viruses

MDPI

Article

Variability in Codon Usage in Coronaviruses Is Mainly Driven by Mutational Bias and Selective Constraints on CpG Dinucleotide

Josquin Daron [1,*] and Ignacio G. Bravo [1,2]

[1] Laboratoire MIVEGEC (CNRS, IRD, Université de Montpellier), 34394 Montpellier, France; Ignacio.bravo@cnrs.fr
[2] Center for Research on the Ecology and Evolution of Diseases (CREES), 34394 Montpellier, France
* Correspondence: josquin.daron@ird.fr

Abstract: The *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) is the third human-emerged virus of the 21st century from the *Coronaviridae* family, causing the ongoing coronavirus disease 2019 (COVID-19) pandemic. Due to the high zoonotic potential of coronaviruses, it is critical to unravel their evolutionary history of host species breadth, host-switch potential, adaptation and emergence, to identify viruses posing a pandemic risk in humans. We present here a comprehensive analysis of the composition and codon usage bias of the 82 *Orthocoronavirinae* members, infecting 47 different avian and mammalian hosts. Our results clearly establish that synonymous codon usage varies widely among viruses, is only weakly dependent on their primary host, and is dominated by mutational bias towards AU-enrichment and by CpG avoidance. Indeed, variation in GC3 explains around 34%, while variation in CpG frequency explains around 14% of total variation in codon usage bias. Further insight on the mutational equilibrium within *Orthocoronavirinae* revealed that most coronavirus genomes are close to their neutral equilibrium, the exception being the three recently infecting human coronaviruses, which lie further away from the mutational equilibrium than their endemic human coronavirus counterparts. Finally, our results suggest that, while replicating in humans, SARS-CoV-2 is slowly becoming AU-richer, likely until attaining a new mutational equilibrium.

Keywords: virus evolution; SARS-CoV2; host switch; codon usage bias

check for updates

Citation: Daron, J.; Bravo, I.G. Variability in Codon Usage in Coronaviruses Is Mainly Driven by Mutational Bias and Selective Constraints on CpG Dinucleotide. *Viruses* **2021**, *13*, 1800. https://doi.org/10.3390/v13091800

Academic Editors: Rachele Cagliani and Manuela Sironi

Received: 26 July 2021
Accepted: 31 August 2021
Published: 10 September 2021

1. Introduction

The *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) is the cause of respiratory disease COVID-19, occasionally leading to acute respiratory distress syndrome and eventually death [1]. With no antiviral drugs nor vaccines initially available, and with the presence of asymptomatic carriers, the COVID-19 outbreak turned into a public health emergency of international concern. Before the 2019 zoonotic spillover, viruses closely related to SARS-CoV-2 had circulated probably for decades in bats as well as in other intermediate hosts, such as the Sunda pangolin *Manis javanica* [2]. Cross-species transmissions are common among coronaviruses (CoVs) [3,4], and they are an important mechanism driving the evolution of bat-CoVs in nature [5,6]. In humans, all CoVs are likely linked to zoonotic events, mostly from bats or rodents, with occasionally domestic animals playing the role of intermediate hosts [7]. Consequently, given the high zoonotic potential of the *Orthocoronavirinae* [8,9] and the vast repertoire of mammalian and avian hosts they infect, there is an urgent need to evaluate the potential zoonotic risk for each individual CoV. This knowledge will guide virus discovery, surveillance and research to identify for each virus the differential risk of efficiently infecting humans and cause a new pandemic.

None of the viruses infecting vertebrates encodes for any tRNA nor for any ribosomal protein, and CoVs are no exception. As a result, the translation of viral proteins relies exclusively on the host tRNA repertoire and translational machinery [10]. In order to efficiently support the production of viral proteins, it has been proposed that viruses would evolve to use the set of synonymous codons found overrepresented in their hosts, as the result of an adaptation to their host cellular environment [11–14]. This hypothesis is based on the fascinating discovery that codon usage is subject to natural selection [15–19]. This adaptive hypothesis, called translational selection, proposes that the non-random usage of synonymous codons and the abundance of tRNAs have co-evolved to optimize translation efficiency. Experimental analyses have been successful at characterising how syno-nymous substitutions influence the cellular fitness of an organism, by acting on a broad range of cellular processes, including changes in transcription [20], translation initiation [21,22], translation elongation [23], translation accuracy [17,24], RNA stability [25], and splicing [26].

Numerous studies conducted essentially in certain phages and their bacterial hosts have reported a strong match between the codon usage bias (CUB) of viral genes with respect with their hosts [27–29]. Similarly, the CUB of human papillomaviruses has been associated with differential clinical presentations of the infections [30], with a stronger virus–host match in CUB for human papillomaviruses causing productive lesions than for those causing asymptomatic infections. However, such a trend has not been observed for SARS-CoV-2: most of the highly frequent codons in SARS-CoV-2 are A- and U-ending codons [31,32] overall showing a poor match with the average human CUB [33]. This mismatch has raised important questions about the nature of the CUB in CoVs, the underlying mechanisms involved, and the impact on virus longer-term evolution. Furthermore, investigating the CUB match between viral and cellular genes constitutes a challenge to supporting the use of synonymous mutations as biotechnological tools to develop live attenuated vaccines [34].

In addition to being driven by constraints on tRNA abundance, the non-random usage of synonymous codons in viruses is also shaped by dinucleotide abundance. It has been long established that many RNA viruses infecting mammals have evolved CpG and UpA deficiency [35–38]. In the mammalian RNA Echovirus 7, it has been formally demonstrated that the severe attenuation of viral replication can be attributed to the increase in CpG and UpA dinucleotides frequencies rather than to the use of disfavored codon pairs [39]. Viruses with high CpG or UpA frequencies may be more likely to be recognized by pathogen innate immune sensors, preventing them from replication initiation [40]. Functionally, the inhibition of viral replication and the degradation of the viral genome have been attributed to the zinc finger antiviral protein (ZAP), a powerful restriction factor that specifically binds to CpG motifs [39]. It is hence critical when investigating CUB to disentangle confounding effects of CUB and dinucleotide abundance.

Finally, non-adaptive evolutionary forces can also affect the viral genome composition. Because fitness differences associated with individual codons are very small, it requires very large population sizes (i.e., in the case of viruses highly productive and highly prevalent viral infections) for natural selection to act and lead to a significant impact on the global genomic CUB. This trend is verified in large organisms with small population sizes, such as most mammals, where natural selection on CUB is weak [41,42] and CUB is instead primarily shaped by mutational biases [43] and GC-biased gene conversion [44]. Mutations are the fundamental substrate for genotypic diversity, leading to phenotypic diversity upon which natural selection can act. Point mutations are stochastic processes but they concur with certain deterministic and directional biases, in bacteria [45] as well as in eukaryotes [46–49]. Previous studies suggest that mutations occurring during genome replication are universally biased towards AT. Similarly, for SARS-CoV-2, analysis of mutational profiles indicates a strong mutation bias towards U [50,51]; however, the impact of this bias on CUB variation has not been characterized yet at the scale of the whole *Orthocoronavirinae*.

Given the impact of the match between virus and host CUB in viral gene expression, the key challenge is to determine the impact of a specific genome composition or CUB in the initial zoonotic spillover from animals towards humans that may eventually modulate the risk of stable human-to-human transmission. Here, we investigated the CUB variability in *Orthocoronavirinae*, with an emphasis on CoVs infecting humans. We aimed at determining whether CUB in CoVs is actively selected according to the codon preferences of their hosts or whether it reflects instead the action of other evolutionary forces. We show that variation in CUB in *Orthocoronavirinae* mainly results from differences in GC3-content and in CpG and UpA abundance, independently of the host infected. Variation in GC3-content is primarily dictated by mutation biased towards AU, a trend universally observed in all *Coronavirinae* genera, regardless of the host. Finally, selection against CpG and UpA dinucleotides strongly impacts the CUB of CoVs. Altogether, we conclude that variation in CUB plays a minor role, if any, on the probability of the establishment of a zoonotic spillover towards humans.

2. Results

2.1. Variation in Codon Usage Bias among Orthocoronavirinae Is Not Dependent on the Host

To better understand the causes of CUB differences between CoVs, we investigated a total of 82 complete CoVs genomes. Our sample spans viral and host taxonomical diversity, covering the four viral genera within *Orthocoronavirinae*, and embracing a total of 47 different vertebrate host species within nine mammalian and five avian orders (Supplementary Table S1). First, we explored the variation in the 59 synonymous codons frequencies through a Principal Component Analysis (PCA) (Figure 1A, Figure S1A). The PCA efficiently reduced information dimensionality as the first two components captured, respectively, 34.9% and 20.1% of the variance, and the first four dimensions contributed with above 70% of explanatory power. Interestingly, the 82 CoVs were distributed scattered along the two first PCA axes without any obvious stratification as a function neither of the viral taxonomy nor the host infected. This result contradicts our initial hypothesis of a host-specific CUB signature in CoVs and suggests that translational selection for convergence towards the host's CUB is presumably weak.

The PCA results contribute further with information about the spreading of the individual codons as well as their contribution to the total variance. The first PCA axis sharply splits the codons depending on their nucleotide in the third codon position (often referred to as GC3), with the exception of UUG-Leu codon, which stands alone among all other A- or U-ending codons (Supplementary Figure S1B). Strikingly, we found that variation in genomic GC3 content strongly correlated with variation on the first PCA axis (adjusted $R^2 = 0.94$, *p* value $< 1 \times 10^{-50}$, Figure 1B). Note that variation in GC3 did not show any correlation with any other of the main PCA axes (Supplementary Figure S2). In-depth analysis of the frequency patterns for the 18 synonymous codons families showed that A- or U-ending codons are systematically preferred over the G- or C-ending ones (Figure 1C). This trend was especially true for amino acids with multiplicity two (i.e., encoded by two codons), but also confirmed for amino acids with multiplicity four: U-ending codons were systematically the most used among the four codons, while A-ending codons were preferred over the G-ending ones. For amino acids with multiplicity six, the overall scheme corresponds to the combined patterns of a family of multiplicity four and a family of multiplicity two. Altogether, our observations show that variation in GC3 is the main explanatory factor for CUB variation between CoVs, which strongly suggests a universal AT-ending over GC-ending synonymous codons enrichment in CoVs genomes.

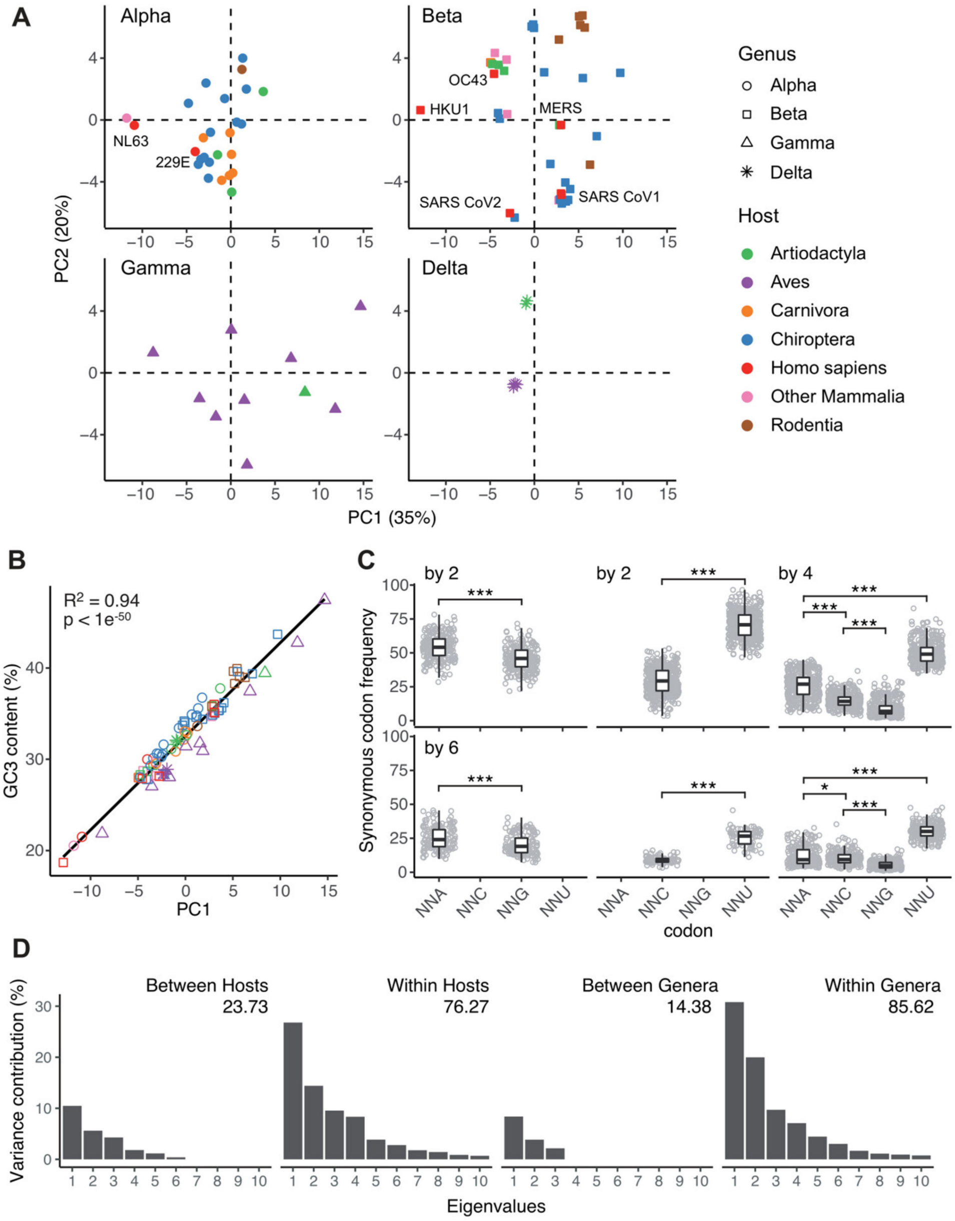

Figure 1. Variation in synonymous codon usage in *Coronavirinae*. (**A**) Principal component analysis of synonymous codon usage among CoVs. Each dot corresponds to an individual CoV, for which the frequencies of the 59 synonymous codons have been calculated. CoVs are stratified after the corresponding genera (Alpha-Delta), and host (color code in the inset). The percentage of the variance explained by the first and second axes is given in parenthesis. (**B**) Correlation between the projection on the first of principal components shown in (**A**) and the average GC-content at third codon position (GC3) for the corresponding viral genome. Symbols and color code are the same as in (**A**). The results for a Pearson's correlation test are shown in the inset. (**C**) Synonymous codon usage frequency stratified by amino acids of multiplicity 2, 4, or 6 synonymous codons. Differences in median codon usage frequency values were assessed by a paired Wilcoxon signed rank test (code for *p*-value: * <0.05; ** <0.01; *** <0.001). (**D**) Results of an internal correspondence analysis for the contribution of the levels "viral taxonomy" (four levels, corresponding to viral genera) and "host taxonomy" (fourteen levels, corresponding to host orders) to explain variability among the eigenvalues of viral genomes on the first 10 principal components in the PCA shown in (**A**). The values for expected explanatory power for a homogeneous distribution were and 3.72% (p-value $< 9 \times 10^{-4}$) for viral taxonomy and 7.59% (p-value $< 9 \times 10^{-4}$) for host taxonomy.

We aimed at further quantifying the proportion of the global variability in CUB that is explained by the host and by the virus taxonomic stratifications. Through a correspondence analysis, we associated a subset of 75 different viruses of our codon usage table into blocks corresponding to the seven host taxonomic orders (only host taxonomic orders with at least five viruses were considered; Supplementary Table S1), allowing us to split the total variability into between-category and within-category variability. The top 10 eigenvalues of this decomposition are represented in Figure 1D showing that within-host differences in CUB explain four times more of the overall variability than differences between-hosts (respectively 76.3% vs. 23.7%); the explanatory power of both levels is larger than the randomly expected one from a homogeneous distribution (8.1%, p value < 9×10^{-4}). A similar analysis was reproduced associating the viruses into blocks corresponding to the four viral genera within *Coronavirinae*. We observed that within-genus differences in CUB explain over five times more of the overall variability than between-genera differences (respectively 85.6% vs. 14.4%); again, the explanatory power of both levels is larger than the randomly expected one from a homogeneous distribution (3.7%, p value < 9×10^{-4}). Together, our results are consistent with our initial PCA analysis, and suggest that both host and virus taxonomy variables are not the main factors that drive variation in CUB within *Orthocoronavirinae*.

2.2. The Mutational Spectrum in Orthocoronavirinae Is AU-Biased

Since variation in GC3 content is the main individual driver of the variation in CUB between CoVs, we investigated next whether mutational biases are the underlying main force driving nucleotide content. Population genetics studies show that in order to explore mutational biases independently of the effect of natural selection, one should work at shallow, short-term evolutionary periods, where natural selection is less powerful to impact nucleotide polymorphism patterns [45]. Consequently, we estimated the frequencies of individual transition and transversions by narrowing down our analysis to a subset of 16 different CoV metapopulations. From the public databases (see Methods), we downloaded a total of 12,102 different viral genome entries to construct these 16 different metapopulations (Supplementary Table S2). Prior to the identification of SNPs for each metapopulation, we carefully preprocessed each dataset by removing putative recombinant sequences (see Methods, Supplementary Figure S3A), as recombinant sequences violate the assumptions for phylogenetic inference methods used downstream. For each viral metapopulation, we removed further the effect of population structure by phylogeny-driven selection of one single homogeneous population per viral species (Supplementary Figure S3B). Our final dataset was composed of a total of 10,373 viral genomes for the 16 metapopulations (median number of genomes per metapopulation 31; range 9–9588), for which a total of 57,059 SNPs were called (median number of SNPs called per metapopulation 3037; range 253–7778; Supplementary Table S2). Despite the heterogeneous sampling size across our metapopulations, the number of SNPs called was highly comparable, the exceptions being SARS-CoV-1 sequences from humans and MERS sequences from humans and from camels, which exhibited a low number of sequence polymorphisms, albeit large enough to grant sound analyses (respectively 716, 621 and 253). There was no obvious correlation between the number of genomes in each metapopulation and the number of sequence polymorphisms retrieved (R^2 = 0.183, $F_{(1,14)}$ = 3.14, p = 0.098), even less when removing the SARS-CoV-2 metapopulation (R^2 = 0.054, $F_{(1,13)}$ = 0.74, p = 0.405), which was an outlier in number of genomes considered (n = 9588).

Under a maximum likelihood framework, we separately estimated for the synonymous and non-synonymous compartments the frequencies of transitions and transversions using the generalized time-reversible model for the phylogenetic reconstruction fitting the genomic data best. This analysis allows us to estimate the frequencies of individual transitions and transversions after correcting for nucleotide composition in each compartment. Our results show that for the synonymous substitutions compartment and for all viral metapopulations studied, U<>C transitions were more frequent than A<>G transitions, and for most metapopulations C->U transitions were more frequent than the reverse U->C (Figure 2A). These remarkable differences were of a larger magnitude for substitutions occurring within the synonymous compartment than within non-synonymous compartment, where the C->U changes were only slightly more common than the reverse substitution. Consistent with these findings, stratification of substitutions into GC-enriching (i.e., AU->GC) or AU-enriching (i.e., GC->AU) categories shows that for all metapopulations AU-enriching substitutions occur at much higher rate than GC-enriching ones in both synonymous and nonsynonymous compartments (respectively mean fold change = 2.49, p-value < 7.6×10^{-6}; and mean fold change = 1.38, p-value < 7.6×10^{-6}; paired Wilcoxon signed rank test). Individual departures from this trend are observed and may deserve future, focused research, the most obvious case being the avian gamma-coronavirus *Infectious bronchitis virus*, for which G->A and U->C transitions predominate. Altogether our results suggest that the mutational spectrum in CoVs is biased towards AU-enrichment and that the nucleotide content of the viral genomes is primarily determined by mutational biases.

2.3. Recent Human Coronaviruses Display Greater Mutational Disequilibrium Than Endemic Human Coronaviruses

Humans are the host to endemic CoVs, responsible for common respiratory diseases, but it has been proposed that these endemic hCoVs have an ancient zoonotic origin. Molecular dating studies have tried to evaluate the timing for host shift events for endemic hCoVs, estimating an old emergence of several hundreds of years ago for NL63 and a more recent one in the 19th–20th century for 229E, OC42 and HKU1 [52–55]. This epidemiological stratification among hCoVs, differentiating recently zoonotic hCoVs and ancient zoonosis that have become endemic in humans, provided us with the unique opportunity to characterize the impact that spillover and subsequent establishment in the new host has on the mutational equilibrium of the virus genomes. Previous studies have demonstrated that a high-resolution estimation of the mutational equilibrium can be gained from the simple property of the folded site frequency spectrum (SFS) of AU-to-GC mutations, in which SNPs are not polarized [56]. The theory proposes that for populations at demographic equilibrium the GC-content converges towards the mutational equilibrium, yielding a perfect U-shape in the AU-GC allele frequency distribution, resulting in a symmetrically folded SFS. Deviations from the U-shape distribution would thus reflect that the genomic GC-content is not at its mutational equilibrium, a negative (or positive) skewness indicating higher (or lower) GC-content than expected under the mutational equilibrium. We applied this framework, computed the folded SFS for the seven hCoVs species (Figure 2A) and quantified the deviation from the expected U-shape distribution by calculating the skewness of each folded SFS as a proxy for the departure of the GC-content from the expected mutational equilibrium (Figure 2B). Our results show that for all four endemic hCoVs, the observed skewness was not different from the null expectation, while for the three recently emerged hCoVs (SARS-CoV-1, MERS-CoV and SARS-CoV-2) we observed a significant departure from the expected GC content at equilibrium. Both MERS-CoV and SARS-CoV-2 displayed a negative skewness, meaning that their genomes were slightly more GC-rich than the anticipated mutational equilibrium, while SARS-CoV-1 displayed a positive skewness, reflecting an AU-richer genomic content than expected for the mutational equilibrium. Because the SFS represents population summary statistics, potentially sensitive to population size changes (growth or contraction), we aimed to determine whether the exponential growth of SARS-CoV2 was modulating our estimation of the SFS-skewness. Interestingly, we did not observe any impact of the exponential growth of SARS-CoV2 on the estimation of the skewness (Supplementary Figure S5). While the exponential growth of SARS-CoV2 changes the shape of the SFS by increasing the proportion of rare variants, this should symmetrically affect AU to GC and GC to AU mutations. Consequently, as theoretically demonstrated by Glemin and coworkers [56], the result of an SFS skewness value significantly different from the null expectation is robust to demographical changes.

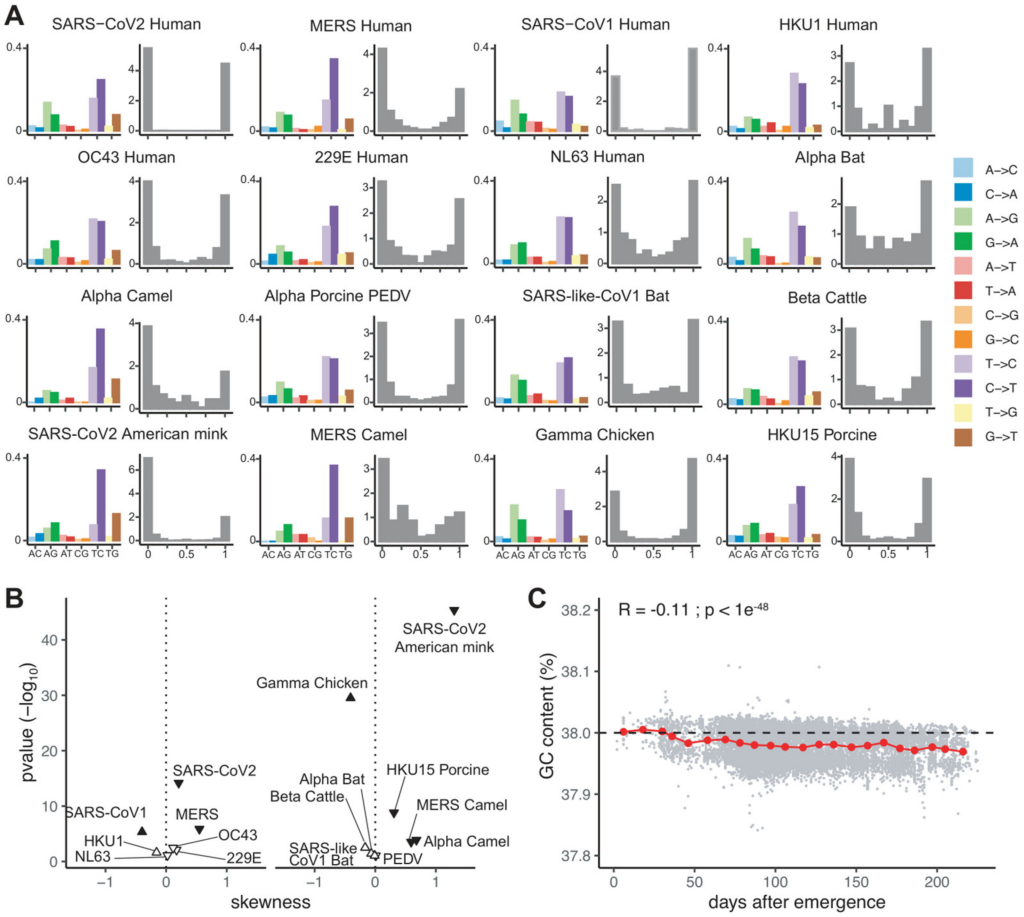

Figure 2. Mutational bias and equilibrium among endemic and recently zoonotic coronaviruses. (**A**) (left side of each panel) Relative substitution rates for all twelve transition and transversion events, calculated from the total number of SNPs called on the synonymous compartment of a metapopulation composed by all available sequences of the corresponding virus, curated, filtered and aligned. In the x axis, for each of the nucleotide pairs indicated, the frequency of the two substitutions are indicated, e.g.,: A->G in light green and G->A in dark green for the AG pair, or T->C in light purple and C->T in dark purple for the TC pair. (right side of each panel) Folded Site Frequency Spectrum (SFS) of G/C-to-A/T substitutions. For each viral metapopulation, the frequencies of SNPs involving a change in GC content called are plotted, the *x*-axis progressing from GC-enriching to AT-enriching substitutions. (**B**) Volcano plot of the folded SFS skewness values presented in (**A**), plotted against the probability that such a value could have been obtained if the corresponding metapopulation were at its mutational equilibrium. The left panel displays CoVs infecting humans while the right panel displays CoVs infecting other mammalian hosts. Downward and upward triangles, respectively, refer to SFS-negative and SFS-positive skewness values. Filled triangles indicate metapopulations with a skewness significantly different from the expected one at mutational equilibrium. (**C**) Temporal evolution of the SARS-CoV-2 GC-content in the coding genome during the spread of the pandemic. The GC-content of each individual SARS-CoV-2 genome is represented by a gray dot. Red dots represent the average GC-content by time window of 10 days. The results of a Pearson correlation test of GC content over time for the complete data set (n = 81,963) are displayed in the inset.

Finally, benefiting from the wealth of sequence data generated on SARS-CoV-2, we investigated the dynamics of the genomic GC-content over the spread of the COVID-19 pandemic (Figure 2C). We observed that the GC-content of SARS-CoV-2 is very slowly albeit significantly decreasing with the progression of the pandemic (Figure 2C, R = −0.11, *p*-value < 1×10^{-49}). This result of a trend towards an overall AU-enrichment is consistent with the SFS observation of a viral genome far from compositional equilibrium and suggests that during the expansion of SARS-CoV-2 in humans, the viral population is experiencing a GC->AU mutational bias resulting in a slow decrease in the GC genomic content compared to the viral genomes retrieved from humans at the origin of the pandemic.

2.4. CpG Dinucleotides Are Selected against in Orthocoronavirinae Genomes

Our initial characterization demonstrated that variation in GC3 content is the prevailing force shaping CUB among CoVs, explaining one-third of the total variation in CUB. We thus tried to identify other evolutionary forces further shaping CUB in CoVs. Previous works have identified that in several RNA viruses infecting humans certain dinucleotides are under-represented, notably CpG and UpA, and that this low dinucleotide frequency has a strong impact on codon-pair bias [57]. We investigated thus the ratio of the observed over the expected dinucleotide frequency in the CoVs coding sequences. Figure 3A shows that the observed abundance of the CpG dinucleotide in all CoVs lineages is significantly lower than the null expectation based on the individual nucleotide frequency, while this is not the case for the UpA dinucleotide. To a lesser extent, we identify CpA and UpG to be marginally more frequent than expected, which can be linked to the strong decrease in CpG, as they correspond, respectively, to the transitions CpG->CpA and CpG->UpG. We further state that the observed/expected ratios for CpG and UpA correlated well with the second and third dimension of our PCA analysis, respectively (R^2 adjusted = 0.67 and 0.44; Figure 3B,C). It is important to remember that these dinucleotide frequency values have been estimated for the complete viral coding sequence and are not limited to the codons themselves, i.e., we also considered the presence of CpG and UpA dinucleotides in the codon boundary context, so that this impact is not simply related to higher frequency of CG-rich or AT-rich codons. This specific point will be addressed below. Overall, our results show that variation in CUB between CoVs is associated with variation in CpG and UpA dinucleotide content.

In order to formally demonstrate the impact that CpG and UpA dinucleotide frequencies have on CUB, we compared the synonymous codon frequencies among codon families either containing or lacking a CpG- or UpA-ending codon. First, only the Arginine amino acid allows for synonymous mutations between the first and second codon position that lead to CpG changes, as it is encoded by CGN and by AGY. A CGY->AGY transition thus allows for a synonymous substitution that results in the loss of a CpG motif. Such a depletion is present in the human genome, in which ca. 42.6% of the Arginine residues are encoded by AGY, but is cogent in the SARS-CoV-2 genome, where this value amounts to ca. 57.8%, the third highest value among the analysed CoVs genomes (Supplementary Table S3). This value is particularly high when compared to other human-infecting CoVs, as it reaches 49.7% for the zoonotic SARS-CoV-1 genome, and amounts only to between 47.7% and 37.7% for the four endemic human CoVs (Supplementary Table S3).

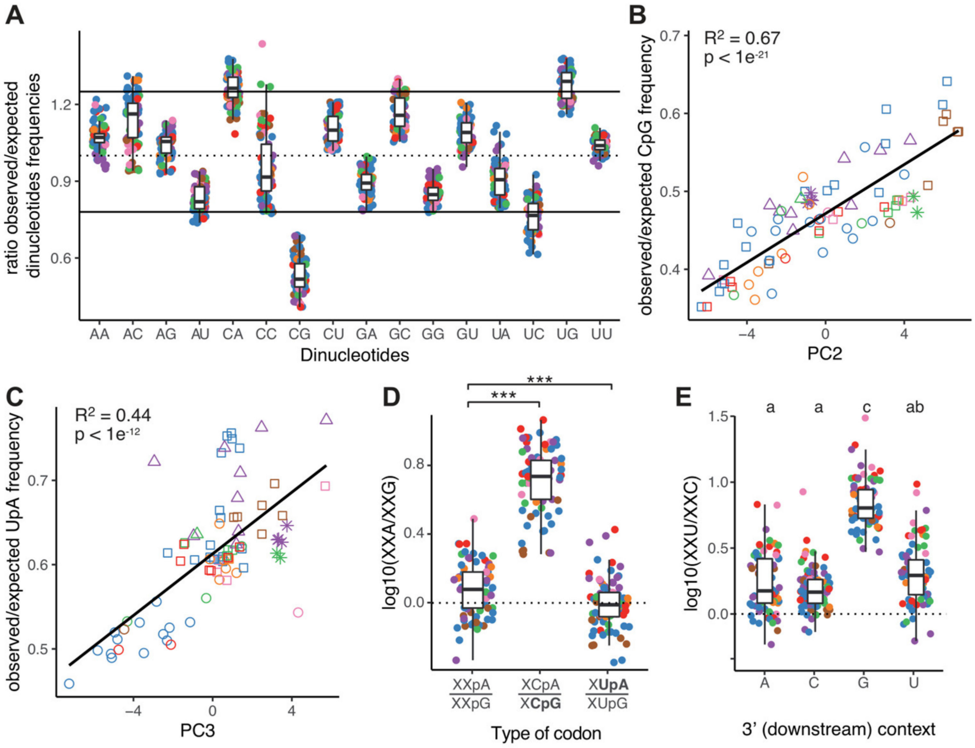

Figure 3. Dinucleotide Bias in *Coronavirinae.* (**A**) Box (median, 1st and 3rd quartiles) and whiskers (95% CI) plot for the ratio of the observed over the expected frequency (dinucleotide relative abundance) for all 16 dinucleotides in the coding sequences of 82 CoVs. Continuous lines indicate the thresholds for considering over/under representation of the corresponding dinucleotide over the expected values given the frequencies of the individual nucleotides. The threshold values at 0.78 and 1.25 have been phenomenologically determined following [57]. Colors correspond to the different host taxonomy orders. (**B**) Linear regression between the projections of each individual viral genome on the second dimension of the synonymous codon usage Principal Component Analysis (Figure 1A) and the observed/expected frequency of the CpG dinucleotide. Shapes correspond to viral taxonomy genera and colors correspond to host taxonomy orders. (**C**) Linear regression between the projections of each individual viral genome on the third dimension of the synonymous codon usage Principal Component Analysis and the observed/expected frequency of the UpA dinucleotide. Shapes correspond to viral taxonomy genera and colors correspond to host taxonomy orders. (**D**) Box (median, 1st and 3rd quartiles) and whiskers (95% CI) plot for the ratio in synonymous codon frequency of A- over G-ending codons, calculated for codon families with multiplicity two lacking CpG and UpA dinucleotides (Gln, Lys and Glu, indicated as “XXpA/XXpG”, constituting the reference set for comparison), amino acids encoded by CpG-ending codons (Ser4, Pro, Thr, Ala, indicated as “XCpA/XCpG”), and amino acids encoded by UpA-ending codons (Leu2, Leu4, Val, indicated as “XUpA/XUpG”). Colors correspond to the different host taxonomy orders. Differences within amino categories were assessed by a paired Wilcoxon signed rank test (Bonferroni correction, code for *p*-value: *** <0.001). (**E**) Box (median, 1st and 3rd quartiles) and whiskers (95% CI) plot for the ratio in synonymous codon frequency in U- over C-ending codons, calculated for codon families with multiplicity four, and stratified by the nature of the 3′ downstream nucleotide. Differences depending on 3′ base context were assessed by a paired Wilcoxon signed rank test (Bonferroni correction) and summarized among sets of groups statistically different one from another. Colors correspond to the different host taxonomy orders.

Second, we focused on the synonymous changes between second and third codon position. Having previously established that mutational biases modulate frequencies among synonymous codons, we accounted for this confounding effect by calculating an expected ratio in synonymous codon frequencies for the codon pairs in the form of

XXA/XXG. To this end, we defined as our reference set for comparison the three codon families with multiplicity two lacking CpG and UpA dinucleotides and ending by either A or G (i.e., Gln-CAA/G-, Lys-AAA/G- and Glu-GAA/G-, indicated as "XXpA/XXpG" in Figure 3D). For this reference set, we calculated an overall fold change XXA/XXG ratio of 1.26, consistent with the AT-enriching mutational bias described above. For amino acids encoded by CpG-ending codons (Ser4, Pro, Thr, Ala, indicated as "XCpA/XCpG" in Figure 3D), we observed a 5.33-fold change XCA/XCG ratio, significantly higher than the corresponding one for the reference amino acids set (within-genome paired Wilcoxon signed rank test, p-value $< 1 \times 10^{-14}$). This difference indicates that regardless of the amino acid encoded, CpA-ending codons are systematically preferred over their CpG-ending synonymous counterparts, at higher proportion than expected under the pressure of A->G mutational bias alone. In the case of UpA-ending codons (Leu2, Leu4, Val), we observed an overall XXA/XXG ratio of 1.04, slightly but statistically significantly lower than the corresponding one for the reference amino acid set (within-genome paired Wilcoxon signed rank test, p-value $< 1 \times 10^{-6}$). Thus, we interpret that although the UpA-dinucleotide is not significantly depleted in CoV genomes, UpA-ending codons are less frequent than their synonymous UpG counterparts. Consequently, the genomic UpA dinucleotide-content is shaped by two antagonistic evolutionary forces: on the one hand a mutational bias promoting the overall excess of G->A transitions and on the other hand a selection for XUpA->XUpG transitions in synonymous codons. Indeed, this interpretation also supports the finding that the UpG dinucleotide is borderline significantly more frequent than expected, as it also corresponds to the transition UpA->UpG.

Finally, we aimed at characterizing the impact that the pressure against CpG and UpA dinucleotides imposes on CUB, when these dinucleotides are located at the codon pair boundary, such as XXC-GXX or XXU-AXX. We thus computed the frequency ratio XXU/XXC for synonymous codons depending on the nature on the downstream first nucleotide: if a selection again CpG nucleotides existed, we would expect the XXU/XXC ratio to be higher when the downstream codon starts with G; similarly, if a selection against UpA dinucleotides existed, the XXU/XXC ratio should be lower when the downstream codon starts with A. Our results indeed confirmed the depletion for CpG, as we observe a significantly higher XXU/XXC ratio upstream a codon starting by a G compared to codon starting with any other base (Figure 3E), avoiding the creation of CpG at the overlap between two codons. Regarding the hypothesis of UpA depletion, the median XXU/XXC ratio upstream a codon starting by A was the lowest of all bases, albeit with a large variance and not different from the values for codons starting by C or U. This result mitigates our previous finding of a slight intra-codon avoidance of the UpA dinucleotide, but is concordant with the overall absence of significant avoidance of UpA within the coding sequence.

2.5. Mutational Bias and CpG/UpA Depletion Explain Most of the Variation in Synonymous Codon Usage of Coronaviruses

Having identified GC->AT mutational bias as well as CpG and possibly UpA dinucleotide depletion as variables impacting CUB in *Coronavirinae*, we aimed at further quantifying the differential contributions of each of these factors to the overall CUB variation. We therefore built a linear model quantifying the relative contribution of the different variables (GC3, CpG, UpA, host taxonomy at the level of order, and virus taxonomy at the level of genus) and their corresponding interactions with the covariance of the projections of the first four main dimensions of the PCA for CUB. The analysis of variance demonstrated that variation in GC3, CpG and UpA are, respectively, the best predictors of the first three dimensions of the PCA (Figure 4), respectively explaining 94%, 67% and 56% of the variation in each dimension. By contrast, the integration of host and virus taxonomy stratification levels and their interactions with the compositional variables did not contribute with any further improvement of the model fit to the data. Lastly, the fourth PC dimension was only marginally explained by UpA content and by the *UpA*viral taxonomy* interaction, which explained 19.3% and 18.2% of the PC4 dimension, respectively.

Given these results, and considering the power of each PCA axis for explaining variation in CUB, we conclude that variation in GC3 explains around 34%, variation in CpG frequency explains around 14%, and variation in UpA frequency explains around 7% of the total variation in CUB in *Coronavirinae*. On the contrary, the contribution of viral taxonomy and of host taxonomy diversity to explain variation in CUB is negligible.

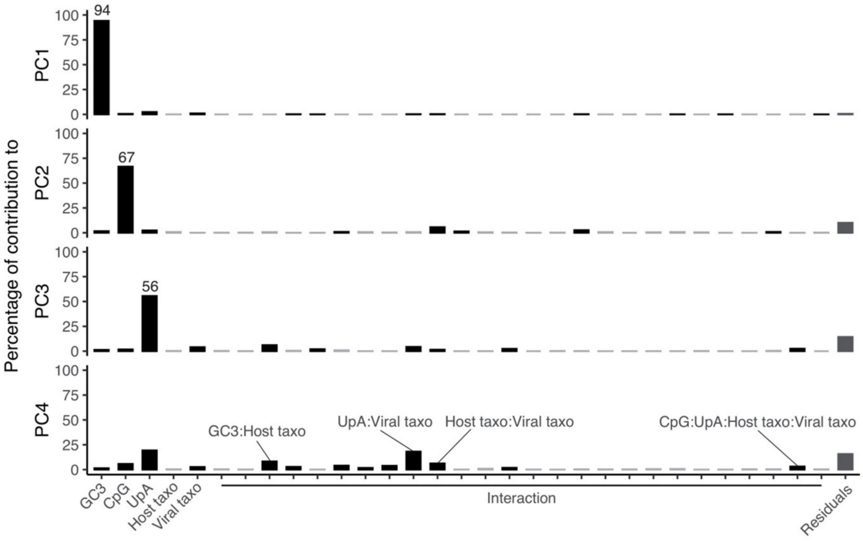

Figure 4. Linear model quantifying the relative contribution of the different variables to the projections of the first four main dimensions of the synonymous codon usage PCA (Figure 1A). Variables included in the linear model are: GC3-content, observed/expected frequency of the CpG dinucleotide, observed/expected frequency of the UpA dinucleotide, host taxonomy at the level of family, viral taxonomy at the level of genus, and all their pairwise interactions. Statistically significant contributions are shown as black bars, stating the percentage of the total explanatory contribution to the correspondent PCA dimension.

3. Discussion

With the COVID-19 pandemic, the unprecedented wealth of sequence data generated on human SARS-CoV-2 and its close relatives infecting pangolin and bats provides a unique opportunity to investigate the variability and the biological role of synonymous codon usage among *Orthocoronavirinae*. Previous works in the field have been successful at inventorying CUB in SARS-CoV-2 [31,32,58], as well as at reporting the poor match in the observed CUB between SARS-CoV-2 and humans [59]. Together, those studies raised important questions about the nature of the CUB in CoVs, the underlying evolutionary mechanism involved, and their impact on the longer-term evolution of the virus. Here, we explore those questions, by meticulously disentangling the effects that natural selection, dinucleotide avoidance and neutral forces have on the variability in CUB in CoVs.

3.1. Mutational Bias and CpG Avoidance Shape Codon Usage Bias in CoVs

We report a strong heterogeneity in CUB among CoVs, mainly driven by GC->AU biased mutations. In analyses restricted to the SARS-CoV-2 genomes, previous studies had identified C->U deamination as the main mutational contributor to spontaneous mutation [50,51,60]. Our work conducted on the largest diversity of CoVs investigated so far is a novel compelling piece of evidence suggesting that mutational bias towards AU is a universal feature among CoVs, regardless of the viral taxonomy and of the host infected. We further report that for all CoVs, C->U substitutions occur at a much higher frequency than G->A substitutions. This asymmetric substitution pattern has been proposed to be driven by a host APOBEC3-like editing process, rather than occurring during virus replication and being associated with mutational biases of the coronavirus RNA-dependent RNA polymerase [51,61]. The *APOBEC3* locus in mammals consists of a series of tandem copies of different *APOBEC3* genes that have undergone a complex evolutionary history of duplications, deletions and fusions, further complicated at the transcriptome level with a large diversity of mRNA [62]. The evolution of the *APOBEC3* locus is tightly linked to viral infections via a different arms race [63,64]. The structure and synteny in the *APOBEC3* locus and the protein repertoire therein encoded are extremely lineage-specific (see for instance [65,66] for Chiroptera or [67,68] for Primates). It is thus important to note that the mutagenic role of host APOBEC3 enzymes acting onto viral genomes and exerting strong evolutionary pressures in a particular virus–host interaction may fuel viral adaptations, which facilitate viral transmission and colonisation of novel host species, as suggested for lentiviruses infecting Primates [69].

In addition to the increased GC->AU substitutions, we remark that CoVs genomes are strongly depleted in the CpG dinucleotide, resulting in the avoidance of synonymous codons and of codon pairs containing this motif. Our results are consistent with previous works demonstrating that the experimental increase in CpG frequencies impaired virus fitness [37,70–72]. CpG depletion has been observed across CoVs infecting a whole range of mammals and birds, indicating that the attenuation mechanism might be fundamental to vertebrate eukaryotic antiviral defense, evolutionarily conserved and active over three hundreds of millions of years of evolution [73]. Surprisingly, we did not observe any statistically significant depletion of the UpA dinucleotide in CoV genomes (Figure 3A), while underrepresentation of UpA is a common trend among human RNA virus [57,74]. Analogous to the host immune response against the presence of CpG in the viral genomes, the increased presence of UpA dinucleotides has been formally identified as reducing virus replication [75], potentially through the cleavage of viral RNAs by RNase L [71,76]. Our investigation of the impact of UpA dinucleotide frequency on the synonymous codon frequency displayed mitigated results: avoidance of UpA dinucleotide was observed when the UpA dinucleotide is located at the end of a codon (Figure 3D) but not when the UpA dinucleotide is located at the codon pair boundary (Figure 3E). Hence, we hypothesize that UpA sites in CoVs are at the center of the antagonist effects of mutation bias (promoting the formation of new UpA sites through GC->AU mutational bias) and selection (directly acting against UpA sites). However, testing this mutation-selection trade-off hypothesis goes beyond the power of our bioinformatics analyses and should be properly addressed by actually modifying the UpA content of CoV genomes, quantifying changes in viral fitness and following the evolutionary trajectories of the modified genomes by means of experimental evolution. Altogether, our work fits well with the mounting evidence supporting the biotechnological application of modifications in CpG and UpA dinucleotide frequencies in viral genomes for the production of efficient, safe and evolution-proof vaccines [71].

3.2. Lack of Evidence for Translational Selection Acting on CoVs

Mismatches between the mRNA characteristics and the translation machinery can have a strong impact on the quality and quantity of protein production. Selection acting on mutations leading to streamlined translation is known as translational selection. However, studies on the distribution of fitness effect show that the selective advantage associated with beneficial individual synonymous mutations is very low [77–81], implying that translational selection could only be efficient in the case of organisms with large effective population sizes. Indeed, strong phenotypic effects of codon usage on expression levels of single genes have been experimentally reported in some metazoan species with large population sizes (such as drosophila or nematode) [42,82,83]. Similarly, studies in vertebrates have established that translational selection is much weaker in large-sized vertebrates compared to small-sized ones, because the former are slow-growing organisms with a small effective population size [84]. Additionally, the physicochemical organization of vertebrate chromosomes may further hamper translational selection from playing a strong role in vertebrates: chromosomes in many vertebrates are organized in long, consecutive regions enriched in AT or in GC nucleotides, known as isochores [85], so that the factor with the largest individual impact on CUB of vertebrate genes is its physical location along the chromosome [86].

As with cellular genes, viral genes rely completely on the host cell apparatus for translation. It has thus been proposed that selection could act to optimize the match between the virus CUB and tRNA repertoire and abundance in its host, leading to increased efficiency and accuracy of viral protein synthesis [14]. In the case of viruses, the large within-host population sizes that many of them generate during productive infections could allow translational selection to shape CUB. For bacteriophages, CUB has been shown to match codon preferences in the bacterial host [28], which in their turn match the most abundant cognate tRNAs available in the cell [87]. In viruses infecting vertebrates, mounting experimental evidence suggests that modification of viral codon usage leads to sharp changes in viral fitness, the necessary condition for natural selection to act upon viral CUB [88]. However, the results presented here show that host stratification explains only a minor fraction of the global variability of CUB in CoVs, suggesting that for these viruses the variability of synonymous codons is largely unrelated to the type of host infected. Thus, we conclude that translational selection based on a differential tRNA abundance as a function of the host is not the main evolutionary driver of CUB in CoVs. Experimental results on ribosome profiling for *Murine coronavirus* genes have reported that despite the poor match in CUB between this virus and the mouse host, coronavirus mRNAs were translated with similar efficiency as the host mRNAs [89]. Thus, the high frequency of U- and A-ending synonymous codon in CoVs genomes, systematically departing from the CUB of their host [59], would not negatively impact the synthesis of viral proteins.

In metazoans the interplay between the main factors shaping CUB, mutational forces, GC-biased gene conversion and translational selection, have been finely analyzed and interpreted as a function of effective population size [84,90]. Notwithstanding, the same studies identify CUB trends, such as the systematic preference of pyrimidines over purines at the third codon position, for which we still lack an interpretative framework [84]. In our analyses for CoVs, despite the strong explanatory power of variation in nucleotide composition and CpG avoidance to explain variation in CUB, over 40% of the overall variation in CUB still remains to be explained. Given the impact of effective population size on the efficiency of translational selection to shape CUB in metazoans, could translational selection contribute to explain the proportion of the variation in CUB in CoVs that has not been attributed to any variable? Here, we report that the between-host variability of CUB in CoVs accounts for 19.7% of the total variability in CUB, a fraction two times larger than the random expectation. Although statistically significant, this contribution does not necessarily prove that differential adaptation to the translational machinery of the different hosts is shaping CUB in CoVs. Our analyses suggest instead that significant differences in GC3 and UpA occur between CoVs infecting different hosts (Supplementary Figures S7 and S8), so that between-hosts differences in CUB in CoVs could actually also reflect the impact of compositional variation at shaping CUB. A proper answer to the question of the actual role of translational selection in CoVs will require experimental work identifying differences in the natural history of the infection (e.g., target tissue, productivity, clinical presentation) of genetically close viruses upon host switch.

We want to point out a potential limitation for our study, namely that we focused on a nucleotide and CUB analysis for the complete genome, instead of having focused on the individual genes. Furthermore, as detailed in the methods and to prevent biases introduced by annotation errors in the SARS-COV-2 genomes, we limited ourselves to the best-annotated open reading frames, namely ORF1ab, S, E, M and N. These global analyses may have erased any potential signal for differential composition and/or CUB between open reading frames. It is conceivable that different open reading frames in the CoV genomes may present differences in their nucleotide composition and CUB, as is the case in many other viruses [30,91]. This is actually more relevant in the case of CoVs, because of the (+)ssRNA nature of the viral genome. Indeed, transcription in CoVs generates molecules of sub-genomic length, differing in the 5′ starting open reading frame [92], and among all proteins potentially encoded by a CoV genome or sub-genome, only the one corresponding to the open reading frame located in the 5′ end is actually translated [93,94]. Thus, in CoVs, the central mechanism for gene expression regulation is transcriptional regulation, which may evidently facilitate, for instance, the selection for different coding trends between open reading frames located in the 5′ and in the 3′ ends of the viral genome. This question can be properly addressed by differentially analyzing the translation efficiency of each viral open reading frame during the course of the individual cell infection, relating them to the putative CUB differences between the proteins encoded in the genome 5′ and involved in the replication-transcription complex on the one hand, and the structural proteins encoded in the genome 3′ on the other hand.

3.3. Composition of CoVs Genomes Tends to Reach Their Mutational Equilibria

In order to investigate mutational bias in CoVs and to disentangle it from the effect of natural selection, we analyzed mutations occurring during recent evolution in 16 viral metapopulations at different taxonomic host and viral levels. With this approach, we aimed at comparing mutational biases, while minimizing the impact of natural selection. When considering highly accurate folded site-frequency spectrum at the population level, we observe that endemic hCoVs are closer to an equilibrium in the AU—GC allele distribution than zoonotic, recently acquired hCoVs (Figure 2A,B). This suggests that endemic hCoVs may have undergone a compositional drift towards a novel equilibrium under the novel mutational pressures in the human upon host switch. Indeed, at the short time scale that our analyses can explore for the SARS-CoV-2 epidemics in humans, we verify a small albeit significant trend towards GC reduction (Figure 2C). Similar trends have also been reported when analyzing different datasets [95,96]. Future work will benefit from accessing a sufficient number of SARS-CoV-2 non-human strains collected from different hosts (bats and pangolins in first instance, given the current knowledge) to assess the population-level characteristics of the direction and intensity of mutational bias in endemic host species. Additionally, endemic hCoVs exhibit a large variation in observed synonymous GC3-content, ranging from 18.7% to 30.6%, which suggests that the strength of the mutational bias causing C->U changes is not similar among all hCoVs. We hypothesize that these differences in mutational signature could be related to the differences in the host mutator APOBEC3 repertoires, which vary between hosts, but also between cell types within an organism, so that changes in the host tropism will have an impact on the actual mutational intensity and direction that the viral genomes experience. It is interesting to note that when performing a similar analysis for non-human CoVs, a number of viruses display a significant departure from neutrality in SFS skewness, mostly flagrant for avian gamma-coronavirus *Infectious bronchitis virus* as well as for the SARS-CoV-2 sequences from American mink *Neovison vison* isolates, but also evident in MERS sequences from camel isolates, *Camel alphacoronavirus* and *Porcine coronavirus HKU15*. Should our interpretation of the SFS skweness for hCoVs be correct, we could speculate that these viral lineages might have undergone a recent shift in mutational bias, compatible with a host switch or tropism shift event.

3.4. Conclusion: CUB Is a Poor Proxy to Predict Zoonotic Infection in CoVs

SARS-CoV-2 is the third *Coronavirinae* zoonotic spillover of the 21-century, and the associated COVID-19 poses an unprecedented burden on human public health. Understanding the drivers and facilitators of interspecies viral transmission in CoVs is a key public health and fundamental research priority. Nucleotide composition and CUB are distinctive characteristics of a virus, shaped by the global action of different pressures, including host adaptation [29,97,98]. In the case of SARS-CoV-2, the viral CUB is closer to that of humans than other hCoVs [58]. Although one could interpret that this "closer-to-human" CUB may have facilitated the establishment in humans following the zoonotic transmission event, our overall results show that the particular match between CoVs and host CUB alone plays a minor role in the chances for a zoonotic spillover to establish in humans. Instead, we have shown here that the main drivers of CUB (low GC3 content, strong CpG avoidance and slight UpA avoidance) are common to all CoVs, independently of viral taxonomy and of the hosts they infect. We interpret thus that the role of CUB at modulating the chances of a CoV zoonotic spillover to thrive in humans is negligible. Notwithstanding, the available data suggest that upon colonization of humans, endemic hCoVs have experienced a compositional drift towards a novel compositional equilibrium in the new hosts, and that this could also be the case for SARS-CoV-2 if the current pandemic evolves towards an endemic circulation in humans.

4. Material and Methods

4.1. Data Collection and Processing

To access the diversity of coronavirus species we downloaded sequences classified within the *Coronaviridae* families from the virus–host database (https://www.genome.jp/virushostdb/, accessed on 19 March 2021). Because the 212 viral entries present in the database are composed of a mixture of sequence classified at different taxonomy levels, i.e., species (HKU1, HKU4, HKU5, . . .) or strains (HKU4-1, HKU4-2, HKU4-3, . . .), we used CD-HIT (option–c 0.95–n 5) to cluster together sequences sharing over 95% of identify. A total of 82 distinct groups of complete genomes were identified, from which we manually choose one single representative member, preferentially selecting the NCBI manually curated genomic sequence.

For the intra-species diversity, we downloaded strain sequences from the Virus Pathogen Resource database (VIRP). Unusually long or short sequences (>130% or <70% of the median length of the reference sequence of a species) were filtered out. In addition, the complete genomic sequences of SARS-CoV-2 isolates were obtained from GISAID (available at https://www.gisaid.org/epiflu-applications/nexthcov-19-app/), accessed on 6 May 2020 to collect the sequences for our metapopulation analysis. SARS-CoV2 sequences bellow 29,000 bp were automatically discarded. Detailed accession ID for both datasets are provided in Supplementary Tables S1 and S2.

4.2. Nucleotide Composition Analysis

From the NCBI Genbank files of the previously selected 82 complete genomes, we used BIOPYTHON to concatenate the coding sequences of each CoV genome, and then compute the total synonymous codon frequencies for each genome. Only ORFs from the orf1ab, spike glycoprotein (S), envelope small membrane protein (E), membrane protein (M), and nucleoprotein (N) were considered for our downstream analysis, as those ORFs were presumably the best annotated ones across CoVs. We want to raise a word of caution, as we have analyzed the complete coding genomes of the different viral taxa without individual focus on the different genome regions that may be differentially transcribed and replicated and present local peculiarities in nucleotide composition and/or synonymous codon usage. Thus, our results refer to overall trends for the complete genomes and do not exclude the possibility of the presence of local compositional anomalies. A matrix of 82 (CoVs) × 59 (synonymous codons) was created, which served as input for either our PCA analysis (FactoMineR) and correspondence analysis (run with R package ade4). In parallel, we calculated the dinucleotide observed/expected abundances in coding sequences by calculating the ratio rXY = fXY/fXfY, where fXY denotes the observed frequency of the dinucleotide XY and fXfY the product of the individual frequencies of the nucleotides X and Y in a sequence. We used as significant lower and upper boundaries the thresholds of 0.78 and 1.25, corresponding to $p < 0.001$ for sufficiently long (>20 kb) random sequences [99].

4.3. SNP Calling

To identify intra-species polymorphisms, we used MUMMER [100] to perform a genome-to-genome alignment of our 16 metapopulations against its corresponding reference sequence, to then call SNPs using the program SHOW-SNPS from the MUMMER TOOLS suite. Multi-allelic sites and structural variants were filtered out to only consider bi-allelic variants. In order to differentiate synonymous to non-synonymous mutations, using BIOPYTHON we annotated each variant based on the CDS annotation provided within the corresponding Genbank reference sequence.

4.4. Assessment Mutation Profiles

For the SARS-CoV2 genome, accessing the mutation profiles at synonymous positions is trivial since one SARS-CoV2 sequence sampled from the early pandemic has been sequenced. Hence, to estimate the whole-genome nucleotide flux, we need to count the frequency of each type of mutation with respect to this quasi-ancestral genome. However, for the other CoVs genomes, such a sequence has not been characterized yet, and we need to develop a strategy to determine which allele is ancestral and which is derived. Here, we used a conservative approach, considering variants with a minor allele frequency lower than 10%, with the ancestral state was considered for the major allele while the minor allele was considered as the derivate state.

4.5. Site Frequency Spectrum Assessment Mutational Equilibrium

Prior to the inference the site frequency spectrum (SFS), the population structure of each metapopulation was characterized. Using BIOPYTHON, we first transformed our multi-sample VCF file into a multi-alignment (MSA) file, which served as input file for the identification of population sub-structure and recombinant sequences using FASTGEAR [101]. FASTGEAR presents the advantage of identifying population genetic structure from an alignment, as well as detecting recombination between the inferred lineages. Complete genomes presenting long recombining segments (>200 bp) were removed from each dataset, while short recombination segments were masked for the downstream analysis.

From the VCF file of each metapopulation, the allele frequency of each variant was computed using VCFTOOLS (–freq). In order to work in a framework where population structure of our metapopulation does not interfere with the shape of the derived allele frequency counts, we focused for each viral entry on the single lineage per metapopulation, having the highest number of complete genomes (as previously identified by FASTGEAR). For each of the 18 viral metapopulations we calculated the derived allele frequency spectrum by comparing each extant sequence with the ancestral one. Previous studies have shown that the analysis of the derived allele frequency spectrum are sufficiently robust to demographic and/or sampling effects [102,103]. We constructed then the folded SFS by polarizing the derived allele frequency as AU-enriching or GC-enriching with respect to the ancestral sequence, as in [56] using the R script kindly provided by Dr. S. Glémin. The folded SFS is symmetrical if and only if GC content is at mutation balance equilibrium, and this result is robust for populations outside demographic equilibrium [56]. Indeed, analyses of the folded SFS can be used to infer demographic histories, reflecting bottlenecks and population expansions (see for instance [104]).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/v13091800/s1, Supplementary Table S1: *Orthocoronaviridae* sequence data sets used for the study. Supplementary Table S2: Summary statistics for the 16 *Orthocoronaviridae* metapopulations. Supplementary Figure S1: (A) Bar chart representing the percentage of variance explained by each principal component. (B) Projection of the individual codon on the first two principal components. PC1 sharply splits A- and T-ending codons (on the left) from C- or G- ending codons (on the right), the only exception being TTG (Leucine) codon, framed in purple. Supplementary Figure S2: Spearman correlation between GC3 content (%) and coordinate of the first 4 principal components. Supplementary Figure S3: Representation of the steps we used to select strains for our metapopulation analysis through the example of the *Infectious bronchitis virus*. (A) Representation of the recent recombination events happening within strains inferred by fastGear. Rows correspond to viral strains sequences, the columns to positions in the alignment (a total of 27,685 positions), and colors show the membership of each portion of sequences to lineage detected by fastGear (in this case four lineages, summarised on the colored bar). Sequences identified as of admixed nature and thus potential recombinants are labelled in white on the right bar. Sequences used for subsequent phylogenetic inference are labelled in black on this same right bar. (B) Phylogenetic tree of the relationship among viral strains. The tree is rooted based on an outgroup (grey) as indicated on Supplementary Table S2. Supplementary Figure S4: Scaterplot of the relationship between skweness of the folded site frequency spectrum (SFS) and coronavirus emergence time. The downward and upward triangles represent negative and positive skweness, respectively, and filled triangles represent skweness values significantly different from the null expectation. Supplementary Figure S5: Number of SARS-CoV2 sequenced genomes since the beginning of the outbreak, represented by bin of 10,000 sequences (top panel). Estimation of the SFS skweness, by subsampling five replicates of 1000 sequences within each bin (bottom panel). Supplementary Figure S6: Spearman correlation between relative abundance of CpG dinucleotide and coordinate of the first 4 principal components. Supplementary Figure S7: Spearman correlation between relative abundance of UpA dinucleotide and coordinate of the first 4 principal components. Supplementary Figure S8: Comparison of the CG3 content, CpG, and UpA dinucleotide relative abundances across CoVs hosts. Variance analysis of normally distributed data was carried out using ANOVA followed by a post doc Tukey test showing further individual pairwise differences. Non-normal data were processed using Kruskal–Wallis test follow by a rank Wilcoxon test with a Bonferroni correction (code for *p*-value: * <0.05, ** <0.01, *** <0.001). Supplementary Figure S9: Comparison of the CG3 content, CpG, and UpA dinucleotide relative abundances across CoVs genera. Variance analysis of normally distributed data was carried out using ANOVA followed by a post doc Tukey test showing further individual pairwise differences. Non-normal data were processed using Kruskal–Wallis test follow by a rank Wilcoxon test with a Bonferroni correction (code for *p*-value: * <0.05, ** <0.01, *** <0.001).

Author Contributions: Conception: J.D. and I.G.B.; funding acquisition: J.D. and I.G.B.; method development and data analysis: J.D.; interpretation of the results: J.D. and I.G.B.; drafting of the manuscript: J.D. and I.G.B. Both authors have read and agreed to the published version of the manuscript.

Funding: J.D. was supported by the Fondation pour la recherche Médicale (FRM, ARF20170938823) as well as by the Marie-Curie EU Horizon 2020 Marie-Skłodowska-Curie research and innovation program grant METHYVIREVOL (contract number 800489). I.G.B. was supported by the European Union Horizon 2020 research and innovation program under the grant agreement CODOVIREVOL (ERC-2014-CoG-647916).

Data Availability Statement: We acknowledge the three different sources of genomes that we employed: the virus–host database (https://www.genome.jp/virushostdb/, version of 19 March 2020), the Virus Pathogen Resource database (VIRP, https://www.viprbrc.org/brc/home.spg?decorator=vipr, version of 26 April 2020), and the GISAID database (https://www.gisaid.org/, version of 19 August 2020).

Acknowledgments: The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/-http://www.southgreen.fr. We warmly thank Sylvain Glémin for sharing with us R script function and helpful suggestions regarding our estimation of the mutational equilibrium.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [CrossRef] [PubMed]
2. Boni, M.F.; Lemey, P.; Jiang, X.; Lam, T.T.-Y.; Perry, B.W.; Castoe, T.A.; Rambaut, A.; Robertson, D.L. Evolutionary Origins of the SARS-CoV-2 Sarbecovirus Lineage Responsible for the COVID-19 Pandemic. *Nat. Microbiol.* **2020**, *5*, 1408–1417. [CrossRef] [PubMed]
3. Graham, R.L.; Baric, R.S. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* **2010**, *84*, 3134–3146. [CrossRef]
4. Forni, D.; Cagliani, R.; Clerici, M.; Sironi, M. Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol.* **2017**, *25*, 35–48. [CrossRef] [PubMed]
5. Anthony, S.J.; Johnson, C.K.; Greig, D.J.; Kramer, S.; Che, X.; Wells, H.; Hicks, A.L.; Joly, D.O.; Wolfe, N.D.; Daszak, P.; et al. Global Patterns in Coronavirus Diversity. *Virus Evol.* **2017**, *3*, vex012. [CrossRef]

6. Leopardi, S.; Holmes, E.C.; Gastaldelli, M.; Tassoni, L.; Priori, P.; Scaravelli, D.; Zamperin, G.; De Benedictis, P. Interplay between Co-Divergence and Cross-Species Transmission in the Evolutionary History of Bat Coronaviruses. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2018**, *58*, 279–289. [CrossRef] [PubMed]
7. Cui, J.; Li, F.; Shi, Z.-L. Origin and Evolution of Pathogenic Coronaviruses. *Nat. Rev. Microbiol.* **2019**, *17*, 181–192. [CrossRef]
8. Olival, K.J.; Hosseini, P.R.; Zambrana-Torrelio, C.; Ross, N.; Bogich, T.L.; Daszak, P. Host and Viral Traits Predict Zoonotic Spillover from Mammals. *Nature* **2017**, *546*, 646–650. [CrossRef] [PubMed]
9. Dhama, K.; Patel, S.K.; Sharun, K.; Pathak, M.; Tiwari, R.; Yatoo, M.I.; Malik, Y.S.; Sah, R.; Rabaan, A.A.; Panwar, P.K.; et al. SARS-CoV-2 Jumping the Species Barrier: Zoonotic Lessons from SARS, MERS and Recent Advances to Combat This Pandemic Virus. *Travel Med. Infect. Dis.* **2020**, *37*, 101830. [CrossRef] [PubMed]
10. Albers, S.; Czech, A. Exploiting TRNAs to Boost Virulence. *Life* **2016**, *6*, 4. [CrossRef] [PubMed]
11. Franzo, G.; Tucciarone, C.M.; Cecchinato, M.; Drigo, M. Canine Parvovirus Type 2 (CPV-2) and Feline Panleukopenia Virus (FPV) Codon Bias Analysis Reveals a Progressive Adaptation to the New Niche after the Host Jump. *Mol. Phylogenet. Evol.* **2017**, *114*, 82–92. [CrossRef]
12. Simón, D.; Fajardo, A.; Sóñora, M.; Delfraro, A.; Musto, H. Host Influence in the Genomic Composition of Flaviviruses: A Multivariate Approach. *Biochem. Biophys. Res. Commun.* **2017**, *492*, 572–578. [CrossRef]
13. Rahman, S.U.; Yao, X.; Li, X.; Chen, D.; Tao, S. Analysis of Codon Usage Bias of Crimean-Congo Hemorrhagic Fever Virus and Its Adaptation to Hosts. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2018**, *58*, 1–16. [CrossRef]
14. Tian, L.; Shen, X.; Murphy, R.W.; Shen, Y. The Adaptation of Codon Usage of +ssRNA Viruses to Their Hosts. *Infect. Genet. Evol.* **2018**, *63*, 175–179. [CrossRef] [PubMed]
15. Ikemura, T. Codon Usage and TRNA Content in Unicellular and Multicellular Organisms. *Mol. Biol. Evol.* **1985**, *2*, 13–34. [CrossRef]
16. Kanaya, S.; Yamada, Y.; Kinouchi, M.; Kudo, Y.; Ikemura, T. Codon Usage and TRNA Genes in Eukaryotes: Correlation of Codon Usage Diversity with Translation Efficiency and with CG-Dinucleotide Usage as Assessed by Multivariate Analysis. *J. Mol. Evol.* **2001**, *53*, 290–298. [CrossRef] [PubMed]
17. Drummond, D.A.; Wilke, C.O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **2008**, *134*, 341–352. [CrossRef]
18. Hershberg, R.; Petrov, D.A. Selection on Codon Bias. *Annu. Rev. Genet.* **2008**, *42*, 287–299. [CrossRef] [PubMed]
19. Dos Reis, M.; Wernisch, L. Estimating Translational Selection in Eukaryotic Genomes. *Mol. Biol. Evol.* **2009**, *26*, 451–461. [CrossRef]
20. Zhou, Z.; Dang, Y.; Zhou, M.; Li, L.; Yu, C.; Fu, J.; Chen, S.; Liu, Y. Codon Usage Is an Important Determinant of Gene Expression Levels Largely through Its Effects on Transcription. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E6117–E6125. [CrossRef]
21. Kudla, G.; Murray, A.W.; Tollervey, D.; Plotkin, J.B. Coding-Sequence Determinants of Gene Expression in Escherichia Coli. *Science* **2009**, *324*, 255–258. [CrossRef]
22. Goodman, D.B.; Church, G.M.; Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **2013**, *342*, 475–479. [CrossRef] [PubMed]
23. Sørensen, M.A.; Kurland, C.G.; Pedersen, S. Codon Usage Determines Translation Rate in Escherichia Coli. *J. Mol. Biol.* **1989**, *207*, 365–377. [CrossRef]
24. Akashi, H. Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. *Genetics* **1994**, *136*, 927–935. [CrossRef] [PubMed]
25. Presnyak, V.; Alhusaini, N.; Chen, Y.-H.; Martin, S.; Morris, N.; Kline, N.; Olson, S.; Weinberg, D.; Baker, K.E.; Graveley, B.R.; et al. Codon Optimality Is a Major Determinant of MRNA Stability. *Cell* **2015**, *160*, 1111–1124. [CrossRef]
26. Pagani, F.; Raponi, M.; Baralle, F.E. Synonymous Mutations in CFTR Exon 12 Affect Splicing and Are Not Neutral in Evolution. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6368–6372. [CrossRef] [PubMed]
27. Bahir, I.; Fromer, M.; Prat, Y.; Linial, M. Viral Adaptation to Host: A Proteome-Based Analysis of Codon Usage and Amino Acid Preferences. *Mol. Syst. Biol.* **2009**, *5*, 311. [CrossRef]
28. Lucks, J.B.; Nelson, D.R.; Kudla, G.R.; Plotkin, J.B. Genome Landscapes and Bacteriophage Codon Usage. *PLoS Comput. Biol.* **2008**, *4*, e1000001. [CrossRef]
29. Wong, E.H.; Smith, D.K.; Rabadan, R.; Peiris, M.; Poon, L.L. Codon Usage Bias and the Evolution of Influenza A Viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **2010**, *10*, 253. [CrossRef] [PubMed]
30. Félez-Sánchez, M.; Trösemeier, J.-H.; Bedhomme, S.; González-Bravo, M.I.; Kamp, C.; Bravo, I.G. Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses. *Genome Biol. Evol.* **2015**, *7*, 2117–2135. [CrossRef] [PubMed]
31. Gu, H.; Chu, D.K.W.; Peiris, M.; Poon, L.L.M. Multivariate Analyses of Codon Usage of SARS-CoV-2 and Other Betacoronaviruses. *Virus Evol.* **2020**, *6*, veaa032. [CrossRef]
32. Tort, F.L.; Castells, M.; Cristina, J. A Comprehensive Analysis of Genome Composition and Codon Usage Patterns of Emerging Coronaviruses. *Virus Res.* **2020**, *283*, 197976. [CrossRef] [PubMed]
33. Gong, Y.; Wen, G.; Jiang, J.; Xie, F. Codon Bias Analysis May Be Insufficient for Identifying Host(s) of a Novel Virus. *J. Med. Virol.* **2020**, *92*, 1434–1436. [CrossRef]
34. Lauring, A.S.; Jones, J.O.; Andino, R. Rationalizing the Development of Live Attenuated Virus Vaccines. *Nat. Biotechnol.* **2010**, *28*, 573–579. [CrossRef] [PubMed]

35. Yap, Y.L.; Zhang, X.W.; Danchin, A. Relationship of SARS-CoV to Other Pathogenic RNA Viruses Explored by Tetranucleotide Usage Profiling. *BMC Bioinform.* **2003**, *4*, 43. [CrossRef]
36. Greenbaum, B.D.; Levine, A.J.; Bhanot, G.; Rabadan, R. Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLoS Pathog.* **2008**, *4*, e1000079. [CrossRef] [PubMed]
37. Atkinson, N.J.; Witteveldt, J.; Evans, D.J.; Simmonds, P. The Influence of CpG and UpA Dinucleotide Frequencies on RNA Virus Replication and Characterization of the Innate Cellular Pathways Underlying Virus Attenuation and Enhanced Replication. *Nucleic Acids Res.* **2014**, *42*, 4527–4545. [CrossRef] [PubMed]
38. Takata, M.A.; Gonçalves-Carneiro, D.; Zang, T.M.; Soll, S.J.; York, A.; Blanco-Melo, D.; Bieniasz, P.D. CG Dinucleotide Suppression Enables Antiviral Defence Targeting Non-Self RNA. *Nature* **2017**, *550*, 124–127. [CrossRef]
39. Tulloch, F.; Atkinson, N.J.; Evans, D.J.; Ryan, M.D.; Simmonds, P. RNA Virus Attenuation by Codon Pair Deoptimisation Is an Artefact of Increases in CpG/UpA Dinucleotide Frequencies. *eLife* **2014**, *3*, e04531. [CrossRef] [PubMed]
40. Kumagai, Y.; Takeuchi, O.; Akira, S. TLR9 as a Key Receptor for the Recognition of DNA. *Adv. Drug Deliv. Rev.* **2008**, *60*, 795–804. [CrossRef] [PubMed]
41. Duret, L. Evolution of Synonymous Codon Usage in Metazoans. *Curr. Opin. Genet. Dev.* **2002**, *12*, 640–649. [CrossRef]
42. Chamary, J.V.; Parmley, J.L.; Hurst, L.D. Hearing Silence: Non-Neutral Evolution at Synonymous Sites in Mammals. *Nat. Rev. Genet.* **2006**, *7*, 98–108. [CrossRef] [PubMed]
43. Lynch, M. Rate, Molecular Spectrum, and Consequences of Human Mutation. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 961–968. [CrossRef]
44. Duret, L.; Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genom. Hum. Genet.* **2009**, *10*, 285–311. [CrossRef]
45. Hershberg, R.; Petrov, D.A. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genet.* **2010**, *6*, e1001115. [CrossRef]
46. Petrov, D.A.; Hartl, D.L. Patterns of Nucleotide Substitution in Drosophila and Mammalian Genomes. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 1475–1479. [CrossRef] [PubMed]
47. Haddrill, P.R.; Charlesworth, B. Non-Neutral Processes Drive the Nucleotide Composition of Non-Coding Sequences in Drosophila. *Biol. Lett.* **2008**, *4*, 438–441. [CrossRef] [PubMed]
48. Denver, D.R.; Dolan, P.C.; Wilhelm, L.J.; Sung, W.; Lucas-Lledó, J.I.; Howe, D.K.; Lewis, S.C.; Okamoto, K.; Thomas, W.K.; Lynch, M.; et al. A Genome-Wide View of Caenorhabditis Elegans Base-Substitution Mutation Processes. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 16310–16314. [CrossRef]
49. Ossowski, S.; Schneeberger, K.; Lucas-Lledó, J.I.; Warthmann, N.; Clark, R.M.; Shaw, R.G.; Weigel, D.; Lynch, M. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana. *Science* **2010**, *327*, 92–94. [CrossRef] [PubMed]
50. Rice, A.M.; Castillo Morales, A.; Ho, A.T.; Mordstein, C.; Mühlhausen, S.; Watson, S.; Cano, L.; Young, B.; Kudla, G.; Hurst, L.D. Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design. *Mol. Biol. Evol.* **2021**, *38*, 67–83. [CrossRef]
51. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **2020**, *5*, e00408-20. [CrossRef] [PubMed]
52. Vijgen, L.; Keyaerts, E.; Moës, E.; Thoelen, I.; Wollants, E.; Lemey, P.; Vandamme, A.-M.; Ranst, M.V. Complete Genomic Sequence of Human Coronavirus OC43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event. *J. Virol.* **2005**, *79*, 1595–1604. [CrossRef] [PubMed]
53. Pfefferle, S.; Oppong, S.; Drexler, J.F.; Gloza-Rausch, F.; Ipsen, A.; Seebens, A.; Müller, M.A.; Annan, A.; Vallo, P.; Adu-Sarkodie, Y.; et al. Distant Relatives of Severe Acute Respiratory Syndrome Coronavirus and Close Relatives of Human Coronavirus 229E in Bats, Ghana. *Emerg. Infect. Dis. J.* **2009**, *15*, 1377–1384. [CrossRef] [PubMed]
54. Huynh, J.; Li, S.; Yount, B.; Smith, A.; Sturges, L.; Olsen, J.C.; Nagel, J.; Johnson, J.B.; Agnihothram, S.; Gates, J.E.; et al. Evidence Supporting a Zoonotic Origin of Human Coronavirus Strain NL63. *J. Virol.* **2012**, *86*, 12816–12825. [CrossRef]
55. Al-Khannaq, M.N.; Ng, K.T.; Oong, X.Y.; Pang, Y.K.; Takebe, Y.; Chook, J.B.; Hanafi, N.S.; Kamarulzaman, A.; Tee, K.K. Molecular Epidemiology and Evolutionary Histories of Human Coronavirus OC43 and HKU1 among Patients with Upper Respiratory Tract Infections in Kuala Lumpur, Malaysia. *Virol J.* **2016**, *13*, 33. [CrossRef]
56. Glémin, S.; Arndt, P.F.; Messer, P.W.; Petrov, D.; Galtier, N.; Duret, L. Quantification of GC-Biased Gene Conversion in the Human Genome. *Genome Res.* **2015**, *25*, 1215–1228. [CrossRef] [PubMed]
57. Kunec, D.; Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep.* **2016**, *14*, 55–67. [CrossRef]
58. Dilucca, M.; Forcelloni, S.; Georgakilas, A.G.; Giansanti, A.; Pavlopoulou, A. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* **2020**, *12*, 498. [CrossRef] [PubMed]
59. Ji, W.; Wang, W.; Zhao, X.; Zai, J.; Li, X. Cross-Species Transmission of the Newly Identified Coronavirus 2019-NCoV. *J. Med. Virol.* **2020**, *92*, 433–440. [CrossRef]
60. De Maio, N.; Walker, C.R.; Turakhia, Y.; Lanfear, R.; Corbett-Detig, R.; Goldman, N. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **2021**, *13*, evab087. [CrossRef]
61. Di Giorgio, S.; Martignano, F.; Torcia, M.G.; Mattiuz, G.; Conticello, S.G. Evidence for Host-Dependent RNA Editing in the Transcriptome of SARS-CoV-2. *Sci. Adv.* **2020**, *6*, eabb5813. [CrossRef]

62. Münk, C.; Willemsen, A.; Bravo, I.G. An Ancient History of Gene Duplications, Fusions and Losses in the Evolution of APOBEC3 Mutators in Mammals. *BMC Evol. Biol.* **2012**, *12*, 71. [CrossRef] [PubMed]
63. Harris, R.S.; Anderson, B.D. Evolutionary Paradigms from Ancient and Ongoing Conflicts between the Lentiviral Vif Protein and Mammalian APOBEC3 Enzymes. *PLoS Pathog.* **2016**, *12*, e1005958. [CrossRef]
64. Ito, J.; Gifford, R.J.; Sato, K. Retroviruses Drive the Rapid Evolution of Mammalian APOBEC3 Genes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 610–618. [CrossRef] [PubMed]
65. Hayward, J.A.; Tachedjian, M.; Cui, J.; Cheng, A.Z.; Johnson, A.; Baker, M.L.; Harris, R.S.; Wang, L.-F.; Tachedjian, G. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. *Mol. Biol. Evol.* **2018**, *35*, 1626–1637. [CrossRef]
66. Jebb, D.; Huang, Z.; Pippel, M.; Hughes, G.M.; Lavrichenko, K.; Devanna, P.; Winkler, S.; Jermiin, L.S.; Skirmuntt, E.C.; Katzourakis, A.; et al. Six Reference-Quality Genomes Reveal Evolution of Bat Adaptations. *Nature* **2020**, *583*, 578–584. [CrossRef]
67. Garcia, E.I.; Emerman, M. Recurrent Loss of APOBEC3H Activity during Primate Evolution. *J. Virol.* **2018**, *92*, e00971-18. [CrossRef] [PubMed]
68. Yang, L.; Emerman, M.; Malik, H.S.; McLaughlin, R.N. Retrocopying Expands the Functional Repertoire of APOBEC3 Antiviral Proteins in Primates. *eLife* **2020**, *9*, e58436. [CrossRef] [PubMed]
69. Nakano, Y.; Yamamoto, K.; Ueda, M.T.; Soper, A.; Konno, Y.; Kimura, I.; Uriu, K.; Kumata, R.; Aso, H.; Misawa, N.; et al. A Role for Gorilla APOBEC3G in Shaping Lentivirus Evolution Including Transmission to Humans. *PLoS Pathog.* **2020**, *16*, e1008812. [CrossRef]
70. Burns, C.C.; Campagnoli, R.; Shaw, J.; Vincent, A.; Jorba, J.; Kew, O. Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. *J. Virol.* **2009**, *83*, 9957–9969. [CrossRef]
71. Gaunt, E.; Wise, H.M.; Zhang, H.; Lee, L.N.; Atkinson, N.J.; Nicol, M.Q.; Highton, A.J.; Klenerman, P.; Beard, P.M.; Dutia, B.M.; et al. Elevation of CpG Frequencies in Influenza A Genome Attenuates Pathogenicity but Enhances Host Response to Infection. *eLife* **2016**, *5*, e12735. [CrossRef]
72. Antzin-Anduetza, I.; Mahiet, C.; Granger, L.A.; Odendall, C.; Swanson, C.M. Increasing the CpG Dinucleotide Abundance in the HIV-1 Genomic RNA Inhibits Viral Replication. *Retrovirology* **2017**, *14*, 49. [CrossRef]
73. Ibrahim, A.; Fros, J.; Bertran, A.; Sechan, F.; Odon, V.; Torrance, L.; Kormelink, R.; Simmonds, P. A Functional Investigation of the Suppression of CpG and UpA Dinucleotide Frequencies in Plant RNA Virus Genomes. *Sci. Rep.* **2019**, *9*, 18359. [CrossRef] [PubMed]
74. Simmonds, P.; Xia, W.; Baillie, J.K.; McKinnon, K. Modelling Mutational and Selection Pressures on Dinucleotides in Eukaryotic Phyla –Selection against CpG and UpA in Cytoplasmically Expressed RNA and in RNA Viruses. *BMC Genom.* **2013**, *14*, 610. [CrossRef] [PubMed]
75. Fros, J.J.; Dietrich, I.; Alshaikhahmed, K.; Passchier, T.C.; Evans, D.J.; Simmonds, P. CpG and UpA Dinucleotides in Both Coding and Non-Coding Regions of Echovirus 7 Inhibit Replication Initiation Post-Entry. *eLife* **2017**, *6*, e29112. [CrossRef]
76. Cooper, D.A.; Banerjee, S.; Chakrabarti, A.; García-Sastre, A.; Hesselberth, J.R.; Silverman, R.H.; Barton, D.J. RNase L Targets Distinct Sites in Influenza A Virus RNAs. *J. Virol.* **2015**, *89*, 2764–2776. [CrossRef] [PubMed]
77. Sanjuán, R.; Moya, A.; Elena, S.F. The Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in an RNA Virus. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8396–8401. [CrossRef] [PubMed]
78. Peris, J.B.; Davis, P.; Cuevas, J.M.; Nebot, M.R.; Sanjuán, R. Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in Bacteriophage F1. *Genetics* **2010**, *185*, 603–609. [CrossRef]
79. Jacquier, H.; Birgy, A.; Le Nagard, H.; Mechulam, Y.; Schmitt, E.; Glodt, J.; Bercot, B.; Petit, E.; Poulain, J.; Barnaud, G.; et al. Capturing the Mutational Landscape of the Beta-Lactamase TEM-1. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13067–13072. [CrossRef]
80. Fragata, I.; Matuszewski, S.; Schmitz, M.A.; Bataillon, T.; Jensen, J.D.; Bank, C. The Fitness Landscape of the Codon Space across Environments. *Heredity* **2018**, *121*, 422–437. [CrossRef]
81. Williams, M.J.; Zapata, L.; Werner, B.; Barnes, C.P.; Sottoriva, A.; Graham, T.A. Measuring the Distribution of Fitness Effects in Somatic Evolution by Combining Clonal Dynamics with DN/DS Ratios. *eLife* **2020**, *9*, e48714. [CrossRef]
82. Plotkin, J.B.; Kudla, G. Synonymous but Not the Same: The Causes and Consequences of Codon Bias. *Nat. Rev. Genet.* **2011**, *12*, 32–42. [CrossRef]
83. Mordstein, C.; Savisaar, R.; Young, R.S.; Bazile, J.; Talmane, L.; Luft, J.; Liss, M.; Taylor, M.S.; Hurst, L.D.; Kudla, G. Codon Usage and Splicing Jointly Influence MRNA Localization. *Cell Syst.* **2020**, *10*, 351–362.e8. [CrossRef]
84. Galtier, N.; Roux, C.; Rousselle, M.; Romiguier, J.; Figuet, E.; Glémin, S.; Bierne, N.; Duret, L. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Mol. Biol. Evol.* **2018**, *35*, 1092–1103. [CrossRef]
85. Caspersson, T.; Farber, S.; Foley, G.E.; Kudynowski, J.; Modest, E.J.; Simonsson, E.; Wagh, U.; Zech, L. Chemical Differentiation along Metaphase Chromosomes. *Exp. Cell Res.* **1968**, *49*, 219–222. [CrossRef]
86. Holmquist, G.P. Evolution of Chromosome Bands: Molecular Ecology of Noncoding DNA. *J. Mol. Evol* **1989**, *28*, 469–486. [CrossRef] [PubMed]
87. Rocha, E.P.C. Codon Usage Bias from TRNA's Point of View: Redundancy, Specialization, and Efficient Decoding for Translation Optimization. *Genome Res.* **2004**, *14*, 2279–2286. [CrossRef]

88. Martínez, M.A.; Jordan-Paiz, A.; Franco, S.; Nevot, M. Synonymous Virus Genome Recoding as a Tool to Impact Viral Fitness. *Trends Microbiol.* **2016**, *24*, 134–147. [CrossRef] [PubMed]
89. Irigoyen, N.; Firth, A.E.; Jones, J.D.; Chung, B.Y.-W.; Siddell, S.G.; Brierley, I. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. *PLoS Pathog.* **2016**, *12*, e1005473. [CrossRef] [PubMed]
90. Ratnakumar, A.; Mousset, S.; Glémin, S.; Berglund, J.; Galtier, N.; Duret, L.; Webster, M.T. Detecting Positive Selection within Genomes: The Problem of Biased Gene Conversion. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 2571–2580. [CrossRef]
91. Shapiro, M.; Krug, L.T.; MacCarthy, T. Mutational Pressure by Host APOBEC3s More Strongly Affects Genes Expressed Early in the Lytic Phase of Herpes Simplex Virus-1 (HSV-1) and Human Polyomavirus (HPyV) Infection. *PLoS Pathog.* **2021**, *17*, e1009560. [CrossRef]
92. Sola, I.; Almazán, F.; Zúñiga, S.; Enjuanes, L. Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu. Rev. Virol.* **2015**, *2*, 265–288. [CrossRef] [PubMed]
93. Kim, D.; Lee, J.-Y.; Yang, J.-S.; Kim, J.W.; Kim, V.N.; Chang, H. The Architecture of SARS-CoV-2 Transcriptome. *Cell* **2020**, *181*, 914–921.e10. [CrossRef] [PubMed]
94. Finkel, Y.; Mizrahi, O.; Nachshon, A.; Weingarten-Gabbay, S.; Morgenstern, D.; Yahalom-Ronen, Y.; Tamir, H.; Achdout, H.; Stein, D.; Israeli, O.; et al. The Coding Capacity of SARS-CoV-2. *Nature* **2021**, *589*, 125–130. [CrossRef] [PubMed]
95. Van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351. [CrossRef]
96. Matyášek, R.; Kovařík, A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes* **2020**, *11*, 761. [CrossRef]
97. Jenkins, G.M.; Holmes, E.C. The Extent of Codon Usage Bias in Human RNA Viruses and Its Evolutionary Origin. *Virus Res.* **2003**, *92*, 1–7. [CrossRef]
98. Cristina, J.; Moreno, P.; Moratorio, G.; Musto, H. Genome-Wide Analysis of Codon Usage Bias in Ebolavirus. *Virus Res.* **2015**, *196*, 87–93. [CrossRef]
99. Burge, C.; Campbell, A.M.; Karlin, S. Over- and under-Representation of Short Oligonucleotides in DNA Sequences. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1358–1362. [CrossRef]
100. Delcher, A.L.; Phillippy, A.; Carlton, J.; Salzberg, S.L. Fast Algorithms for Large-Scale Genome Alignment and Comparison. *Nucleic Acids Res.* **2002**, *30*, 2478–2483. [CrossRef] [PubMed]
101. Mostowy, R.; Croucher, N.J.; Andam, C.P.; Corander, J.; Hanage, W.P.; Marttinen, P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol. Biol. Evol.* **2017**, *34*, 1167–1182. [CrossRef] [PubMed]
102. Eyre-Walker, A.; Woolfit, M.; Phelps, T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics* **2006**, *173*, 891–900. [CrossRef] [PubMed]
103. Muyle, A.; Serres-Giardi, L.; Ressayre, A.; Escobar, J.; Glémin, S. GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (Rice). *Mol. Biol. Evol.* **2011**, *28*, 2695–2706. [CrossRef] [PubMed]
104. Lapierre, M.; Lambert, A.; Achaz, G. Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics* **2017**, *206*, 439–449. [CrossRef] [PubMed]