pandemic thus far.

Methods

Data acquisition

We downloaded the SARS-CoV-2 phylogeny provided by GISAID[5,6] to registered users for data current to 15/04/2021 via the Audacity feature which is built following the pipeline presented at https://github.com/roblanf/sarscov2phylo. We restricted our study to the assemblies available in the Mafft-based multiple sequence alignment available on the GISAID EpiCoV™ database (https://www.gisaid.org) already filtered to exclude samples displaying a total genome length <= 29,000bp, a fraction of 'N' nucleotides >5%, or displaying long series of leading or trailing 'N' nucleotides. We additionally discarded strains displaying spurious alleles at positions of known nucleotide deletions: Δ686-694; Δ1,605-1,607; Δ11,288-11,296; Δ21,765-21,770; Δ21,991-21,993 (corresponding to NSP1 Δ141-143; NSP2 Δ267-268; NSP6 Δ106-108; S Δ69-70 and S Δ144, respectively). All ambiguous sites in the alignment were set to 'N'. This resulted in an alignment and a maximum likelihood phylogeny both comprising 550,743 SARS-CoV-2 assemblies for downstream analysis. A full metadata table, acknowledgement of data generators and submitters and accession exclusions is provided in Supplementary Table S5.

Detection of recurrent mutational events

We filtered the alignment to only include variable positions for which the most common alternate allele was present in at least 0.1% of accessions. The sometimes heterogeneous quality of assemblies submitted to GISAID poses challenges to automated calling of deletions. We therefore restricted our study to the five deletions listed above and checked them manually in the alignment. We additionally masked sequencing error prone and/or back-mutation prone nucleotide positions 11,083 and 21,575[11] as well as the first 150 and last 300 bp because of their higher proportion of missing data. We applied HomoplasyFinder v0.0.0.9[69] to this alignment and the GISAID Audacity tree to quantify the number of independent emergences of all mutations and deletions considered and to identify the parental node of every recurrent mutation/deletion in the dataset. This resulted in detection of 1,552 total homoplastic positions. We discarded all homoplasies for which we did not have at least five independent emergences supported by nodes obeying the following rules: at least ten descending tips carrying either allele and with no children node embedded carrying a subsequent mutation at the same site. We previously showed that such filtering was necessary to obtain robust and symmetrically distributed scores[9]. The homoplasy detection and filtration procedure has been thoroughly described in our first implementation of a Ratio of Homoplastic Offspring (RoHO) scoring method (see Methods in van Dorp and Richard et al 2020). Following each of these steps 625 homoplasies passed our filtering criterion. While this does not include all mutations flagged as of concern, for instance in known VoCs and VoIs, further mutations may pass our stringency filters as additional genomes are sequenced.

Computation of the Coefficients of Exponential Growth Increase (CEGAs)

Our previous work[9] considered the ratio of homoplastic offspring ('RoHO') of two lineages descending from a given ancestor (the number of genomes carrying and not carrying the considered mutation). While this RoHO statistic can capture accurately the increased success of a lineage compared to its sister lineage, it fails to account for the fact that under an epidemic growth model, the RoHO is not expected to be constant over time. We herein aimed for a metric that considers variation through time. The approach, which we term the Coefficients of Exponential Growth Alteration (CEGA) aims to normalise the excess of the number of offspring per generation under exponential growth (Supplementary Figure S5).

Consider a recurrent mutation or deletion *G*. We assume that the effect of *G* is to increase the growth rate to *r*+*s* compared to a basal growth rate *r* for lineages without *G*. This means that *s* is the additive growth rate due to *G* (units of per time), which can be positive or negative. When *s*>0, the mutated lineage produces a larger number of offspring so RoHO will grow with time (and vice versa for *s*<0).

We assume that the ratio $\alpha(t)$ of the mutated lineage to the sister lineage at a time *t* after the emergence of *G* at the ancestral homoplastic node can be expressed as the ratio of exponential growth of the two lineages:

$$\alpha(t) = \frac{\alpha_0 e^{(r+s)t}}{e^{rt}} = \alpha_0 \times e^{st}$$

Here α_0 denotes the ratio of the initial numbers of infections of the mutated lineage to the sister lineage, and *t* is the number of generations separating the homoplastic node and the time of observation. This assumes that generations do not overlap, that all individuals are sampled at the same final time *t*, and that *r* remains constant over the time period considered. These simplifying assumptions are reasonable since we are considering short periods of time near the tips of the phylogeny. Rearranging, we have

$$s = \frac{1}{t}\ln[\frac{\alpha(t)}{\alpha_0}]$$

From the phylogeny of observed genomes, we can compute the ratio of homoplastic offspring (RoHO: ρ). We assume this is an approximation to $\frac{\alpha(t)}{\alpha_0}$, giving us

$$s \approx \frac{\ln(\rho)}{t}$$

We approximate *t* as t^{days} / gen^{days} where t^{days} is the timespan between the oldest and the youngest offspring of the considered node and gen^{days} the generation interval. We use five days for gen^{days} based on the mean and median estimates provided by four independent studies[70–73]. Values of *s* ('CEGA scores') presented in the manuscript are therefore given per generation interval. We obtain multiple CEGA scores for a given mutation or deletion from different homoplastic nodes in the phylogeny. We summarise these using the mean and median of these CEGA scores (provided in Supplementary Table S3). Averaging *s* over different nodes in this way implicitly assumes that each mutation or deletion has a consistent effect in time and space.

Our aim here is to attempt to compare the effect of different recurrent mutations or deletions in a relative sense. We stress that different assumptions could give rise to different expressions for s.

Estimation of the transmissibility of major SARS-CoV-2 clades

Our method estimates a CEGA value for each recurrent mutation and deletion under scrutiny (passing filters) in the alignment. It follows that the estimated transmissibility of any individual genome is influenced by the multiple mutations or deletions it carries. To test a multilocus model we estimated the transmissibility score for each genome in our dataset which we term poly-CEGA. To do so we employed a simple model, where the effect of mutations and deletions on transmissibility are considered as independent, by computing the sum of mean CEGA scores of a genome made up of mutations $1 \ldots n$ as poly-CEGA= $\sum_{i=1}^{n} CEGA_i$, where sites for which no CEGA score was computed (did not pass our filters) are considered as having a CEGA of 0. By extension, we assessed the combined transmissibility change estimate– the poly-CEGA – of all SARS-CoV-2 isolates assessed over the global population and by NextStrain clades.

Influence of phylogenetic misplacement and sample representation on CEGA estimates

Given recurrent mutations may be observed due to poor phylogenetic placement of accessions, we assessed the influence of phylogenetic uncertainty on the CEGA scores recovered. We applied our method to a subsampled dataset of 130,000 assemblies selected at random from the global dataset. The Audacity phylogeny provided by GISAID was subsampled accordingly, but with random misplacement of 0, 1, 2, 3, 5 and 10% of the 130,000 tips (Figure S12). In addition, to assess the robustness of our scores to the high representation of 501Y.v1 in our dataset, we applied the same CEGA scoring pipeline as implemented on the global dataset after removal of all 501Y.v1 genomes, resulting in a CEGA value for 308 of the 652 recurrent mutations and deletions that could be scored on the whole dataset (Figure S13).

Acknowledgements

This analysis was enabled by the GISAID EpiCoV™ database (https://www.gisaid.org) and Audacity releases. We gratefully acknowledge all contributing and submitting laboratories around the globe including the COG-UK consortium (https://www.cogconsortium.uk) who have openly shared large numbers of UK SARS-CoV-2 assemblies. A full list of acknowledgements of submitting and originating laboratories is provided in Supplementary Table S5. We are also grateful to developers and maintainers of CoVariants (https://covariants.org) and Outbreak.info (https://outbreak.info) for information on mutation and deletion carriage in VoCs and VoIs.

References

1. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
2. Kumar, S. *et al.* An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. *bioRxiv* (2020) doi:10.1101/2020.09.24.311845.
3. Rambaut, A. Phylogenetic analysis of 23 nCoV-2019 genomes, 2020-01-23 - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological. *Virological* https://virological.org/t/phylogenetic-analysis-of-23-ncov-2019-genomes-2020-01-23/335 (2020).
4. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
5. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
6. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46 (2017).
7. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. (2020) doi:10.20944/PREPRINTS202006.0225.V1.
8. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science (80-.).* **371**, 1139–1142 (2021).
9. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
10. Kemp, S. *et al.* Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. *bioRxiv* 2020.12.14.422555 (2020) doi:10.1101/2020.12.14.422555.
11. Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLOS Genet.* **16**, e1009175 (2020).
12. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv* (2021) doi:10.1101/2021.02.23.21252268.
13. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276 (2008).
14. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, (2020).
15. Giorgio, S. Di, Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, eabb5813 (2020).
16. De Maio, N. *et al.* Mutation rates and selection on synonymous mutations in SARS-CoV-2. *bioRxiv Prepr. Serv. Biol.* 2021.01.14.426705 (2021) doi:10.1101/2021.01.14.426705.
17. Mourier, T. *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.* (2020) doi:10.1016/j.bbrc.2020.10.092.
18. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* (2021) doi:10.1126/science.abg0821.
19. Richard, D., Owen, C. J., Dorp, L. van & Balloux, F. No detectable signal for ongoing

genetic recombination in SARS-CoV-2. *bioRxiv* 2020.12.15.422866 (2020) doi:10.1101/2020.12.15.422866.

20. VanInsberghe, D., Neish, A. S., Lowen, A. C. & Koelle, K. Recombinant SARS-CoV-2 genomes are currently circulating at low levels. *bioRxiv* (2021).
21. Varabyou, A., Pockrandt, C., Salzberg, S. L. & Pertea, M. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *bioRxiv* 2020.09.21.300913 (2020) doi:10.1101/2020.09.21.300913.
22. Jackson, B. *et al.* Recombinant SARS-CoV-2 genomes involving lineage B.1.1.7 in the UK - SARS-CoV-2 coronavirus / SARS-CoV-2 Molecular Evolution - Virological. *Virological* https://virological.org/t/recombinant-sars-cov-2-genomes-involving-lineage-b-1-1-7-in-the-uk/658 (2021).
23. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
24. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).
25. Rambaut, A. *et al. 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 (2020).
26. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* 2020.12.21.20248640 (2020) doi:10.1101/2020.12.21.20248640.
27. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science.* eabh2644 (2021) doi:10.1126/science.abh2644.
28. Tchesnokova, V. *et al.* Acquisition of the L452R mutation in the ACE2-binding interface of Spike protein 1 triggers recent massive expansion of SARS-Cov-2 variants 2. *bioRxiv* 2021.02.22.432189 (2021) doi:10.1101/2021.02.22.432189.
29. Zhang, W. *et al.* Emergence of a novel SARS-CoV-2 strain in Southern California, USA. *medRxiv* 2021.01.18.21249786 (2021) doi:10.1101/2021.01.18.21249786.
30. Piccoli, L. *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024-1042.e21 (2020).
31. Voloch, C. M. *et al.* Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **95**, (2021).
32. Kemp, S. A. *et al.* Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion. *medRxiv* 2020.12.05.20241927 (2020) doi:10.1101/2020.12.05.20241927.
33. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912.e9 (2020).
34. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* NEJMc2031364 (2020) doi:10.1056/NEJMc2031364.
35. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
36. Zahradník, J. *et al.* SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv* 2021.01.06.425392 (2021) doi:10.1101/2021.01.06.425392.

37. Nelson, G. *et al.* Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escap. *bioRxiv* 2021.01.13.426558 (2021) doi:10.1101/2021.01.13.426558.
38. Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv* 2020.12.30.20249034 (2021) doi:10.1101/2020.12.30.20249034.
39. Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2372-2383.e9 (2021).
40. Cele, S. *et al.* Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* **593**, 142–146 (2021).
41. Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* 1–4 (2021) doi:10.1038/s41591-021-01285-x.
42. Wu, K. *et al.* mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. *bioRxiv Prepr. Serv. Biol.* 2021.01.25.427948 (2021) doi:10.1101/2021.01.25.427948.
43. Hoffmann, M. *et al.* SARS-CoV-2 variants B.1.351 and B.1.1.248: Escape from therapeutic antibodies and antibodies induced by infection and vaccination. *bioRxiv* 2021.02.11.430787 (2021) doi:10.1101/2021.02.11.430787.
44. Bager, P. *et al.* Increased Risk of Hospitalisation Associated with Infection with SARS-CoV-2 Lineage B.1.1.7 in Denmark. *SSRN Electron. J.* (2021) doi:10.2139/ssrn.3792894.
45. Davies, N. G. *et al.* Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* 1–5 (2021) doi:10.1038/s41586-021-03426-1.
46. Graham, M. S. *et al.* Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Heal.* (2021) doi:10.1016/s2468-2667(21)00055-4.
47. Santos De Oliveira, M. H., Lippi, G. & Henry, B. M. Sudden rise in COVID-19 case fatality among young and middle-aged adults in the south of Brazil after identification of the novel B.1.1.28.1 (P.1) SARS-CoV-2 strain: analysis of data from the state of Parana. *medRxiv* 2021.03.24.21254046 (2021) doi:10.1101/2021.03.24.21254046.
48. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science (80-.).* **10**, eabf2946 (2021).
49. Duchêne, S. & Lanfear, R. Phylogenetic uncertainty can bias the number of evolutionary transitions estimated from ancestral state reconstruction methods. *J. Exp. Zool. Part B Mol. Dev. Evol.* **324**, 517–524 (2015).
50. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171-1187.e20 (2021).
51. Hodcroft, E. B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv* vol. 12 16 (2020).
52. Yatish Turakhia, *et al.* Title: Ultrafast Sample Placement on Existing Trees (UShER) Empowers Real-Time Phylogenetics for the SARS-CoV-2 Pandemic. *bioRxiv* 2020.09.26.314971 (2020) doi:10.1101/2020.09.26.314971.
53. Alaa Abdel Latif *et al. N:D3L Mutation Report Outbreak.info*. (2021).
54. Matthew Parker, A. D. *et al.* Altered Subgenomic RNA Expression in SARS-CoV-2 B.1.1.7. *bioRxiv* 2021.03.02.433156 (2021) doi:10.1101/2021.03.02.433156.

55. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
56. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
57. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
58. Hasegawa, K. *et al.* Affinity Thresholds for Membrane Fusion Triggering by Viral Glycoproteins. *J. Virol.* **81**, 13149–13157 (2007).
59. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44-57.e9 (2021).
60. Dicken, S. J. *et al.* Characterisation of B.1.1.7 and Pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. *bioRxiv* 2021.03.22.436468 (2021) doi:10.1101/2021.03.22.436468.
61. Eguia, R. *et al.* A human coronavirus evolves antigenically to escape antibody immunity. *PLOS Pathog.* **17**, e1009453 (2021).
62. Kistler, K. E. & Bedford, T. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229E. *Elife* **10**, 1–35 (2021).
63. Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep.* **33**, 108234 (2020).
64. Guo, K., Barrett, B. S., Mickens, K. L., Hasenkrug, K. J. & Santiago, M. L. Interferon Resistance of Emerging SARS-CoV-2 Variants. *bioRxiv* 2021.03.20.436257 (2021) doi:10.1101/2021.03.20.436257.
65. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **1**, 2–9 (1964).
66. Felsenstein, J. The evolution advantage of recombination. *Genetics* **78**, 737–756 (1974).
67. Bishop, K. N., Holmes, R. K., Sheehy, A. M. & Malim, M. H. APOBEC-mediated editing of viral RNA. *Science.* **305**, 645 (2004).
68. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
69. Crispell, J., Balaz, D. & Gordon, S. V. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb. genomics* **5**, (2019).
70. Griffin, J. *et al.* Rapid review of available evidence on the serial interval and generation time of COVID-19. *BMJ Open* **10**, 40263 (2020).
71. Ganyani, T. *et al.* Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* **25**, 1 (2020).
72. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science.* **368**, (2020).
73. Lehtinen, S., Ashcroft, P. & Bonhoeffer, S. On the relationship between serial interval, infectiousness profile and generation time. *J. R. Soc. Interface* **18**, 20200756 (2021).