No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2

Lucy van Dorp[1*+]
Damien Richard[2,3*]
Cedric CS. Tan[1]
Liam P. Shaw[4]
Mislav Acman[1]
François Balloux[1+]

[1] UCL Genetics Institute, University College London, London WC1E 6BT, UK
[2] Cirad, UMR PVBMT, F-97410 St Pierre, Réunion, France
[3] Université de la Réunion, UMR PVBMT, F-97490 St Denis, Réunion, France
[4] Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK

*contributed equally
+ corresponding; lucy.dorp.12@ucl.ac.uk (Lucy van Dorp) and f.balloux@ucl.ac.uk (François Balloux)

Abstract

The COVID-19 pandemic is caused by the coronavirus SARS-CoV-2, which jumped into the human population in late 2019 from a currently uncharacterised reservoir. Due to this extremely recent association with humans, SARS-CoV-2 may not yet be fully adapted to its human host. This has led to speculations that some lineages of SARS-CoV-2 may be evolving towards higher transmissibility. The most plausible candidate mutations under putative natural selection are those which have emerged repeatedly and independently (homoplasies). Here, we formally test whether any of the recurrent mutations that have been observed in SARS-CoV-2 to date significantly alter viral transmission. To do so, we developed a phylogenetic index to quantify the relative number of descendants in sister clades with and without a specific allele. We apply this index to a carefully curated set of recurrent mutations identified within a dataset of over 15,000 SARS-CoV-2 genomes isolated from patients worldwide. We do not identify a single recurrent mutation convincingly associated with increased viral transmission. Instead, recurrent SARS-CoV-2 mutations currently in circulation appear to be either neutral or weakly deleterious. These mutations seem primarily induced by the human immune system via host RNA editing, rather than being signatures of adaption to the novel human host. There is no evidence at this stage for the emergence of more transmissible lineages of SARS-CoV-2 due to recurrent mutations.

Keywords
Betacoronavirus; Homoplasies; Mutation; Phylogenetics; Transmission

Introduction

Severe acute respiratory coronavirus syndrome 2 (SARS-CoV-2), the causative agent of Covid-19, is a positive single-stranded RNA virus that jumped into the human population towards the end of 2019 [1-4] from a zoonotic reservoir [5]. Since then, the virus has gradually accumulated mutations leading to patterns of genomic diversity which can be leveraged to inform on the spread of the disease and to identify sites putatively under selection as SARS-CoV-2 may adapt to its new human host. Large-scale efforts from the research community during the ongoing Covid-19 pandemic have resulted in an unprecedented number of SARS-CoV-2 genome assemblies available for downstream analysis. To date (21 May 2020), the Global Initiative on Sharing All Influenza Data (GISAID) [6, 7] repository has nearly 20,000 complete high-quality genome assemblies available. This is being supplemented by increasing raw sequencing data available through the European Bioinformatics Institute (EBI) and NCBI Short Read Archive (SRA), together with data released by specific genome consortiums including COVID-19 Genomics UK (COG-UK) (https://www.cogconsortium.uk/data/). Research groups around the world are continuously monitoring the genomic diversity of SARS-CoV-2, with a focus on the distribution and characterisation of emerging mutations.

Mutations within coronaviruses, and indeed all RNA viruses, can arrive as a result of three processes. First, mutations arise intrinsically as copying errors during viral replication, a process which may be reduced in SARS-CoV-2 relative to other RNA viruses, due to the fact that coronavirus polymerases include a proof-reading mechanism [8, 9]. Second, genomic variability may also arise as a result of recombination or reassortment. Third, mutations can be induced by host RNA editing systems, which form part of natural host immunity [10-12]. While the majority of all mutations are expected to be neutral [13], some may be advantageous or deleterious to the virus. Mutations which are highly deleterious, such as those preventing virus host invasion, will be rapidly purged from the population; mutations that are only slightly deleterious may be retained, if only transiently. Conversely, neutral and in particular advantageous mutations can reach higher frequencies.

Mutations in SARS-CoV-2 have already been scored as putatively adaptive relying on a range of population genetics methods [1, 14-17], and there have been suggestions that some mutations are associated with increased transmission and/or virulence [14, 15]. Early flagging of such adaptive mutations could arguably be useful to control the pandemic. However, distinguishing neutral mutations, whose frequencies have increased through demographic processes, from those directly increasing the virus' transmission can be difficult [18]. The most plausible candidate mutations under putative natural selection are therefore those that have emerged repeatedly and independently within the global viral phylogeny. Such sites (homoplasies) may arise convergently as a result of the virus responding to adaptive pressures.

We previously identified and catalogued homoplasies in SARS-CoV-2 assemblies, of which approximately 200 could be considered as warranting further inspection following stringent filtering [1]. A crucial next step is to test the potential impact of homoplasies on transmission. For a virus, transmission fitness can be considered as a proxy for overall fitness [19, 20]. This transmission fitness can be estimated by the relative fraction of descendants produced by an ancestral virion genotype. Such an approach feels warranted, in this case, given the unprecedented number of available SARS-CoV-2 isolates and the lack of strong structure in the global distribution of genetic diversity caused by the large number of independent introduction and transmission events in most densely sampled countries [1]. The fairly homogeneous global distribution pattern of SARS-CoV-2 genetic diversity with 'everything being everywhere' limits the risk that a homoplasic mutation could be deemed to provide a fitness advantage to its viral

carrier simply because it is overrepresented, by chance, in regions of the world more conducive to transmission.

In this work, we make use of a larger, curated alignment comprising 15,691 SARS-CoV-2 assemblies to formally test whether any identified recurrent mutation is involved in altering viral fitness. We find that none of the recurrent SARS-CoV-2 mutations in circulation to date are associated with increased viral transmission. Instead, recurrent mutations seem to be primarily induced by host immunity through RNA editing mechanisms, and likely tend to be weakly deleterious to the virus.

Results

Global diversity of SARS-CoV-2

The global genetic diversity of 15,691 SARS-CoV-2 whole genome assemblies is presented as a maximum likelihood phylogenetic tree in **Figure 1A**. No assembly deviates by more than 29 SNPs from the reference genome, Wuhan-Hu-1, which is consistent with the relatively recent emergence of SARS-CoV-2 towards the latter portion of 2019 [1-5]. We informally estimate the mutation rate over our alignment to 9.7×10^{-4} substitutions per site per year, which is consistent with previous rates estimated for SARS-CoV-2 [1-4] (**Figure S1-S2**). This rate also falls in line with those observed in other coronaviruses [21, 22], and is fairly unremarkable relative to other positive single-stranded RNA viruses, which do not have a viral proof-reading mechanism [23, 24].

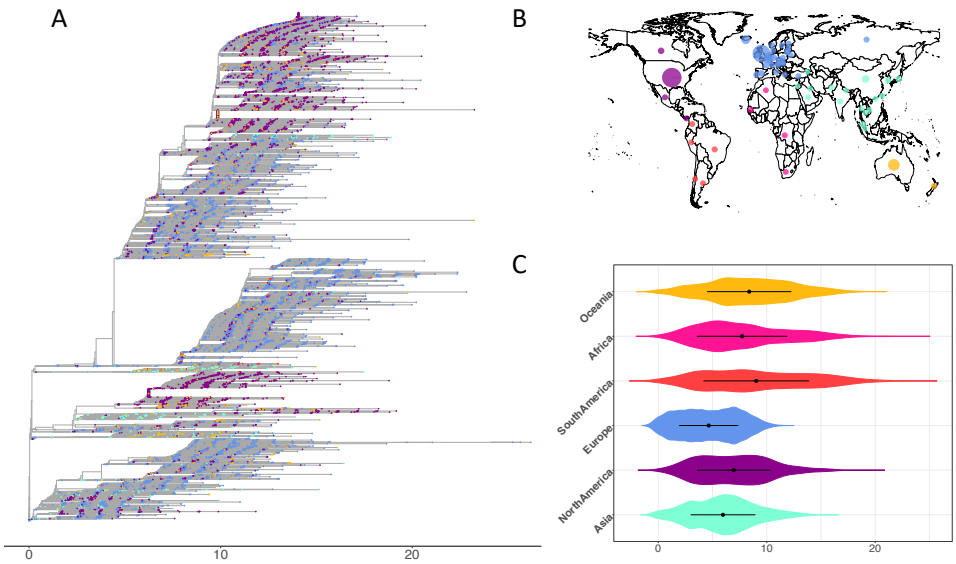

Figure 1 Overview of the global genomic diversity across 15,691 SARS-CoV-2 assemblies (sourced 15 May 2020) coloured as per continental regions**. A.** Maximum Likelihood phylogeny for complete SARS-CoV-2 genomes. **B.** Viral assemblies available from 75 countries. C. Within-continent pairwise genetic distance on a random subsample of 76 assemblies from each continental region to match the lowest number of samples (South America). Colours in all three panels represent continents where isolates were collected. Magenta: Africa; Turquoise: Asia; Blue: Europe; Purple: North America; Yellow: Oceania; Orange: South America according to metadata annotations available on GISAID (https://www.gisaid.org).

Genetic diversity in the SARS-CoV-2 population remains moderate with an average pairwise SNP difference across isolates of 9.62 (9-13). This low number of mutations between any two viruses currently in circulation means that, to date, SARS-CoV-2 can be considered as a single essentially clonal lineage, notwithstanding taxonomic efforts to categorise extant diversity into sublineages [25]. Our dataset includes viruses sequenced from 75 countries (**Figure 1B, Table S1**), with a good temporal coverage (**Figure S1B)**. While some countries are far more densely sequenced than others (**Figure 1B**), the emerging picture is that fairly limited geographic structure is observed in the viruses in circulation in any one region. All major clades in the global diversity of SARS-CoV-2 are represented in various regions of the world (**Figure 1A, Figure S3**), and the genomic diversity of SARS-CoV-2 in circulation in different continents is fairly uniform (**Figure 1C**).

Distribution of recurrent mutations

Across the alignment we detect 6,822 variable positions resulting in an observed genomewide ratio of non-synonymous to synonymous substitutions of 1.97 (calculated from **Table S2**). Following masking and homoplasy detection we observe over 2,000 homoplasic positions (2,200 and 2,133 respectively using two different masking criteria, see Methods) (**Figures S4-S5, Table S3**). However, recurrent mutations may arise as a result of sequencing or genome assembly artefacts. In line with our previous work ([1]; see Methods) we therefore applied a set of stringent filters to delineate a set of well supported homoplasies. This resulted in 273 and 238 homoplasies respectively (**Figures S4-S5, Table S3**).

The current distribution of genomic diversity identified across the alignment, together with identified homoplasic positions is available as an open access and interactive web-resource available at: https://macman123.shinyapps.io/ugi-scov2-alignment-screen/.

As identified by previous studies [26-31], we recover evidence of strong mutational biases across the SARS-CoV-2 genome. A remarkably high proportion of C→U changes was observed relative to other types of SNPs and this pattern was observed at both non-homoplasic and homoplasic sites (**Figures S6-S8**). Additionally, mutations involving cytosines were almost exclusively C→U mutations (98%) and the distributions of *k*-mers for homoplasic sites appeared markedly different compared to that across all variable positions (**Figures S9-S10**). In particular, we observe an enrichment in CCA and TCT 3-mers containing a variable base in their central position, which are known APOBEC targets [32].

Signatures of transmission

In order to test for an association between individual homoplasies and transmission, we defined a novel phylogenetic index designed to quantify the fraction of descendent progeny produced by any ancestral virion having acquired a particular mutation. We term this index the Ratio of Homoplasic Offspring (RoHO). In short, the RoHO index computes the ratio of the number of descendents in sister clades with and without a specific mutation over all independent emergences of a homoplasic allele (shown in red in **Figure 2**). Of our filtered list, we considered homoplasies determined to have arisen at least n=10 times independently, after discarding from this analysis the comparison of all clades including a secondary homoplasic emergence for the same allele and which had fewer than three descendants (**Figure 2**).

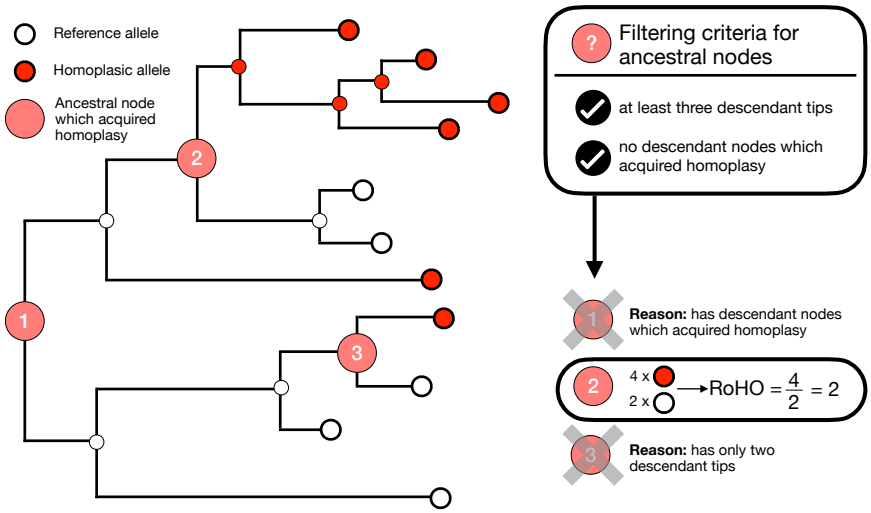

Figure 2 Schematic depicting the rationale behind the Ratio of Homoplasic Offspring (RoHO) scoring metric. White tips correspond to an isolate carrying the reference allele and red tips correspond to the homoplasic allele. This schematic phylogeny comprises three highlighted internal nodes annotated by HomoplasyFinder as corresponding to an ancestor that acquired a homoplasy. Node 3 is not considered because it fails our first criterion of having at least three descendant tips. Node 1 is not considered because it fails our second criterion of having no children nodes themselves annotated as carrying the homoplasy. Node 2 meet both our criterions: its RoHO score is 4/2 = 2. Our third quality criterion comprises two stringency levels and is not illustrated in the figure. In order to consider RoHO scores for an homoplasic position, it requires that at least n=5 or n=10 nodes satisfy the two first criteria.

None of the 31 detected recurrent mutations having emerged independently a minimum of ten times were statistically significantly associated with an increase in viral transmission (Sign test with Bonferroni correction for multiple testing; **Figure 3 and Table S4**). The only weak positive association with transmission was observed for the homoplasy at position 3037 (F924F; ORF1a) in the alignment (p=0.027; Sign test prior to correction for multiple testing). As the association is fairly weak and corresponds to a synonymous mutation in *nsp3*, this mutation is unlikely to represent a strong candidate for a mutation providing increased transmission.

Conversely, three recurrent mutations were significantly associated to decreased transmission (positions 11,083, 21,575, 26,681; p<0.05, Sign test corrected for multiple testing). The association to decreased transmission for the mutation at position 21,575 (L5F spike protein) is particularly noticeable (adjusted p-value=5.07×10^{-7}). This suggests that the L5F signal peptide mutation in the Spike protein is deleterious to the virus. Interestingly, there is a strong trend in the distribution of log(RoHO) indices towards negative values with 30/31 homoplasies with ≥10 comparisons falling below zero. This suggests that the majority of homoplasic mutations are weakly deleterious (**Figure 3**).

We also computed RoHO indices for 118 homoplasic mutations for which we only enforced a lower minimal number of five independent emergences. Even with this less stringent threshold, we did not identify any additional recurrent mutation statistically significantly associated with increased transmission (**Table S4**). The same analysis was also replicated on a smaller alignment of 3,441 genomes for which short-read sequences were available on the NCBI Sequence Read Archive (SRA) (**Table S6**). Analysis of the SRA assemblies yielded similar results, which were not statistically tested because no homoplasy had more than n=4 independent emergences in the dataset (**Figure S11, Table S4**).

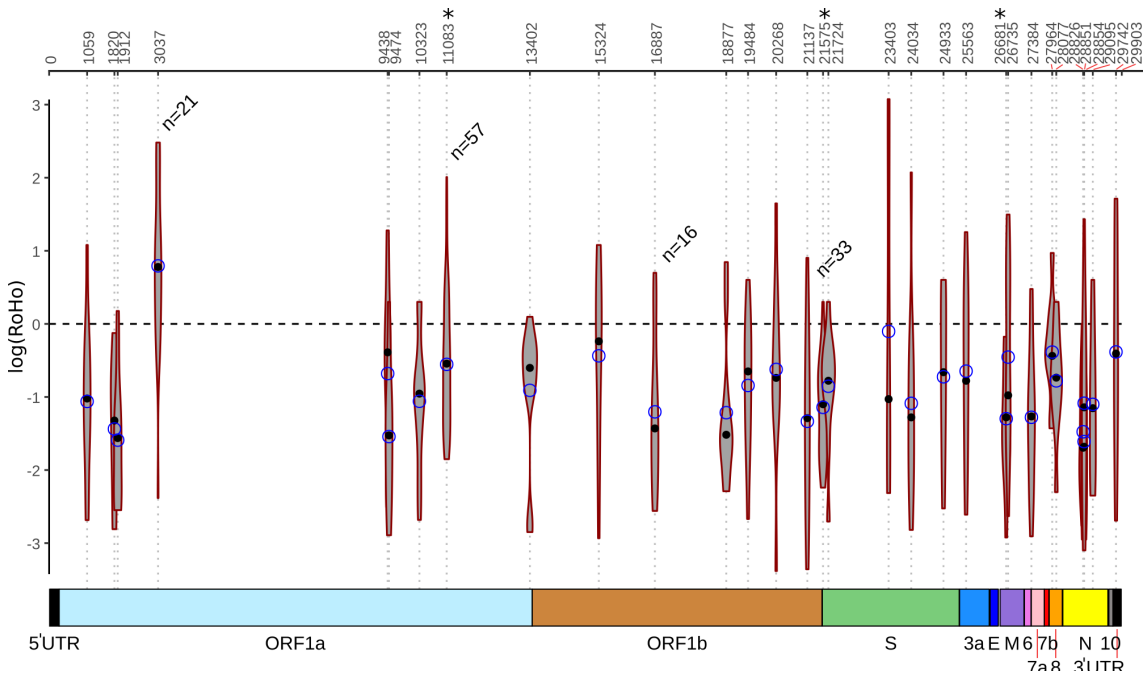

Figure 3. Genome-wide Ratio of Homoplasic Offspring (RoHO) scores. Violin plots show log10(RoHO) scores for homoplasies that arose in at least 10 filtered nodes in the Maximum Parsimony phylogeny of 12,626 SARS-CoV-2 isolates. Black dot: median RoHO value; blue circle: mean RoHO value. Number of replicates ("n=") values comprised between 10 and 15 were masked for readability. Top scale provides positions of the homoplasies on Wuhan-Hu-1 reference genome and the bottom coloured boxes correspond to encoded ORFs. Distribution of RoHO index for positions followed by a star (*) are significantly different from zero (Bonferroni corrected Sign test, alpha=0.05).

To summarise, we detected no convincing evidence for any of the recurrent mutations tested being associated to an increase in viral transmission. Conversely, there is a strong overall trend towards recurrent mutations being neutral or weakly deleterious, with a few mutations apparently significantly detrimental to SARS-CoV-2 transmissibility.

Discussion

In this work, we analysed a dataset of over 15,000 SARS-CoV-2 assemblies sampled across 75 different countries and all major continental regions. Current patterns of genomic diversity highlight multiple introductions in all continents (**Figure 1, Figures S1-S3**) since the host-switch to humans in late 2019 [1-4]. While SARS-CoV-2 still represents a single, essentially clonal lineage, the gradual accumulation of mutations in viral genomes in circulation may offer early clues to adaptation to its novel human host. Across our dataset we identified a total of 6,822 mutations, heavily enriched in C→U transitions, of which we identified 273 strongly supported recurrent mutations (**Table S3, Figures S4-S5**). Employing a newly devised index (Ratio of Homoplasic Offspring; RoHO) to test whether any of these mutations contribute to a change in transmission, we found no mutation was convincingly associated with a significant increase in transmission (**Figures 2-3, Table S4**).

Given the importance of monitoring putative changes in virus transmission, several studies have attempted to associate the presence of particular sets of mutations in SARS-CoV-2 to changes in transmission and virulence [14, 15]. We strongly caution that efforts to determine if any specific mutation contributes to a change in viral phenotype, using solely genomic approaches, relies on the ability to adequately distinguish between changes in allele frequency due to demographic or epidemiological processes, compared to those driven by selection [18]. A convenient and powerful alternative is to focus on sites which have emerged repeatedly

independent of the phylogeny (homoplasies), as we do here. While this is obviously restricted to recurrent mutations, it reduces the effect of demographic confounding problems such as founder bias.

A much discussed mutation in the context of demographic confounding is D614G, a nonsynonymous change in the SARS-CoV-2 Spike protein. D614G emerged early in the pandemic and is found at high frequency globally, with 4,744 assemblies carrying the associated mutation in the data we analysed (**Table S3**). Korber *et al*. suggest that D614G increases transmissibility, and reported experimental evidence consistent with higher viral loads but with no measurable effect on patient infection outcome [14]. In our analysis D614G (nucleotide position 23,403) has at least 12 independent emergences. However, in line with the vast majority of other recurrent mutations we analysed, it does not appear to be associated with increased viral transmission. Consistent with our findings for D614G, our results support a wider narrative where the vast majority of the nearly 7,000 mutations we detect in SARS-CoV-2 are either neutral or even weakly deleterious.

Notably 66% of the detected mutations comprise nonsynonymous changes of which 39% derive from C→U transitions. This high compositional bias, as also detected in other studies [29-31], as well as in other members of the Coronaviridae [26-28], suggests that mutations observed in the SARS-CoV-2 genome are not solely the result of errors by the viral RNA polymerase during virus replication [29, 30]. One possibility is the action of human RNA editing systems which have been implicated in innate and adaptive immunity. These include the AID/APOBEC family of cytidine deaminases which catalyse deamination of cytidine to uridine in RNA or DNA and the ADAR family of adenosine deaminases which catalyse deamination of adenosine to inosine (recognized as a guanosine during translation) in RNA [33, 34].

The exact targets of these RNA editing systems of the immune system of the host are not fully characterized but comprise viral nucleotide sequence target motifs whose editing may leave characteristic biases in the viral genome [32, 35, 36]. For example, detectable depletion of the preferred APOBEC3 target dinucleotides sequence TC have been reported in Papillomavirus [37]. In the context of SARS-CoV-2, Simmonds [30] and Di Giorgio et al. [29] both highlight the potential of APOBEC-mediated cytosine deamination as an underlying biological mechanism driving the over-representation of C→U mutations. However, APOBEC3 was shown to result in cytosine deamination but not hypermutation of HCoV-NL63 *in vitro* [38], which may suggest that additional biological processes play a role.

Our proposed RoHO index provides an intuitive metric to quantify the association between a given mutation and viral transmission. However, we acknowledge this approach has limitations. First, we have relied on admittedly arbitrary choices concerning the number of minimal observations and nodes required to conduct statistical testing. It seems unlikely this would change our overall conclusions, but results for particular mutations should be considered in light of this caveat. Second, our approach entails a loss of information and therefore statistical power. This is because our motivation to test *independent* occurrences means that we do not handle "nested homoplasies" explicitly: we simply discard them (**Figure 2**). The size of the dataset means this still leaves a great deal of signal to test, but we are undoubtedly discarding some information through this procedure. Third, our approach is deliberately simple and makes minimal assumptions. More sophisticated approaches for phylodynamic modelling of viral fitness do exist [20, 39, 40], however, these are not directly portable to SARS-CoV-2 and would be too computationally demanding for a dataset of this size. Fourth, while our approach is undoubtedly more robust to demographic confounding (such as founder bias), it is impossible to completely remove all the sources of bias that come with using available public genomes.

Our results do not point to any increase in transmissibility of SARS-CoV-2 at this stage and highlight that the genomic diversity of the global SARS-CoV-2 population is currently still very limited. It is to be expected that SARS-CoV-2 will diverge into phenotypically different lineages as it establishes itself as an endemic human pathogen. However, there is no *a priori* reason to believe that this process will lead to the emergence of any lineage with increased transmission ability in its human host.

Methods

Data acquisition

16,924 SARS-CoV-2 assemblies were downloaded from GISAID on 15/05/2020 selecting only those marked as 'complete', 'low coverage exclude' and 'high coverage only'. To this dataset, all assemblies of total genome length less than 29,700bp were removed, as were any with a fraction of 'N' nucleotides >5%. In addition, all animal isolates strains were removed, including those from bat, pangolin, mink and tiger. All samples flagged by NextStrain as 'exclude' (https://github.com/nextstrain/ncov/blob/master/config/exclude.txt) as of 15/05/2020 were also removed. Thirteen further accessions were also removed from our analysis as they appeared as major outliers following phylogenetic inference despite passing other filtering checks. This left 15,691 assemblies for downstream analysis. A full metadata table and list of acknowledgements is provided in **Table S1**.

Multiple sequence alignment and maximum likelihood tree

All 15,691 assemblies were aligned against the Wuhan-Hu-1 reference genome (GenBank NC_045512.2, GISAID EPI_ISL_402125) using MAFFT [41] implemented in the rapid phylodynamic alignment pipeline provided by Augur (github.com/nextstrain/augur). This resulted in a 29,903 nucleotide alignment comprising 6,822 variable sites. As certain sites in the alignment have been flagged up as putative sequencing errors (http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473, accessed 08/05/2020), we followed two separate masking strategies. The first approach follows NextStrain guidance and masks positions 18,529, 29,849, 29,851 and 29,853 as well as the first 130bp and last 50bp of the alignment. The second strategy is more stringent, designed to test the impact of the inclusion of putative sequencing errors in phylogenetic inference, masking several further sites together with the first 55 and last 100 sites of the alignment (the list of sites flagged as 'mask' is available at https://raw.githubusercontent.com/W-L/ProblematicSites_SARS-CoV2/master/problematic_sites_sarsCov2.vcf accessed 15/05/2020). A complete list of masked positions is provided in **Table S6**. This resulted in two alignments of 15,691 assemblies with 6,674 and 6,655 total SNPs respectively.

Resulting alignments were manually inspected in UGene (http://ugene.net). Subsequently, for both alignments, a maximum likelihood phylogenetic tree was built using the Augur tree implementation selecting IQ-TREE as the tree-building method [42]. The resulting phylogenies were viewed and annotated using ggtree [43] (**Figure 1, Figure S1**). Site numbering and genome structure are provided for available annotations (non-overlapping open reading frames) using Wuhan-Hu-1 (NC_045512.2) as reference.

Phylogenetic dating

We informally estimate the substitution rate and time to the most recent common ancestor of both alignments by computing the root-to-tip temporal regression implemented in BactDating

[44]. Both alignments exhibit a significant correlation between the genetic distance from the root and the time of sample collection following 10,000 random permutations of sampling date (**Figure S2**).

Maximum parsimony tree and homoplasy screen

In parallel Maximum Parsimony trees were built for both the masked and more stringently masked alignments using MPBoot [45], specifying 1000 boot-strap replicates (*-bb 1000*) and excluding exact duplicate sequences. The resulting Maximum Parsimony treefiles for 12,626 and 12,020 sequences respectively were used, together with the input alignment, to rapidly identify recurrent mutations (homoplasies) using HomoplasyFinder [1, 46]. HomoplasyFinder employs the method first described by Fitch [47], providing, for each site, the site specific consistency index and the minimum number of changes invoked on the phylogenetic tree. For this analysis all ambiguous sites in the alignment were set to 'N'. HomoplasyFinder identified a total of 2,200 homoplasies, which were distributed over the SARS-CoV-2 genome (**Figure S4**). For the more stringently masked alignment, HomoplasyFinder identified a total of 2,133 homoplasies (**Figure S5**).

As previously described, we filtered both sets of identified homoplasies using a set of thresholds attempting to circumvent potential assembly/sequencing errors (filtering scripts are available at https://github.com/liampshaw/CoV-homoplasy-filtering and see reference [1]). This resulted in 273 filtered sites (238 following more stringent masking) of which 229 overlap (**Table S3**).

In addition, we considered an additional filtering criterion to identify homoplasic sites falling close to homopolymer regions, which may be prone to sequencing error (see Methods and **Table S3**). We defined homopolymer regions as positions on the Wuhan-Hu-1 reference with at least four repeated nucleotides. While homopolymer regions can arise through meaningful biological mechanisms, for example polymerase slippage, such regions have also been implicated in increased error rates for both nanopore [48] and Illumina sequencing [49]. As such, homoplasies detected near these regions (± 1 nt) could have arisen due to sequencing error rather than solely as a result of underlying biological mechanisms. If this were true, we would expect the proportion of homoplasic sites near these regions to be greater than that of homopolymeric positions across the entire genome. We tested this by identifying homopolymer regions using a python script (https://github.com/cednotsed/genome_homopolymer_counter) and performing a binomial test on the said proportions. A list of homopolymer regions across the genome is provided in **Table S7**. 23 of the 273 (8.4%) filtered homoplasies were within ± 1 nt of homopolymer regions but this proportion did not differ significantly from that of homopolymeric positions across the reference (9.7%; $p = 0.2718$). As such, we did not exclude homopolymer-associated homoplasies and suggest that these sites are likely to be biologically meaningful.

In addition, to determine if systematic biases were introduced in our filtering steps, we performed a principal component analysis (PCA) on the unfiltered list of homoplasies obtained from HomoplasyFinder (n = 2,200). The input space of the PCA included 11 variables, of which eight were dummy-coded reference/variant nucleotides and a further three corresponded to the minimum number of changes on tree, SNP count and consistency index output by HomoplasyFinder. Visualisation of PCA projections (**Figure S8A**) suggested that there was no hidden structure introduced by our homoplasy filtering steps. The first two principle components accounted for 54.5% of the variance and were mostly loaded by the variables encoding the reference and variant nucleotides (**Figure S8B**).

To further validate our 273 detected homoplasies, we obtained the 5,411 short-read datasets available on the NCBI Sequence Read Archive (SRA) as of 11th May 2020. Mapping to Wuhan-Hu-1 was performed using a BWA-MEM [50]. We retained the 3,441 genomic samples that covered at least 99% of the Wuhan-Hu-1 reference genome with a depth of five (**Table S5**). After PCR-duplicates removal using PicardTools MarkDuplicates v.2.7.0, SNPs were called using Freebayes v.0.9.21 [51]. SNPs displaying an intra-individual alternate allele frequency < 0.65 and/or a phred quality score < 20 were discarded. In addition, any position displaying more than 300 'N' across all isolates was not considered. This resulted in 3,147 filtered SNPs. Following the masking procedure used in the multi-sequence alignment the first 130 and final 50 base pairs were masked together with positions 18,529, 29,849, 29,851, 29,853. HomoplasyFinder [46] was run on a maximum parsimony tree built using MPboot [45]. Of the unfiltered 193 detected homoplasies, 184 were also detected in the unfiltered GISAID dataset. 84 out of the 273 filtered homoplasies of the GISAID dataset were validated by the SRA dataset (**Table S3)**.

Annotation and characterisation of homoplasic sites

All variable sites across the coding regions of the genomes were identified as synonymous or non-synonymous. This was done by retrieving the amino acid changes corresponding to all SNPs at these positions using a custom Biopython (v.1.76) script (https://github.com/cednotsed/nucleotide_to_AA_parser.git). The ORF coordinates used (including the ORF1ab ribosomal frameshift site) were obtained from the associated metadata according to Wuhan-Hu-1 (NC_045512.2).

To determine if certain types of SNPs are overrepresented in homoplasic sites, we computed the base count ratios and cumulative frequencies of the different types of SNPs across all SARS-CoV-2 genomes at homoplasic and/or non-homoplasic sites (**Figures S6-S7**). In addition, we identified the sequence context of all variable positions in the genome (± 1 and ± 2 neighbouring bases from these positions) and computed the frequencies of the resultant 3-mers (**Figure S9**) and 5-mers (**Figure S10**).

Quantifying pathogen fitness (transmission)

Under random sampling we expect that a mutation that positively affects a pathogen's transmission fitness will be represented in proportionally more descendent nodes. As such a pathogen's fitness can be expressed simply as the number of descendent nodes from the direct ancestor of the strain having acquired the mutation, relative to the descendants without the mutation (schematic **Figure 2**). We define this as the Ratio of Homoplasic Offspring (RoHO) index (https://github.com/DamienFr/RoHO).

HomoplasyFinder [46] flags all nodes of a phylogeny corresponding to an ancestor that acquired an homoplasy. We only considered nodes with at least three descending tips and whereby no children nodes are themselves annotated as carrying the homoplasy. For each such node of the tree we counted the number of isolates of each allele and computed the RoHO score. We finally restricted our analysis to homoplasies having at least $n = 5$ and $n = 10$ RoHO scores (i.e. for which five or ten independent lineages acquired the mutation) and obtained two datasets of varying stringency. Sign tests were computed for each homoplasy to test whether RoHO scores were significantly different from zero. The resultant p-values were adjusted with the Bonferroni procedure to account for multiple testing. To validate the methodology, this analysis was carried on GISAID with both masking strategies as well as the SRA dataset (**Table S4, Figure S11**).

Acknowledgments and Funding

L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). Computational analyses were performed on UCL Computer Science cluster and the South Green bioinformatics platform hosted on the CIRAD HPC cluster. We additionally wish to acknowledge the very large number of scientists in originating and submitting labs who have readily made available SARS-CoV-2 assemblies to the research community.

Author Contributions

L.v.D. and F.B. conceived and designed the study; L.v.D., M.A, D.R, L.P.S., C.C.S.T., analysed data and performed computational analyses; L.v.D., and F.B. wrote the paper with inputs from all co-authors.

Competing Interests

The authors have no competing interests to declare.

References

1. van Dorp, L., et al., *Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.* Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2020: p. 104351.
2. Li, X.G., et al., *Transmission dynamics and evolutionary history of 2019-nCoV.* Journal of Medical Virology, 2020. **92**(5): p. 501-511.
3. Giovanetti, M., et al., *The first two cases of 2019-nCoV in Italy: Where they come from?* Journal of Medical Virology, 2020. **92**(5): p. 518-521.
4. Lu, J., et al., *Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China.* medRxiv, 2020: p. 2020.04.01.20047076.
5. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-+.
6. Elbe, S. and G. Buckland-Merrett, *Data, disease and diplomacy: GISAID's innovative contribution to global health.* Global Challenges, 2017. **1**(1): p. 33-46.
7. Shu, Y.L. and J. McCauley, *GISAID: Global initiative on sharing all influenza data - from vision to reality.* Eurosurveillance, 2017. **22**(13): p. 2-4.
8. Snijder, E.J., et al., *Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage.* Journal of Molecular Biology, 2003. **331**(5): p. 991-1004.
9. Minskaia, E., et al., *Discovery of an RNA virus 3 '-> 5 ' exoribonuclease that is critically involved in coronavirus RNA synthesis.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(13): p. 5108-5113.
10. Mangeat, B., et al., *Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts.* Nature, 2003. **424**(6944): p. 99-103.
11. Harris, R.S., et al., *DNA determination mediates innate immunity to retroviral infection.* Cell, 2003. **113**(6): p. 803-809.
12. Harris, R.S. and J.P. Dudley, *APOBECs and virus restriction.* Virology, 2015. **479**: p. 131-145.

13. Kimura, M. and T. Ohta, *On the Rate of Molecular Evolution.* Journal of Molecular Evolution, 1971. **1**: p. 1-17.
14. Korber, B., et al., *Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2.* bioRxiv, 2020: p. 2020.04.29.069054.
15. Tang, X., et al., *On the origin and continuing evolution of SARS-CoV-2.* National Science Review, 2020.
16. Cagliani, R., et al., *Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2.* Journal of Virology, 2020: p. JVI.00411-20.
17. Li, X., et al., *Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection.* bioRxiv, 2020: p. 2020.03.20.000885.
18. MacLean, O.A., et al., *No evidence for distinct types in the evolution of SARS-CoV-2.* Virus Evolution, 2020. **6**(1).
19. Wertheim, J.O., et al., *Transmission fitness of drug-resistant HIV revealed in a surveillance system transmission network.* Virus Evolution, 2017. **3**(1).
20. Kühnert, D., et al., *Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics.* PLOS Pathogens, 2018. **14**(2): p. e1006895.
21. Zhao, Z., et al., *Moderate mutation rate in the SARS coronavirus genome and its implications.* BMC Evolutionary Biology, 2004. **4**(1): p. 21.
22. Dudas, G., et al., *MERS-CoV spillover at the camel-human interface.* eLife, 2018. **7**: p. e31257.
23. Domingo-Calap, P., et al., *An unusually high substitution rate in transplant-associated BK polyomavirus in vivo is further concentrated in HLA-C-bound viral peptides.* Plos Pathogens, 2018. **14**(10): p. 18.
24. Holmes, E.C., et al., *The evolution of Ebola virus: Insights from the 2013-2016 epidemic.* Nature, 2016. **538**(7624): p. 193-200.
25. Rambaut, A., et al., *A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology.* bioRxiv, 2020: p. 2020.04.17.046086.
26. Woo, P.C.Y., et al., *Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses.* Virology, 2007. **369**(2): p. 431-442.
27. Pyrc, K., et al., *Genome structure and transcriptional regulation of human coronavirus NL63.* Virology Journal, 2004. **1**(1): p. 7.
28. Grigoriev, A., *Mutational patterns correlate with genome organization in SARS and other coronaviruses.* Trends in Genetics, 2004. **20**(3): p. 131-135.
29. Di Giorgio, S., et al., *Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2.* bioRxiv, 2020: p. 2020.03.02.973255.
30. Simmonds, P., *Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories.* bioRxiv, 2020: p. 2020.05.01.072330.
31. Rice, A.M., et al., *Evidence for strong mutation bias towards, and selection against, T/U content in SARS-CoV2: implications for attenuated vaccine design.* bioRxiv, 2020: p. 2020.05.11.088112.
32. Salter, J.D. and H.C. Smith, *Modeling the Embrace of a Mutator: APOBEC Selection of Nucleic Acid Ligands.* Trends in Biochemical Sciences, 2018. **43**(8): p. 606-622.
33. Hamilton, C.E., F.N. Papavasiliou, and B.R. Rosenberg, *Diverse functions for DNA and RNA editing in the immune system.* Rna Biology, 2010. **7**(2): p. 220-228.

34. Lamers, M.M., B.G. van den Hoogen, and B.L. Haagmans, *ADAR1: "Editor-in-Chief" of Cytoplasmic Innate Immunity.* Frontiers in Immunology, 2019. **10**: p. 11.
35. Lerner, T., F.N. Papavasiliou, and R. Pecori, *RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease.* Genes, 2019. **10**(1): p. 19.
36. Salter, J.D., R.P. Bennett, and H.C. Smith, *The APOBEC Protein Family: United by Structure, Divergent in Function.* Trends in Biochemical Sciences, 2016. **41**(7): p. 578-594.
37. Warren, C.J., et al., *Role of the host restriction factor APOBEC3 on papillomavirus evolution.* Virus Evolution, 2015. **1**(1).
38. Milewska, A., et al., *APOBEC3-mediated restriction of RNA virus replication.* Scientific reports, 2018. **8**(1): p. 5960-5960.
39. Rasmussen, D.A. and T. Stadler, *Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models.* eLife, 2019. **8**: p. e45562.
40. Maddison, W.P., P.E. Midford, and S.P. Otto, *Estimating a Binary Character's Effect on Speciation and Extinction.* Systematic Biology, 2007. **56**(5): p. 701-710.
41. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
42. Minh, B.Q., et al., *IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.* Molecular Biology and Evolution, 2020. **37**(5): p. 1530-1534.
43. Yu, G.C., et al., *GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.* Methods in Ecology and Evolution, 2017. **8**(1): p. 28-36.
44. Didelot, X., et al., *Bayesian inference of ancestral dates on bacterial phylogenetic trees.* Nucleic Acids Research, 2018. **46**(22): p. 11.
45. Hoang, D.T., et al., *MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation.* Bmc Evolutionary Biology, 2018. **18**: p. 11.
46. Crispell, J., D. Balaz, and S.V. Gordon, *HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny.* Microbial Genomics, 2019. **5**(1): p. 10.
47. Fitch, W.M., *Toward defining course of evolution - minimum change for a specific tree topology.* Systematic Zoology, 1971. **20**(4): p. 406–416.
48. Cretu Stancu, M., et al., *Mapping and phasing of structural variation in patient genomes using nanopore sequencing.* Nature Communications, 2017. **8**(1): p. 1326.
49. Schirmer, M., et al., *Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data.* BMC bioinformatics, 2016. **17**: p. 125-125.
50. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
51. Garrison, E. and G. Marth, *Haplotype-based variant detection from short-read sequencing.* arXiv, 2012. **1207.3907**.