bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The butterfly effect: mutational bias of SARS-CoV-2 affects its pattern of molecular evolution on synonymous and nonsynonymous levels

Alexandr Voronka*[1], Bogdan Efimenko*[1], Sergey Oreshkov[1,2], Melissa Franco[3], Zoe Fleischmann[3], Valerian Yurov[1], Arina Trufanova[1], Valeria Timonina[1,4,6], Natalia Ree[1], Emma Penfrat[4], Thomas Junier[4,6], Alexey Agranovsky[5], Konstantin Khrapko[3], Konstantin Gunbin[1], Jacques Fellay[4,6#], Konstantin Popadin[1,4,6#]

[1] Center for Mitochondrial Functional Genomics, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation.
[2] University of Lausanne, Lausanne, Switzerland.
[3] Northeastern University, Massachusetts, USA.
[4] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
[5] Faculty of Biology, Lomonosov Moscow State University, Moscow, Russian Federation.
[6] Swiss Institute of Bioinformatics, Lausanne, Switzerland.

*equal contribution
#equal contribution

Abstract

Evolution is a function of mutagenesis and selection. To analyse the role of mutagenesis on the structure of the SARS-CoV-2 genome, we reconstructed the mutational spectrum, which was highly C>U and G>U biased. This bias suggests that, in weakly constrained regions, the SARS-CoV-2 genome becomes increasingly U-rich. We analysed the consequences of this bias on the composition of the most neutral (four-fold degenerate synonymous substitutions) and the least neutral positions (nonsynonymous substitutions). The neutral nucleotide sites of the genome are already saturated by U, suggesting that in the past, an ancestral virus was exposed to similar mutational pressure, and in the future, we don't expect any significant changes since the system is at equilibrium. However, nonsynonymous changes continue slowly evolve towards equilibrium substituting CG-rich amino-acids ("losers") with U-rich ones ("gainers"). Comparing Coronaviridae with all other positive-stranded RNA viruses, we observed an excess of gainers and a deficit of losers in the former, suggesting that the coronavirus-specific mutational bias affected the amino-acid content of the entire family on the long-term evolutionary scale. We propose the butterfly effect - a tuning of a protein space through permissive amino-acid trajectories by mutational bias, which can be common in species with biased mutagenesis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Molecular evolution is a function of mutagenesis and selection. Deconvolution of these two factors is an essential step for a deep understanding of the whole process. Mutagenesis is far from random: a set of tissue and species-specific mutagens determines highly nonuniform probabilities of nucleotide substitutions. The mutational spectrum, defined here as a relative rate of single nucleotide substitutions in the neutral region of nucleic acid, allows characterising and comparing the mutational processes. Depending on the depth of input data, the mutational spectrum can be reconstructed with various precision: 2 components, the ratio of transitions to transversions (1, 2); 6 components, single nucleotide substitutions with symmetrical mutagenesis on both strands (where for example C>T is equivalent to G>A) (3); 12 components, single nucleotide substitutions with strand-specific mutagenesis (where C>T is not equivalent to G>A) (4); 96 components, the most commonly accepted spectrum in the cancer field, assumes symmetrical single nucleotide substitutions with trinucleotide context (4x6x4=96) (5, 6); 192 components, strand-specific single nucleotide substitutions with trinucleotide context (4*12*4) (7, 8) or even more.

The mutational spectrum, after reconstruction, can be used for various purposes. The most important implication is tracing potential mutagens through their mutational signatures (9). The mutational spectrum, interpreted as a new molecular phenotype of cancer patients, can have several clinical applications soon (10). Although historically, the progress in the analysis of the deep mutational spectra was associated with human cancers, currently, it goes beyond it and expands to all species. For example, several recent studies of SARS-CoV-2 mutational spectrum allowed to understand its evolution better: high similarities between the mutation spectra of SARS-CoV-2 and the bat coronavirus RaTG13, suggested the natural origin of SARS-CoV-2 (11); asymmetry and directionality of the mutational spectrum observed in SARS-CoV-2 enabled authors to trace the root of the tree (8); the mutational spectrum helped better to infer selection and suggest a way to design more robust vaccines (12).

Long-term consequences of the mutational spectrum lead to several fundamental questions: (i) how strongly do mutational spectra affect the nucleotide composition in neutral (for example, in synonymous positions) and not neutral (for example, nonsynonymous) sites; (ii) how far different species (genomes and genic regions) are from their mutation-selection equilibrium; (iii) how often mutational spectra is changing along the phylogeny? The reconstruction of the mitochondrial (mtDNA) mutational spectrum of hundreds of vertebrate species allowed to address some of these questions and demonstrated that synonymous positions of mammalian mtDNA (4) and mtDNA of Actinopterygii (13) are strongly affected by the mutational spectrum and close enough to equilibrium. One of the most famous and still unsolved discrepancies between the mutational spectrum and the genome-wide nucleotide content is the increased GC content of aerobic versus anaerobic bacteria (14–16): according to the well-known hallmark of oxidative damage, G>T is expected to be higher in aerobic versus anaerobic bacteria, making genomes of the former more GC-poor and AT-rich; however the observed trend is opposite: aerobic bacteria have a higher GC content as compared to anaerobic (14–16). It is important to emphasise also that mutational bias can change nucleotide composition not only in neutral sites but also in nonneutral ones - affecting the amino-acid composition of species (17–19).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here to analyse the consequences of the mutational spectrum on neutral and non-neutral sites of a genome, we focused on SARS-CoV-2, which is advantageous because of: (i) strongly asymmetric mutational spectrum; (ii) high absolute mutational rate, typical for all RNA viruses; (iii) an expansion phase during the pandemic, which is known to be associated with inefficient selection and accumulation of slightly-deleterious substitutions (20, 21); (iv) large scale dataset, allowing to reconstruct detailed mutational spectrum (22). First, we confirmed the highly biased towards U mutational spectrum; we demonstrated that synonymous positions are very close to the compositional equilibrium and nonuniformity in codon usage is shaped by the mutational spectrum. Second, we observed an effect of the mutational spectrum on the most common amino-acid substitutions, which are predominantly happen from CG-rich aminocadis (“losers”) to U-rich aminoacids (“gainers”). Finally, analysing the amino-acid composition of all single-stranded positive RNA viruses, we observed that all coronaviridae have an excess of gainers versus losers, assuming that, the coronaviris-specific mutational spectrum shapes its amino-acid content. We propose the butterfly effect scenario when little changes in the mutational spectrum (which presumably occurred in ancestral coronaviridae) in a very long-term perspective lead to drastic changes in the amino-acid composition. We hypothesise that the butterfly effect can be the strongest in species with the asymmetric mutational spectrum, high absolute mutation rate and relaxed purifying selection. We believe that the progress in the understanding the permissive (weekly selected) amino-acid trajectories (23) will help deconvolute the mutational bias's effect on amino-acid composition.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

Data collection and processing

A set of 4,339,984 full-length SARS-CoV-2 genome sequences was downloaded on 14.10.21 from GISAID (https://www.epicov.org). Duplicates and sequences with low quality, containing uncertain nucleotides (N, K, H etc.) or less than 29001 nucleotides, were removed. Filtered 1,139,387 sequences were aligned to Wuhan-Hu-1 reference genome (NC_045512.2) with bwa bwasw command (v.0.7.17-r1188) (24). Alignment was used for phylogenetic tree building and different descriptive statistics. The phylogenetic tree was constructed using FastTree 2.1.10 (25) on a sample set of aligned 203,045 genomes. The complete tree was pruned to 54,521 nodes with bottom-up approach to conserve more phylogenetic differences in the tree. Briefly, pruning consists of iterative deleting of nodes in terminal polytomies and random terminal nodes until the number of tree nodes was less than 55000. Ancestral genomes of pruned tree internal nodes were reconstructed using PRANK v.170427 (26) using the default model of DNA sequence evolution (HKY model). PRANK uses BppAncestor library (27) for the fast marginal maximum likelihood reconstruction of the most likely ancestral sequence at each node. This type of maximum likelihood ancestral reconstruction goes from the descendants to progressively assign the most likely character state to each ancestor based on only its immediate descendants and taking into account the user defined model of sequence evolution. This approach does not take into consideration the ancestral states' probability joint distribution in each alignment position and each node of the tree, and, therefore often converging to a local optima (28), which in turn, leads to overestimation of substitution number. However, this approach is fast enough for feasible computations of alignment consisting of several tens of thousands of viral genomes.

Single nucleotide substitutions from each node in the tree were called taking into account that upstream and downstream two-nucleotide neighbour positions were the same and did not contain deletions.

Mutational spectrum reconstruction

Mutation spectrum of SARS-CoV-2 was based on a sample of substitutions collected from the phylogenetic tree. We calculated mutational profiles of sample dataset of mutations, i.e. counted all possible variants of single nucleotide substitutions (12 numbers or 192 numbers in case of using the neighbouring nucleotides) and adjusted them to expected numbers of substitutions that were collected from the reference sequence Wuhan-Hu-1. When a mutation sample was collected from a distinct genome region or was based, for example, on synonymous substitutions, we adjusted the calculated mutational profile to corresponding expected substitutions numbers.

Comparison of mutation ratios in mutational spectra

Pairs of normalized values of 12-component spectra were compared using a binomial test to test imbalances in mutational spectra (29). We compared only substitutions that could be assumed as mutually exclusive, i.e. reciprocal pairs (C>U and U>C etc.) and complementary pairs (C>U and G>A etc.), consequently numbers of mutations in such pairs might be considered as binomially distributed. P-values were corrected with the Benjamini and Hochberg procedure.

A sampling of late and early genomes

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

From the full dataset of genome sequences downloaded from GISAID we extracted 3 groups of sequences splitted by date: from December 2019 to March 2020 (early), from the beginning of October 2020 till the end of October 2020 (intermediate) and from September 2021 to October 2021 (late). From each group we sampled 10000 sequences and used them in the analyses to compare the nucleotide frequencies.

Simulation of the nucleotide content

An expected neutral nucleotide composition, evolved under 12-component mutational spectrum, was obtained by two methods: computational simulations and differential equations. Both approaches gave very similar results. Details are available in our recent manuscript (4).

Analysis of codon usage of human single-stranded positive RNA viruses.

From the NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) we extracted RefSeqID of all human, completely sequenced single-stranded positive RNA viruses (N = 149 - segmented viruses are counted with each separate segment), downloaded all viral RefSeqs from ftp ncbi (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) and estimated their genome-wide codon usage.

Analysis of amino acid substitutions

To analyse how changes in amino acid composition depend on the mutation spectrum we used nonsynonymous substitutions. We separated all amino acid substitutions by mutational spectrum: direct substitutions (caused by the main substitutions C>U and G>U) and opposite substitutions (caused by U>C and U>G). Then to count whether the amino acid composition changes, we took the number of direct substitutions, for example, Val>Phe (GUU>UUU or GUC>UUC), and the number of opposite substitutions, Phe>Val (UUU>GUU or UUC>GUC). Using these numbers, we calculated how much the fractions of direct substitutions differ from the random value (0.5) using a binomial test.

RdRp dataset

From the Serratus virus database (30), Trees and Alignments section (https://serratus.io/trees), we downloaded alignments of RNA-dependent RNA polymerase sequences of (+)ssRNA viruses. For each sequence, we estimated amino acid usage and performed PCA on the obtained dataset.

Data access

Data and scripts are available on our GitHub: https://github.com/mitoclub/SARS-CoV-2-MutSpec.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

1. The mutational spectrum of SARS-CoV-2 demonstrates strong strand-asymmetry and directionality both leading to an increase in U

To uncover the main patterns of molecular evolution of SARS-CoV-2 it is important to distinguish mutagenesis from a selection. In order to do it, we reconstructed the mutational spectrum of SARS-CoV-2 by the following pipeline (see methods for details):
(i) sampled 1,139,387 whole-genome sequences out of 4,339,984 available at GISAID at 14.10.2021;
(ii) aligned genomes against the Wuhan-Hu-1 reference genome and pruned the set to 54,521 sequences to make the following steps computationally feasible;
(iii) reconstructed phylogenetic tree and ancestral sequences in each internal node;
(iv) counted all observed single nucleotide substitutions and normalized them on nucleotide content (for each type of substitution we divided the number of observed mutations by the number of expected ones). This normalization step eliminates the effect of different nucleotide compositions, which helps to uncover asymmetry and directionality, as well as makes the mutational spectra comparable between different organisms and signatures.

As a result of our pipeline we got 542 768 high-quality (see methods) single-nucleotide substitutions among which 314 538 were nonsynonymous, 206 319 synonymous and 92734 synonymous at four-fold positions. Next, we reconstructed three types of the mutational spectrum of SARS-CoV-2: (i) “ALL” is a spectrum based on all mutations, which assures the highest sample size with the cost of potential selection signal; (ii) “SYN” is a spectrum based on synonymous mutations only, which is more neutral but less numerous and (iii) “SYN4F” is a spectrum based on synonymous mutations in four-fold degenerate sites which is the most neutral subset of mutations but, in parallel, is the least numerous one (some contexts are missing due to the structure of the genetic code).

All three mutational spectra showed (i) strong excess of C>U substitutions, (ii) strand-asymmetry and (iii) directionality. C>U substitutions contributed the majority of substitutions: 37% for ALL (Fig 1A), 25% for SYN and 35% for SYN4F (Fig. S1a-b). The mutational spectrum was strand-asymmetrical in terms of the complementary base pairing: if a probability of a given nucleotide mutating into another one (for example C>U) is the same on all strands, frequencies of C>U on the positive strand would be similar to C>U on the negative strand, which is equivalent to G>A on the positive strand. However, the observed frequencies of such complementary substitutions on positive strand were different with a strong excess of C>U as compared to G>A and G>U as compared to C>T (Fig 1A, Fig. S1a-b), suggesting that there is either a strand difference in mutagenesis or reparation. The mutational spectrum was also directional in terms of the rates of reciprocal mutation, i.e. mutated and substituted bases: for example, C>U is higher than U>C and G>U is higher than U>G (ALL - Fig 1A; SYN and SYN4F - Fig. S1a-b; Supplementary Tables 2-4, results of the binomial test) suggesting that, in the absence of selection, an increase in U is expected.

To check the stability of the mutational spectra along the viral genome we reconstructed the mutational spectrum for each separate gene. We observed that the mutational spectrum was rather similar along the

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

viral genome (Fig. S1c), suggesting that the key mutational process is not gene-specific but rather universal for the whole genome.

To check the stability of the mutational spectrum along the time of the pandemic, we reconstructed mutational spectra from nine time intervals of several months. In order to do it, we focused on genomes, sequenced within those time intervals and used only terminal branches of the phylogenetic tree for the spectra reconstruction (see methods). We observed that (Fig 1B) ALL mutational spectrum changed through the pandemic: it became even more directional (C>U increased while U>C decreased; G>U increased while U>G decreased) and even more strand-asymmetrical (C>U increased while G>A didn't change; G>U increased while C>A didn't change). A similar pattern was observed also for SYN and SYN4F spectra (Fig. S1d-e). This finding couldn't be explained by the graduate and low changes in the nucleotide composition of the viral genome and thus it requires a special explanation, based either on changes in error-associated techniques, viral mutagenic properties or host mutagenic environment.

To evaluate the effect of the nucleotide context, which typically has a strong impact on mutagenesis (10), we analyzed the three-nucleotide context, resulting in the 192-component spectrum (4 neighbour nucleotides at 5' x 12 types of substitutions x 4 neighbour nucleotides at 3'). ALL spectrum (Fig 1C) showed that C>U occurred preferentially in the context of [A/C/U]CG, while G>U substitution took place in the context [G/U]GU and GGA. Analysis of the 192-component spectra, reconstructed for SYN and SYN4F sets, demonstrated similar mutagenic motives (Fig. S1f-g).

2. The mutagenesis shapes the SARS-CoV-2 nucleotide content

Both strand-asymmetry and directionality of the SARS-CoV-2 mutational spectra (Fig 1, Fig. S2) are expected to increase the frequency of U unless selection is strong enough to purge this effect. Do we observe an increase in U content during the pandemic or the mutational increase in U is completely erased by a selection, decreasing frequency of U? To evaluate the effect of the mutational bias on the integral molecular evolution of the viral genome, we analyzed the changes in the nucleotide composition of the complete genome along the time of the pandemic. We randomly sampled genomes from three different time points (see fig 2A) and used a reference genome as a proxy for the most ancestral state. In line with our expectations, we observed an accumulation of U (an increase in 10 nucleotides between the first and the last time point), and a reduction of all other nucleotides (A, G and C) on the scale of the complete viral genome (dataset ALL, Fig 2A). It is important to emphasise that this integral analysis - a comparison of the whole-genome nucleotide content of random genomes, sequenced at certain time intervals - depended on a full mix of evolutionary forces, such as mutagenesis, selection and drift. Despite this, we observed that the changes in the whole-genome nucleotide composition follow the mutational bias, suggesting that mutagenesis is one of the strongest determinants of the current molecular evolution of SARS-CoV-2.

Interestingly, analyses of nucleotide composition based on the SYN4F set demonstrated no trend during the pandemic (Fig. S2a). This might be due to either the low number of such sites and weak statistical signal or the proximity of these sites to the equilibrium (i.e. saturation effect). To test the potential proximity of the sites to the nucleotide compositional equilibrium we compared four categories of sites characterized by different selective constraints and mutation rate: ALL NpN - all sites in the genome

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

irrespectively of the nucleotide context (N is any nucleotide); ALL NpG - all sites in the genome preceding G (which is the hot spot for mutations C>U, see fig 1C); SYN4F NpN - four-fold degenerated synonymous sites irrespectively of the nucleotide context; SYN4F NpG - four-fold degenerated synonymous sites in the genome, preceding G. We observed a significant excess of U fraction in the least constrained (SYN4F versus ALL) and the most mutagenic sites (NpG versus NpN context, where N is any nucleotide) (Fig. 2B, Fig. S2b-c), suggesting that the nucleotide composition of the neutral and highly mutated sites is the closest to the expected equilibrium, which can be explained by the long-term evolution of a SARS-CoV-2 (even in previous hosts) under similar mutational pressure.

To estimate quantitatively the expected neutral (if we do not take into account the selection) nucleotide composition of the SARS-CoV-2 we performed a computer simulation, based on a 12-component mutational spectrum of synonymous substitutions SYN4F and random initial nucleotide compositions (see methods). Interestingly, we observed that an equilibrium solution precisely reflects the observed nucleotide composition at SYN4F sites, (Fig 2C), confirming that SYN4F sites in SARS-CoV-2 genome has already been affected by the very similar mutational spectrum and this process was strong enough to shape the nucleotide composition towards the equilibrium. The analytical exact solution of the nucleotide compositional equilibrium, obtained with an input 12-component mutational spectrum, was completely in line with the simulation data and confirmed that SYN4F nucleotide composition is very close to equilibrium.

3. The codon usage of SARS-CoV-2 is affected by the mutational bias

The mutational bias affects the global nucleotide content of synonymous sites (Fig 2), which can be traced to the level of codon usage for individual amino acids. We plotted a standard genetic code as a heatmap with colour-coded numbers of codons from all protein-coding genes of the SARS-CoV-2 reference genome (Fig 3A). On the heatmap there are red horizontal stripes emphasising an increased number of NNU codons compared to NNC, an increased number of NNU codons compared to NNG and an increased number of NNA codons as compared to NNG. All these patterns can be explained by two of the most common asymmetrical mutations: C>U and G>U. The strongly asymmetrical codon usage of the reference genome, which is in line with the effect of the described mutational spectrum (Fig 1) provides additional lines of evidence that SARS-CoV-2 before the entering of the human host has evolved under a very similar mutational spectrum, suggesting that the key mutagens of SARS-CoV-2 are not host-specific, but rather viral-specific.

Changes in the codon usage between early and late genomes during the pandemic did not show any significant trends (Fig. S3), which is in line with the equilibrium state in neutral positions (Fig 2C).

4. Amino acid substitutions are affected by the mutational bias

If selection is weak and the mutational spectrum is asymmetrical or directional, it may affect the pattern of amino acid substitutions. To test this, on the standard genetic code table, we plotted the expected trajectories of amino acid substitutions taking into account two of the most frequent asymmetrical substitutions: C>U and G>U (see the “butterfly scheme” in Fig 4A, named so due to a similarity with a

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

buttefly). According to these trajectories, we categorised all amino acids into four groups: losers (Pro, Ala, Arg, Gly, Thr, Ser (AGX), Arg, His, Gln, Asp, Glu), gainers (Phe, Leu (UUX), Ile, Met, Tyr), intermediate (which are in the middle of the trajectory: Leu (CUX), Cys, Trp, Ser, Val) and neutral to the mutational bias (Asn, Lys). Next, we empirically tested these predictions using two types of data: (i) all amino acid substitutions reconstructed from the phylogenetic tree, (ii) changes in the amino acid composition between the late and early genomes.

Among all amino acid substitutions reconstructed on the phylogenetic tree (314538 mutations), the majority (193102 mutations) were located along the hot mutational trajectories (forward or backwards to the red arrows in Fig 4A). Among them, mutations concordant with the directionality of the mutagenesis (99689 mutations, ~ 32% from all amino-acid substitutions, from losers to gainers or from losers to intermediate or from intermediate to gainers) were in 6.7% excess as compared to mutations, discordant with the direction of the mutational trajectories (93413 mutations from gainers to losers or from gainers to intermediate or from intermediate to losers). The difference in the directionality was statistically significant (p = 5.1e-9, binomial test). Interestingly, the directional effect was also significant for most individual pairs of amino acid substitutions (Fig 4C). For example, we observed 4019 mutations from Val to Phe and only 3280 from Phe to Val (Supplementary Table 5). Moreover, the same trend is visible not only at the level of all protein-coding genes but also separately within the ORF1AB region and structural protein-coding regions (Fig. S4B-C). An additional observation is that amino-acid substitutions shaped by C>U mutations (red color in Fig 4C) are less biased as compared to substitutions, shaped by G>U mutations (blue color in Fig 4C), which can be explained by the saturation effect: the majority of effectively-neutral amino-acid substitutions driven by the most common C>U substitutions already happened and the de novo ones are under purifying selection, while substitutions driven by G>U are far from saturation and still are accumulating. For example, Arg(CGX) shows no significant trend towards Cys and Trp (Fig 4C), probably because it is already close to the minimum frequency due to the mutational hot spot: CpG context is one of the strongest contexts to mutate towards UpG (Fig 1C).

After the origin of a mutation, its fate depends on the reproductive success of a carrier. In our previous “phylogeny tree” analyses (Fig 4C), we didn’t consider the frequency of descendants of each mutation, thus diminishing a potential effect of selection. Here, to trace the effect of the mutational bias on the global evolution of SARS-CoV-2, we run an integral analysis by sampling late and early genomes (see methods) and comparing their amino acid composition (Fig 4B). We observed that Pro, Thr and Asp demonstrated the strongest decrease during the pandemic, while Leu (UUX) and Ile demonstrated the strongest increase in their numbers (Fig 4B). Interestingly, the dynamics of these five amino acids are consistent with the mutational trajectories: those with decreasing trends are losers, and those with increasing trends are gainers. (Fig 4A). Moreover, if we take into account eleven expected losers (Fig 4A) we observe that eight of them (all except Arg(CGN), Ser(AGN) and Gln) indeed increase in frequency at least moderately. Similarly, five of the six expected gainers show an increase (all except Phe). Altogether, we observe that even at the amino acid level, asymmetrical mutagenesis shapes the pattern of amino-acid changes between early and late SARS-CoV-2 genomes.

The recent and highly transmissible line of SARS-CoV-2, omicron, is characterised by the numerous nucleotide changes from the Wuhan reference sequence. To test whether the amino-acid substitutions associated with the evolution of omicron are in line with our hypothesis, we analysed all amino acid

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

changes between the Wuhan and omicron sequences. Interestingly, among all amino acid changes (Supplementary Table 6), we observed that 12 amino acid substitutions were in line with the mutational bias. In contrast, only 3 - in the opposite direction (binomial test, the success rate is 0.75, p-value < 0.01). An excess of loser to gainer substitutions in the evolution of the omicron line shows that the mutational spectrum has a predictive strength even on the amino-acid level.

5. Mutational spectra have long-term amino-acid consequences for all single-stranded positive RNA viruses

The codon usage of SARS-CoV-2 reference genome was affected by the mutational bias (Fig 3A) and thus the codon usage of other viruses may shed light on their mutational processes. To test it we analysed genome-wide codon usage of all completely sequenced single-stranded positive RNA viruses affecting humans (N = 149, including separate segments for segmented viruses). Next, we estimated the nucleotide composition in SYN4F sites of these viruses and performed the principal component analysis based on the frequency of A, U, G and C in SYN4F of these 149 viruses. We observed that all coronaviridae form a separate cluster, characterised by the increased fraction of U (Fig 5A, segmented viruses like influenza are presented on this plot with several segments).

If mutational bias affects amino acid composition (Fig 4), we can use also comparative species data on amino acid content to proxy the mutational process. We expect that an excess of U in neutral positions (in coronaviridae, Fig 5A) will be linked with an excess of U in non-neutral sites (i.e. excess of gainers in coronaviridae). Oppositely, U deficiency in neutral positions of other species (for example in Bastrovirus BAS−1, Hepacivirus C, Zika virus and others, see Fig 5A) will be associated with U deficiency in non-neutral position of these species (excess of losers). To test this, we performed two analyses. First, we estimated the whole-genome amino acid composition of all human single-stranded positive RNA viruses and derived for each species a fraction of gainers and losers. We observed an expected strong negative correlation between gainers and losers, with coronaviridae characterised by an excess of gainers, while viruses such as Bastrovirus BAS−1, Hepacivirus C, Zika virus (with low frequency of U in neutral positions) - by an excess of losers (Fig 5B). This suggests that the global amino acid composition of human single-stranded positive RNA viruses is affected by the mutational process. Second, taking into account an excess of U among neutral (four fold degenerated) sites of coronaviridae (Fig 5A) and an excess of gainers among all amino-acids of coronaviridae (Fig 5B) we expect to observe a positive correlation between the frequency of U in neutral positions and the frequency of gainers (which is essentially a frequency of U in not-neutral positions). Indeed the observed positive correlation (Fig 5C) confirmed our hypothesis about a link between mutational spectrum and amino-acid composition.

In the analyses above (Fig 5B, Fig 5C) we used the whole genome amino-acid composition of each viral genome, ignoring potential gene-specificity (different viruses have different sets of genes and different genes have different amino-acid compositions). To overcome this limitation we analysed a recent database (30) which focused on the RNA-dependent RNA polymerase (RdRP) gene in more than 10^5 RNA viruses (species-like operational taxonomic units). In the scope of the current project, we focused on all (N = 60793) positive single-stranded RNA viruses (+ssRNA) and estimated an RdRP-specific amino-acid composition of each virus. First, we observed, that coronaviridae as compared to other Nidovirales and all +ssRNA viruses are characterized by an excess of gainers and a deficit of losers (Fig 5D). Second,

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

performing PCA, we observed that, both cumulative vectors of all gainers and all losers have high enough projections on pc1 and pc2, suggesting that the mutational spectrum is rather an important factor of the molecular evolution of the RdRP gene (Fig 5E). Interestingly, some amino acids, which formally do not support our expectations, still can be driven by mutagenesis. For example, three of four loser amino-acids, showing strong excess in coronaviridae (see four blue arrows with up-right direction Fig 5E): His, Gln and Asp (all except Gly) are in line with the G>A substitution which is the fourth most common substitution in the 12-component mutational spectrum (Fig 1A). This suggests that a more detailed analysis of the effect of the mutational spectrum on amino-acid composition can uncover an even stronger effect of mutagenesis on molecular evolution. Also, despite the observation that Coronaviridae are among the viruses with the highest frequency of gainers, another order - Potatovirales are even higher enriched in gainers. This suggests that a link between life-history traits of different viruses and their specific mutational spectra can be an interesting gap to fill in the future. Altogether, our results are in line with our hypothesis and show that species-specific mutational spectrum and species-specific amino-acid composition can be tightly linked with each other. Future analyses are needed to shed a light on the effect of the mutational spectrum on the amino-acid composition.

DISCUSSION

Analyzing more than 542768 phylogenetically polarized single-nucleotide substitutions, accumulated during almost two years of the pandemic (from 24.12.2019 to 14.10.2021), we revealed detailed SARS-CoV-2 mutational spectra based on the reconstructed phylogenetic tree. We hypothesized that the majority of analyzed mutations (especially from SYN and SYN4F sets) were selectively neutral and thus the reconstructed mutational spectra reflect the mutagenic process rather than selection. The congruency of the three spectra (ALL - for all mutations, SYN - synonymous substitutions and SYN4F - synonymous mutations in four-fold degenerate sites) with each other (Fig 1, Fig. S1) suggests that indeed all of them are mainly shaped by the mutagenesis. It is important to emphasize that in our approach we count each new mutation only once - at the time of the origin of a given mutation on a given branch on the tree. Even if this mutation was beneficial and led to the strong expansion of the descendant genomes, it will contribute only one event of mutagenesis to the total mutational spectrum. In the case of deleterious variants, a given genome will not propagate and thus does not contribute to our datasets. Thus, the total burden of mutations in our dataset is mainly neutral or effectively neutral. The high similarity of the observed and expected nucleotide compositions (Fig 2C) and nonuniform codon usage of the reference sequence (Fig 3A) suggest that the ancestral viral genome was affected by a very similar mutational process before the entering the human host.

In line with other papers (8, 31) we confirmed strong strand-asymmetry and directionality of the 12-component spectra (Fig 1A, 1C), corroborating that an excess of C>U and G>U substitutions is a hallmark of SARS-CoV-2 mutagenesis. In the case of 192-component spectra, our results show slightly different motifs due to our normalisation. According to several authors (8, 31, 32), the most mutagenic context for C>U was ACU, ACA, UCA, UCU and GCU. More detailed analysis showed that normalization on the trinucleotide content of the SARS-CoV-2 genome wasn't taken into account and thus their results were biased by the nonuniform background nucleotide content (excess of [AU]C[AU] motifs in the SARS-CoV-2 genome). Our results show that the most C to U substitution occurred in the

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[A/C/U]CG context. We would like to emphasize the importance of the normalization procedure for future works: it significantly facilitates a comparison of mutational spectra obtained from different organisms with various nucleotide content. Historically, the main progress in the reconstruction and interpretation of the mutational spectra has been done in the field of human cancers (33, 34), where normalization was not important and thus, not used, since the background always was the human genome. Currently, in the era of reconstruction of mutational spectra for different organisms, normalisation becomes more and more important. It will facilitate not only the comparison of the mutational spectra but also their deconvolution into signatures of different mutagens using large-scale databases such as COSMIC and SIGNAL.

Mutational bias, if stronger than selection, can affect amino acid substitutions (Fig 4A). The first results, in line with the effect of the mutational spectrum on the amino-acid composition of SARS-CoV-2, were suggested by Simmond (31), which we took into account and investigated in more detail. The butterfly scheme (Fig 4A) helps to formalize our expectations and run tests. Analyzing amino acid substitutions reconstructed on the SARS-CoV-2 phylogenetic tree (Fig 4C), comparing amino acid frequencies between early and late genomes (Fig 4B) and comparing Wuhan and omicron genomes we observed that the majority of substitutions indeed follow the predicted mutational trajectory: from losers to gainers. An extension of this logic on all +ssRNA viruses (Fig 5) suggests the long-term importance of the coronaviridae-specific mutational spectrum on the coronaviridae-specific amino-acid composition. Despite the fact that we can’t eliminate selective argument, the mutational explanation is the most parsimonious one for our main findings. Thus we propose that viral-specific amino acid content can be associated with the viral-specific mutational spectrum as a result of the accumulation of slightly-deleterious variants (losers to gainers) during the effectively neutral molecular evolution. It is important to emphasize that the large-scale accumulation of slightly-deleterious variants is typically expected for genomes with low effective population size, such as the mitochondrial genome of mammals (35). Here we show that the same effect is visible in viral species with very high, but strongly fluctuating, effective population sizes. An additional reason might be that an expansion of population, happening during the pandemic, is associated with an increased mutational burden (20).

Our results are in line with a classical “direction mutation pressure” hypothesis by Sueoka (17, 18) and the seminal paper describing correlations between the nucleotide and amino-acid content (19). In our project, we added one more important piece of information - a deeply reconstructed mutational spectrum, which was missing in the previous classical papers (due to the absence of large-scale data). It adds a strong argument that all observed correlations between nucleotide and amino-acid content are caused by the mutational bias and corroborate the “direction mutation pressure” hypothesis. RNA viruses are among the fastest evolving creatures on our planet and thus uncovering a detailed mutational spectrum and consequences of the spectrum on nucleotide, codon and amino acid content is of high importance. Recent large-scale mutational datasets allow uncovering the permissive trajectories of amino-acid substitutions (23) and understand better the fitness landscape of different proteins (36). Combined with a deep understanding of the mutational spectrum such studies can predict the global direction of amino-acid changes. This can help to better predict the spectrum of the most probable amino acid substitutions, significantly decreasing the dimensionality of expected variants. Understanding the mutational spectrum can help to uncover also evolutionary constrained regions, since ‘under all else equal’ they are expected to have an excess of losers and a deficit of gainers. Oppositely, some regions which look constrained (frozen

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequence) in reality can be just a mutational cold spot (gainers). The proposed “butterfly effect” when the tiny changes in the pattern of mutagenesis can affect the amino-acid content in effectively-neutral regions of different genomes over millions of years can be a very common evolution scenario in species with the asymmetric mutational spectrum, low effective population size and/or high rate of population expansion.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ACKNOWLEDGMENTS

The authors acknowledge Arthur Zalevsky for valuable comments on the protein flexibility and permissive trajectories of amino-acid substitutions in light of the mutational spectrum.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURES

1. Mutational spectrum of SARS-CoV-2 based on an ALL set of mutations.

All substitutions are presented in the positive strand notation, which is a reference sequence of the SARS-CoV-2. The colour code (12 types of substitutions) is the same in all panels.

1A The 12-component mutational spectrum of SARS-CoV-2. For the figure, we used the set of mutations ALL.

1B Changes in the 12-component mutational spectrum of SARS-CoV-2 over pandemic. For the figure, we used the set of mutations ALL.

1C 192-component mutational spectrum of SARS-CoV-2 reconstructed for all nucleotides. For the figure, we used the set of mutations ALL.

2. Changes in the nucleotide composition of SARS-CoV-2 over pandemic

The colour code (4 types of nucleotides) is the same in all panels.

2A Changes in the nucleotide composition of SARS-CoV-2 in all sites of the genome during the pandemic. The most left horizontal interval corresponds to the nucleotide composition of the reference sequence.

2B Nucleotide composition in the reference sequence of SARS-CoV-2. An excess of U is more prominent in sites with low selective constraints (SYN4F versus ALL) and a high substitution rate (G versus X).

2C Observed and expected nucleotide compositions are very similar. SYN4F dataset.
Coloured shadows reflect changes in the nucleotide frequencies with the simulation time: they start at random positions and converge to their equilibrium state. Dash lines reflect the nucleotide composition of SYN4F in the reference sequence. Squares reflect an exact equilibrium state derived by the analytical approach with differential equations.

3. The codon usage of SARS-CoV-2 is affected by the mutational bias

3A A heatmap of the codon usage in the SARS-CoV-2 reference genome. All codons of all protein-coding genes of the SARS-CoV-2 reference sequence were plotted as a heatmap. Horizontal stripes (increased frequency of NNU and NNA) can be explained by the mutational bias C>U, G>U and probably G>A.

4 Amino acid substitutions of SARS-CoV-2 are affected by the mutational bias: a butterfly effect

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4A A "butterfly scheme" - expected amino acid trajectories driven by the C>U and G>U mutations in the space of the standard genetic code (indicated by red arrows). Arrows representing amino acid trajectories are drawn from the expected losers (Pro, Ala, Arg, Gly, Thr, Ser (AGX), His, Gln, Asp, Glu) to the expected gainers (Phe, Leu (UUX), Ile, Met, Tyr).

4B Average changes in amino acid composition over the course of the pandemic: increase (in red) in the frequency of gainers and decrease (in blue) in the frequency of losers in the late versus earlier genomes (see methods).

4C The pattern of the amino acid substitutions reconstructed from the phylogenetic tree. All substitutions are ranked correspondingly to the Grantham distance: from the amino acids closest to each other in physico-chemical terms, at the top, to the most distant ones at the bottom. Red colour marks substitutions shaped by C>U mutations, and blue - mutations shaped by G>U mutations. *_UG and *_UA denote stop codons UGA and UAA/UAG correspondingly. Asterisks mark statistically significant deviations from 0.5: ** p value < 0.01, *** p-value < 0.001 (binomial test, see methods and main text; all p-values are nominal).

5 Mutational spectra have long-term amino-acid consequences for all single-stranded positive RNA viruses

All viruses belonging to Coronaviridae family are marked in red colour.

5A Principal component analysis of the neutral (SYN4F) nucleotide content of all completely-sequenced human single-stranded positive RNA viruses.

5B Whole-genome frequencies of gainers and losers among all completely-sequenced human single-stranded positive RNA viruses.

5C Positive correlation of gainers frequency with neutral U frequency (SYN4F) in all completely-sequenced human single-stranded positive RNA viruses.

5D Frequencies of gainers and losers in RNA-dependent RNA polymerase gene of 60793 single-stranded positive RNA viruses. Coronaviridae as compared to other Nidovirales and all +ssRNA viruses are characterized by an excess of gainers and a deficit of losers.

5E Principal component analysis of amino-acid content of RNA-dependent RNA polymerase gene from 60793 single-stranded positive RNA viruses (species-like operational taxonomic units from serratus database). Arrows show a loading of each amino-acid on the first and the second principal components: red marks gainers, blue marks losers, black marks intermediate and neutral amino-acids including leucine (since without the codon structure we can't distinguish Leu(CUV) from Leu(UUX)).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Belle,E.M.S., Piganeau,G., Gardner,M. and Eyre-Walker,A. (2005) An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene*, **355**, 58–66.

2. Seplyarskiy,V.B., Kharchenko,P., Kondrashov,A.S. and Bazykin,G.A. (2012) Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol. Biol. Evol.*, **29**, 1943–1955.

3. Montooth,K.L. and Rand,D.M. (2008) The spectrum of mitochondrial mutation differs across species. *PLoS Biol.*, **6**, e213.

4. Mikhaylova,A.G., Mikhailova,A.A., Ushakova,K., Tretiakov,E.O., Iliushchenko,D., Shamansky,V., Iurchenko,A., Zazhytska,M., Kozenkova,E., Zdobnov,E., *et al.* (2022) A mitochondria-specific mutational signature of aging: increased rate of A>G substitutions on a heavy chain. *bioRxiv*, 10.1101/2021.12.03.460832.

5. Zou,X., Koh,G.C.C., Nanda,A.S., Degasperi,A., Urgo,K., Roumeliotis,T.I., Agu,C.A., Badja,C., Momen,S., Young,J., *et al.* (2021) A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat Cancer*, **2**, 643–657.

6. Kucab,J.E., Zou,X., Morganella,S., Joel,M., Nanda,A.S., Nagy,E., Gomez,C., Degasperi,A., Harris,R., Jackson,S.P., *et al.* (2019) A Compendium of Mutational Signatures of Environmental Agents. *Cell*, **177**, 821–836.e16.

7. Klimczak,L.J., Randall,T.A., Saini,N., Li,J.-L. and Gordenin,D.A. (2020) Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. *PLOS ONE*, **15**, e0237689.

8. Yi,K., Kim,S.Y., Bleazard,T., Kim,T., Youk,J. and Ju,Y.S. (2021) Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp. Mol. Med.*, **53**, 1229–1237.

9. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Tian Ng,A.W., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A., Bergstrom,E.N., *et al.* (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.

10. Koh,G., Degasperi,A., Zou,X., Momen,S. and Nik-Zainal,S. (2021) Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer*, **21**, 619–637.

11. Shan,K.-J., Wei,C., Wang,Y., Huan,Q. and Qian,W. (2021) Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. *The Innovation*, **2**, 100159.

12. Johnson,N. (2021) Faculty Opinions recommendation of Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*, 10.3410/f.738360033.793581756.

13. Mikhailova,A.G., Iliushchenko,D., Shamansky,V., Mikhailova,A.A., Ushakova,K., Tretyakov,E., Oreshkov,S., Knorre,D., Polishchuk,L., Lawless,D., *et al.* (2021) A mitochondrial mutational signature of temperature in ectothermic and endothermic vertebrates. *bioRxiv*,

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10.1101/2020.07.25.221184.

14. Naya,H., Romero,H., Zavala,A., Alvarez,B. and Musto,H. (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.*, **55**, 260–264.

15. Romero,H., Pereira,E., Naya,H. and Musto,H. (2009) Oxygen and guanine-cytosine profiles in marine environments. *J. Mol. Evol.*, **69**, 203–206.

16. Aslam,S., Lan,X.-R., Zhang,B.-W., Chen,Z.-L., Wang,L. and Niu,D.-K. (2019) Aerobic prokaryotes do not have higher GC contents than anaerobic prokaryotes, but obligate aerobic prokaryotes have. *BMC Evol. Biol.*, **19**, 35.

17. Sueoka,N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.*, **85**, 2653–2657.

18. Sueoka,N. (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. *J. Mol. Evol.*, **37**, 137–153.

19. Singer,G.A. and Hickey,D.A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.*, **17**, 1581–1588.

20. Peischl,S., Dupanloup,I., Kirkpatrick,M. and Excoffier,L. (2013) On the accumulation of deleterious mutations during range expansions. *Mol. Ecol.*, **22**, 5972–5982.

21. Bosshard,L., Dupanloup,I., Tenaillon,O., Bruggmann,R., Ackermann,M., Peischl,S. and Excoffier,L. (2017) Accumulation of Deleterious Mutations During Bacterial Range Expansions. *Genetics*, **207**, 669–684.

22. Shu,Y. and McCauley,J. (2017) GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.*, **22**.

23. Dunham,A.S. and Beltrao,P. (2021) Exploring amino acid functions in a deep mutational landscape. *Mol. Syst. Biol.*, **17**, e10305.

24. Löytynoja,A. (2014) Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*, **1079**, 155–170.

25. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

26. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

27. Dutheil,J. and Boussau,B. (2008) Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.*, **8**, 255.

28. Joy,J.B., Liang,R.H., McCloskey,R.M., Nguyen,T. and Poon,A.F.Y. (2016) Ancestral Reconstruction. *PLoS Comput. Biol.*, **12**, e1004763.

29. Abascal,F., Harvey,L.M.R., Mitchell,E., Lawson,A.R.J., Lensing,S.V., Ellis,P., Russell,A.J.C., Alcantara,R.E., Baez-Ortega,A., Wang,Y., *et al.* (2021) Somatic mutation landscapes at single-molecule resolution. *Nature*, **593**, 405–410.

30. Edgar,R.C., Taylor,J., Lin,V., Altman,T., Barbera,P., Meleshko,D., Lohr,D., Novakovsky,G.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Buchfink,B., Al-Shayeb,B., *et al.* (2022) Petabase-scale sequence alignment catalyses viral discovery. *Nature*, **602**, 142–147.

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

32. Ratcliff,J. and Simmonds,P. (2021) Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology*, **556**, 62–72.

33. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

34. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Campbell,P.J. and Stratton,M.R. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.

35. Popadin,K., Polishchuk,L.V., Mamirova,L., Knorre,D. and Gunbin,K. (2007) Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 13390–13395.

36. Sarkisyan,K.S., Bolotin,D.A., Meer,M.V., Usmanova,D.R., Mishin,A.S., Sharonov,G.V., Ivankov,D.N., Bozhanova,N.G., Baranov,M.S., Soylemez,O., *et al.* (2016) Local fitness landscape of the green fluorescent protein. *Nature*, **533**, 397–401.

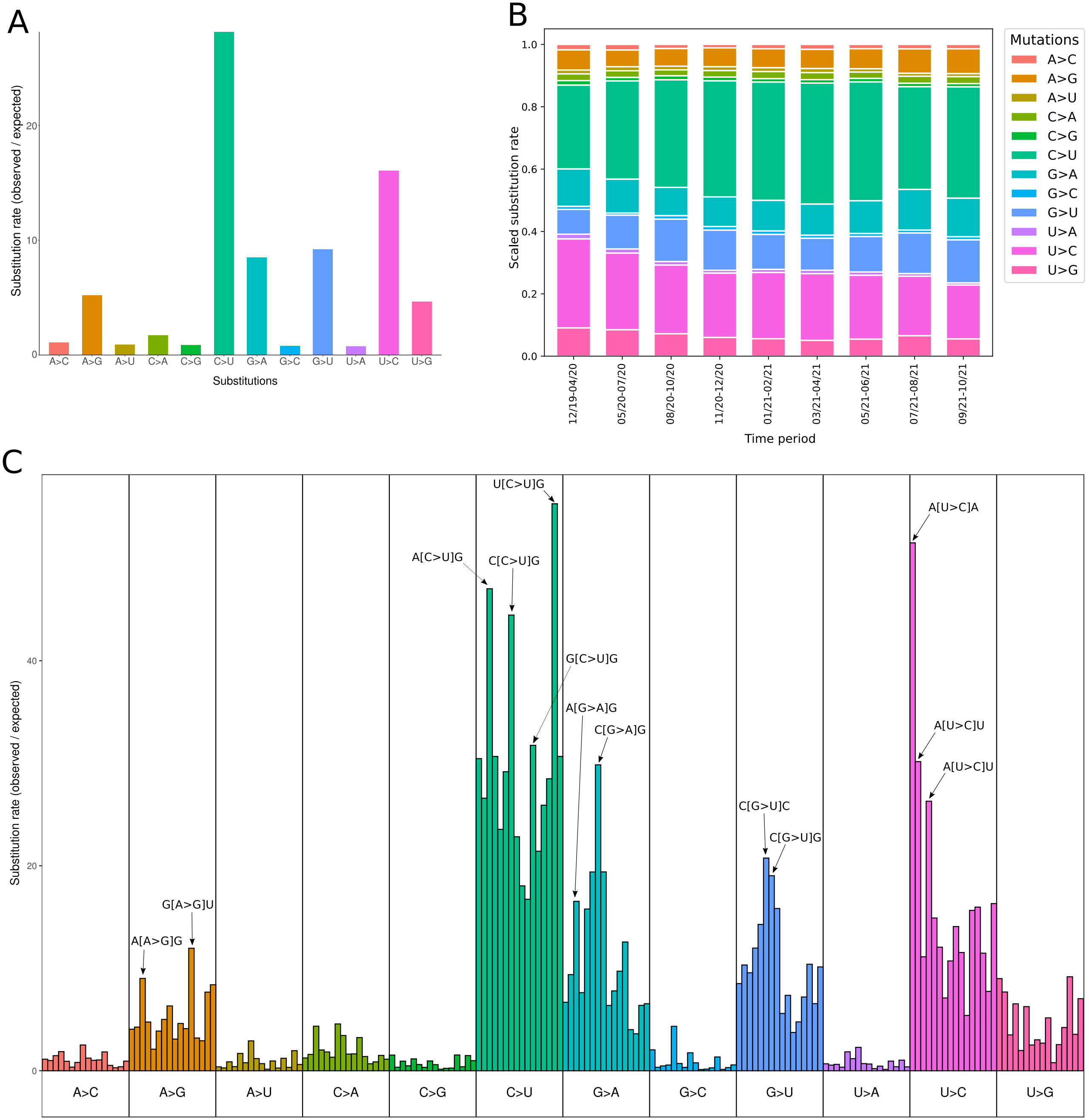

A
Substitution rate (observed / expected)
0
10
20
A>C
A>G
A>U
C>A
C>G
C>U
G>A
G>C
G>U
U>A
U>C
U>G
Substitutions
B
Scaled substitution rate
0.0
0.2
0.4
0.6
0.8
1.0
12/19-04/20
05/20-07/20
08/20-10/20
11/20-12/20
01/21-02/21
03/21-04/21
05/21-06/21
07/21-08/21
09/21-10/21
Time period
Mutations
A>C
A>G
A>U
C>A
C>G
C>U
G>A
G>C
G>U
U>A
U>C
U>G
C
Substitution rate (observed / expected)
0
20
40
A[A>G]G
G[A>G]U
A[C>U]G
C[C>U]G
U[C>U]G
G[C>U]G
A[G>A]G
C[G>A]G
C[G>U]C
C[G>U]G
A[U>C]A
A[U>C]U
A[U>C]U
A>C
A>G
A>U
C>A
C>G
C>U
G>A
G>C
G>U
U>A
U>C
U>G

Figure 1

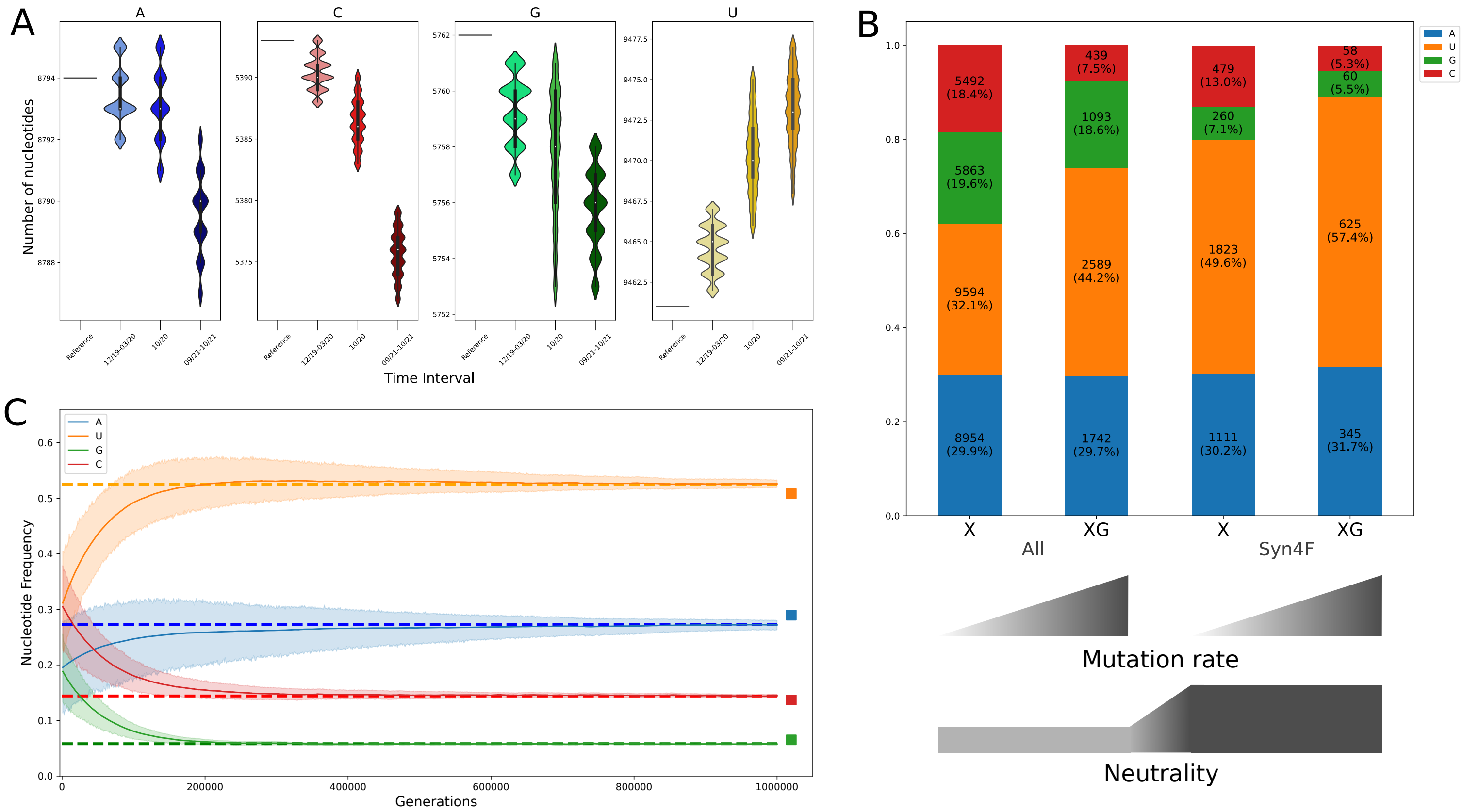

A
A
C
G
U
Number of nucleotides
8794
8792
8790
8788
5390
5385
5380
5375
5762
5760
5758
5756
5754
5752
9477.5
9475.0
9472.5
9470.0
9467.5
9465.0
9462.5
Reference
12/19-03/20
10/20
09/21-10/21
Time Interval
B
1.0
0.8
0.6
0.4
0.2
0.0
A
U
G
C
5492 (18.4%)
5863 (19.6%)
9594 (32.1%)
8954 (29.9%)
439 (7.5%)
1093 (18.6%)
2589 (44.2%)
1742 (29.7%)
479 (13.0%)
260 (7.1%)
1823 (49.6%)
1111 (30.2%)
58 (5.3%)
60 (5.5%)
625 (57.4%)
345 (31.7%)
X
XG
X
XG
All
Syn4F
Mutation rate
Neutrality
C
A
U
G
C
Nucleotide Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
200000
400000
600000
800000
1000000
Generations

Figure 2

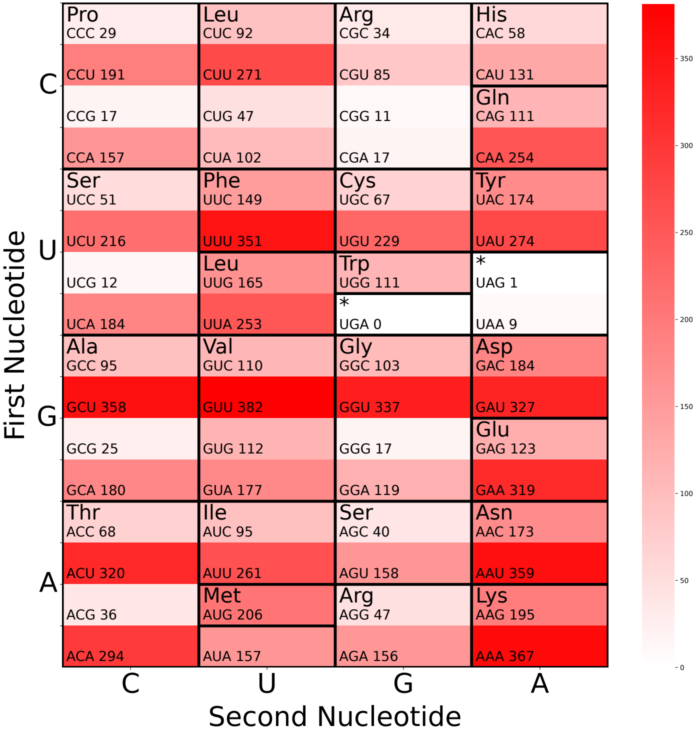
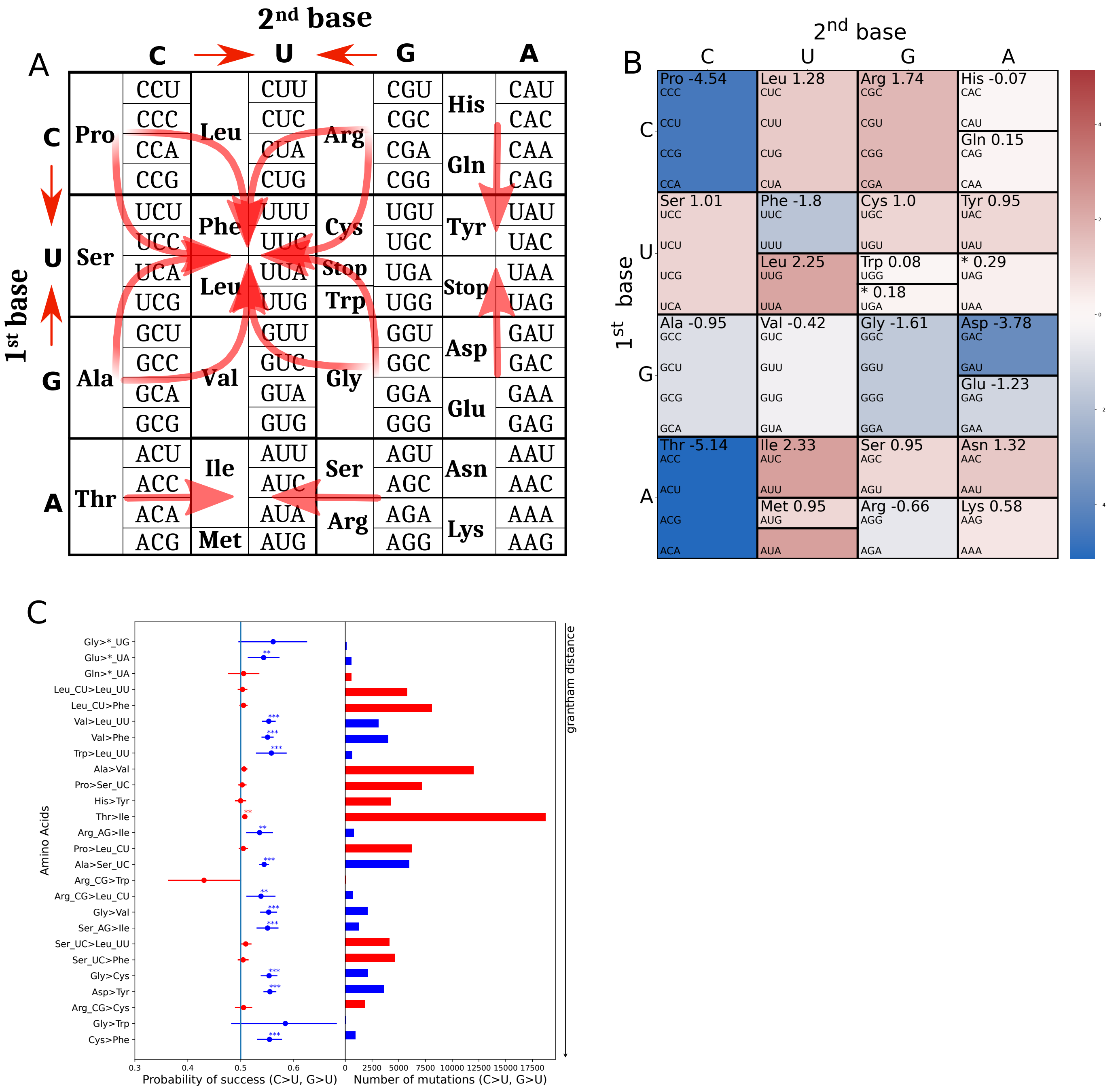

Pro
CCC 29
CCU 191
CCG 17
CCA 157
Leu
CUC 92
CUU 271
CUG 47
CUA 102
Arg
CGC 34
CGU 85
CGG 11
CGA 17
His
CAC 58
CAU 131
Gln
CAG 111
CAA 254
Ser
UCC 51
UCU 216
UCG 12
UCA 184
Phe
UUC 149
UUU 351
Leu
UUG 165
UUA 253
Cys
UGC 67
UGU 229
Trp
UGG 111
*
UGA 0
Tyr
UAC 174
UAU 274
*
UAG 1
UAA 9
Ala
GCC 95
GCU 358
GCG 25
GCA 180
Val
GUC 110
GUU 382
GUG 112
GUA 177
Gly
GGC 103
GGU 337
GGG 17
GGA 119
Asp
GAC 184
GAU 327
Glu
GAG 123
GAA 319
Thr
ACC 68
ACU 320
ACG 36
ACA 294
Ile
AUC 95
AUU 261
Met
AUG 206
AUA 157
Ser
AGC 40
AGU 158
Arg
AGG 47
AGA 156
Asn
AAC 173
AAU 359
Lys
AAG 195
AAA 367
C
U
G
A
First Nucleotide
C
U
G
A
Second Nucleotide
350
300
250
200
150
100
50
0

Figure 3



1st base	C	U	G	A
C	Pro CCU CCC CCA CCG	Leu CUU CUC CUA CUG	Arg CGU CGC CGA CGG	His CAU CAC Gln CAA CAG
U	Ser UCU UCC UCA UCG	Phe UUU UUC Leu UUA UUG	Cys UGU UGC Stop UGA Trp UGG	Tyr UAU UAC Stop UAA UAG
G	Ala GCU GCC GCA GCG	Val GUU GUC GUA GUG	Gly GGU GGC GGA GGG	Asp GAU GAC Glu GAA GAG
A	Thr ACU ACC ACA ACG	Ile AUU AUC AUA Met AUG	Ser AGU AGC Arg AGA AGG	Asn AAU AAC Lys AAA AAG

1st base	C	U	G	A
C	Pro -4.54 CCC CCU CCG CCA	Leu 1.28 CUC CUU CUG CUA	Arg 1.74 CGC CGU CGG CGA	His -0.07 CAC CAU Gln 0.15 CAG CAA
U	Ser 1.01 UCC UCU UCG UCA	Phe -1.8 UUC UUU Leu 2.25 UUG UUA	Cys 1.0 UGC UGU Trp 0.08 UGG * 0.18 UGA	Tyr 0.95 UAC UAU * 0.29 UAG UAA
G	Ala -0.95 GCC GCU GCG GCA	Val -0.42 GUC GUU GUG GUA	Gly -1.61 GGC GGU GGG GGA	Asp -3.78 GAC GAU Glu -1.23 GAG GAA
A	Thr -5.14 ACC ACU ACG ACA	Ile 2.33 AUC AUU Met 0.95 AUG AUA	Ser 0.95 AGC AGU Arg -0.66 AGG AGA	Asn 1.32 AAC AAU Lys 0.58 AAG AAA

Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504819; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

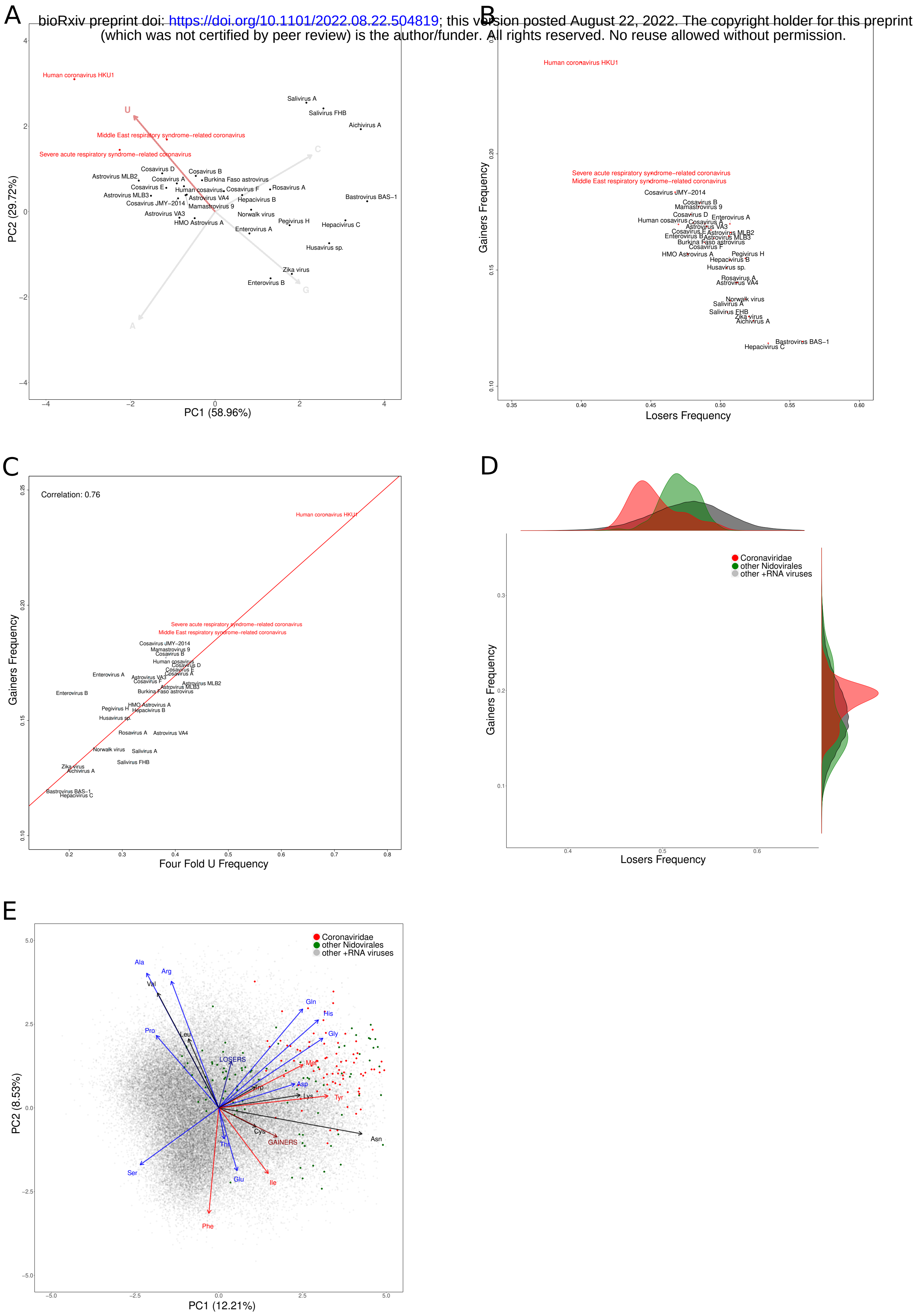


Figure 5