scientific reports

OPEN

Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design

Jacob Kames[1], David D. Holcomb[1], Ofer Kimchi[2], Michael DiCuccio[3], Nobuko Hamasaki-Katagiri[1], Tony Wang[4], Anton A. Komar[5], Aikaterini Alexaki[1]✉ & Chava Kimchi-Sarfaty[1]✉

As the SARS-CoV-2 pandemic is rapidly progressing, the need for the development of an effective vaccine is critical. A promising approach for vaccine development is to generate, through codon pair deoptimization, an attenuated virus. This approach carries the advantage that it only requires limited knowledge specific to the virus in question, other than its genome sequence. Therefore, it is well suited for emerging viruses, for which we may not have extensive data. We performed comprehensive in silico analyses of several features of SARS-CoV-2 genomic sequence (e.g., codon usage, codon pair usage, dinucleotide/junction dinucleotide usage, RNA structure around the frameshift region) in comparison with other members of the coronaviridae family of viruses, the overall human genome, and the transcriptome of specific human tissues such as lung, which are primarily targeted by the virus. Our analysis identified the spike (S) and nucleocapsid (N) proteins as promising targets for deoptimization and suggests a roadmap for SARS-CoV-2 vaccine development, which can be generalizable to other viruses.

The recent emergence of the 2019 novel coronavirus (SARS-CoV-2) has gained worldwide attention and sparked an international effort to develop treatments and vaccines. As of June 9, 2020 there have been 6,931,000 confirmed cases and 400,857 deaths from COVID-19 worldwide[1]. Given the urgency to combat this pandemic, multiple efforts to develop effective vaccines are underway. A relatively recent approach for vaccine development, first proposed by Coleman et al. in 2008 for the attenuation of poliovirus[2], has been used for the attenuation of dozens of viruses, and more recently for bacteria[3]. This approach accomplishes viral attenuation through codon pair deoptimization and appears to be promising for vaccine development, particularly against emerging viruses, as it does not require extensive virus-specific knowledge. It does, however, require knowledge of the viral genome sequence and extensive characterization of its codon and codon pair usage attributes.

Codon usage is biased across all domains of life, i.e., synonymous codons occur at different frequencies in different organisms[4,5]. It is thought that preferred codons correspond to more abundant tRNAs, and therefore, are translated more efficiently[6]. Similarly, there is bias in codon pair usage, with certain codon pairs occurring at a much different frequency than would be expected based on the codon usage[5]. Codon pair usage also appears to affect translation efficiency[6], although the mechanism is not entirely clear, and it has been argued that dinucleotide usage, particularly CpG dinucleotides, may be the driving force in determining viral sequence fitness, while codon pair bias may be a secondary effect of altered dinucleotide frequency[7]. CpG dinucleotides are known to stimulate immune responses and inhibit virion production through their interaction with toll like receptor-9 (TLR9)[8] and the zinc-finger antiviral protein (ZAP)[9]. For these reasons they are commonly used as vaccine adjuvants for many viruses including coronaviruses[10].

Considering that viruses are obligate intracellular parasites and rely on the host-cell machinery for proper expression of their genes, it is expected that their tropism would be affected to some extent by their codon usage, because having codon usage similar to their host would allow them to replicate faster due to better translation efficiency of their mRNA(s)[11]. However, it is worth noting that viral codon usage often does not closely resemble

[1]Division of Plasma Protein Therapeutics, Office of Tissue and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA. [2]Harvard University School of Engineering and Applied Sciences, Cambridge, MA, USA. [3]National Center of Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. [4]Division of Viral Products, Office of Vaccines Research and Review, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA. [5]Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH, USA. ✉email: Aikaterini.alexaki@fda.hhs.gov; Chava.kimchi-sarfaty@fda.hhs.gov

the codon usage of their hosts[12,13], a phenomenon that is not well understood. In this regard, a thorough characterization of codon, codon pair and dinucleotide usage of SARS-CoV-2 can provide useful information regarding expression potential of the viral genes and the fitness of the virus in its human or other hosts. Furthermore, it has been shown that viral attenuation can be achieved through extensive changes in codon pair usage of viral genes[2]. Since the mechanism of viral attenuation through codon pair deoptimization is not entirely clear, this in-depth analysis is necessary to guide the development of new vaccines.

Coronaviruses (CoVs) are enveloped, positive-stranded RNA viruses with a large genome of about 30 kb encoding multiple proteins[14]. Translation of a positive-stranded RNA from the initial infectious virus particles generates (among other proteins) a virally encoded RNA-dependent RNA polymerase (replicase). This replicase is necessary for viral replication and subsequent generation of viral subgenomic RNAs (sgRNAs), from which the synthesis of structural and accessory proteins occurs[14]. ORF1ab, which encodes the replicase polyprotein (among other proteins) occupies about two thirds of the 5′ prime end of this genome[14,15]. A − 1 programmed ribosomal frameshift (PRF) occurs half-way through ORF1ab, allowing the translation of ORF1b[14]. The efficiency of the frameshift thus modulates the relative ratios of proteins encoded by ORF1b and the upstream ORF1a and is critical for coronavirus propagation. Frameshift efficiency (ranging from 15 to 60%) in − 1 PRFs is commonly regulated by pseudoknotted mRNA structures following the frameshift, and the conservation of a three-stem pseudoknot in coronaviruses has been previously characterized[16]. Following ORF1ab, are the spike (S), ORF3a, envelope (E), membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N) and ORF10 genes. The S protein promotes attachment and fusion to the host cell, during infection[17]. In the case of SARS-CoV-2, S binds to the human angiotensin-converting enzyme 2 (ACE2)[15,18,19]. The E protein is an ion channel and regulates virion assembly[20]. The M protein also participates in virus assembly and in the biosynthesis of new virus particles[21], while the N protein forms the ribonucleoprotein complex with the virus RNA[22] and has several functions, such as enhancing transcription of the viral genome and interacting with the viral membrane protein during virion assembly[23]. Many of the other ORFs have unknown functions or are not well characterized[24], as their presence is not consistent across all coronaviruses.

We have conducted a thorough analysis of the codon, codon pair and dinucleotide usage of the SARS-CoV-2 and have assessed how it relates to other coronaviruses, its hosts, and to the tissues that SARS-CoV-2 has been reported to infect[25–27]. We have taken advantage of our recently published databases, which include genomic codon usage statistics for all species with available sequence data, and transcriptomic codon usage statistics from several human tissues[4,5,28]. We further analyzed each viral gene in terms of its codon characteristics and used an array of codon usage metrics that informed us of the potential of each gene sequence to contribute to the deoptimization of the virus. In the case of ORF1ab, we further examined the structure of the mRNA in the region following the frameshift, finding the SARS-CoV-2 mRNA to exhibit a similar pseudoknotted structure to known coronaviruses. We identified two viral genes, S and N, that represent valuable targets for deoptimization to generate an attenuated virus. In the future we plan to continue this in silico study with an experimental investigation of deoptimized virus and test whether the deoptimized S and/or N proteins are immunogenic, produce epitopes that are neutralizing, and result in antibodies that are escape-resistant. We believe, that our combined analysis can be used as a pipeline to guide codon pair deoptimization for viral attenuation and vaccine development or a posteriori to evaluate the effectiveness of the attenuation of a viral sequence.

Results

SARS-CoV-2 proximity to coronaviruses, host genomes and tissue transcriptomes. Since the end of last year when it first emerged, SARS-CoV-2 has been mutating and spreading around the world. Over 5,000 complete or near-complete SARS-CoV-2 genomes are currently accessible in GenBank, with various mutations. To determine which SARS-CoV-2 sequence was most appropriate to use, we retrieved all the published sequences of the virus available in NCBI's SARS-CoV-2 data hub (5,064 complete SARS-CoV-2 genomes); after excluding incomplete and low-quality sequences and CDSs with insertions or deletions, we calculated the percent difference in codon usage between these and the reference sequence. The average percent difference in codon usage was ~ 0.08%, or ~ 8 codons/10,000, clearly showing that variation in sequences is not significantly affecting overall codon usage. This degree of mutation between strains is corroborated by a recently published study[29], and is encouraging as it suggests that escape mutants are unlikely to develop, even for viral genes that have the highest selection pressure such as the S protein. Furthermore, we examined genetic diversity data from Nextstrain[30], accounting for 4,675 SARS-CoV-2 genomes. From these sequences, there are 293 codon positions (~ 23% of the S gene) with reported Shannon entropies, i.e. with a documented mutation at that position. Among all other genes, there are 2,328 such positions (~ 27% of all non-S genes), indicating a higher percentage of mutated codons outside of the S gene.

A discrepancy between virus and host codon and codon pair usage bias has been observed across a range of viruses[12,13,31,32], therefore we examined whether this was true for SARS-CoV-2 and its current host and to other Coronaviruses. Codon pair data inherently contain the codon usage data and therefore are better suited than codon usage data for this type of comparison. As expected, SARS-CoV-2 codon pair usage closely resembles the codon usage of the coronoviridae family, while it is quite distinct from the codon pair usage of the human genome (Table 1). Bat (*Chiroptera*) and pangolin (*Pholidota*) from which the virus may have been transmitted to humans, as well as dog (*Canis lupus familiaris*) to which the virus is feared may be transmitted next, were included in the analysis. We find that these species have a similar codon usage when compared with human; therefore, viral tropism cannot be inferred based on codon usage data alone (Table 1). Since SARS-CoV-2 infects bronchial epithelial cells and type II pneumocytes and our recent findings show that transcriptomic tissue-specific codon pair usage can vary greatly from genomic codon pair usage[28], we also examined the transcriptomic codon pair usage of the lung and how it compares with the SARS-CoV-2 codon pair usage. Rather surprisingly, the codon

	SARS-CoV-2	All CoV	*H. sapiens* genomic	Lung	Kidney	Small intestine	Pholidota	Chiroptera
All CoV	12.75							
H. sapiens genomic	23.20	20.79						
Lung	26.18	23.52	5.80					
Kidney	26.12	23.43	6.00	1.56				
Small intestine	25.93	23.26	5.54	1.73	1.85			
Pholidota	24.69	22.16	2.76	4.18	4.46	4.09		
Chiroptera	24.00	21.54	1.75	4.70	4.91	4.48	1.76	
Canis lupus familiaris	23.68	21.24	1.28	5.18	5.37	4.92	2.18	1.04

Table 1. Euclidean distances between codon pair usage frequencies. Euclidean distance (scaled/1,000) between codon pair usage frequencies of SARS-CoV-2, Coronaviridae (All CoV), *Homo sapiens* (genomic), lung, kidney (cortex), small intestine (terminal ileum), *Pholidota* (pangolins), *Chiroptera* (bats) and *Canis lupus familiaris*.

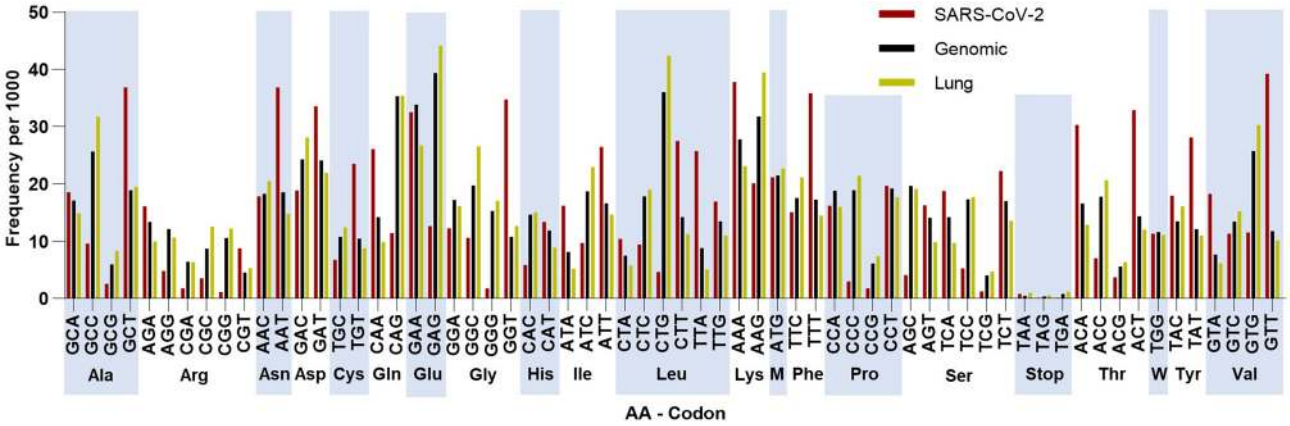

Figure 1. Codon frequencies per 1,000 for SARS-CoV-2 (Red), *Homo sapiens* Genomic (Black) and *Homo sapiens* Lung (Yellow). Codons are grouped by the amino acid they encode (alternating light blue columns, Met (M) and Trp (W) represented as single letter).

pair usage in the lung was more distinct from SARS-CoV-2 codon pair usage than the *Homo sapiens* genomic codon pair usage. The transcriptomic codon pair usage of kidney and small intestine, tissues that are also susceptible to the infection, are similarly distant from SARS-CoV-2 (Table 1). Recently, it was argued that some degree of dissimilation in codon usage between the virus and the host may be beneficial to the virus, as it does not severely impede host gene translation[11].

Codon, codon pair and dinucleotide usage of SARS-CoV-2. To inspect the sequence features of SARS-CoV-2 in more detail, we plotted its codon usage per amino acid and compared it with the human genome and lung transcriptome (Fig. 1). SARS-CoV-2 clearly exhibits a preference in codons ending in T and A (71.7%), which is not observed in the human genome (44.9% ending in T or A) and lung transcriptome (37.6% ending in T or A). Similarly, the kidney and small intestine transcriptome show a preference for codons ending in C and G (62.5% in the kidney and 61.8% in the small intestine, Supplemental Figure 1). The codon pair usage of SARS-CoV-2 was also examined in juxtaposition with the human codon pair usage (Fig. 2A,B). The differences in codon pair usage of the two genomes are highlighted in Fig. 2C.

Since the mechanism of viral attenuation through codon pair deoptimization is not entirely clear, and it has been argued that it is an indirect result of increased CpG content, we further investigated the dinucleotide and junction dinucleotide profile of the SARS-CoV-2 as it compares with *Homo sapiens* genome and lung transcriptome (Fig. 3). Clearly, CpG dinucleotides are avoided in the SARS-CoV-2 genome, and to a lesser extent CC and GG dinucleotides are too. This provides an opportunity to increase immunogenicity of a potential attenuated virus vaccine by increasing its CpG content.

RNA folding. The genome sequence determines not only the amino acid sequence, but also the structure of the mRNA. The mRNA structure following the frameshift site is expected to be especially biologically relevant, as pseudoknots following programmed ribosomal frameshifts have been found to regulate the efficiency of the frameshift[33,34]. We therefore sought to study the similarity of the SARS-CoV-2 mRNA structure compared with the structures of different coronavirus mRNAs in the region following the ORF1ab frameshift.

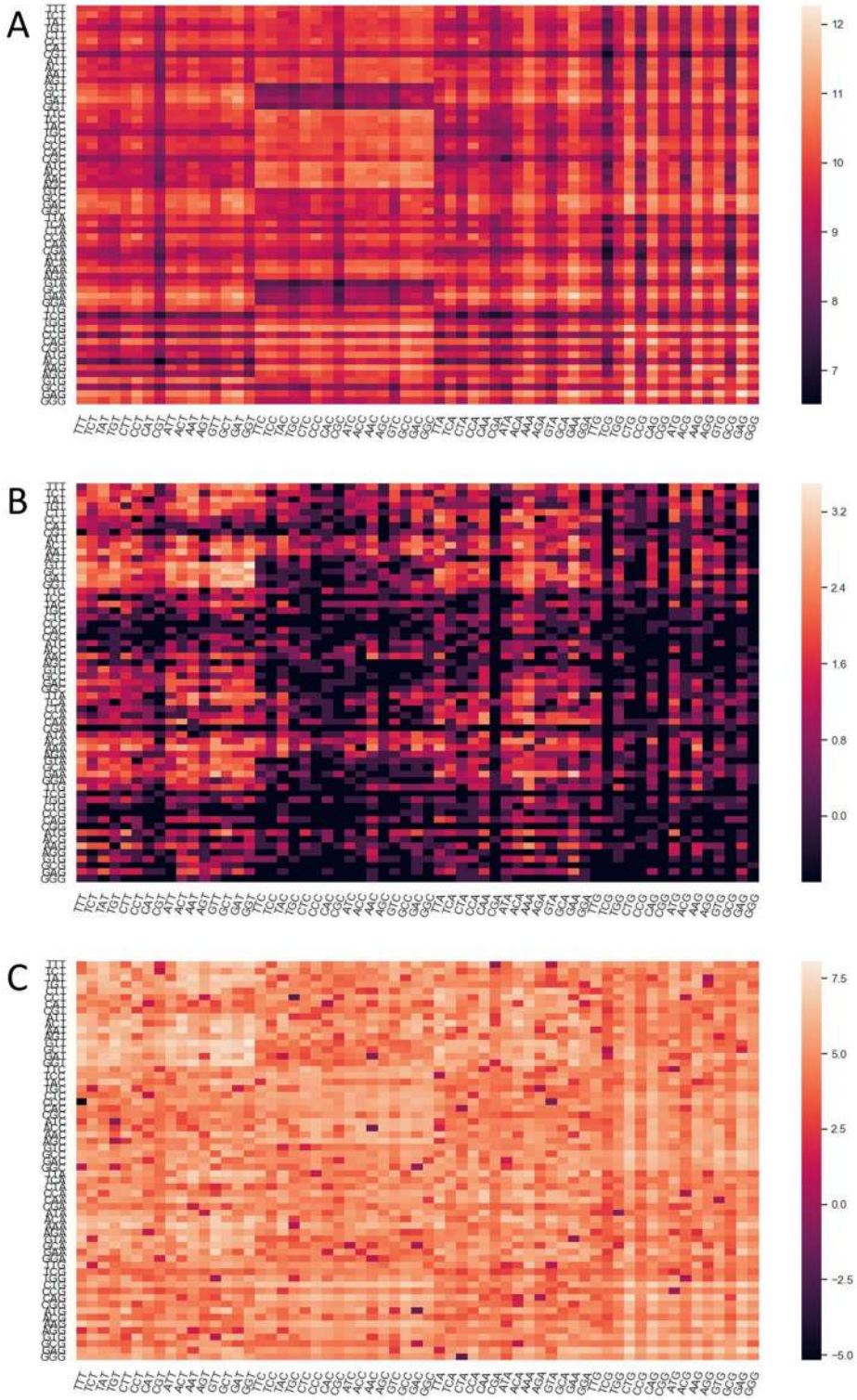

Figure 2. Heat maps of log transformed codon pair frequencies per 1 M for *Homo sapiens* Genomic (**A**), SARS-CoV-2 (**B**) and the absolute value of difference between the two (**C**). Codon pairs increase in frequency from dark to light.

RNA structures were predicted using two distinct secondary structure prediction algorithms, LandscapeFold[36] and NuPack[35,37]. Of the top 10 coronaviruses whose predicted minimum free energy (MFE) structures best aligned to that of SARS-CoV-2, seven matched among the two algorithms, showing a high degree of agreement

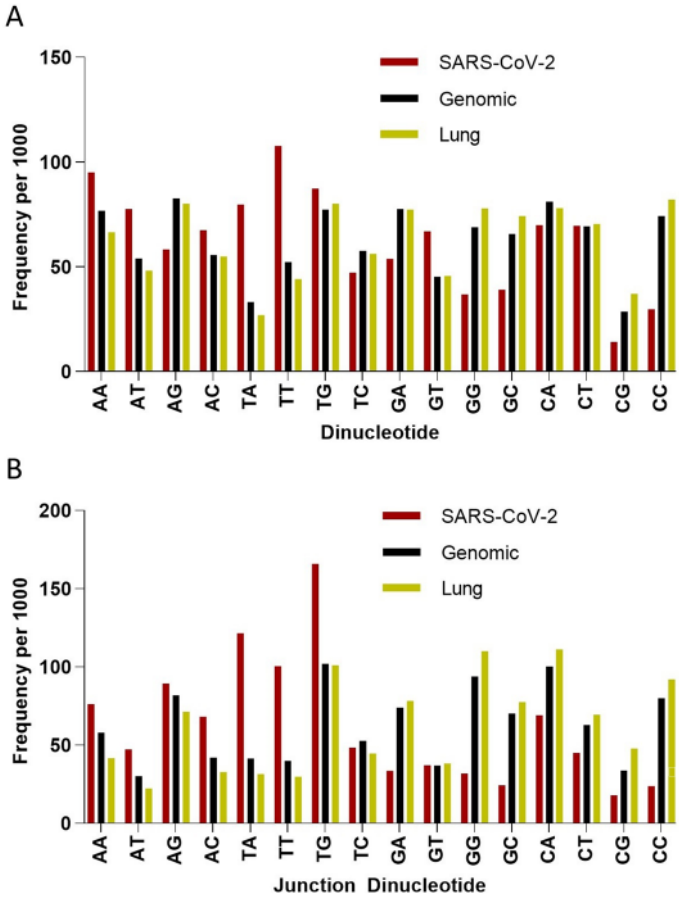

Figure 3. Dinucleotide (**A**) and junction dinucleotide (**B**) frequencies per 1,000 for SARS-CoV-2 (Red), *Homo sapiens* Genomic (Black) and *Homo sapiens* Lung (Yellow).

among the two sets of structure predictions. Those seven consensus best-aligned structures are shown, alongside the novel coronavirus post-frameshift structure, in Fig. 4A–H. The similarity of two of these structures to SARS-CoV-2 can be explained by a high degree of sequence similarity to the SARS-CoV-2 mRNA (a SARS-related coronavirus and a bat coronavirus, shown in Fig. 4B, C). However, the other five—all belonging to avian coronaviruses, which are part of the group of the so-called gammacoronaviruses, causing highly contagious diseases of chickens, turkey and other birds—were not in the top 10 sequences most closely aligned to the SARS-CoV-2 mRNA on the basis of sequence. It should be noted that, of the 5,064 SARS-CoV-2 sequences analyzed, 3,978 had a complete ORF1ab with the exact 'UUUAAAC' frameshift sequence in the annotated position. Of these, 3,951 share the same sequence in the 100 nts downstream of the signal, indicating a high degree of conservation in this region.

Finally, we used LandscapeFold[36] to study the RNA folding beyond the MFE structures. We find that even those coronaviruses whose MFE structure does not contain a pseudoknot will fold into a pseudoknot in a relatively high fraction of cases, and that most coronaviruses have a relatively high probability of the initial stem following the frameshift folding into part of a 3-stem pseudoknot like the one exhibited by the SARS-CoV-2 MFE structure (Fig. 4I).

Viral gene codon usage properties. We next sought to examine each viral gene separately in terms of their codon and codon pair usage. Relative synonymous codon usage (RSCU) and codon pair score (CPS) are commonly used metrics to describe the codon and codon pair usage bias, respectively. RSCU expresses the observed over expected synonymous codon usage ratio, while CPS is the natural log of the observed over expected synonymous codon pair ratio using observed individual codon usage[2,38]. In our analyses, RSCU and CPS are derived from human genomic codon and codon pair usage frequencies. For ease of comparison, we used ln(RSCU) to measure the codon usage bias. The average CPS across a gene is referred to as codon pair bias (CPB) of the gene[2]. The average ln(RSCU) and CPB of each viral gene was calculated and compared with host genes average ln(RSCU) and CPB (Fig. 5). The average RSCU, ln(RSCU) and CPB of each viral gene appear in Table 2. ORF10 was strikingly the least similar gene to the human genome in terms of both its codon and codon pair usage, followed by the E gene. These genes provide little opportunity for deoptimization, since their sequence is already far from optimal. On the other hand, genes S and N are more similar to human in terms

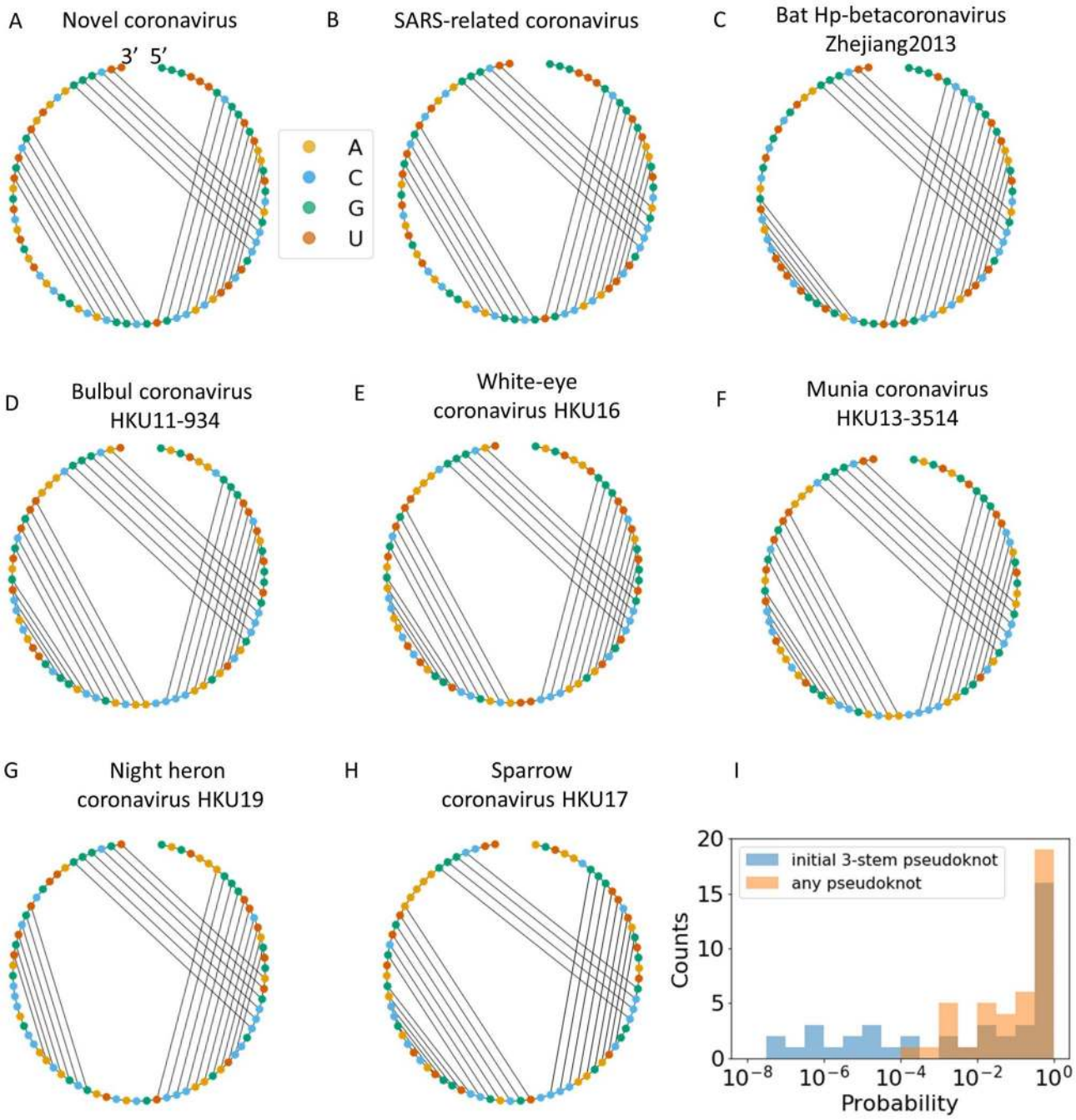

Figure 4. (**A**) The predicted minimum free energy (MFE) secondary structure of the novel coronavirus RNA in the 75 nts following the frameshift. All MFE structures displayed are those predicted by LandscapeFold; results discussed were found to be insensitive to prediction algorithm by comparison to NuPack. (**B**,**C**) Known coronaviruses with high degree of sequence and structure similarity to the novel coronavirus. (**D**–**H**) Known coronaviruses with a high degree of structure similarity to the novel coronavirus, but less sequence similarity. See main text for further discussion. (**I**) In addition to examining the predicted MFE structures, we considered the full free-energy landscapes. The probability of each coronavirus to form a pseudoknot in the 75 nts following the frameshift (orange), and the probability of the first stem to be part of a 3-stem pseudoknot (blue), are histogrammed.

of their codon pair usage. To explore further the potential for codon pair deoptimization, we plotted their CPS across their sequence (Supplemental Figure 2 and Fig. 6). As seen in these figures all viral genes use mostly rare codons (ln(RSCU) < 0); however, it is striking that ORF6 and ORF10 use almost exclusively rare codons, while ORF3a and M and ORF10 have some of the lowest ln(RSCU) values. Regarding codon pair usage, S stands out as the gene that uses frequent codon pairs more often (peaks with relatively high CPS scores), while N, ORF6 and ORF7b are genes that do not use very rare codon pairs (CPS values are only moderately negative).

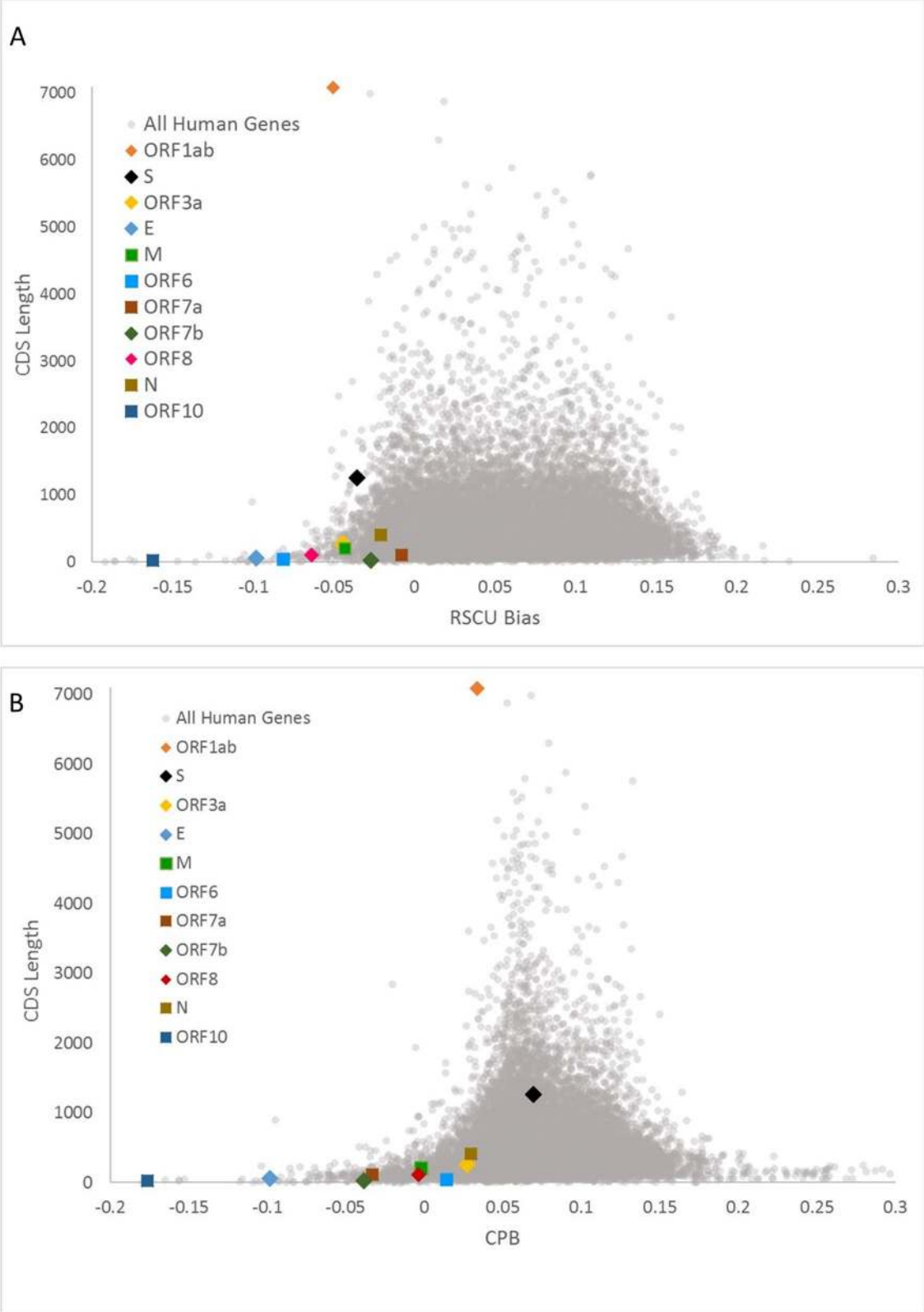

Figure 5. Scatterplots of RSCU bias [average ln(RSCU)] (**A**) and codon pair bias (CPB) (**B**) by CDS length of human and viral genes. Human genes appear as grey dots and viral genes appear with different colored markers.

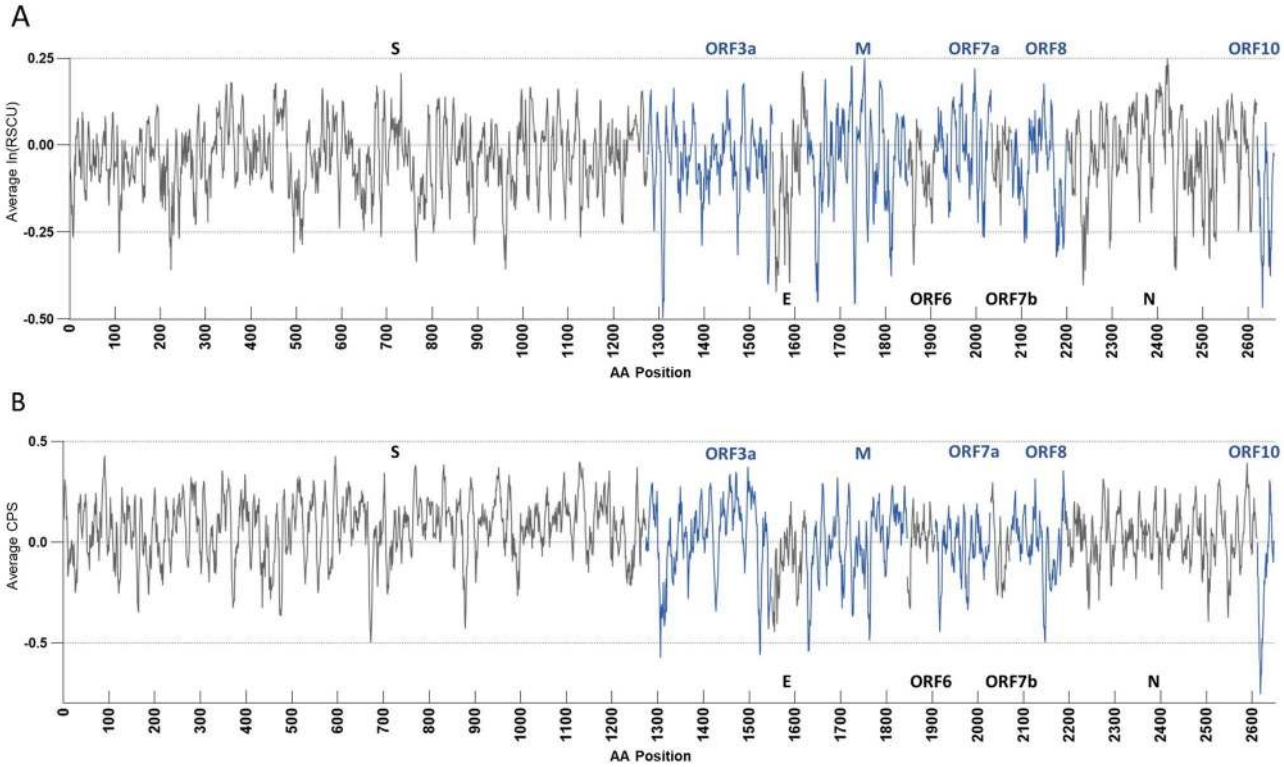

Figure 6. Seven codon sliding window average of ln(RSCU) (**A**) and codon pair score (CPS) (**B**) of structural SARS-CoV-2 genes. Genes are shown in the order they appear in the viral genome, but gaps between open reading frames have been removed. Genes alternate in colors black and blue for clarity, with the gene name in the corresponding color appearing above or below the window. RSCU and CPS are calculated based on *Homo sapiens* genomic codon and codon pair usage.

	ORF1ab	S	ORF3a	E	M	ORF6	ORF7a	ORF7b	ORF8	N	ORF10
Avg RSCU	0.98	1.00	0.99	0.97	1.01	0.95	1.04	1.02	0.98	1.02	0.88
Avg ln(RSCU)	− 0.05	− 0.04	− 0.05	− 0.10	− 0.04	− 0.08	0.00	− 0.03	− 0.07	− 0.02	− 0.17
CPB	0.03	0.07	0.03	− 0.10	0.00	0.02	− 0.03	− 0.04	0.00	0.03	− 0.17

Table 2. Codon and codon pair metrics of SARS-CoV-2 genes. Relative synonymous codon usage (RSCU), ln(RSCU) and codon pair bias (CPB) for 11 viral genes.

Discussion

We performed a comprehensive characterization of the codon, codon pair and dinucleotide usage of the SARS-CoV-2 genome with the intention to identify the best targets for codon pair deoptimization in order to design an attenuated virus for vaccine development. Genes N and S were singled out as the best potential targets for deoptimization, due to the relatively high CPB among the viral genes. Furthermore, they are structural proteins with known functions, which will facilitate subsequent studies.

Currently, most published attempts of viral attenuation through codon pair deoptimization do not discuss the strategy for selecting which genes to deoptimize. Although codon pair deoptimization has been proven successful for viral attenuation[2,39,40], the mechanism is not clear. In addition, it is reasonable to assume that for every successful published deoptimization attempt, there may be several unsuccessful and therefore unpublished ones. Similarly, there have been successful attempts to generate attenuated viruses through codon deoptimization[41–44]. Understanding the mechanism that leads to viral attenuation requires a thorough characterization of the viral sequence and of the consequences of sequence changes. There are several factors that may contribute to the efficacy of deoptimization strategies. In changing the codon pair usage, the dinucleotide frequency and the GC content are altered; mRNA secondary structure and translational kinetics are also perturbed. Further, the CpG content is changing, leading potentially to altered immunogenicity. It is likely that codon pair (or codon) deoptimization leads to reduced expression, either due to changes in transcription, mRNA stability, or translation efficiency[45,46]. Alternatively, it is possible that deoptimization may lead to perturbed cotranslational folding[47], resulting in altered protein conformation. In the case of the S protein, this may lead to decreased binding affinity for the ACE2 protein, thus affecting viral fitness.

A number of parameters were considered in determining which proteins could be targets for codon pair deoptimization. It has been shown that deoptimizing one third or less of the virus is sufficient to attenuate the virus, and more extensive deoptimization may lead to a completely inactive virus[48]. ORF1ab takes up about two thirds of the virus; therefore, its size may make it an unsuitable target. Furthermore, altering its RNA sequence is likely to disturb the pseudoknot that is responsible for frameshifting. Since we have identified the sequence that is responsible for the frameshift, a partial codon pair deoptimization is possible. However, ORF1ab is essential for genome replication, which also does not support its capacity as a codon deoptimization target.

ORF10 has strikingly low CPB and RSCU; given that it is at the very end of the viral genome, there may be a structural reason for its nucleotide sequence. The E gene also has a very low CPB; interestingly, although ORF10 has both positive and negative CPS across its sequence, E has mostly negative CPSs, which make it unsuited for codon pair deoptimization. ORF7a is unusual, as it has the highest RSCU of the viral genes, but a rather low CPB (it uses preferred codons in unusual combinations). Although ORF7a is not the most compelling target for codon pair deoptimization, if a codon deoptimization strategy is attempted, this gene should be considered. It should, however, be noted that since it overlaps for a few nucleotides with ORF7b, any sequence changes should be considered in coordination for both genes.

Our sequence analysis pointed to the S and N genes as potential targets for codon pair deoptimization. The S protein, which binds to the cell-surface receptor and induces virus-cell membrane fusion, has the highest CPB score of all viral proteins, leading to significant flexibility for codon pair deoptimization. It should be noted that although protein S is a surface protein and is expected to be under pressure to avoid the immune system, an examination of the variants that have emerged over the past few months indicates that the S gene does not appear to mutate at a faster rate than other SARS-CoV-2 genes. The N protein, which forms the ribonucleoprotein complex with the virus RNA, is the most conserved and stable protein among the coronavirus structural proteins. It uses mostly codon pairs with intermediate frequency; thus, it could be substantially codon pair deoptimized.

The next step, which is beyond the scope of this in silico study, is to construct the deoptimized virus and test its infectivity, ability to replicate, and whether it retains any pathogenicity. The study should investigate whether deoptimized S and/or N proteins are immunogenic, produce epitopes that are neutralizing, and result in antibodies that are escape-resistant. While testing the fitness of the virus, our strategy of selecting targets would be tested and validated, which could lead to better understanding of the factors that make codon pair deoptimization successful in generating attenuated viruses. While these planned steps are essential in the experimental validation of our strategy, we aimed to provide the scientific community with the framework that would allow anyone to independently proceed towards engineering a codon deoptimized SARS-CoV-2 for vaccine development.

The risk of a new emerging virus is always present, and this has been poignantly highlighted by the current SARS-CoV-2. The current work could be used for the quick generation of a SARS-CoV-2 vaccine but also as a pipeline to facilitate vaccine development when the next virus is presented.

Materials and methods

Sequence accession and codon comparison. The complete reference sequences for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 , accession NC_045512.2) were downloaded from NCBI RefSeq[49] on June 8, 2020. CDS sequences from 5,064 complete SARS-CoV-2 isolates were downloaded using NCBI SARS-CoV-2 data hub on June 8, 2020[50]. Sequences of poor quality or with CDS lengths that did not match those of the reference sequence due to deletion or insertion were removed. To calculate percent difference in codon usage, each CDS was compared at the codon level to that of the reference sequence. Codons containing nucleotides where a base call could not be made ("N") were removed from the calculation. All scripts for this calculation were written in Python 3.7.4.

Comparison of codon and codon pair usage in host species. Codon, codon pair and dinucleotide usage data for *Homo sapiens*, *Canis lupus familiaris*, *Chiroptera* (bats) and *Pholidota* (pangolins) were downloaded from the CoCoPUTs database[5] on March 13, 2020. Likewise, human lung, kidney (cortex) and small intestine (terminal ileum) tissue-specific codon, codon pair and dinucleotide usage data were accessed from the TissueCoCoPUTs database[28] on March 13, 2020. Codon, codon pair and dinucleotide usage data for SARS-CoV-2 was calculated from the reference sequence (accession NC_045512.2) using scripts written in Python 3.7.4. Euclidean distances between codon pair usage frequencies were calculated using the *dist* function from the *stats* package in R 3.6.1.

RSCU, CPS and CPB. RSCU was calculated as defined in Sharp et al.[38] based on *Homo sapiens* genomic codon usage data accessed from the CoCoPUTs database[5] on March 13, 2020. CPSs for all 4,096 codon pairs were calculated as described in Coleman et al.[2] using *Homo sapiens* genomic codon pair usage data accessed from the CoCoPUTs database[5] on March 13, 2020. CPB of a gene is the arithmetic mean of all CPSs throughout the gene, as defined in Coleman et al.[2].

RNA folding. To ensure our results are robust to the prediction algorithm chosen as well as to the size of the window examined, we used two secondary structure prediction algorithms on two window sizes. We used NuPack to predict the minimum free energy (MFE) secondary structure on the 100 nucleotides (nts) following the frameshift[35], and our own recently published free energy landscape enumeration algorithm, LandscapeFold, to examine the full structure landscape of the 75 nts following the frameshift[36]. For the latter, we employed the heuristic that the minimum stem length was set to 4. Aside from this heuristic, the two algorithms differ primarily in the loop entropy calculation, which especially affects the probability of pseudoknot formation.

Sequence alignment was measured using MatLab's Needleman-Wunsch sequence alignment implemented on the 100 nts following the frameshift using default parameters. The parameters employed were the defaults: the NUC44 scoring matrix and a gap penalty of 8 for all gaps.

Structure alignment was measured using a method similar to our previously-studied "per-base topology" score[36]. Taking the dot-bracket representation of each secondary structure, we summed the number of positions containing identical elements. Employing an alignment model allowing for gaps, with a gap penalty and a misalignment penalty of − 1, did not change our results[51].

Data availability

The datasets analyzed during the current study are available in the figshare repository, accessible by the DOIs https://doi.org/10.6084/m9.figshare.12094116 and https://doi.org/10.6084/m9.figshare.12469148.v1.

Received: 31 March 2020; Accepted: 19 August 2020
Published online: 24 September 2020

References

1. Coronavirus disease 2019 (COVID-19) Situation Report–140. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200608-covid-19-sitrep-140.pdf?sfvrsn=2f310900_2 (World Health Organization, Geneva, 2020).
2. Coleman, J. R. *et al.* Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**, 1784–1787 (2008).
3. Coleman, J. R., Papamichail, D., Yano, M., Garcia-Suarez, M. D. M. & Pirofski, L. A. Designed reduction of *Streptococcus pneumoniae* pathogenicity via synthetic changes in virulence factor codon-pair bias. *J. Infect. Dis.* **203**, 1264–1273 (2011).
4. Athey, J. *et al.* A new and updated resource for codon usage tables. *BMC Bioinform.* **18**, 391 (2017).
5. Alexaki, A. *et al.* Codon and codon-pair usage tables (CoCoPUTs): Facilitating genetic variation analyses and recombinant gene design. *J. Mol. Biol.* **431**, 2434–2441 (2019).
6. Komar, A. A. The Yin and Yang of codon usage. *Hum. Mol. Genet.* **25**, R77–R85 (2016).
7. Kunec, D. & Osterrieder, N. Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep.* **14**, 55–67 (2016).
8. Zakhartchouk, A. N. *et al.* Immunogenicity of a receptor-binding domain of SARS coronavirus spike protein in mice: Implications for a subunit vaccine. *Vaccine* **25**, 136–143 (2007).
9. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124–127 (2017).
10. Lan, J. *et al.* Tailoring subunit vaccine immunity with adjuvant combinations and delivery routes using the Middle East respiratory coronavirus (MERS-CoV) receptor-binding domain as an antigen. *PLoS ONE* **9**, e112602 (2014).
11. Chen, F. *et al.* Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat. Ecol. Evol.* **4**, 589–600 (2020).
12. Holcomb, D. D., Alexaki, A., Katneni, U. & Kimchi-Sarfaty, C. The Kazusa codon usage database, CoCoPUTs, and the value of up-to-date codon usage statistics. *Infect. Genet. Evol.* **73**, 266–268 (2019).
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.* **58**, 1–16 (2018).
14. Lim, Y. X., Ng, Y. L., Tam, J. P. & Liu, D. X. Human coronaviruses: A review of virus–host interactions. *Diseases* **4**, 26 (2016).
15. Guo, Y. R. *et al.* The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—An update on the status. *Mil. Med. Res.* **7**, 11 (2020).
16. Plant, E. P. *et al.* A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.* **3**, e172 (2005).
17. Belouzard, S., Millet, J. K., Licitra, B. N. & Whittaker, G. R. Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses* **4**, 1011–1033 (2012).
18. Jia, H. P. *et al.* ACE2 receptor expression and severe acute respiratory syndrome coronavirus infection depend on differentiation of human airway epithelia. *J. Virol.* **79**, 14614–14621 (2005).
19. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
20. Ruch, T. R. & Machamer, C. E. The coronavirus E protein: Assembly and beyond. *Viruses* **4**, 363–382 (2012).
21. Neuman, B. W. *et al.* A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* **174**, 11–22 (2011).
22. Risco, C., Anton, I. M., Enjuanes, L. & Carrascosa, J. L. The transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins. *J. Virol.* **70**, 4773–4777 (1996).
23. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).
24. Fehr, A. R. & Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23 (2015).
25. Zhang, T., Wu, Q. & Zhang, Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol.* **30**, 1346–1351.e2 (2020).
26. Luan, J., Lu, Y., Jin, X. & Zhang, L. Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. *Biochem. Biophys. Res. Commun.* **526**, 165–169 (2020).
27. Tilocca, B. *et al.* Molecular basis of COVID-19 relationships in different species: A one health perspective. *Microbes Infect.* **22**, 218–220 (2020).
28. Kames, J. *et al.* TissueCoCoPUTs: Novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. *J. Mol. Biol.* **432**, 3369–3378 (2020).
29. Lv, L., Li, G., Chen, J., Liang X. & Li, Y. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. Preprint at https://www.biorxiv.org/content/10.1101/2020.02.27.969006v1 (2020).
30. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
31. Zhou, T., Gu, W., Ma, J., Sun, X. & Lu, Z. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. *Biosystems* **81**, 77–86 (2005).
32. van Hemert, F., van de Kuyl, A. C. & Berkhout, B. Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage. *J. Gen. Virol.* **97**, 2608–2619 (2016).
33. Namy, O., Moran, S. J., Stuart, D. I., Gilbert, R. J. & Brierley, I. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* **441**, 244–247 (2006).
34. Baranov, P. V. *et al.* Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology* **332**, 498–510 (2005).
35. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
36. Kimchi, O., Cragnolini, T., Brenner, M. P. & Colwell, L. J. A polymer physics framework for the entropy of arbitrary pseudoknots. *Biophys. J.* **117**, 520–532 (2019).
37. Dirks, R. M. & Pierce, N. A. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.* **25**, 1295–1304 (2004).

38. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
39. Kaplan, B. S. *et al.* Vaccination of pigs with a codon-pair bias de-optimized live attenuated influenza vaccine protects from homologous challenge. *Vaccine* **36**, 1101–1107 (2018).
40. Mueller, S. *et al.* Live attenuated influenza virus vaccines by computer-aided rational design. *Nat. Biotechnol.* **28**, 723–726 (2010).
41. Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S. & Wimmer, E. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* **80**, 9687–9696 (2006).
42. Manokaran, G., McPherson, K. G. & Simmons, C. P. Attenuation of a dengue virus replicon by codon deoptimization of nonstructural genes. *Vaccine* **37**, 2857–2863 (2019).
43. Cai, Y. *et al.* A lassa fever live-attenuated vaccine based on codon deoptimization of the viral glycoprotein gene. *mBio* **11**, e00039–20. https://doi.org/10.1128/mBio.00039-20 (2020).
44. Tsai, Y. H. *et al.* Enterovirus A71 containing codon-deoptimized VP1 and high-fidelity polymerase as next-generation vaccine candidate. *J. Virol.* **93**, e02308–18. https://doi.org/10.1128/JVI.02308-18 (2019).
45. Le Nouen, C., Collins, P. L. & Buchholz, U. J. Attenuation of human respiratory viruses by synonymous genome recoding. *Front. Immunol.* **10**, 1250 (2019).
46. Groenke, N. *et al.* Mechanism of virus attenuation by codon pair deoptimization. *Cell Rep.* **31**, 107586 (2020).
47. Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A. & Clark, P. L. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. USA* **117**, 3528–3534 (2020).
48. Wimmer, E., Mueller, S., Tumpey, T. M. & Taubenberger, J. K. Synthetic viruses: A new opportunity to understand and prevent viral disease. *Nat. Biotechnol.* **27**, 1163–1172 (2009).
49. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
50. Hatcher, E. L. *et al.* Virus Variation Resource—Improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482–D490 (2017).
51. LingPy. A Python library for historical linguistics v. 2.6.5 (2019).

Acknowledgements
O.K. acknowledges funding from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard award number 1764269, and the Harvard Quantitative Biology Initiative. This research was supported by the Intramural Research Program of the National Library of Medicine at the NIH (M.D.).

Author contributions
J.K., D.D.H. and O.K. conducted analysis, verified analytic methods, prepared figures and assisted in preparing the original manuscript. M.D., N.H.-K., T.W. and A.A.K. suggested analyses, conducted critical review of the data and assisted in preparing the original manuscript. A.A. and C.K.-S. conceived the original idea, suggested analyses, conducted critical review of the data and prepared the original manuscript.

Funding
This work was supported by funds from the U.S. Food and Drug Administration CBER Coronavirus (COVID-19) Supplemental Funding, CBER operating funds and in part supported by the National Institutes of Health grant HL151392 (A.A.K.) and the Harvard Quantitative Biology Initiative (O.K.).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-72533-2.

Correspondence and requests for materials should be addressed to A.A. or C.K.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2020