Biased Mutation and Selection in RNA Viruses

Talia Kustin[1] and Adi Stern *[,1,2]

[1]The Shmunis School of Biomedicine and Cancer Research, Tel Aviv University, Tel Aviv, Israel
[2]Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel

*Corresponding author: E-mail: sternadi@tauex.tau.ac.il.
Associate editor: Talia Kustin

Abstract

RNA viruses are responsible for some of the worst pandemics known to mankind, including outbreaks of Influenza, Ebola, and COVID-19. One major challenge in tackling RNA viruses is the fact they are extremely genetically diverse. Nevertheless, they share common features that include their dependence on host cells for replication, and high mutation rates. We set out to search for shared evolutionary characteristics that may aid in gaining a broader understanding of RNA virus evolution, and constructed a phylogeny-based data set spanning thousands of sequences from diverse single-stranded RNA viruses of animals. Strikingly, we found that the vast majority of these viruses have a skewed nucleotide composition, manifested as adenine rich (A-rich) coding sequences. In order to test whether A-richness is driven by selection or by biased mutation processes, we harnessed the effects of incomplete purifying selection at the tips of virus phylogenies. Our results revealed consistent mutational biases toward U rather than A in genomes of all viruses. In +ssRNA viruses, we found that this bias is compensated by selection against U and selection for A, which leads to A-rich genomes. In −ssRNA viruses, the genomic mutational bias toward U on the negative strand manifests as A-rich coding sequences, on the positive strand. We investigated possible reasons for the advantage of A-rich sequences including weakened RNA secondary structures, codon usage bias, and selection for a particular amino acid composition, and conclude that host immune pressures may have led to similar biases in coding sequence composition across very divergent RNA viruses.

Key words: **virus evolution, phylogeny, RNA viruses.**

Introduction

Genomes of all replicating entities, including viruses and cellular hosts, have been shaped by millions of years of evolution. The rapid progress of genomics in the past few decades has brought about enormous amounts of genomic information, and today there are thousands of genomes of viruses available, which allow studying the processes that govern the evolution of these genomes (Belshaw et al. 2008; Duffy et al. 2008; Pybus and Rambaut 2009). RNA viruses are an extremely diverse collection of entities, spanning a diverse range of hosts, morphologies, genome organizations, and genetic composition. Nevertheless, RNA viruses do share several common features that drive their evolution: 1) their ultimate dependence on the cell; 2) their high mutation rates; 3) strong purifying selection derived from constraints operating on a small and densely coding genome, and 4) sporadic but powerful positive selection driven by an evolutionary arms race with the host they infect. We hence reasoned that we may find common genomic signatures shared by RNA viruses, which in turn may allow us to learn more about the drivers of virus evolution.

One example of a process that may affect viral genomes is host editing by cellular enzymes. Two notable examples are adenosine deaminases acting on RNA (ADAR), which promotes A > G mutations, and APOBEC3 (A3), which promotes C > U mutations in single-stranded DNA, manifested as G > A mutations on the coding RNA strand of HIV (Bishop et al. 2004; Samuel 2012). In principle, A3 promotes hypermutated viral genomes, which are unlikely to be capable of replicating, and hence undergo purifying selection. However, there has been extensive debate whether A3 may sometimes operate in a suboptimal manner, leading to genomes that are "viable", that is, replication competent (Jern et al. 2009; Sadler et al. 2010; Cuevas et al. 2015; Delviks-Frankenberry et al. 2016). If indeed A3 or ADAR enzymes lead to viable replicating genomes, we would expect to see footprints of their activity in contemporary virus genomes.

Another notable example of a shared common signature across RNA viruses is the depletion of CG and UA dinucleotides across almost all known RNA viruses (Karlin et al. 1994; Greenbaum et al. 2009; Cheng et al. 2013; Tulloch et al. 2014). This under-representation is shared by viral hosts as well; TA dinucleotides (UA in RNA) are under-represented in most organisms, likely due to RNA-degrading enzymes located in the cytoplasm, and CG are under-represented in plants and vertebrates, likely due to deamination processes (Burge et al. 1992; Karlin et al. 1994). It appears that cells have evolved mechanisms to detect foreign genetic material bearing high levels of CG dinucleotides: Recently, it has been shown that

Article

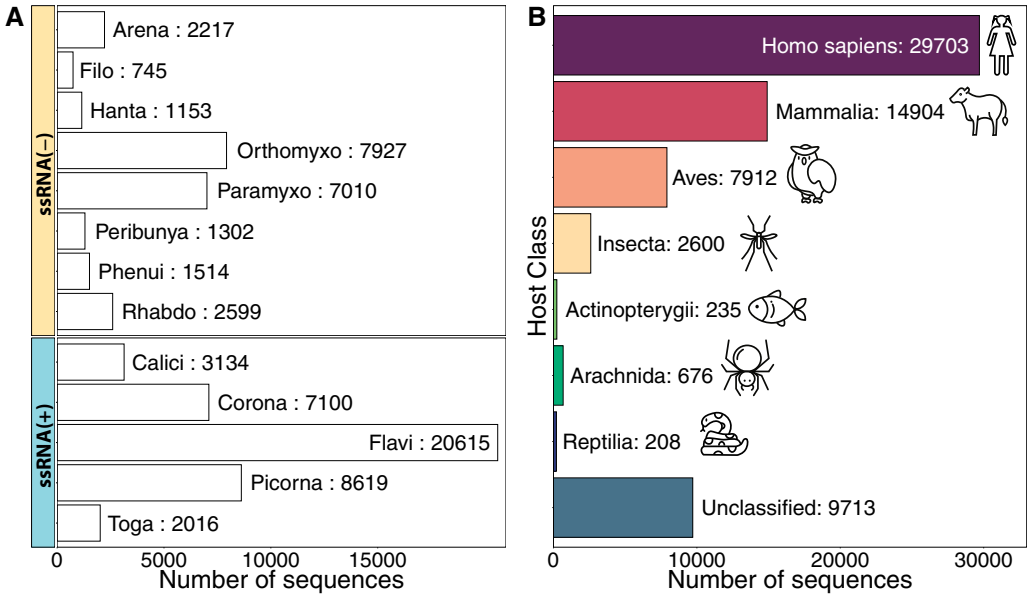

Fig. 1. PhyVirus data set. Approximately, 65,900 coding sequences of pathogenic animal viruses were assembled into multiple sequence alignments and phylogenies. Breakdown of numbers of sequences by (*A*) virus family and (*B*) virus host.

the cellular enzyme zinc finger antiviral protein (ZAP) restricts HIV genomes bearing RNA with multiple CGs (Takata et al. 2017), and we and others have shown strong selection against introduction of CG in HIV and RNA viruses in general (Burns et al. 2009; Atkinson et al. 2014; Stern et al. 2017; Theys et al. 2018; Caudill et al. 2020).

We thus set out to test if there are additional shared genomic and evolutionary features in RNA viruses and compiled a large data set of sequences from pathogenic single-stranded RNA viruses from Baltimore classes IV and V (+ssRNA and −ssRNA viruses, respectively). We focus on these classes in order to 1) avoid the confounding effects of double-stranded viruses such as stableness of double-stranded DNA/RNA, and 2) avoid reverse-transcribing viruses, whose replication cycle is unique compared with other RNA viruses, and includes a DNA stage. One of the key challenges in our study was to disentangle the roles of mutation and selection. Any extant sequence is a product of evolution from an ancestral sequence, and this process includes the action of both mutation and natural selection, occurring repetitively. Indeed, in the examples above we see that either increased introductions of mutations (via A3 enzymes or ADAR) or selection (mediated by ZAP restriction of CG-rich sequences) may both lead to unique genomic signatures. To disentangle the effects of mutation versus selection, we harness the notion of incomplete purifying selection operating on viral genomes, whereby selection is relaxed at the tips of phylogeny (Fitch et al. 1997; Pybus et al. 2007; Strelkowa and Lassig 2012; Gire et al. 2014). By contrasting between rates of substitutions at internal versus external branches of phylogenies we were able to test for the presence of mutational biases (i.e., mutations that are biased toward specific nucleotides) or for selection for specific types of mutations. Overall, our results suggest a consistent selective advantage for the abundance of the A nucleotide across almost all vertebrate RNA viral genomes.

Results

Compilation of the PhyVirus Data Set

We first generated an extensive data set of ∼65,900 full coding sequences from pathogenic single-stranded RNA viruses, which we name PhyVirus (fig. 1; supplementary table S1, Supplementary Material online). Hosts include a wide array of animals, spanning from arachnids, to birds, to fish, to mammals. As expected, the data set contained a disproportionate number of human viruses; yet reassuringly, hundreds to thousands of sequences were available from other phylogenetic clades. We implemented an automated process to generate multiple sequence alignments and their corresponding phylogenetic trees (see Materials and Methods). We focused only on alignments of coding sequences (rather than longer genomic alignments) so as to mitigate as much as possible the effects of recombination. We further iteratively ensured that phylogenies are limited to sequences with a high degree of homology, by focusing only on phylogenies where any given branch length is smaller than 0.5 substitutions/site (see Materials and Methods). Finally, we also ensured that phylogenies were not dramatically affected by mutational saturation (supplementary fig. S1, Supplementary Material online). We have made the PhyVirus data set available online at https://www.sternadi.com/phyvirus (last accessed October 6, 2020), where all alignments, phylogenies, as well as metadata files are accessible to the wide public.

Nucleotide Composition

We first calculated the fraction of A, C, G, and U in the *coding* sequences of all viral families. To our surprise we found an over-representation of A across literally all viral families, accompanied by a strong diminution of C (fig. 2A). The fraction of A ranged from ∼28% to ∼40% in most sequences, reaching a high of 49% in *VPg* sequences of Rhinovirus. The exceptions were the positive single-stranded RNA (+ssRNA)

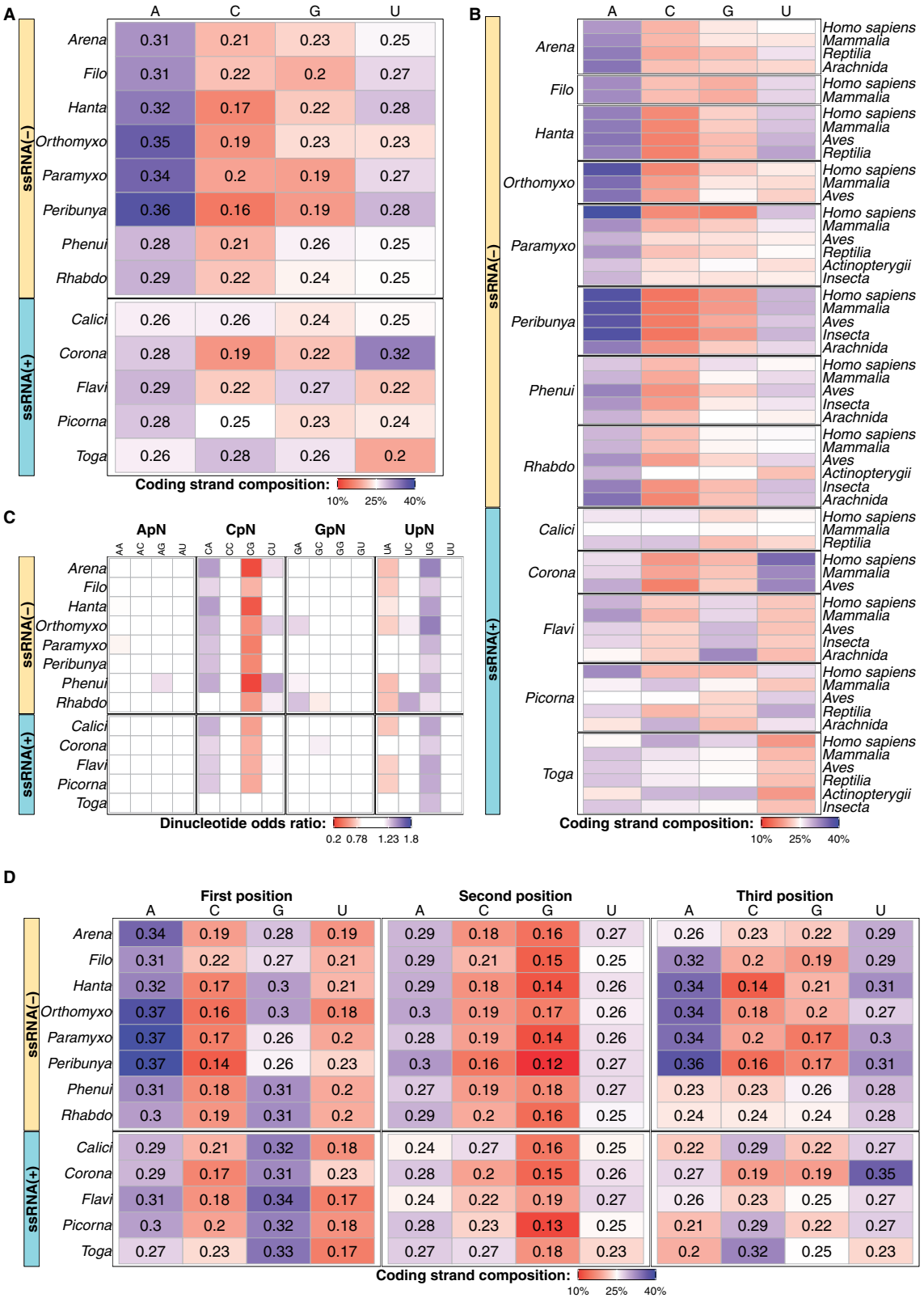

Fig. 2. Nucleotide composition of single-stranded RNA viruses. (*A*) Fraction of each nucleotide across all coding sequences belonging to a particular virus family. In all families a significant departure from a uniform distribution was observed (χ^2, $P < 10^{-3}$). (*B*) Breakdown of nucleotide composition based on phylogenetic assignment of host. (*C*) Odd ratios for dinucleotide composition across viral families. (*D*) Nucleotide composition stratified based on the first, second or third codon positions.

families togaviruses, where more C was observed, and caliciviruses, which were relatively homogenous in nucleotide content. As well, some abundance of U was noted in coronaviruses and some −ssRNA families and of G in flaviviruses. We next examined the nucleotide composition after breaking down by host classification, to test whether composition was dependent on the host. In general, we did *not* notice nucleotide composition dependence on the host, with some minor exceptions, mainly in the Picornaviridae family (fig. 2*B*). Finally, when analyzing coding sequences of double-stranded DNA and RNA viruses, RNA bacteriophages, or coding sequences of hosts, we did not find any consistent preference for A (supplementary fig. S2*A*−*C*, Supplementary Material online), and we note that mixed evidence exists regarding A-richness in retroviruses (van Hemert and Berkhout 1995).

We next went on to examine the nucleotide composition across the three codon positions. This analysis revealed an interesting and consistent pattern: First codon positions were found to be enriched for A and G, second codon positions were found to be enriched for A and U, whereas the third codon positions were enriched for A and U in −ssRNA viruses, and for U only in +ssRNA viruses (fig. 2*D*). Once again this was in stark contrast to high GC content at third codon positions of host coding sequences (Kudla et al. 2006). The pattern at the codon level also led to a particular pattern of amino acid frequencies, with some differences between the host and viral amino acid frequencies observed (supplementary fig. S23, Supplementary Material online).

We went on to test if a similarly consistent pattern is found in noncoding regions. In general, RNA viruses are devoid of noncoding sequences, and thus these sequences are quite short. Our analysis revealed a base composition that was quite different from that in the coding regions, with no consistent enrichment for A or any other nucleotide (supplementary fig. S2*D*, Supplementary Material online). It seemed that a different set of "rules" apply to the noncoding regions, likely driven by the regulatory roles of noncoding RNA in RNA viruses. Most often these sequences are under strong purifying selection to maintain particular RNA structures (Robertson 1979; Desselberger et al. 1980; Le et al. 1992; Thurner et al. 2004), and this most likely leads to a base composition that is specific to every virus and its noncoding region.

Finally, we examined whether we find longer patterns of biased composition, and focused on the frequencies of dinucleotides. As has been noted previously (Karlin et al. 1994; Greenbaum et al. 2009; Cheng et al. 2013; Tulloch et al. 2014), we observed a strong and consistent depletion of CG and to a lesser extent a depletion of UA dinucleotides across all viruses except for togaviruses (fig. 2*C*). Conversely, we saw an enrichment for CA and UG. This may be explained by ADAR editing (UA > UG or reverse complement of UA > CA), although other ADAR editing products were less observed (AG and CU). Alternatively, CA and UG may compensate for the lack of CG and UA, since they are both one transition mutation away (the most frequent mutation that occurs naturally in viruses) from either CG or UA (but see Di Giallonardo et al. 2017). Returning to our observations of A-richness, we observed no enrichment for longer patterns that include A, suggesting that A in itself is the unique factor in the virus coding sequences. Moreover, a phylogenetic analysis of substitution patterns revealed that all three types of to-A substitution (C > A, G > A, and U > A) are high (supplementary fig. S4, Supplementary Material online).

Mutation Bias or Selection?

Our results showed A-richness for almost all coding sequences of ssRNA viruses, yet this pattern was different at third codon positions of +ssRNA viruses. We set out to understand whether the patterns we observed were due to a biased mutation process, or selection. To resolve this enigma, we utilized the concept of incomplete purifying selection (also known as deleterious mutation load), which has been shown to be prevalent in RNA viruses (Fitch et al. 1997; Pybus et al. 2007; Strelkowa and Lassig 2012; Gire et al. 2014). Accordingly, at the external branches of the phylogeny, selection is less stringent, and deleterious mutations may be more prevalent than in internal branches of the phylogeny, where there has been a longer time for purifying selection to exert its effect. Thus, by contrasting the proportion of to-A substitutions at the internal branches (mutation & selection) and the external branches (mutation, less selection), we can tease apart the roles of mutation and selection (Fig. 3A-B). If mutation is the reason for A-richness we expect to see the same ratio of to-A mutation throughout the phylogeny, but if selection is the underlying reason we expect to see higher to-A at the internal branches than the external branches. The analysis we perform is essentially the same as Hudson-Kreitman-Aguadé (HKA) methods and its derivatives, that contrast between the rate of preferred/non-preferred mutations that segregate in a population, and preferred/non-preferred mutations that have been fixed in a population (Hudson et al. 1987; Mcdonald and Kreitman 1991; Eyre-Walker 1997). The external branches of the phylogeny are hence analogous to a viral population, where we observe "polymorphisms", whereas substitutions in internal branches reflecting fixation events in a virus "species". We thus search for a significantly higher rate of preferred (to-A)/non-preferred (to-A/C/G) substitutions at internal versus external branches.

The hallmark of incomplete purifying selection is higher dN/dS at external nodes (tips) as compared with internal nodes (Zhang et al. 2005). We thus first tested for the presence of incomplete purifying selection across all our data sets, by contrasting the rate of nonsynonymous to synonymous (dN/dS) substitutions between the internal branches and the external branches (see Materials and Methods). Notably, external branch lengths may dramatically differ, and depend heavily on density of sampling. We thus tested various branch length cutoffs to define inclusion or exclusion of a branch as internal or external. We found a higher dN/dS ratio at the tips in 64% of the alignments in our data set. However only 42% of the data sets displayed significant support using a likelihood ratio test ($P < 0.05$, after false discovery correction; Benjamini and Hochberg 1995) for the two-rate model that allows for different dN/dS ratios at different branches (here, internal vs.

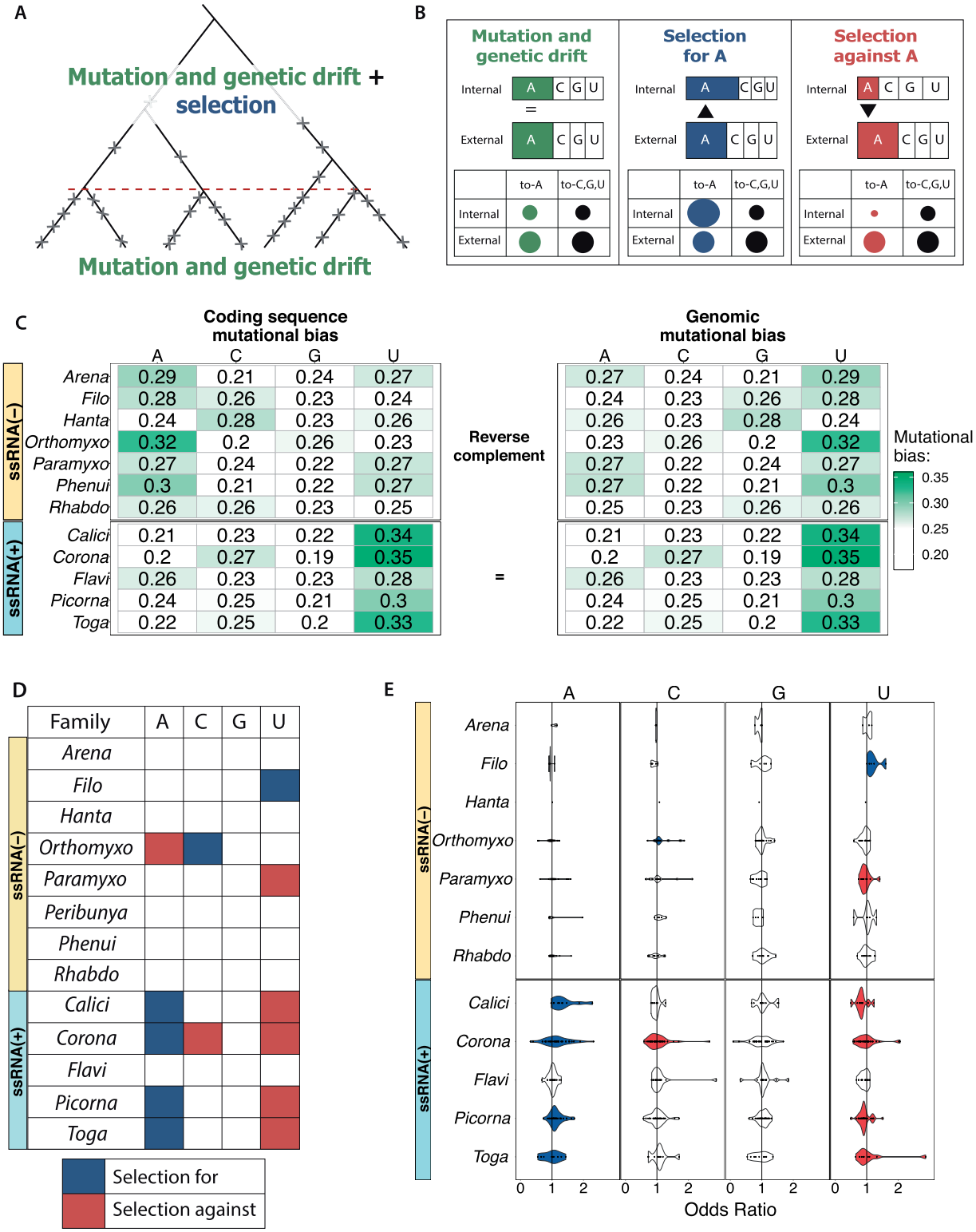

FIG. 3. Use of incomplete purifying selection reveals mutational biases and selection. (*A*) Illustration of incomplete purifying selection: At the external branches of the tree selection is relaxed, and observed substitutions (marked by *x*) are mostly a consequence of mutation and genetic drift. At internal branches substitutions are a composite process of mutation and both selection and genetic drift. (*B*) Illustration of contingency tables used to test for an association between internal/external branches and the type of substitutions observed, with three possible interpretations regarding observed abundance of A. (*C*) Inferred mutational biases at coding sequences, based on substitutions observed at external branches of the phylogeny. Rates to each nucleotide are normalized so as to sum up to one (see Materials and Methods). See supplementary fig. S7, Supplementary Material online for comparison between internal and external biases. (*D*) Families that displayed a significant association between branch location and type of substitution based on a Mantel—Hansel test (see Materials and Methods). (*E*) Violin plots of inferred odds ratios of to-X/to-Y (where X stands for a given nucleotide and Y stands for any other nucleotide apart from X) for each nucleotide at internal versus external branches across all viral families. Panels (*C*—E) show only virus phylogenies where incomplete purifying selection was observed to be significantly supported.

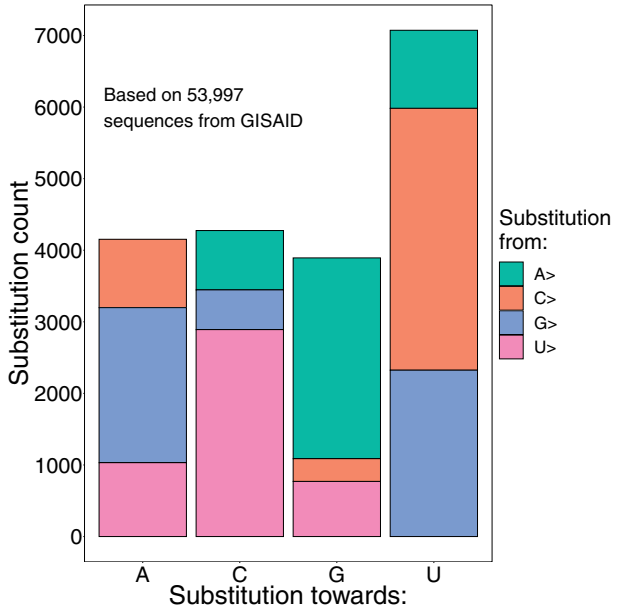

FIG. 4. Mutational bias toward U of SARS-CoV-2 virus. Counts of substitutions from the ancestral Wuhan sequence (see Materials and Methods) toward each available sequence. The colors represent the reference nucleotide.

external branches). In general, data sets that did not display significant support often contained fewer or less divergent sequences, or had longer branch lengths (less dense sampling). We conclude that incomplete purifying selection is probably pervasive, but significant support requires more data and denser sampling. Importantly, the alignments that did pass significance testing were a faithful taxonomical representation of our full data set (supplementary fig. S5, Supplementary Material online) and were not significantly different in terms of estimated age (supplementary fig. S1, Supplementary Material online). We continued our analysis only with the alignments where we observed significant support for incomplete purifying selection.

We next set out to contrast between the rate of to-A substitution at the internal branches and the external branches. Modeling of directional selection for A resulted in incorrect inference (supplementary fig. S6, Supplementary Material online), and to overcome this we performed mutational mapping along the phylogeny (see Materials and Methods). We first inferred the extent to which mutation rates are biased in our data sets by focusing only on substitutions at the external branches, where selection is relaxed. We found that +ssRNA viruses display a strong mutational bias toward U, in strong contrast to the overall A-rich genome they present, but in line with their content at third codon positions (fig. 3C). −ssRNA viruses maintain a mutational bias toward A in their coding sequences (with the exception of hantaviruses), which is effectively a bias toward U on their genomic strand. Interestingly, our data contains two families with an ambisense coding strategy, Arenaviridae and Phenuiviridae, in which proteins are encoded from both the positive and negative strands. By default, these families are characterized as −ssRNA viruses, since only one strand (the negative one) is packed. We tested whether the differences in mutational biases between −ssRNA and +ssRNA viruses hold when separating the coding sequences of the ambisense viruses based on the strand they reside on. We observed that regardless of the strand, all coding sequences of ambisense viruses are A-rich (supplementary fig. S8A, Supplementary Material online). However, the mutational bias was different based on the strand of the coding sequence: Coding sequences on positive strands displayed a mutational bias toward U, whereas coding sequences on negative strands displayed a mutational bias toward A (supplementary fig. S8B, Supplementary Material online).

We were intrigued by the finding of the mutational bias toward U and sought to test if this phenomenon presents itself in the recent COVID-19 epidemic caused by the SARS-CoV-2 virus, a +ssRNA from the Coronaviridae family. Uniquely, SARS-CoV-2 evolution should reflect short-term evolution since the virus has been spreading for merely a few months, and hence observed diversity reflects for the most part mutational biases rather than selection. Extensive sequencing of the virus around the globe allowed us to analyze mutational patterns that show a strong abundance of substitutions toward U (fig. 4) (see also Simmonds 2020), further supported by within host diversity analyses (supplementary fig. S9, Supplementary Material online). We note that sequencing errors (and in particular deamination/oxidation) may lead to an increase in C > U and G > U, however, this should rarely affect the consensus sequence of a virus, which is typically based on dozens to hundreds of sequencing reads (see also supplementary fig. S9, Supplementary Material online; Materials and Methods). All in all, the SARS-CoV-2 sequences support our observation of to-U mutational bias in viral genomes. We discuss the finding of mutations toward U in RNA viruses more in depth below.

We next created contingency tables of inferred to-X/to-Y (where X stands for a given nucleotide and Y stands for any other nucleotide apart from X) substitutions at internal branches/external branches, allowing us to test for an association between branch location and direction of substitution (fig. 3B). Our results showed a highly consistent pattern across almost all +ssRNA viruses, supporting selection for A and selection against U (fig. 3D and E). In −ssRNA viruses the pattern was mixed, and we did not see any consistent signs of selection for or against any nucleotide. To conclude, our analysis shows 1) a consistent mutational bias toward U in genomes of all viruses, which leads to a mutational bias toward A in coding strands of −ssRNA viruses, and 2) selection for A in most +ssRNA viruses, which presumably compensates for the bias toward U caused by the mutational process.

Underlying Reasons for Selection for A

We put forth three possible explanations why A may be selected for in viruses. First, A is a weak RNA binder: It base-pairs only with U, whereas all other three nucleotides may base-pair with the other two via Watson−Crick base-pairing or

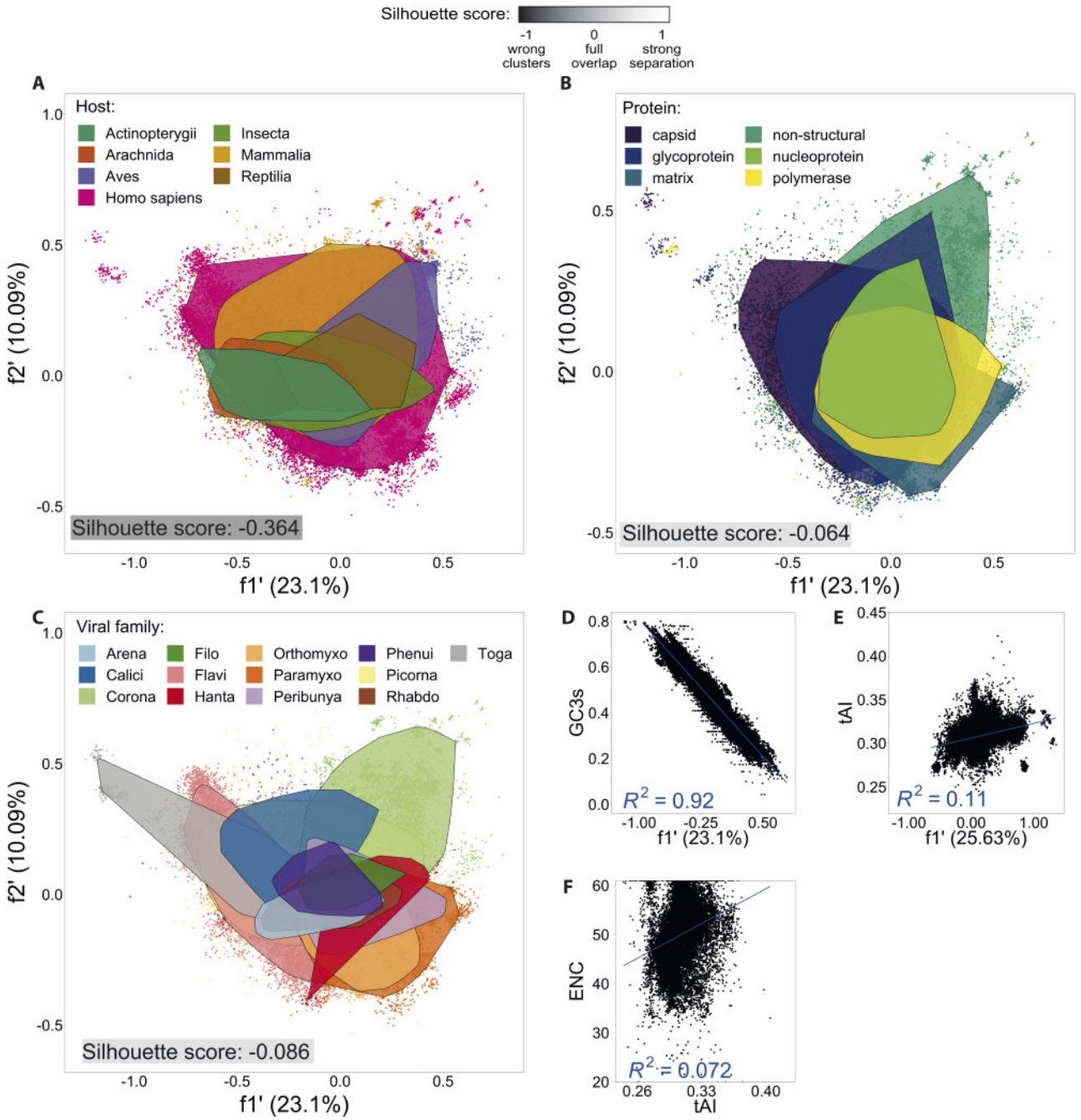

FIG. 5. Codon usage bias analysis of PhyVirus sequences. CA analyses showing the first and second axes explaining the variability in RSCU values across viral coding sequences, with sequences color coded by (*A*) host that the virus infects, (*B*) encoded protein classification (see Materials and Methods), and (*C*) viral families. Each polygon is drawn to include 90% of the points and silhouette scores are given below for each of the clustering variables (see Materials and Methods). (*D*–*F*) Pearson's correlations between: (*D*) the first CA axis and GC_{3s} defined as the frequency of GC content at synonymous third codon positions, (*E*) the first CA axis of human viral sequences and tAI values, and (*F*) tAI and ENC values.

noncanonical G-U pairing. This means that A promotes less RNA secondary structures, as has been previously suggested for HIV, which is also an A-rich virus (Keating et al. 2009; van der Kuyl and Berkhout 2012; van Hemert et al. 2013). Since double-stranded RNA elicits an antiviral response in cells (de Faria et al. 2013), viruses should be under selection to avoid secondary structures. If avoidance of secondary structures is the main reason for selection toward A we would expect to see higher A content at the third codon positions, since this position is generally under weaker protein-associated selection. Although most −ssRNA viruses (except Phenuiviridae and Rhabdoviridae) are A-rich at the third codon positions we note that none of the +ssRNA viruses are A-rich at their third codon positions (fig. 2*D*). It thus seems unlikely that avoidance of RNA secondary structures is the only driving force of selection toward A.

Second, it is possible that codon usage bias and translational optimization have led to the particular sequence

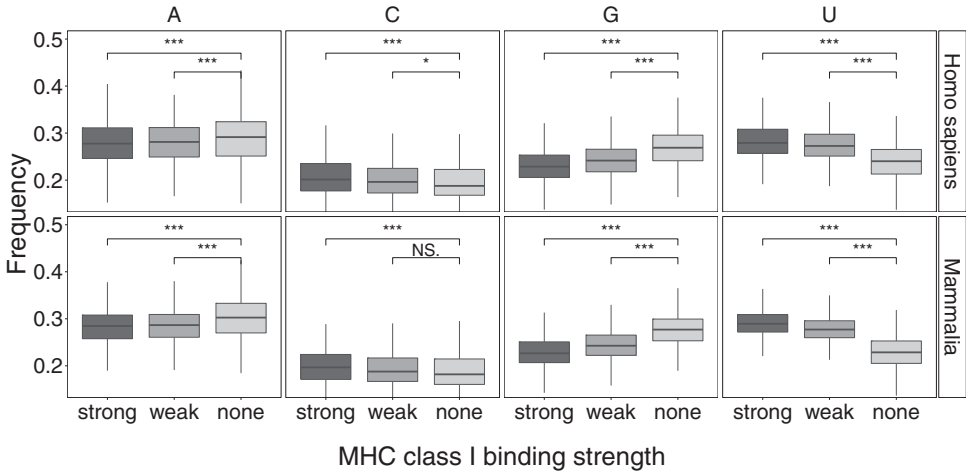

FIG. 6. Nucleotide content composition of inferred MHC peptides in the PhyVirus data set. MHC epitope prediction was run on all translated PhyVirus coding sequences across a variety of MHC alleles, including human, chimp, gorilla, rhesus macaque, bovine, porcine, and mouse alleles (supplementary table S2, Supplementary Material online). Peptides were classified into peptides that would be strongly detected by the MHC system ("strong"), weakly detected ("weak"), or not at all ("none") (see also supplementary fig. S11, Supplementary Material online).

composition observed herein. Accordingly, if codons with more A are associated with more abundant tRNAs, viral genes should be translated more efficiently. Varying codon usage has been reported for many different viruses, yet the underlying reasons for this variation remain obscure (Jenkins and Holmes 2003; Gu et al. 2004; Kryazhimskiy et al. 2008; Wong et al. 2010; Belalov and Lukashev 2013; Cardinale et al. 2013; Tian et al. 2018; Chen et al. 2020). The breadth of the PhyVirus data set allows probing codon bias in depth; we calculated the relative synonymous codon usage (RSCU), the effective number of codons (ENC) (supplementary fig. S10*A*, Supplementary Material online), and the tRNA adaptation index (tAI) (supplementary fig. S10*B*, Supplementary Material online; see Materials and Methods), focusing on human viruses for the latter. Our correspondence analysis (CA) analysis showed that RSCU differences between viral sequences are not attributed to viral host classification, type of protein, or viral family, as reflected in the silhouette values (fig. 5*A*–*C*). Silhouette scores below zero reflect bad clustering, where clusters are embedded within each other, whereas values around zero reflect an almost complete overlap of clusters, suggesting that the clustering variables do not explain differences in RSCU values. When probing which factors are responsible for the first and second components of the CA (23% and 10% of the variability in the data, respectively), we observed a strong correlation of the first axis with the synonymous nucleotide content of the third codon position GC_{3s}, but very weak to no correlations of both axes with tAI (fig. 5*D*), and of tAI with ENC (fig. 5*F*). If codon usage had been driven by selection for enhanced translation of proteins, we would have expected one or more of the following: higher correlation with tAI, low ENC (supplementary fig. S10*A*, Supplementary Material online), unique RSCU profile of genes known to be highly expressed in viruses, such as capsid products (fig. 5*B*) or unique RSCU profile based on host/type of virus (fig. 5*A* and *C*). We do not observe any of these phenomena, whereas the correlation with GC_{3s} suggests that other forces drive the codon composition. We conclude that translational optimization is not likely the driver of the sequence composition observed herein, and that the particular codon composition of a viral sequence is likely a by-product of other factors.

We continue to a third possible explanation for selection toward A: Selection for amino acids encoded by A-rich codons. Although there are many possible reasons leading to selection for specific amino acids, we speculated that the major histocompatibility complex (MHC) class I system may play a role in selection for specific peptides, as it is has been shown to drive the evolution of many vertebrate viruses (Kuntzen et al. 2007; Foll et al. 2014; Carlson et al. 2015; Kløverpris et al. 2015). To test the effect of MHC on composition of viral genomes, we predicted which peptides derived from virus genomes would be weakly or strongly detected by the MHC system. Remarkably, peptides preferentially displayed by MHC systems were found to be encoded by A/G-poor and C/U-rich sequences (fig. 6). In other words, there should be a selective advantage for A/G-rich and C/U poor sequences that would allow escape from the MHC system. Although clearly, this subject merits further in-depth investigation, selection due to the MHC class I system would explain our results, in particular the selection against U and for A.

Discussion

We have found that the vast majority of single strand RNA viruses examined herein have skewed nucleotide composition in their coding sequences, with most viruses bearing A-rich and C-poor sequences. This pattern appears to be quite consistent across hosts ranging from fish, to insects, to mammals, with the caveat that the largest number of sequences in our data was from mammalian viruses (including human viruses). The A-rich pattern disappeared when analyzing viruses of bacteria, viral noncoding regions, or coding sequences of hosts (supplementary fig. S2*B*–*D*, Supplementary Material online).

One of the original goals of this analysis was to test whether we observe the presence of signatures of RNA editing by A3 enzymes or ADAR enzymes, or restriction by cellular enzymes such as ZAP in long-term evolution of viruses. Interestingly, we did not find any such evidence for A3, at least not in a widespread manner, found partial evidence for possible ADAR editing, and although we found a decreased CG presence in all viral families excluding Togaviridae we cannot prove that ZAP is the underlying reason for this phenomenon. In any case, it is not likely that restriction by ZAP would explain the A-richness observed in all single-stranded viruses.

Two nonmutually exclusive hypotheses may be put forth to explain the consistent pattern of A-richness that we observe: There is selection for more A in viral sequences, and/or there is a mutational bias that leads to more A in coding sequences of viruses. In order to tease apart the roles of selection and mutation, we used the notion of incomplete purifying selection, which allows us to separate between recent and nonrecent evolution. Our results revealed that both mutational biases and selection operate in viral genomes. In both +ssRNA and −ssRNA, we observed a mutational bias toward U on the genomic strand of these viruses, which is counteracted by selection against U and toward A in the +ssRNA viruses.

We begin by discussing the mutational biases we observed. In the absence of selection, which is what we measure at external branches of trees, any biased introduction of one nucleotide should lead to its pair being introduced at an equal proportion. For example, if we denote P_A as the probability of erroneously incorporating an A, when this A is reverse-complemented this will lead to $P_A = P_U$ (supplementary fig. S12, Supplementary Material online). Although for some −ssRNA viruses we see a similar mutational bias toward both A and U, this equality in mutational biases does not hold for any of the +ssRNA viruses (fig. 3*C*). Even more intriguingly, we see a mutational bias that differs between negative and positive strands of the ambisense viruses (supplementary fig. S8*B*, Supplementary Material online). Yet if we consider the genomic strand only, this bias collapses to a mutational bias that introduces more U on the genomic strand. This suggests that the mutational process is biased toward one of the strands and acts as a nonsymmetrical process (supplementary fig. S12, Supplementary Material online).

One possible explanation for this directional mutation bias may be genomic damage in the form of spontaneous deamination. Interestingly, this is consistent with a model suggesting that DNA damage is a major source of replication errors in humans (Gao et al. 2019). For single-stranded RNA viruses, it is possible that the genomic damage will affect packaged genomes more than strands replicated within the cells, leading to this nonsymmetrical bias. Another explanation for this mutational bias is host mediated enzymatic deamination of virus genomes that are packed as virions.

We note that the genomic mutational bias toward U we find in viruses falls in line with previously published work that shows that mutation is universally biased toward AT in several species, including human and bacteria (Hershberg and Petrov 2010; Lynch 2010). This suggests that there may be a commonality in the unknown mechanism that creates this bias. Furthermore, in double-stranded genomes it is nearly impossible to determine the specific mutation that causes the bias observed (to-A, to-T or both) due to the complementary nature of the genome. If the mechanism generating biases is the same in viruses and in cells, based on our analysis we speculate that the bias is toward U mutations rather than toward A.

Finally, we turn to examine the underlying reasons for selection toward A. We have proposed three explanations: First, A is a weak RNA binder thus selection to A will promote less RNA secondary structures and will aid viruses in avoiding the host defense mechanisms. Second, translational selection promotes specific codons and causes bias in the nucleotide content and third, there may be selection for amino acids encoded by codons with A. At this stage, our analyses suggest that avoidance of secondary structure or translational selection are most probably not the sole underlying cause for selection toward A, and we show tentative evidence suggesting that the MHC class I system may drive selection for codons with elevated A.

In line with the above, we noted throughout our analysis that flaviviruses were outliers; their sequences were A and G rich, and they are the only +ssRNA viral family where we did not infer selection for A and selection against U (fig. 3*D* and *E*). When probing the sequences in this family, we noted that the majority of sequences were of vector-borne viruses (e.g., dengue virus, Zika virus, and West Nile virus). In general, vector-borne viruses were rare in other virus families, suggesting that our observations regarding selection for A do not hold for vector-borne viruses. Carefully tying this together with our hypotheses in the previous section, we note that insects lack both an MHC system and an interferon response (Flajnik and Kasahara 2001; Secombes and Zou 2017) which augments the immune response to dsRNA, and this might be the underlying reason for the lack of selection against U and toward A observed in flaviviruses. Alternatively, other characteristics of the life cycle and replication of flaviviruses may be responsible for the absence of selection we observed in these viruses.

To conclude, we have found similar patterns of coding sequencing composition across a wide variety of RNA viruses. We found that both mutation toward U and selection for A drive these patterns. In general, we show here that probing viral sequences and phylogenies allows a better understanding of mechanisms that shape the evolution of viruses, and in particular, allows insights into possible footprints of host activity, potentially illuminating the interaction between hosts and viruses.

Materials and Methods

PhyVirus Data Set Curation

Sequences for the PhyVirus data set were primarily obtained from NIAID Virus Pathogen Database and Analysis Resource (ViPR) (Pickett et al. 2012) and were augmented by sequences of Influenza from the NIAID Influenza Research Database

(IRD) (Zhang et al. 2017). The sequences were retrieved as single gene/protein (as opposed to genome segments) during January 2019. Host information was retrieved from ViPR and IRD. Notably, around 9,700 sequences lacked host assignment. We manually sampled a few dozen sequences and checked their host assignment in the associated publication. Most were human viruses but other hosts were present as well. We note that this does not affect any of the analyses in this study, which were almost always agnostic with respect to host.

Our data contain multiple nonduplicated sequences from the same viral species in order to build a comprehensive evolutionary and phylogenetic history as much as possible. These features of the data set are summarized in supplementary table S1, Supplementary Material online. We would like to acknowledge the ViralZone resource (https://viralzone.expasy.org/, last accessed October 6, 2020) (Hulo et al. 2011) for providing comprehensive and accessible information about viral families and genomes.

An in-house computational pipeline was used for clustering the PhyVirus data set into multiple sequence alignments and associated phylogenetic trees. We first used MegaBLAST (Morgulis et al. 2008) to create clusters of homologous sequences, by using each sequence as a query against all sequences of the PhyVirus data set as a database, using an e-value of 10^{-13}. We then aligned sequences using PRANK (Löytynoja 2014) with default settings, and reconstructed phylogenies using the maximum-likelihood method-based PhyML (Guindon and Gascuel 2003), with default settings. We next sought to ensure that phylogenies do not contain sequences that are too remote from each other. To this end we implemented an iterative scheme where we "cut" phylogenies into two or more at branches whose length was larger than 0.5. The 0.5 cutoff was chosen based on manual curation and inspection, and allowed us to avoid grouping together very remotely related sequences. The phylogenies were then rooted using midpoint rooting for analyses that required a rooted tree. Finally, clusters with less than ten sequences were omitted from the analysis. This pipeline resulted in 465 alignments from 13 viral families that contain overall 65,951 sequences (fig. 1). For analyses that required codon-based alignments, we performed codon alignment using PRANK.

Noncoding Sequence Retrieval and Processing

We manually obtained noncoding sequences for a select number of virus families: Picornavirus alignments were obtained from http://www.virology.wisc.edu/acp/Aligns/seq_align.html (last accessed October 6, 2020) (Palmenberg and Sgro 2002; Palmenberg et al. 2009), full genome dengue and ebolavirus sequences identifiers were downloaded from ViPR (Pickett et al. 2012), allowing us to thus obtain complete record and features from NCBI.

dsDNA, dsRNA, and RNA Phages Sequences Retrieval and Processing

dsDNA and dsRNA sequences were obtained from VIPR. Bacteriophage sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/, last accessed October 6, 2020) using the Cystoviridae and Leviviridae taxonomy codes. Coding sequences were extracted using genomic coordinates from NCBI.

Host Nucleotide Content and Amino Acid Statistics

Codon usage data were downloaded from the Codon Usage Database https://www.kazusa.or.jp/codon/ (last accessed October 6, 2020) (Nakamura et al. 2000) and filtered for the host classes in our data, focusing on hosts that have more than 50 coding sequences in the Codon Usage Database, overall we have analyzed 41 mammalian species, ranging from mouse, cow, monkey, and human. Using the codon usage table, we counted the number of nucleotides and amino acids in coding sequences and their frequencies.

PhyVirus Statistics

Nucleotide frequencies were calculated by viral family, by host and by codon position. We first averaged over each alignment, and then averaged by viral family and codon position. This was done to avoid biasing the calculation when a very large number of sequences were available for a particular gene. Dinucleotide odds ratios P_{xy} $(x, y \in \{A, C, G, U\})$ were calculated as described previously (Cheng et al. 2013): $P_{xy} = f_{xy}/f_x f_y$, where f_x and f_y denote the nucleotide frequencies and f_{xy} denotes the frequency of the dinucleotide xy in the sequence.

Substitution Frequency Inference

The BASEML program from the PAML package (Yang 2007) was used to run the unrestricted nonreversible (UNR) and general time reversible (GTR) models in order to infer the frequencies of substitution between all pairs of nucleotides. Since the UNR model requires a rooted tree, we implemented midpoint rooting on all of the phylogenies. About half of the data sets displayed slight significant support for the UNR over the GTR model, yet results of UNR and GTR were very consistent with respect to the frequencies of substitution (data now shown).

Directional Selection

We applied a mutation-selection model of direction selection that we have previously developed (Stern et al. 2017) to test for selection for a specific nucleotide. Briefly, the model allows an increase in the substitution rate at a proportion (P) of sites by rescaling rows and columns of the substitution matrix going to and from the selected nucleotide. The model further accounts for incomplete purifying selection at the external branches of the tree by rescaling the among-site variation distribution (see Stern et al. 2017 for more details). Notably, the original directional model was run on a phylogeny where the root sequence was known. The model was here modified, and assumed a stationary distribution at the root. As commonly practiced, this distribution was inferred based on the distribution of nucleotides at the leaves. Moreover, the original model was agnostic regarding which nucleotide is under selection, and hence assumed that all four nucleotides may be under selection with a probability of $P/4$. We here changed the model to allow for selection for only selected nucleotide

(A, C, G, or U) with a probability of P. The null model ($P = 0$) allowed the use of a likelihood ratio test to assess for significant selection toward a specific nucleotide. Results of this analysis revealed supposedly pervasive selection toward all four nucleotides, and we concluded this is erroneous inference that is due to problematic assumptions of the model that assume a same set of substitution rate parameters across all sites (supplementary fig. S6, Supplementary Material online).

Incomplete Purifying Selection

In order to assess for incomplete purifying selection we used the branch-site model of the CODEML program from the PAML package (Yang et al. 2000). As described previously (Pybus et al. 2007), we compared between a one-ratio model which assumes one ω value for all branches of the phylogeny and a two-ratio model that assumes one value of ω_e for the external branches and another value of ω_i for internal branches. The underlying assumption is that relaxed selection at external branches will lead to an increase in ω at external branches, that is, $\omega_i < \omega_e$.

Since tip lengths vary across the data sets and some external branches are very long (suggesting purifying selection may exert its effect on them), we used different branch length cutoffs to determine which external branches are treated as external branches and which external branches are categorized as internal branches when attempting to detect incomplete purifying selection. The cutoffs that we used were 0.05, 0.04, 0.03, 0.02, 0.01, 0.005, 0.001, and no cutoff (i.e., the classic definition of external branch: A branch that had no progeny branches). To determine which data sets display significant support for incomplete purifying selection we first performed a likelihood ratio test between the one-ratio null model and the two-ratio alternative model. All P-values were corrected for multiple testing using false discovery rate (FDR) (Benjamini and Hochberg 1995). Only data sets showing significant support under one of the cutoffs described above, and where $\omega_i < \omega_e$, were considered as showing evidence of incomplete purifying selection.

Mutational Bias

Ancestral sequence reconstruction was performed using the FastML program under a Jukes and Cantor model for nucleotides and applying joint reconstruction of characters across the phylogenies (Ashkenazy et al. 2012). This enabled us to map the different substitutions across each phylogenetic tree. We then classified substitutions as external or internal, based on the maximal cutoff value that allowed for detection of incomplete purifying selection. The mutational bias of each nucleotide was calculated based on external substitutions only, by calculating the fraction of substitutions toward x ($x \in \{A, C, G, U\}$) divided by the fraction of ancestral nucleotides that are not x. For convenience, the biases were then normalized so they sum up to one.

Calculation of Selection for or Against a Specific Nucleotide

To determine if there is selection for or against a specific nucleotide we constructed a 2×2 contingency table for each viral family, with the type of mutation (e.g., to-A and to-C/G/U) at the columns, and type of branch (external/internal) at the rows. Cells then included the number of substitutions for each intersection of categories. We then used the Mantel–Haenszel (MH) test to test for an association between branch type and substitution type. We performed the MH test for each viral family and for each of the four nucleotides, and corrected for multiple testing using FDR (Benjamini and Hochberg 1995).

Between and Within Host Mutation Counts of the SARS-CoV-2 Virus

For the between host analysis SARS-CoV-2 a multiple sequence alignment containing 53,997 sequences was downloaded from GISAID (https://www.gisaid.org/, last accessed October 6, 2020) on July 7, 2020. For each sequence we counted the numbers of each mutation type relatively to the EPI_ISL_402125 sequence (NCBI accession number: MN908947) from Wuhan, China. We have also reconstructed the most recent ancestor (MRCA) of the SARS-CoV-2 human clade using the bat coronavirus RaTG13 sequence (accession number: MN996532.1) as an outgroup, and the MRCA we obtained was identical to the EPI_ISL_402125 sequence. Each mutation was counted only once, under the assumption that shared mutations were due to shared ancestry. We further discarded positions that have been documented as prone to errors based on the following resource: https://github.com/W-L/ProblematicSites_SARS-CoV2 (last accessed October 6, 2020), although we note that retaining or discarding these positions had almost no effect on the results. For the within host analysis we analyzed deep sequencing data of 212 SARS-CoV-2 samples that we have recently sequenced (Miller et al. 2020). To mitigate sequencing errors, we considered only positions with coverage above 1,000 and mutation frequencies above 5%.

Codon Usage Bias

We have calculated several measures of codon usage bias. First, RSCU was calculated for each gene as previously described (Sharp and Li 1986), where each sequence is represented as a 59D vector. We then performed CA to reduce dimensionality and to detect major trends in codon usage variation among the sequences. In order to asses separation of the sequences on the first two CA axes we calculated the silhouette score (Rousseeuw 1987) based on different clustering categories (host classification, protein type for the six main protein types shared among all viral families depicted in fig. 5, and viral family). Next, we calculated the ENC as previously described (Wright 1990), where ENC values range from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid). We also calculated GC_{3s}, which is the frequency of GC content at the synonymous third codon position. RSCU, CA, ENC, and GC_{3s} were calculated using the

codonW software (Peden JF, unpublished, available at http://codonw.sourceforge.net/, last accessed October 6, 2020). Finally, we calculated the tAI using the tAI package (https://github.com/mariodosreis/tai, last accessed October 6, 2020) with genomic tRNA information from Homo sapiens that was obtained from GtRNAdb (Chan and Lowe 2016).

MHC Nucleotide Prediction

To predict peptides that can serve as epitopes for MHC class I recognition we used the NetMHCpan4 program (Jurtz et al. 2017). We ran the program over all PhyVirus sequences using 249 mammalian alleles from the following organisms: Human, chimpanzee, swine, mouse, gorilla, rhesus macaque, and bovine (supplementary table S2, Supplementary Material online). The prediction was performed for peptide lengths of nine with default parameters. We calculated the nucleotide content for strong and weak binding areas and for areas with no binding prediction. In our calculation we first considered only nucleotides that determine the amino acid type unequivocally (for example, for valine we counted G and U only, since the wobble position can be either one of the four nucleotides). A similar analysis was performed considering all three nucleotides that code for these peptides, yielding essentially the same results (supplementary fig. S11, Supplementary Material online). A *t*-test was performed to determine if the nucleotide content was significantly different between the MHC binding strengths. Multiple test correction was performed using FDR.

Data Availability

The PhyVirus data set is available online at https://www.sternadi.com/phyvirus, and includes all alignments, phylogenies, as well as metadata files. Raw sequencing data for the Miller et al. (2020) SARS-CoV-2 samples are available in the NCBI SRA database under BioProject ID PRJNA647529.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Rasmus Nielsen, Eran Bacharach, Pleuni Pennings, and Tzachi Hagai for stimulating discussions and comments on the manuscript.

This work was supported in part by a fellowship to TK from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, and funding to AS from the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics and from the Israeli Science Foundation (1333/16).

References

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40(W1):W580–W584.

Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P. 2014. 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.* 42(7):4527–4545.

Belalov IS, Lukashev AN. 2013. Causes and implications of codon usage bias in RNA viruses. *PLoS One* 8(2):e56642.

Belshaw R, Gardner A, Rambaut A, Pybus OG. 2008. Pacing a small cage: mutation and RNA viruses. *Trends Ecol Evol.* 23(4):188–193.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300.

Bishop KN, Holmes RK, Sheehy AM, Malim MH. 2004. APOBEC-mediated editing of viral RNA. *Science* 305(5684):645.

Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci U S A.* 89(4):1358–1362.

Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. *J Virol.* 83(19):9957–9969.

Cardinale DJ, DeRosa K, Duffy S. 2013. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. *Viruses* 5(1):162–181.

Carlson JM, Le AQ, Shahid A, Brumme ZL. 2015. HIV-1 adaptation to HLA: a window into virus-host immune interactions. *Trends Microbiol.* 23(4):212–224.

Caudill VR, Qin S, Winstead R, Kaur J, Tisthammer K, Pineda EG, Solis C, Cobey S, Bedford T, Carja O, et al. 2020. CpG-creating mutations are costly in many human viruses. *Evol Ecol.* 34(3):339–359.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* 44(D1):D184–D189.

Chen F, Wu P, Deng S, Zhang H, Hou Y, Hu Z, Zhang J, Chen X, Yang J-R. 2020. Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat Ecol Evol.* 4(4):589–600.

Cheng X, Virk N, Chen W, Ji S, Ji S, Sun Y, Wu X. 2013. CpG usage in RNA viruses: data and hypotheses. *PLoS One* 8(9):e74109.

Cuevas JM, Geller R, Garijo R, Lopez-Aldeguer J, Sanjuan R. 2015. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* 13(9):e1002251.

de Faria IJ, Olmo RP, Silva EG, Marques JT. 2013. dsRNA sensing during viral infection: lessons from plants, worms, insects, and mammals. *J Interferon Cytokine Res.* 33(5):239–253.

Delviks-Frankenberry KA, Nikolaitchik OA, Burdick RC, Gorelick RJ, Keele BF, Hu WS, Pathak VK. 2016. Minimal contribution of APOBEC3-induced G-to-A hypermutation to HIV-1 recombination and genetic variation. *Plos Pathog.* 12(5):e1005646.

Desselberger U, Racaniello VR, Zazra JJ, Palese P. 1980. The 3′ and 5′-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* 8(3):315–328.

Di Giallonardo F, Schlub TE, Shi M, Holmes EC. 2017. Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *J Virol.* 91(8):e02381.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet.* 9(4):267–276.

Eyre-Walker A. 1997. Differentiating between selection and mutation bias. *Genetics* 147(4):1983–1987.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A.* 94(15):7712–7718.

Flajnik MF, Kasahara M. 2001. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* 15(3):351–362.

Foll M, Poh YP, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, Malaspinas AS, Ewing G, Liu P, Wegmann D, et al. 2014. Influenza virus drug resistance: a time-sampled population genetics perspective. *Plos Genet.* 10(2):e1004185.

Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, Amster G, Przeworski M. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Natl Acad Sci U S A*. 116(19):9491–9500.

Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345(6202):1369–1372.

Greenbaum BD, Rabadan R, Levine AJ. 2009. Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. *PLoS One* 4(6):e5969.

Gu W, Zhou T, Ma J, Sun X, Lu Z. 2004. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. *Virus Res*. 101(2):155–161.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52(5):696–704.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*. 6(9):e1001115.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1):153–159.

Hulo C, De Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res*. 39(Suppl 1):D576–D582.

Jenkins GM, Holmes EC. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res*. 92(1):1–7.

Jern P, Russell RA, Pathak VK, Coffin JM. 2009. Likely role of APOBEC3G-mediated G-to-A mutations in HIV-1 evolution and drug resistance. *Plos Pathog*. 5(4):e1000367.

Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 199(9):3360–3368.

Karlin S, Doerfler W, Cardon LR. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol*. 68(5):2889–2897.

Karlin S, Ladunga I, Blaisdell BE. 1994. Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A*. 91(26):12837–12841.

Keating CP, Hill MK, Hawkes DJ, Smyth RP, Isel C, Le SY, Palmenberg AC, Marshall JA, Marquet R, Nabel GJ, et al. 2009. The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. *Nucleic Acids Res*. 37(3):945–956.

Kløverpris HN, Leslie A, Goulder P. 2015. Role of HLA adaptation in HIV evolution. *Front Immunol*. 6:665.

Kryazhimskiy S, Bazykin GA, Dushoff J. 2008. Natural selection for nucleotide usage at synonymous and nonsynonymous sites in influenza A virus genes. *J Virol*. 82(10):4938–4945.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*. 4(6):e180.

Kuntzen T, Timm J, Berical A, Lewis-Ximenez LL, Jones A, Nolan B, Schulze zur Wiesch J, Li B, Schneidewind A, Kim AY, et al. 2007. Viral sequence evolution in acute Hepatitis C virus infection. *J Virol*. 81(21):11658–11668.

Le SY, Chen JH, Sonenberg N, Maizel JV. 1992. Conserved tertiary structure elements in the 5′ untranslated region of human enteroviruses and rhinoviruses. *Virology* 191(2):858–866.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple sequence alignment methods. Totowa (NJ): Humana Press. p. 155—170.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acd Sci U S A*. 107(3):961–968.

Mcdonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.

Miller D, Martin MA, Harel N, Tirosh O, Kustin T, , Meir M, Sorek N, Gefen-Halevi S, Amit S, Vorontsov O, et al. Forthcoming 2020. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat Commun*.

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. 2008. Database indexing for production MegaBLAST searches. *Bioinformatics* 24(16):1757–1764.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 28(1):292–292.

Palmenberg AC, Sgro J-Y. 2002. Alignments and comparative profiles of picornavirus genera. In: Selmer BL, Wimmer E, editors. Molecular biology of picornavirus. Washington: ASM Press. p. 149—155.

Palmenberg AC, Spiro D, Kuzmickas R, Wang S, Djikeng A, Rathe JA, Fraser-Liggett CM, Liggett SB. 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science* 324(5923):55–59.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, et al. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res*. 40(D1):D593–D598.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*. 10(8):540–550.

Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol Biol Evol*. 24(3):845–852.

Robertson JS. 1979. 5′ and 3′ terminal nucleotide sequences of the RNA genome segments of influenza virus. *Nucl Acids Res*. 6(12):3745–3757.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 20:53–65.

Sadler HA, Stenglein MD, Harris RS, Mansky LM. 2010. APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. *J Virol*. 84(14):7396–7404.

Samuel CE. 2012. ADARs: viruses and innate immunity. *Curr Top Microbiol Immunol*. 353:163–195.

Secombes CJ, Zou J. 2017. Evolution of interferons and interferon receptors. *Front Immunol*. 8:

Sharp PM, Li WH. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucl Acids Res*. 14(19):7737–7749.

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(3):e00408.

Stern A, Yeh MT, Zinger T, Smith M, Wright C, Ling G, Nielsen R, Macadam A, Andino R. 2017. The evolutionary pathway to virulence of an RNA virus. *Cell* 169(1):35–46.e19.

Strelkowa N, Lassig M. 2012. Clonal interference in the evolution of influenza. *Genetics* 192(2):671–682.

Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* 550(7674):124–127.

Theys K, Feder A, Gelbart M, Hartl M, Stern A, Pennings PS. 2018. Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. *PLoS Genet*. 14(6):e1007420.

Thurner C, Witwer C, Hofacker IL, Stadler PF. 2004. Conserved RNA secondary structures in Flaviviridae genomes. *J Gen Virol*. 85(5):1113–1124.

Tian L, Shen X, Murphy RW, Shen Y. 2018. The adaptation of codon usage of +ssRNA viruses to their hosts. *Infect Genet Evol*. 63:175–179.

Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *Elife* 3:e04531.

van der Kuyl AC, Berkhout B. 2012. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology* 9(1):92.

van Hemert FJ, Berkhout B. 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. *J Mol Evol*. 41(2):132–140.

van Hemert FJ, van der Kuyl AC, Berkhout B. 2013. The A-nucleotide preference of HIV-1 in the context of its structured RNA genome. *RNA Biol*. 10(2):211–215.

Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM. 2010. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evol Biol*. 10(1):253.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* 87(1):23–29.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Yang ZH, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22(12):2472–2479.

Zhang Y, Aevermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, He S, Kumar S, Larsen CN, Lee AJ, et al. 2017. Influenza research database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res*. 45(D1):D466–D474.