medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Identification of a molnupiravir-associated mutational signature in SARS-CoV-2 sequencing databases

Theo Sanderson[1]
Ryan Hisner[2]
I'ah Donovan-Banfield[3,4]
Hassan Hartman[5]
Alessandra Løchen[5]
Thomas P. Peacock[6]
Christopher Ruis[7,8,9]

[1] Francis Crick Institute, London, UK
[2] Department of Bioinformatics, University of Cape Town, Cape Town, South Africa
[3] Department of Infection Biology and Microbiomes, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK
[4] NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, Liverpool, UK
[5] UK Health Security Agency, London, UK
[6] Department of Infectious Disease, Imperial College London, London, UK
[7] Molecular Immunity Unit, University of Cambridge Department of Medicine, MRC-Laboratory of Molecular Biology, Cambridge, UK
[8] Department of Veterinary Medicine, University of Cambridge, Cambridge, UK
[9] Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Identification of a molnupiravir-associated mutational signature in SARS-CoV-2 sequencing databases

Theo Sanderson [1,✉], Ryan Hisner [2], I'ah Donovan-Banfield [3,4], Hassan Hartman [5], Alessandra Løchen [5], Thomas P. Peacock [6], and Christopher Ruis [7,8,9,✉]

[1]Francis Crick Institute, London, UK; [2]Department of Bioinformatics, University of Cape Town, Cape Town, South Africa; [3]Department of Infection Biology and Microbiomes, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK; [4]NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, Liverpool, UK; [5]UK Health Security Agency, London, UK; [6]Department of Infectious Disease, Imperial College London, London, UK; [7]Molecular Immunity Unit, University of Cambridge Department of Medicine, MRC-Laboratory of Molecular Biology, Cambridge, UK; [8]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK; [9]Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK

Molnupiravir, an antiviral medication that has been widely used against SARS-CoV-2, acts by inducing mutations in the virus genome during replication. Most random mutations are likely to be deleterious to the virus, and many will be lethal, and so molnupiravir-induced elevated mutation rates reduce viral load[2,3]. However, if some patients treated with molnupiravir do not fully clear SARS-CoV-2 infections, there could be the potential for onward transmission of molnupiravir-mutated viruses. Here we show that SARS-CoV-2 sequencing databases contain extensive evidence of molnupiravir mutagenesis. Using a systematic approach, we find that a specific class of long phylogenetic branches, distinguished by a high proportion of G-to-A and C-to-T mutations, appear almost exclusively in sequences from 2022, after the introduction of molnupiravir treatment, and in countries and age-groups with widespread usage of the drug. We identify a mutational spectrum, with preferred nucleotide contexts, from viruses in patients known to have been treated with molnupiravir and show that its signature matches that seen in these long branches, in some cases with onwards transmission of molnupiravir-derived lineages. Finally, we analyse treatment records to confirm the association between these high G-to-A branches and the use of molnupiravir.

Correspondence: ***theo.sanderson@crick.ac.uk cr628@cam.ac.uk***

Molnupiravir is an antiviral drug, licensed in some countries for the treatment of COVID-19. In the body, molnupiravir is ultimately converted into a nucleotide-analog, molnupiravir triphosphate (MTP)[1]. MTP is capable of being incorporated into RNA during strand synthesis by viral RNA-dependent RNA polymerases, where it can result in errors of sequence fidelity during viral genome replication. These errors in RNA replication result in many viral progeny that are non-viable, and so reduce the virus's effective rate of growth – molnupiravir was shown to reduce viral replication during 24 hours by 880-fold in vitro, and to reduce viral load both in animal models[2] as well as in patients sampled on the final day of treatment (day 5)[3]. Molnupiravir initially showed some limited efficacy as a treatment for COVID-19[4,5], but subsequent larger clinical trials found that the treatment did not reduce hospitalisation or death rates in high risk groups[3]. As one of the first orally bioavailable antivirals on the market, molnupiravir has been widely adopted by many countries, most recently China[6]. However, recent trial results and the approval of more efficacious antivirals have since led to several countries recommending against molnupiravir usage on the basis of limited effectiveness[7–9].

MTP appears to be incorporated into nascent RNA primarily by acting as an analogue of cytosine (C), pairing opposite guanine (G) bases (Fig. 1A). However, once incorporated, the molnupiravir (M)-base can transition into an alternative tautomeric form which resembles uracil (U) instead. This means that in the next round of strand synthesis, to give the positive-sense SARS-CoV-2 genome, the tautomeric M base can pair with adenine (A), resulting in a G-to-A mutation, as shown in Fig. 1B. G-to-A mutations arise from incorporation of molnupiravir into the negative-sense genome. Incorporation of MTP can also occur during the synthesis of the positive-sense genome: in this scenario, an initial positive-sense C correctly pairs with a G during negative-sense synthesis, but this G then pairs with an M base during positive-sense synthesis. In the next round of replication this M can then pair with A, which will result in a U in the final positive sense genome, with the overall process producing a C-to-U mutation (Fig. 1C).

The free nucleotide MTP is less prone to tautomerisation to the oxime form than when incorporated into RNA, and so this directionality of mutations is the most likely[10]. However it is also possible for some MTP to bind, in place of U, to A bases and undergo the above processes in reverse, causing A-to-G and U-to-C mutations (Fig. 1C).

It has been proposed that many major SARS-CoV-2 variants emerged from long-term chronic infections. This model explains several peculiarities of variants such as a general lack of genetic intermediates, rooting with much older sequences, long phylogenetic branch

[1]MTP is also known as β-D-N^4-hydroxycytidine triphosphate (NHC-TP).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

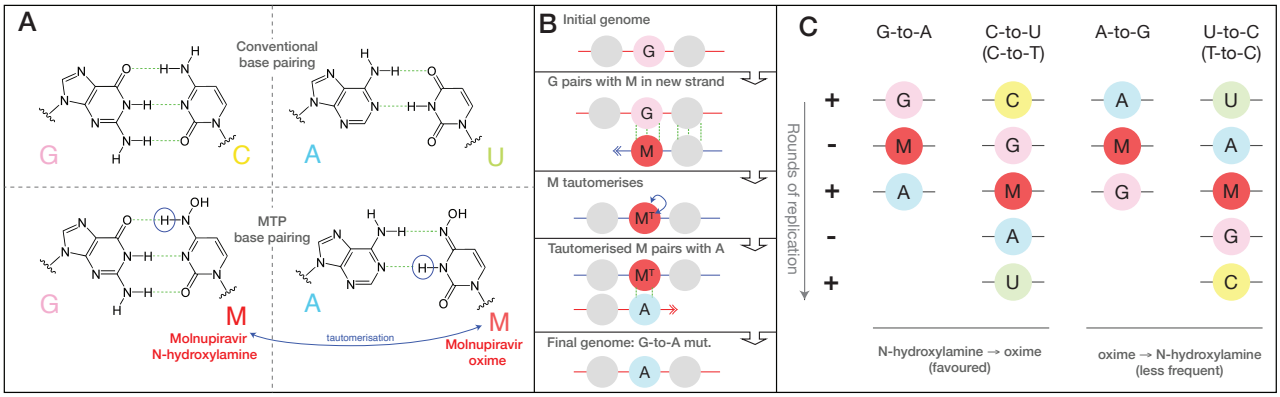


Figure 1. Molnupiravir induces mutations by acting as a nucleotide analogue with multiple tautomeric forms which pair preferentially with different nucleotides. (**A**) Molnupiravir triphosphate can assume multiple tautomeric forms which resemble different nucleotides. The N-hydroxylamine form resembles cytosine (C), and like cytosine can pair with guanine (G) while the oxime form more closely resembles uracil (U), and thus can pair with adenine (A).(Figure adapted in part from Malone et al.[1]) (**B**) In the most common scenario, molnupiravir (M) is incorporated in the N-hydroxylamine form opposite a G nucleotide. It can then tautomerise into the oxime form, which can then pair to an A in subsequent replication, creating a G-to-A mutation. (**C**) Molnupiravir can result in four different mutation types. In the first column a G-to-A mutation is created by M incorporation opposite a positive-sense G, which can then pair with an A in the next positive-sense synthesis. In the second column, the positive-sense genome contains a C, which results in a G in the negative-sense genome. This G can then undergo the same G-to-A mutation, creating an negative-sense A which finally results in a U in the positive sense genome, meaning the entire process results in a C-to-U mutation. Although the biases of tautomeric forms for the free and incorporated MTP nucleotides appear to favour this directionality of mutations, with M incorporated in the N-hydroxylamine form and then transitioning to the oxime form, the reverse can also occur: this results in A-to-G and U-to-C mutations.

lengths, and the level of convergent evolution with known chronic infections[11–14]. During the approval process for molnupiravir, concerns were raised about its potential to increase the rate of evolution of variants of concern[15]. In response, it was noted that no infectious virus had been isolated at or beyond day 5 following molnupiravir treatment, and that mutations recovered following treatment were random with no evidence of selection-based bias[16].

During analysis of long phylogenetic branches in the SARS-CoV-2 tree, recent branches exhibiting potential signs of molnupiravir-driven mutagenesis have been noted, including clusters of sequences indicating onward transmission[17]. We therefore aimed to characterise the mutational profile of molnupiravir and examine the extent to which this signature appeared in global sequencing databases.

Emergence of a molnupiravir-associated mutational profile in global sequencing databases

To establish the mutational profile induced by molnupiravir, we analysed published longitudinal genomic time series that included both untreated patients and patients treated with molnupiravir[18,20], and compared against a typical SARS-CoV-2 mutational spectrum[19]. In agreement with previous findings, we found that molnupiravir treatment led to an 8-fold increase (CI: 2.9-16) in the rate of mutations and that this increase was highly specific to transition mutations (Fig. 2A), especially G-to-A and C-to-T mutations (hereon we use 'T' rather than 'U', as in sequences). While C-to-T mutations are relatively common overall in SARS-CoV-2 evolution[19,21,22], G-to-A mutations occur much less frequently; therefore an elevated G-to-A proportion was especially predictive of molnupiravir treatment (Fig. 2B).

We looked for evidence of such a signal in global sequencing databases by analysing a mutation-annotated tree, derived from McBroome et al.[23], containing >13 million SARS-CoV-2 sequences from GISAID[24] and the INSDC databases[25]. For each branch of the tree we counted the number of each substitution class (A-to-T, A-to-G, etc.). Filtering this tree to branches involving at least 20 substitutions, and plotting the proportion of substitution types revealed a region of this space with higher G-to-A and almost exclusively transition substitutions, that only contained branches sampled since 2022 (Fig. 2C), suggesting some change (either biological or technical) had resulted in a new mutational pressure, with mutational classes consistent with those seen in patients known to be treated with molnupiravir.

We created a criterion for branches of interest, which we refer to as "high G-to-A" branches: we selected branches involving at least 10 substitutions, of which more than 25% were G-to-A, more than 20% were C-to-T and more than 95% were transitions. Simulations predicted that this criterion would have a sensitivity of 51% and a specificity of 99.7% for branches involving 15 substitutions (see methods). Branches satisfying the high G-to-A conditions were almost all sampled after the roll out of molnupiravir in late 2021 and early 2022 (Fig. 2D, Fig. S1). The branches were predominantly sampled from a small number of countries, which could not be explained by differences in sequencing efforts (Fig. 2E-F, Table 1). Many countries which exhibited a high proportion of high G-to-A branches use molnupiravir: >380,000 prescriptions had occurred in Australia by the end of 2022[26], >30,000 in the UK in the same period[3,27], >240,000 in the US within the early months of 2022[28], and >600,000 in Japan by Oct 2022[29]. Countries with high levels of total sequencing but a low number of G-to-A branches (Canada, France, Fig. 2E-F) have not authorised the prescription of mol-

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

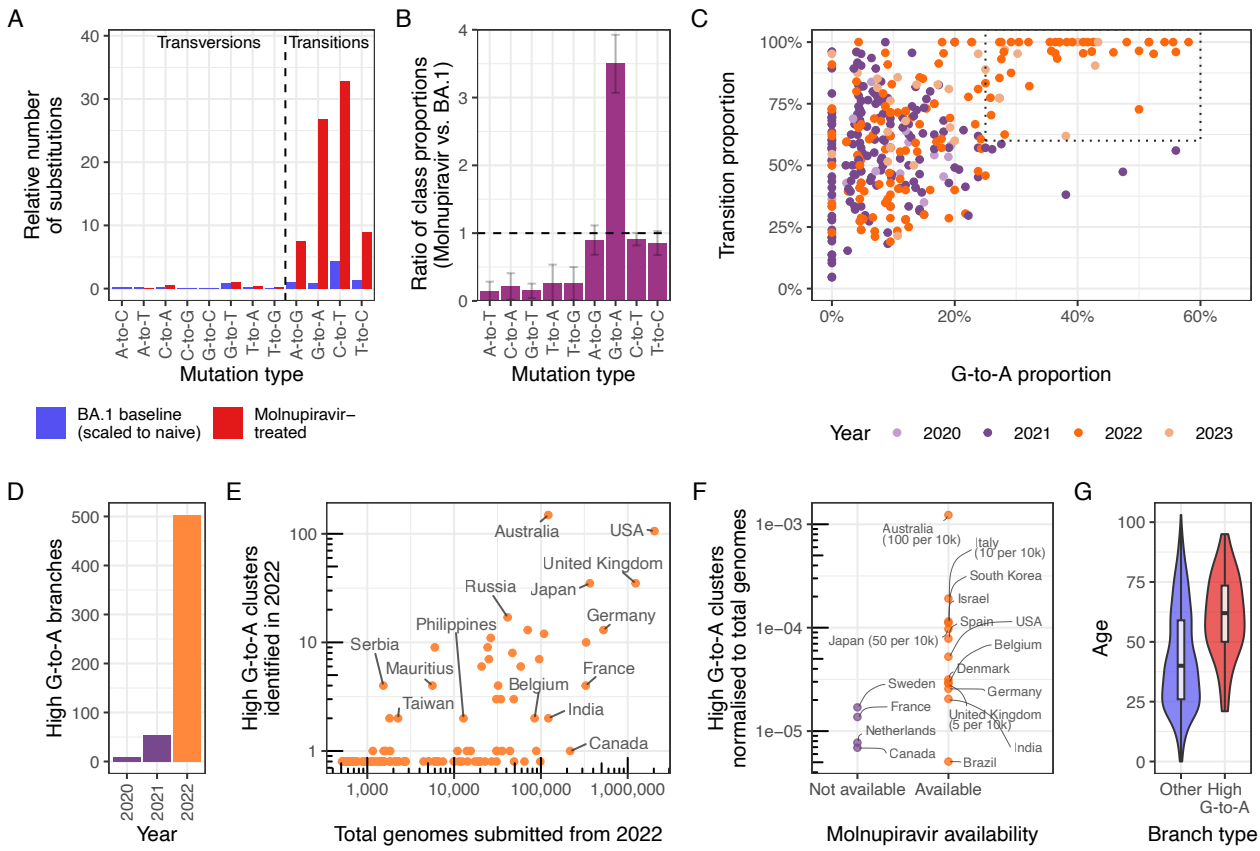


Figure 2. A molnupiravir-associated mutational signature with high G-to-A and high transition ratio emerged in 2022 in some countries in global sequencing databases

(**A**) A comparison the relative rate of different classes of mutations in typical BA.1 mutation dynamics vs. those with molnupiravir treatment (molnupiravir data from Alteri et al. [18]; naive data from Ruis et al. [19], scaled to total mutations in naive individuals from Alteri et al., see Methods) confirms an elevated rate of transitions, and particularly C-to-T and G-to-A mutations. (**B**) Differences in the proportion of mutations due to different mutation classes in molnupiravir-treated individuals (Alteri et al. [18]) vs typical BA.1 mutations (Ruis et al. [19]) highlights elevated G-to-A proportion as especially indicative of molnupiravir (**C**) A scatter plot where each point is a branch with >20 mutations, positioned according to the proportion of these mutations that are G-to-A (x-axis) or transitions (y-axis) reveals a space with elevated G-to-A and transition rate that occurs only with the rollout of molnupiravir in 2022, as also shown in (**D**) for nodes with >10 mutations, G-to-A proportions >25%, C-to-T proportions >20%, and transition proportions >90%. (**E**) Plotting the number of high G-to-A nodes identified in 2022 against the number of total genomes for each country reveals considerable variation. (**F**) Countries confirmed to have made molnupiravir available have more high G-to-A nodes than countries which have not. Numbers in brackets represent number of courses of molnupiravir supplied, normalised to population. (p=0.02 for log-transformed t-test.) (**G**) Age distribution for US nodes, partitioned according whether they satisfy the high G-to-A criterion (p<1e-10). Age metadata are missing for some samples, likely non-randomly.

nupiravir[30,31]. Age metadata from the US showed a significant bias towards patients with older ages for these high G-to-A branches, compared to control branches with similar numbers of mutations but without filtering on substitution-type (Fig. 2G). Where age data was available in Australia it also suggested high G-to-A branches primarily occur in an aged population. This is consistent with the prioritised use of molnupiravir to treat older individuals, who are at greater risk from severe infection, in these countries. In Australia, molnupiravir was preplaced in aged-care facilities, and it was recommended that it be considered for all residents testing positive for COVID-19 aged 70 or older, with or without symptoms[32].

Mutation spectrum analysis supports a a relationship of high G-to-A branches with molnupiravir

To further probe the link between high G-to-A branches and molnupiravir, we used mutation spectrum analysis, which considers both the types of mutations and the genomic context in which the mutations occur (Fig. S4). The spectrum we identified for branches selected by our criteria is, as expected, dominated by G-to-A and C-to-T transition mutations with smaller contributions from A-to-G and T-to-C transitions (Fig. 3A). This pattern is consistent with the known mechanism of action for molnupiravir (Fig. 1C). We similarly calculated spectra both from patients known to be treated with molnupiravir[18,20](Fig. 3B) and from general SARS-CoV-2 evolution[19] (Fig. 3C) .

There was a strong match between the spectrum of known-molnupiravir sequences and that of high G-to-A branches, both in terms of mutation classes and the context preferences within each mutation class (Fig. 3, Fig. S5A). For C-to-T and G-to-A mutations, a comparison of context preferences gave correlation coefficients of 0.97 and 0.90 respectively (p<0.001) (Fig. 3D). Similar results were seen for a spectrum calculated from an additional clinical trial of molnupiravir (Fig. S5B)[20]. The contextual patterns seen in long branches did

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

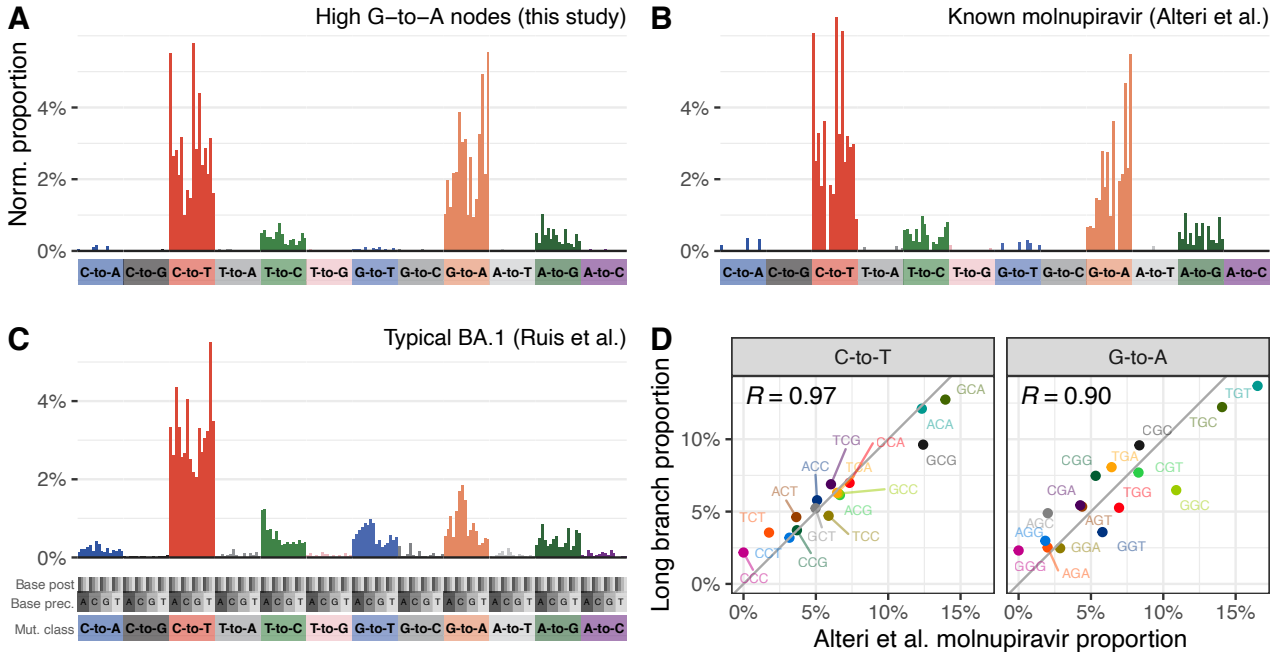


Figure 3. Mutation spectrum analysis supports high G-to-A branches being driven by molnupiravir
(**A,B,C**) Single-base substitution mutation spectra for high G-to-A branches (A), known molnupiravir-treated individuals (B) and typical BA.1 spectra (C). Each individual bar represents a particular type of mutation in a particular trinucleotide context (Fig. S4). Bars are grouped and coloured according to the class of mutation. Within each coloured group, bars are grouped into 4 groups according to the nucleotide preceding the mutated residue, and then each of these groups contains 4 bars according to the nucleotide following the mutated residue. Number of mutations have been normalised to the number of times the trinucleotide occurs in the reference genome, and then normalised so that the entire spectrum sums to 1. (**D**) High correlations between spectra from Alteri et al. from patients known to have been treated with molnupiravir, and the spectra from high G-to-A branches identified in this study. Each point represents the normalised proportion of a particular trinucleotide context. Points are coloured such that a context for C-to-T mutations has the same colour as its reverse complement in G-to-A.

not correlate with typical SARS-CoV-2 mutational processes (Fig. S5C). In high G-to-A branches, G-to-A mutations occurred most commonly in TGT and TGC contexts, which could represent a preference for molnupiravir binding adjacent to particular surrounding nucleotides, a preference of the viral RdRp to incorporate molnupiravir adjacent to specific nucleotides, or a context-specific effect on the viral proof-reading machinery. These correlations between spectra from high G-to-A branches and known molnupiravir-treated individuals strongly support a shared mutational driver, and therefore that the high G-to-A branches are driven by molnupiravir treatment.

Incorporation of molnupiravir during negative strand synthesis will result in G-to-A mutations in the virus sequence while incorporation during positive strand synthesis will manifest as C-to-T mutations in the virus sequence, after a second round of replication (Fig. 1C). Consistent with this, we observe a strong positive correlation between the mutational biases in equivalent contextual patterns within C-to-T and G-to-A mutations (Pearson's $r = 0.86$, $p < 0.001$, Fig. S5D), which are the reverse complements of each other where for example a G-to-A mutation in the TGC context on one strand is the equivalent of a C-to-T mutation in the GCA context on the other strand).

Molnupiravir-associated lineages have formed transmission clusters and are associated with an elevated rate of evolution

Although a majority of the long branches sampled have just a single descendant tip sequence in sequencing databases, in some cases we could see that branches had given rise to clusters with a significant number of descendant sequences. For example, a cluster in Australia in August 2022 involves 20 tip sequences, with distinct age metadata confirming they derive from multiple individuals (Fig. 4A). This cluster involves 25 substitutions in the main branch, of which all are transitions with 44% C-to-T and 36% G-to-A. Closely related outgroups date from July 2022, suggesting that these mutations emerged in a period of 1-2 months. At the typical rate of SARS-CoV-2 evolution, this number of mutations would take years to acquire in an unsampled population with typical dynamics[33]. Overall in the dataset, we observe a systematic accelerated evolutionary rate in high G-to-A branches ($p<0.001$), consistent with the action of a mutagenic drug.

There are many further examples of high G-to-A branches with multiple descendant sequences, including sequence clusters in the United Kingdom, Japan, USA, New Zealand, Slovakia, Denmark, South Korea and Vietnam (Fig. 4B, Appendix).

During the construction of the daily-updated mutation-annotated tree[23], samples highly divergent from the existing tree are excluded. This is a necessary step given

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

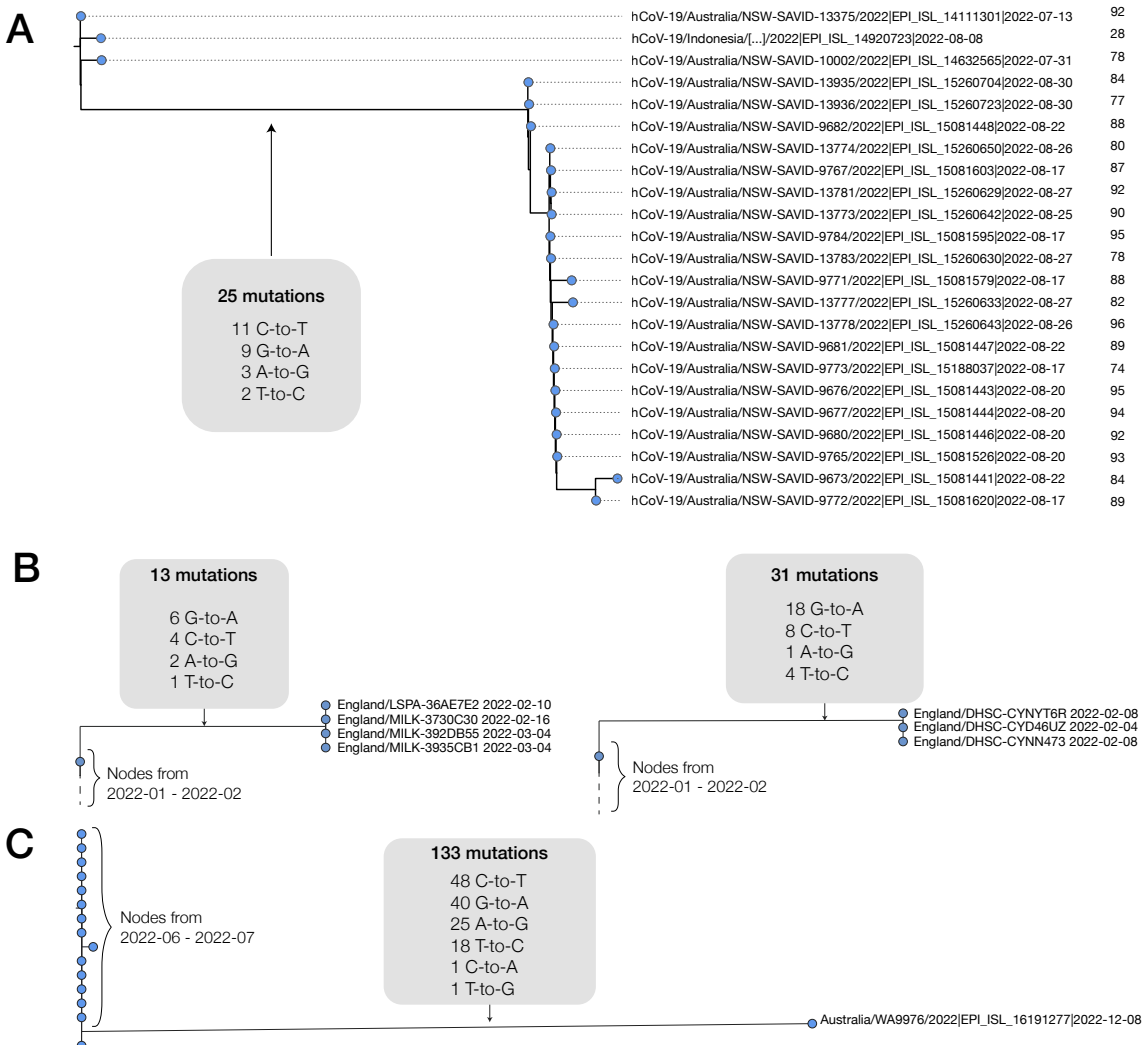


Figure 4. High G-to-A branches can be associated with transmission clusters and, separately, sometimes involve hundreds of mutations
(**A**) A cluster of 20 individuals emerging from a high G-to-A mutation event. This cluster involves a saltation of 25 mutations occurring within a period of around a month, all of which are transition substitutions, with an elevated G-to-A rate. Sequences are annotated with age metadata suggestive of an outbreak in an aged-care facility. Phylogenetic placement within the cluster is affected by missing coverage in some regions. (**B**) Examples of further transmission clusters from the UK, left: four sequences from the UK from Feb-March 2022 with 13 shared muations with the high G-to-A signature, right: a cluster of four sequences from the UK from Feb 2022 with 31 shared muations with the high G-to-A signature. (**C**) A singleton sequence from Australia with a high G-to-A signature and a total of 133 mutations. Just 2 of the 133 mutations observed are transversions and transitions include numerous G-to-A events.

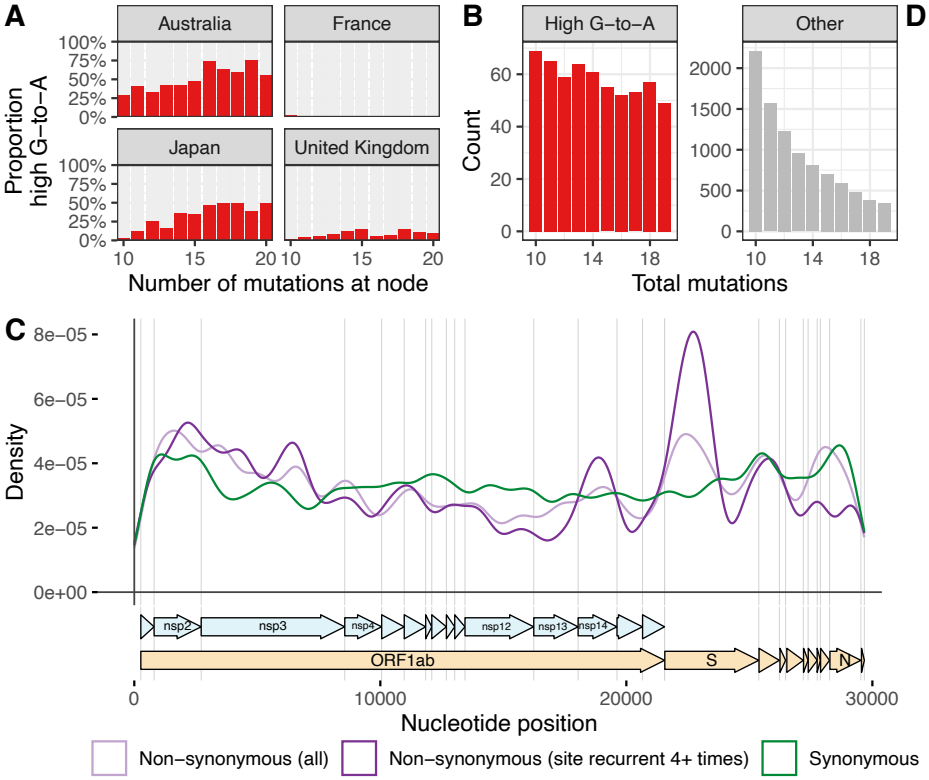
the technical errors in some SARS-CoV-2 sequencing data, but means that highly divergent molnupiravir-induced sequences might be excluded. To search for excluded sequences with a molnupiravir-like pattern of mutations we processed a full sequence dataset with Nextclade and calculated the proportion of each of the mutation classes among the private mutations (see methods) each sequence carried. This analysis allowed the identification of further mutational events, including some involving up to 130 substitutions (Fig. 4C, Fig. S2), with the same signature of elevated G-to-A mutation rates and almost exclusively transition substitutions. The cases we identified with these very high numbers of mutations involved single sequences, and could represent sequences resulting from chronically infected individuals who have been treated with multiple courses of molnupiravir.

Purifying and positive selection in molnupiravir-associated mutational events

High G-to-A branches made up a considerable percentage of branches involving more than 10 substitutions in some countries (Fig. 5A), suggesting that molnupiravir drives a substantial proportion of mutational events resulting in large numbers of mutations. We found that high G-to-A branches had a different distribution of branch lengths from other types of branches. In typical SARS-CoV-2 evolution, the branch length distribution contains many more nodes with shorter branch lengths than with larger branch lengths, however for nodes satisfying the high G-to-A criterion this decline was much less pronounced, with long branch lengths still relatively common (Fig. 5B).

We examined the effects of mutations that made up high G-to-A branches. We calculated dN/dS values for short branches in the tree, and for long branches (>=

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .



D

AA substitution	n	Mut. type
S:P9L	12	C>T
S:D574N	12	G>A
S:E132K	7	G>A
S:E340K	7	G>A
S:A701V	7	C>T
S:A1070T	7	G>A
S:V1122M	7	G>A
S:K147E	6	A>G
S:A262T	6	G>A
S:G446S	6	G>A
S:V6I	5	G>A
S:G184S	5	G>A
S:G252S	5	G>A
S:G257S	5	G>A
S:V289I	5	G>A
S:R403K	5	G>A
S:K440R	5	A>G
S:R493Q	5	G>A
S:V595I	5	G>A
S:T1117I	5	C>T
S:V1128I	5	G>A

Figure 5. High G-to-A branches make up a considerable proportion of long branches in affected countries and include evidence of selection. (**A**) Proportion of branches that are high G-to-A for a range of branch lengths in different countries. Data from collection dates in 2022 and 2023 (submission dates up to June 2023). (**B**) Branch length distributions for high G-to-A and other branches. (**C**) Genomic distribution of mutations in high G-to-A nodes, partitioned into 3 classes: synonymous mutations, non-synonymous mutations, and non-synonymous that occur 4 or more times. (**D**) Table of the most recurrent mutations in *spike* in high G-to-A branches. *n* shows the number of high G-to-A branches that exhibit the mutation. *Mutation type* shows the parental and final nucleotide at the nucleotide position driving the mutation.

10 mutations) – split according to whether the branch had a high G-to-A signature or not. Short branches provided a dN/dS value of 0.45, while long branches without the high G-to-A signature had a higher dN/dS value of 0.73. This is likely to reflect positive selection in individuals with chronic infections. In contrast, long branches with a (molnupiravir-associated) high G-to-A signature had a dN/dS value of 0.44 – similar to that of the short branches. This likely represents the need for purifying selection against the many lethal or deleterious molnupiravir-induced mutations.

Despite this overall indication of purifying selection, consistent with the actions of a mutagenic drug, there was also evidence for positive selection. Even in high G-to-A branches, there was a concentration of non-synonymous mutations in *spike*, especially among the most recurrent mutations (Fig. 5C). Many of the recurrent *spike* mutations, such as S:P9L, S:A701V, S:K147E, S:R493Q, S:N450D, S:G252S and S: R346K, were also mutations that arise in variants of concern and/or chronic infections, including S:E340K which has been associated with sotrovimab resistance (Fig. 5D). This positive selection will be driven by selection by the immune system and other environmental factors.

There was also a relative concentration of recurrent non-synonymous mutations in the exonuclease encoded by nsp14. This proofreading exonuclease functions to correct errors during genome replication, but typically has poor performance in recognising mismatches involving molnupiravir[34]. Future work could examine whether there is a relationship between specific mutations in nsp14 and tolerance to molnupiravir.

Linkage analysis confirms a relationship between high G-to-A branches and molnupiravir treatment

To better test a direct relationship between high G-to-A branches and the use of molnupiravir, linkage analysis was performed at the UK Health Security Agency for high G-to-A branches sampled in England, linking with treatment data in the Blueteq database. This analysis found that 31% of clades descending from a high G-to-A branch involved at least one person prescribed with molnupiravir, representing a more than 500-fold enrichment over the proportion expected for randomly selected sequences (the overall rate of molnupiravir prescription in sequenced individuals in England from 2022 is 0.043%). There are multiple factors likely to have contributed to the fact that not all clusters were linked to a person known to have been prescribed molnupiravir. Some might represent false positives in our analysis, but additionally the Blueteq database does not contain prescription data for people treated as part of clinical trials (which make up around a third of total molnupiravir prescriptions in the UK) or patients who fell outside interim

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

clinical policy. It is also possible that in some cases, we have not sampled the index case treated with molnupiravr, but instead sampled a patient downstream of a treated patient in a transmission chain.

Discussion

We have shown a variety of lines of evidence that together demonstrate that a signature of molnupiravir treatment is visible in global sequencing databases. We identified a set of long phylogenetic branches that exhibit a high number of transition mutations, with a high proportion of G-to-A mutations. The number of these branches increased dramatically in 2022, coinciding with the roll-out of molnupiravir, and they are enriched for countries and age groups in which the drug was prescribed. Substitution rates on these branches were elevated, consistent with a recent study in immunocompromised patients treated with molnupiravir[35]. The branches exhibited a mutational spectrum that is highly similar to that in patients known to be treated with molnupiravir. These sequencing data suggest that in at least some cases, viruses with a large number of molnupiravir-induced mutations have been transmitted to other individuals, at least in a limited manner. Selection analysis performed on these branches shows negative selection against deleterious mutations, but also the positive selection of recurrent mutations in the *spike* gene. Treatment data in England confirms a heavy enrichment for molnupiravir-usage in identified branches.

Molnupiravir's mode of action is often described using the term "error catastrophe" – the concept that there is an upper limit on the mutation rate of a virus beyond which it is unable to maintain self-identity[36]. This model has been criticised on its own terms[37], and is particularly problematic in the case of molnupiravir treatment because a threshold for error catastrophe exists only in a model with a fixed mutation rate for infinite time. We suggest that the use of this term in the context of mutagenic antiviral drugs is unhelpful given that much of the reduction in fitness will be due to "single-hit" lethal events[37]. The "lethal mutagenesis", which concentrates on the number of infective progeny per infected cell, model is more useful in this context[38].

New variants of SARS-CoV-2 are generated through acquisition of mutations that enhance properties including immune evasion and intrinsic transmissibility[39,40]. The impact of molnupiravir treatment on the trajectory of variant generation and transmission is difficult to predict. Molnupiravir increases the amount of genetic diversity in the surviving viral population in the host and this could provide more material for selection to act on during intra-host evolution towards these fitness-enhancing properties. However, a high proportion of induced mutations are likely to be deleterious or neutral, and it is necessary to consider the counterfactual to molnupiravir treatment, as molnupiravir results in a modest reduction in viral load in treated patients[41]. But molnupiravir-induced mutation could also potentially allow infections to persist for longer by creating a more varied target for the immune system – one concerning aspect of the PANORAMIC trial is that while molnupiravir-treated individuals had much lower viral load at day 5, they had slightly higher viral load than placebo-arm individuals at day 14[3]. At the time of writing, the largest clusters satisfying our criterion consisted of ~20 sequenced individuals. We note that neither Omicron, nor any other variant designated by the WHO, has defining mutations indicative of the use of molnupiravir.

There are some limitations of our work. Identifying a particular branch as possessing a molnupiravir-like signature is a probabilistic rather than absolute judgement: where molnupiravir creates just a handful of mutations (which trial data suggests is often the case), branch lengths will be too small to assign the cause of the mutations with confidence. We therefore limited our analyses here to long branches. This approach may also fail to detect branches which feature a substantial number of molnupiravir-induced mutations alongside a considerable number of mutations from other causes (which might occur in chronic infections). We discovered drastically different rates of molnupiravir-associated sequences by country and suggest that this reflects in part whether, and how, molnupiravir is used in different geographical regions; however, there will also be contributions from the rate at which genomes are sequenced in settings where molnupiravir is used. For example, if molnupiravir is used primarily in aged-care facilities and viruses in these facilities are significantly more likely to be sequenced than those in the general community this will elevate the ascertainment rate of such sequences. Furthermore, it is likely that some included sequences were specifically analysed due to representing continued test positivity after molnupiravir treatment as part of specific studies. Such effects are likely to differ based on sequencing priorities in different locations.

We would recommend public health authorities perform continued investigations to establish the link between sequences showing this signature and the use of molnupiravir. These data will be useful for ongoing assessments of the risks and benefits of this treatment, and may guide the future development of mutagenic agents as antivirals, particularly for viruses with high mutational tolerances such as coronaviruses.

Bibliography

1. Malone, B. & Campbell, E. A. Molnupiravir: coding for catastrophe. *Nat Struct Mol Biol* **28**, 706–708 (2021).
2. Rosenke, K. *et al.* Orally delivered MK-4482 inhibits SARS-CoV-2 replication in the syrian hamster model. *Nat. Commun.* **12** (2021).
3. Butler, J. Molnupiravir plus usual care versus usual care alone as early treatment for adults with COVID-19 at increased risk of adverse outcomes (PANORAMIC): an open-label, platform-adaptive randomised controlled trial. *Lancet* (2022).
4. Jayk Bernal, A. *et al.* Molnupiravir for oral treatment of covid-19 in nonhospitalized patients. *N. Engl. J. Med.* **386**, 509–520 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

5. Extance, A. Covid-19: What is the evidence for the antiviral molnupiravir? *BMJ* **377**, o926 (2022).
6. Reuters. China grants conditional approval for merck's COVID treatment. *Reuters* (2022).
7. NICE recommends 3 treatments for COVID-19 in draft guidance. URL https://www.nice.org.uk/news/article/nice-recommends-3-treatments-for-covid-19-in-draft-guidance.
8. NC19CET. Taskforce updates molnupiravir guidance following PANORAMIC trial results. https://clinicalevidence.net.au/news/taskforce-updates-molnupiravir-guidance-following-panoramic-trial-results/ (2022). Accessed: 2023-1-6.
9. Covid-19 antivirals: The role of molnupiravir in new zealand's funded treatments portfolio (2023). URL https://pharmac.govt.nz/news-and-resources/consultations-and-decisions/consultation-2023-04-04-molnupiravir/?keyword=molnupiravir&type=all&page=1.
10. Gordon, C. J., Tchesnokov, E. P., Schinazi, R. F. & Götte, M. Molnupiravir promotes SARS-CoV-2 mutagenesis via the RNA template. *J. Biol. Chem.* **297** (2021).
11. 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 (2020). URL 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.
12. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
13. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol* **8**, veac080 (2022).
14. Harari, S. *et al.* Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat Med* **28**, 1501–1508 (2022).
15. Nelson, C. W. & Otto, S. P. Mutagenic antivirals: the evolutionary risk of low doses. *Virological. https://virological. org/t/mutagenic-antivirals-the-evolutionary-risk-of-low-doses/768 [Google Scholar]* (2021).
16. Presentations to AMDAC Meeting. Antimicrobial drugs advisory committee meeting (AMDAC). https://www.youtube.com/watch?v=fR9FNSJT64M (2021). URL https://www.fda.gov/media/155935/download. Virtual Meeting, November 30, 2021.
17. Hisner, R. RE:Potential BA.2.3 sublineage with many mutations (singleton, Indonesia). https://github.com/cov-lineages/pango-designation/issues/1080#issuecomment-1250412876 (2022).
18. Alteri C *et al.* A proof-of-concept study on the genomic evolution of sars-cov-2 in molnupiravir-treated, paxlovid-treated and drug-naïve patients. *Commun Biol* **5**, 1376 (2022).
19. Ruis, C. *et al.* A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. *Microbial Genomics* **9** (2023).
20. Donovan-Banfield, I. *et al.* Characterisation of SARS-CoV-2 genomic variation in response to molnupiravir treatment in the AGILE Phase IIa clinical trial. *Nat Commun* **13**, 7284 (2022).
21. Bloom, J. D., Beichman, A. C., Neher, R. A. & Harris, K. Evolution of the SARS-CoV-2 Mutational Spectrum. *Molecular Biology and Evolution* **40** (2023).
22. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2 (2021).
23. McBroome, J. *et al.* A Daily-Updated database and tools for comprehensive SARS-CoV-2 Mutation-Annotated trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021).
24. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).
25. Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y. & on behalf of the International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration (2011).
26. https://www.health.gov.au/resources/webinars/covid-19-response-update-for-primary-care-15-december-2022?language=en (2022).
27. COVID Therapeutics Weekly Publication (week ending 1st January 2023). https://www.england.nhs.uk/statistics/statistical-work-areas/covid-therapeutics-antivirals-and-neutralising-monoclonal-antibodies/ (2022).
28. Gold, J. A. W. *et al.* Dispensing of Oral Antiviral Drugs for Treatment of COVID-19 by Zip Code-Level Social Vulnerability - United States, December 23, 2021-May 21, 2022. *MMWR Morb Mortal Wkly Rep* **71**, 825–829 (2022).
29. Ministry of Health, L. & Welfare. Usage status of covid-19 therapeutics (government secured portion) (2023). URL https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000121431_00324.html. Accessed: 2023-06-02; Original title: .
30. Drug and vaccine authorizations for COVID-19: List of authorized drugs, vaccines and expanded indications. https://www.canada.ca/en/health-canada/services/drugs-health-products/covid19-industry/drugs-vaccines-treatments/authorization/list-drugs.html (2022).
31. France cancels order for Merck's COVID-19 antiviral drug. https://www.reuters.com/world/europe/france-cancels-order-mercks-covid-19-antiviral-drug-2021-12-22/ (2021).
32. Use of Lagevrio (molnupiravir) in residential aged care. URL https://www.health.gov.au/sites/default/files/documents/2022/07/coronavirus-covid-19-use-of-lagevrio-molnupiravir-in-residential-aged-care.pdf. Accessed: 2022-01-01.
33. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nature Reviews Microbiology* **21**, 361–379 (2023). URL https://doi.org/10.1038/s41579-023-00878-2.
34. Sheahan, T. P. *et al.* An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Science Translational Medicine* **12** (2020). URL https://doi.org/10.1126/scitranslmed.abb5883.
35. Fountain-Jones, N. M. *et al.* Antiviral treatments lead to the rapid accrual of hundreds of SARS-CoV-2 mutations in immunocompromised patients (2022).
36. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523 (1971).
37. Summers, J. & Litwin, S. Examining the theory of error catastrophe. *J. Virol.* **80**, 20–26 (2006).
38. Bull, J. J., n, R. & Wilke, C. O. Theory of lethal mutagenesis for viruses. *J Virol* **81**, 2930–2939 (2007).
39. Telenti, A., Hodcroft, E. & Robertson, D. The evolution and biology of SARS-CoV-2 variants. *Cold Spring Harbor Perspectives in Medicine* **12**, a041390 (2022).
40. Carabelli, A. M. *et al.* Sars-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* (2023).
41. Khoo, S. H. *et al.* Molnupiravir versus placebo in unvaccinated and vaccinated patients with early SARS-CoV-2 infection in the UK (AGILE CST-2): a randomised, placebo-controlled, double-blind, phase 2 trial. *Lancet Infect Dis* (2022).
42. Petit III, R. A., Hall, M. B., Tonkin-Hill, G., Zhu, J. & Read, T. D. fastq-dl: efficiently download FASTQ files from SRA or ENA repositories. URL https://github.com/rpetit3/fastq-dl.
43. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet* **53**, 809–816 (2021).
44. Sanderson, T. Taxonium, a web-based tool for exploring large phylogenetic trees. *Elife* **11** (2022).
45. McBroome, J. BTE: Cython API for the Mutation Annotated Tree (MAT) Online Phylogenetics Toolkit. https://github.com/jmcbroome/BTE (2023). GitHub repository.
46. Sanderson, T. Chronumental: time tree estimation from very large phylogenies. *bioRxiv* (2021).
47. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).
48. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
49. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).
50. Minh, B. Q. *et al.* Iq-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
51. Sagulenko, P., Puller, V. & Neher, R. A. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution* **4**, vex042 (2018). [Online; accessed 2023-01-13].
52. Rambaut, A. Figtree (2018). URL http://tree.bio.ed.ac.uk/software/figtree/.
53. UK Health Security Agency. ESPAUR report 2021 to 2022: Annexe (2022). URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1118730/ESPAUR-report-2021-2022-annexe.pdf. Accessed: 2023-06-26.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Methods

Processing of pre-existing genomic data from molnupiravir treated individuals

We used three existing sources of genomic data in calculations of the mutational classes (and later the contextual mutational spectra) associated with known use of molnupiravir and with typical SARS-CoV-2 evolution in the absence of molnupiravir. We analysed a dataset from Alteri et al.[18] which contained longitudinal both individuals treated with molnupiravir and untreated individuals. For this we downloaded FASTQ files from BioProject ERP142142 using `fastq-dl`[42]. We mapped these reads to the Hu-1 reference genome using minimap2 and then extracted the number of calls for each base at each position. We identified mutations compared to the day 0 sequence, counting variants where the site had >=100 reads of which >=5% were variant to the day 0 consensus. As a secondary dataset we used data from the AGILE trial (Donovan-Banfield et al.[20], BioProject PRJNA854613). There was general agreement on the nature of molnupiravir mutations between Alteri et al. and the AGILE trial, with the exception of a high G-to-T mutation rate seen only in the AGILE trial. Previous evidence on molnupiravir's mutation classes, as well as the fact that a high G-to-T rate was seen even in untreated individuals in the AGILE data, led us to conclude that this G-to-T signal in the AGILE data represented a technical artifact.

We used the BA.1 mutational spectrum previously calculated by Ruis et al.[19] as an exemplar of the mutation-classes and spectrum under typical SARS-CoV-2 evolution in the relevant time period. To compare mutation burden by mutation-class between molnupiravir-treated and untreated individuals we scaled the Ruis et al. dataset of typical BA.1 evolution to have the same number of total mutations as untreated individuals in the Alteri et al. dataset, and then plotted these against molnupiravir-treated individuals from the Alteri et al. dataset (Fig. 2A).

To identify which mutation-classes were diagnostic of the use of molnupiravir, we first calculated what proportion of mutations came from each mutation class, for both the Alteri et al. molnupiravir dataset and the Ruis et al. BA.1 dataset. We then calculated the ratio of these proportions between the molnupiravir and (naive) BA.1 datasets. To put confidence intervals on this ratio we performed bootstrap resampling from each set of mutations (with 100 bootstrap repeats). These data are presented in Fig. 2B.

Identification of high G-to-A sequences from UShER mutation-annotated tree

To identify sequences in global databases with a molnupiravir-like pattern of mutations, we analysed a regularly-updated mutation-annotated tree built by the UShER team[43] using almost all global data from INSDC and GISAID – a version of the McBroome et al. (2021) tree[23]. We extracted data from using a script initially adapted from TaxoniumTools[44], and later modified to use the Big Tree Explorer (BTE)[45]. The script added metadata from sequencing databases to each node, then passed these metadata to parent nodes using simple heuristics: (1) a parent node was annotated with a year if all of its descendants were annotated with that year, (2) a parent node was annotated with a particular country if all of its descendants were annotated with that country, (3) a parent node was annotated with the mean age of its (age-annotated) descendants. Nodes with descendants spanning multiple years and/or countries were rare. We also calculated a more nuanced time estimate for nodes using Chronumental[46].

We defined "high G-to-A branches" as those with at least 10 mutations, of which >95% were transitions and >25% were specifically G-to-A mutations, with >20% C-to-T. Such a threshold appeared to yield very high specificity, as judged by the ability to detect marked changes in the rate of a rare event (molnupiravir treatment) over time. We also created simulated measures of sensitivity and specificity using the distribution of mutation types from Ruis et al. and Alteri et al, by drawing from each of the distributions and assessing what proportion of draws satisfied the criteria. This yielded 47% and 98.9% for branches with 10 mutations, 51% and 99.6% for branches with 15 mutations and 54% and 99.9% for branches with 20 mutations.

To measure whether high G-to-A branches showed a statistically significant increase in mutation rate we used Chronumental's branch length estimates in time, and performed statistical testing with a t-test on nodes from 2022.

Calculation of mutational spectra

To identify preferred nucleotide contexts for molnupiravir-based mutagenesis we calculated single-base substitution (SBS) spectra. For the high G-to-A branches, we extracted mutation paths from the UShER mutation-annotated tree. The context of each mutation was identified using the Wuhan-Hu-1 genome (accession NC_045512.2), incorporating mutations acquired earlier in the path. Mutation counts were rescaled by genomic content by dividing the number of mutations by the count of the starting triplet in the Wuhan-Hu-1 genome. MutTui (`https://github.com/chrisruis/MutTui`) was used to rescale and plot mutational spectra.

To calculate an SBS spectrum from the Alteri et al. dataset we used the mapped reads from BioProject PRJNA854613, again taking sites which had >=100 reads of which >= 5% were distinct from the day 0 consensus. We rescaled mutation counts to mutational burdens by dividing each mutation count by the number of the starting triplet in the Wuhan-Hu-1 genome (accession NC_045512.2).

We performed a similar analysis for the Donova-Banfield et al.[20]. We used deep sequencing data

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

from samples collected on day one (pre-treatment) and day five (post-treatment) from 65 patients treated with placebo and 58 patients treated with molnupiravir. For each patient, we used the consensus sequence of the day one sample as the reference sequence and identified mutations as variants in the day five sample away from the patient reference sequence in at least 5% of reads at genome sites with at least 100-fold coverage. The surrounding nucleotide context of each mutation was identified from the patient reference sequence.

To ensure that any spectrum differences between placebo and molnupiravir treatments are not due to previously observed differences in spectrum between SARS-CoV-2 variants[19,21], we compared the distribution of variants between the treatments (Fig. S3). The distributions were highly similar.

We compared the contextual patterns within each transition mutation type, assessing the correlation of the 12 mutation classes between values from the high G-to-A phylogenetic branches with those from the Alteri et al. dataset and separately, those from the Donovan-Banfield et al. dataset and the Ruis et al. control spectrum. P-values and R-values were calculated for each mutation class and dataset using Pearson's method. We performed the same correlational analysis within the long branch data, comparing the G-to-A subset with the C-to-T subset, matching the G-to-A context to its reverse-complement in the C-to-T dataset.

Identifying highly divergent molnupiravir-derived sequences excluded from the mutation-annotated tree

Given that in the process of construction of the UShER mutation-annotated tree highly divergent sequences can be excluded, we decided to perform a secondary analysis to identify divergent sequences with a molnupiravir signature. We used Nextclade[47] for this task. We supplied a full dataset of full-length FASTA sequences, and every sequence that could be aligned with Nextclade was included. Nextclade places each sequence onto a sparse tree reference phylognetic tree. Its outputs include a `unlabelled private mutations` column, which contains private mutations at a node with respect to the tree, excluding revertant mutations and mutations that are very common in other clades. We analysed this set of mutations for the presence of molnupiravir-like mutation-class distributions.

We selected sequences that had >=20 mutations of which >=20% where G-to-A, >=20% were C-to-T and >=90% were transitions. Again these were heavily enriched for dates after the roll-out of molnupiravir.

We placed identified sequences onto a downsampled global tree using `usher.bio` and visualised this tree using Nextstrain[48].

Selection analysis

We examined the types of mutations that made up these branches. We used BTE to determine whether each mutation observed was synonymous or not. Mutations were tallied by this status, grouped according to whether the branch was short (<10 mutations) or long (>=10 mutations) and whether it had a high G-to-A signature. We calculated dN/dS values by normalising to the availability of non-synonymous and synonymous sites in the genome.

We plotted the distribution of mutations across the genome for high G-to-A branches, split according to whether the mutations were synonymous or not, and also plotting the distribution specifically for the most recurrent non-synonymous mutations, occurring in four or more high G-to-A branches.

Processing and visualisation of cluster trees

The bulk trees presented in the supplement were plotted from the UShER tree using ggtree[49].

For the cluster of 20 individuals shown in Fig. 4A, we observed small imperfections in UShER's representation of the mutation-annotated tree within the cluster resulting from missing coverage at some positions. We therefore recalculated the tree that we display here. We took the 20 sequences in the cluster, and the three closest outgroup sequences, we aligned using Nextclade[47], calculated a tree using iqtree[50] and reconstructed the mutation-annotated tree using TreeTime[51]. We visualised the tree using FigTree[52].

Linkage analysis to treatment records

49 sequences with high G-to-A signatures from England, which fell into 35 clusters, were analysed by UKHSA. Sequences were linked to Blueteq treatment records[53] based on NHS number. Linkage could be established for all sequences. The analysis found that 11 of the 35 distinct clusters involved a molnupiravir-prescribed individual, giving a cluster hit-rate of 31%. Only sequences sampled after the treatment date were counted, with no upper time limit. The background proportion of sequences from 2022 which could be linked to molnupiravir-prescribed individuals was 0.043%. A rate of 0.043% would imply an expected 0.02 treatments from 49 individuals in the null condition, as compared to the 11 observed, meaning there is a >500-fold enrichment over this null.

ACKNOWLEDGEMENTS

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We are also very grateful to everyone who has contributed to the generation of the genomes that have been deposited in the INSDC databases, on which this research is also based. We thank Angie Hinrichs and colleagues for access to an UShER mutation-annotated tree built with all available genomic data. We would also like to acknowledge NHS England for providing the Blueteq data on treatment records. We thank Jesse Bloom, Michael Lin, Richard Neher, Kelley Harris, Florence Débarre and Meera Chand for useful discussions. This preprint uses a LaTeX template from Stephen Royle and Ricardo Henriques. The opinions are those of the authors and not necessarily those of UKHSA.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

DATA AVAILABILITY

No new primary data was generated for this study. We used data from consensus sequences available through GISAID and the INSDC[24,25], from the AGILE clinical trial[20], where genomic data were obtained from BioProject PRJNA854613 at the SRA, and from Alteri et al[18] from BioProject ERP142142. The AGILE investigators were not involved in the analysis and preparation of this manuscript.

Linkage analysis was performed within UKHSA. Section 251 of the National Health Service Act 2006 permits UKHSA use of patient-level data for specific projects.

The findings of this study are based on metadata associated with 15,572,413 sequences available on GISAID up to June 2023, and accessible at 10.55876/gis8.230110wz and 10.55876/gis8.230110db, 10.55876/gis8.230622mw (see also, Supplemental Tables). The findings of this study are also based on 7,104,124 sequences from INSDC – authors, metadata, and sequences are available here. Data present in both databases are deduplicated during the construction of the mutation-annotated tree on the basis of sequence, name, and metadata. We standardised to GISAID sequence names and accessions for sequences present in both databases.

A version of our analysis using only the INSDC subset of the tree, with INSDC naming conventions, is available at `https://github.com/theosanderson/molnupiravir/tree/main/open_data_version`.

CODE AVAILABILITY

Our GitHub repository is located at `https://github.com/theosanderson/molnupiravir`

AUTHOR CONTRIBUTIONS

RH identified initial branches, and their likely connection to molnupiravir. TS performed analyses of mutation-annotated tree and global metadata. CR performed all mutational spectra analyses. ID-B created bioinformatic pipelines for the AGILE trial data. TP functionally curated mutations identified in long branches. HH and AL performed linkage analyses. All authors participated in manuscript writing.

FUNDING

TS was supported by the Wellcome Trust (210918/Z/18/Z) and the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001043), the UK Medical Research Council (FC001043), and the Wellcome Trust (FC001043). This research was funded in whole, or in part, by the Wellcome Trust [210918/Z/18/Z, FC001043]. For the purpose of Open Access, the authors have applied a CC-BY public copyright licence to any Author Accepted Manuscript resulting from this preprint.

ID-B is supported by PhD funding from the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE) (now UKHSA), in collaboration with Liverpool School of Tropical Medicine and the University of Oxford (award 200907). The views expressed are those of the authors and not necessarily those of the Department of Health and Social Care or NIHR. Neither the funders or trial sponsor were involved in the study design, data collection, analysis, interpretation, nor the preparation of the manuscript.

TP was funded by the G2P-UK National Virology Consortium funded by the MRC (MR/W005611/1).

CR was supported by a Fondation Botnar Research Award (Programme grant 6063), the UK Cystic Fibrosis Trust (Innovation Hub Award 001) and funding from the Oxford Martin School.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Supplementary Information

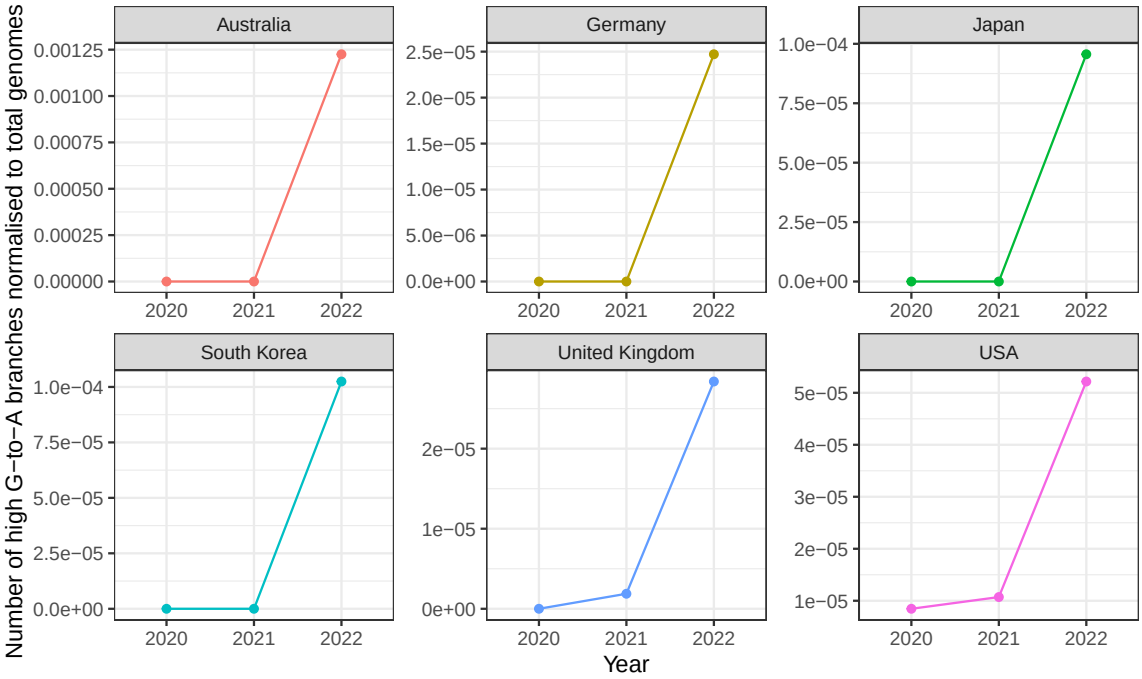


Figure S1. Timeline of number of high G-to-A branches, normalised for sequencing volumes, in 6 countries
The y-axis represents number of high G-to-A branches, divided by total sequencing volume for the year. This analysis demonstrates that the effects seen in raw numbers in Fig. 2D cannot be explained by changes in sequencing volume.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Country	High G-to-A branches in 2022	Total genomes in 2022
Australia	149	121,602
USA	106	2,031,795
United Kingdom	35	1,232,969
Japan	35	366,060
Russia	17	41,416
Germany	13	525,967
Italy	13	70,555
Israel	12	108,770
Slovakia	11	26,461
Denmark	10	332,006
Thailand	9	24,338
Austria	8	46,962
Spain	7	95,635
New Zealand	7	25,170
South Korea	6	58,567
Turkey	6	20,754
France	4	328,527
Czech Republic	4	32,124
Ireland	3	48,704
Indonesia	3	34,499
Luxembourg	3	30,715
India	2	121,841
Belgium	2	84,600
Philippines	2	12,830
Canada	1	217,040
Sweden	1	88,418
Poland	1	44,014
Mexico	1	35,857
Slovenia	1	31,221
Norway	1	30,796
South Africa	1	15,502
Latvia	1	14,039
Hong Kong	1	10,969
Brazil	0	98,346
Netherlands	0	64,614
Switzerland	0	49,062
Peru	0	30,772
Malaysia	0	27,113
Croatia	0	22,786
Chile	0	20,229
Portugal	0	19,483
Finland	0	18,518
Singapore	0	17,353
Greece	0	14,293
Colombia	0	12,302
China	0	11,425
Lithuania	0	10,059

Table 1. Number of high G-to-A branches from 2022 against total number of genomes from 2022 by country. Only countries with >10,000 genomes are included.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

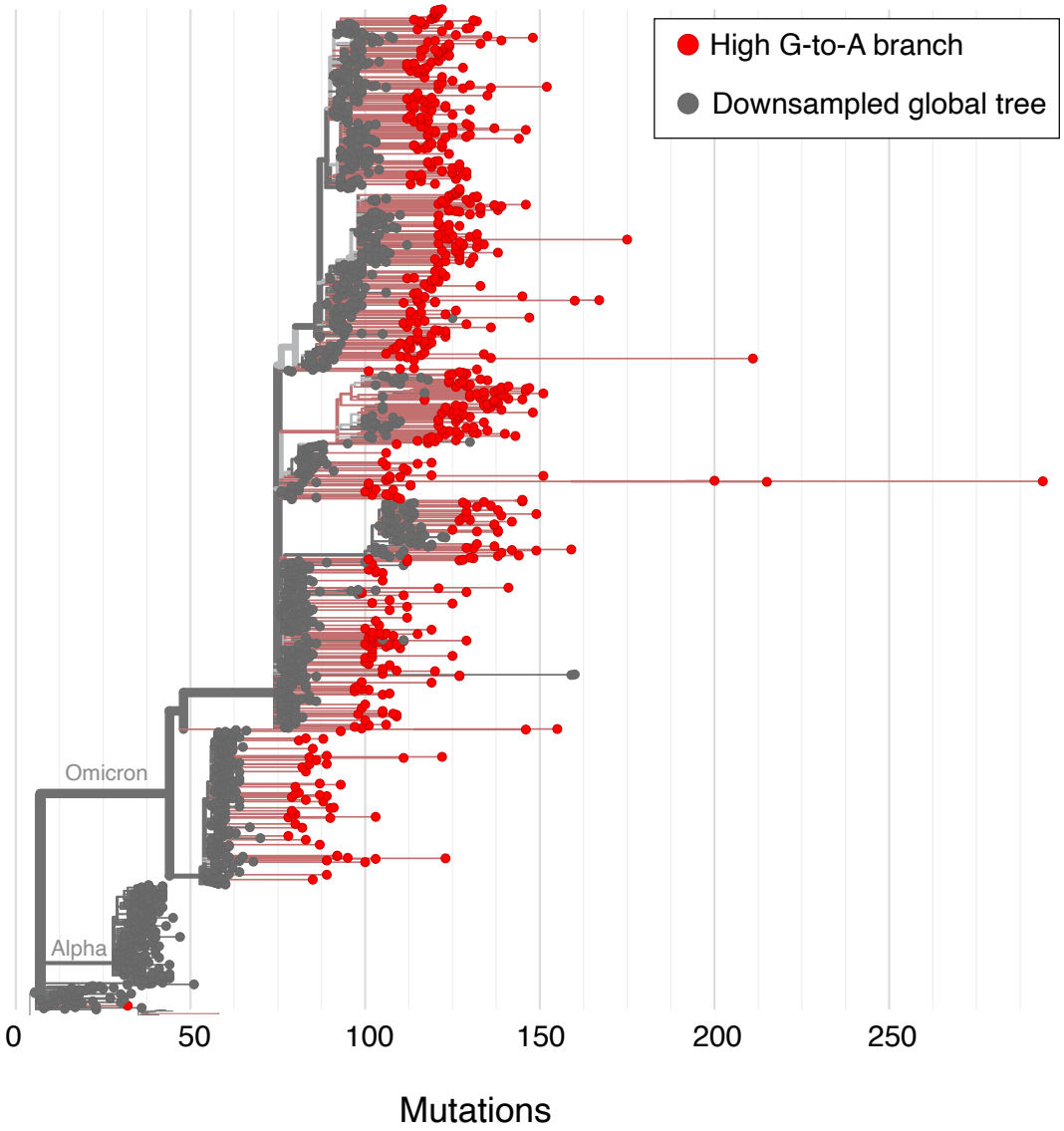


Figure S2. High G-to-A sequences with more than 20 private mutations identified from a Nextclade alignment of all available SARS-CoV-2 sequences

Nextclade was used to align sequences and identify private mutations. High G-to-A branches were identified on the basis of unlabelled private mutations. Usher.bio was then used to create a tree with high G-to-A branches highlighted on a downsampled global tree, with visualisation performed with Nextstrain.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure S3. Distribution of major SARS-CoV-2 variants between placebo and molnupiravir treatments in the AGILE trial dataset.

The proportion of patients infected with each variant is shown. The proportions are similar suggesting that differences between placebo and molnupiravir spectra will not be influenced by previously observed spectrum differences between variants (Ruis et al., Bloom et al.). VOC = variant of concern.

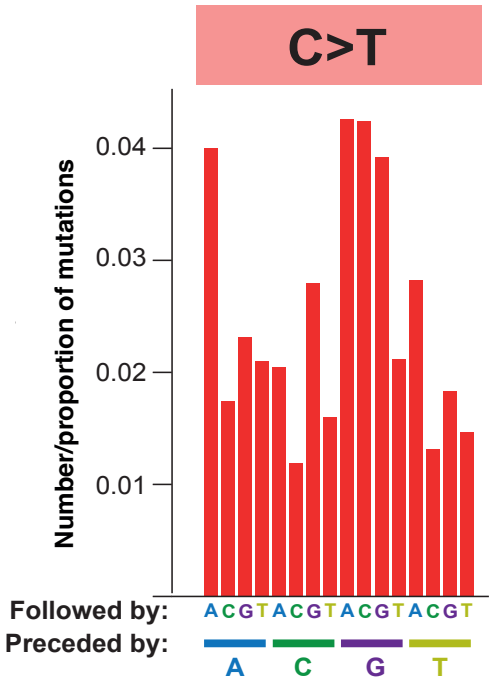


Figure S4. Context locations within the mutational spectrum.

The RNA mutational spectrum contains 12 mutation types, for example C-to-T. The spectrum also captures the nucleotides surrounding each mutation. There are four potential upstream nucleotides and four potential downstream nucleotides. This figure shows the location of each of the 16 contexts within an example mutation type. For example, the leftmost bar represents C-to-T mutations in the ACA context while the second leftmost bar represents C-to-T mutations in the ACC context. The spectrum represented is from AGILE trial data on monlupiravir.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.26.23284998; this version posted June 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

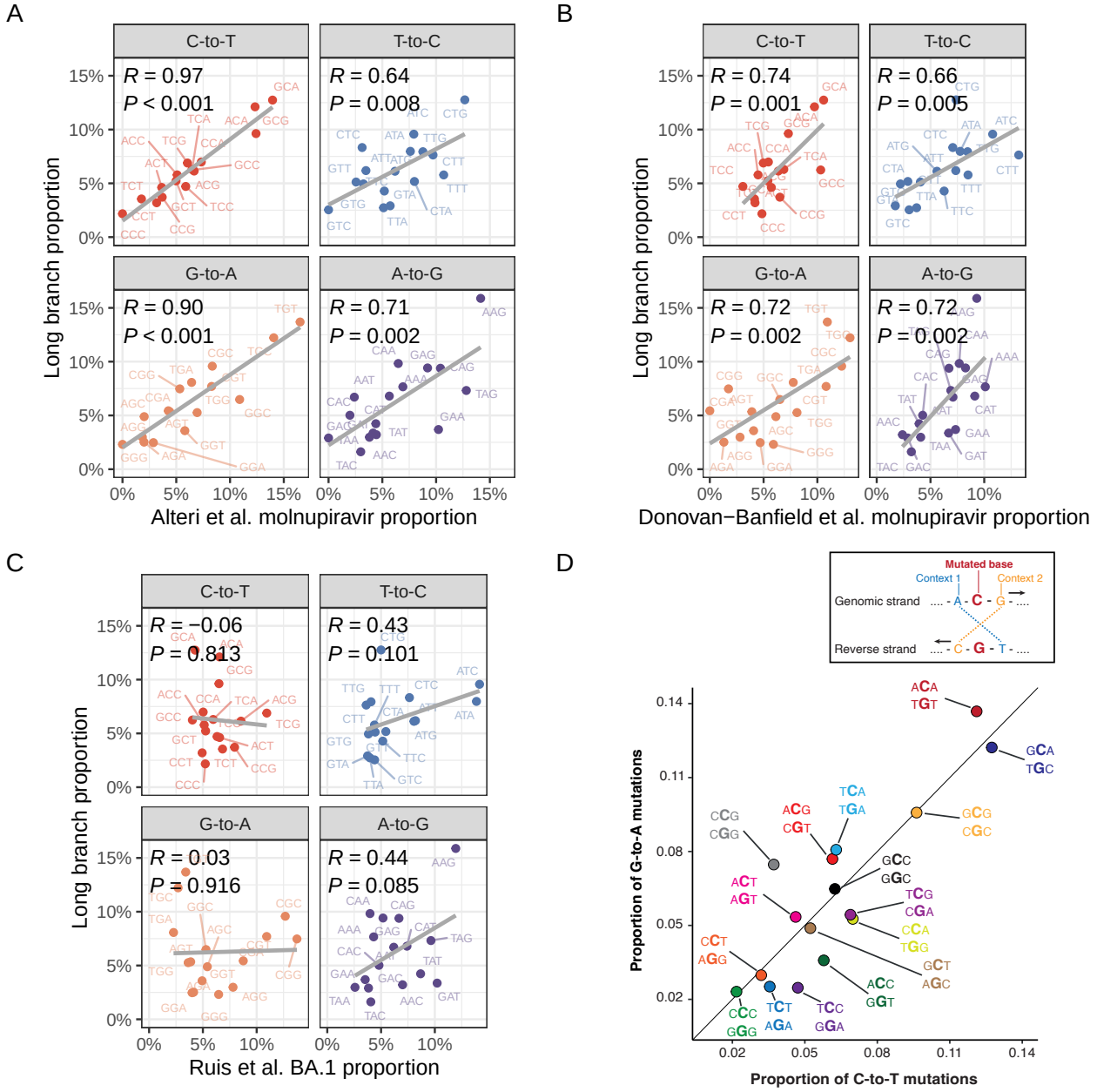


Figure S5. Mutation spectrum analysis supports a molnupiravir origin for high G-to-A nodes

(**A**) Strong correlation for contexts in all transition mutation classes between Alteri et al. molnupiravir-treated patients and high G-to-A long branches. (**B**) Similar analysis, with clear correlation between Donovan-Banfield et al. dataset of molnupiravir treated individuals to long high G-to-A branches. (**C**) Little correlation seen between contexts in typical SARS-CoV-2 evolution (Ruis et al.) and high G-to-A branches. (**D**) In data from long branches, context proportions for G-to-A mutations correlate with context proportions for C-to-T mutations, indicating a common mutational process. Points are labelled with G-to-A context, then C-to-T context.