Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning

Liam Brierley*[1], Anna Fowler[1]

[1] Department of Health Data Science, University of Liverpool, Brownlow Street, Liverpool, L69 3GL, UK.
* Correspondence: liam.brierley@liverpool.ac.uk

Abstract

The COVID-19 pandemic has demonstrated the serious potential for novel zoonotic coronaviruses to emerge and cause major outbreaks. The immediate animal origin of the causative virus, SARS-CoV-2, remains unknown, a notoriously challenging task for emerging disease investigations. Coevolution with hosts leads to specific evolutionary signatures within viral genomes that can inform likely animal origins. We obtained a set of 650 spike protein and 511 whole genome nucleotide sequences from 225 and 187 viruses belonging to the family *Coronaviridae*, respectively. We then trained random forest models independently on genome composition biases of spike protein and whole genome sequences, including dinucleotide and codon usage biases in order to predict animal host (of nine possible categories, including human). In hold-one-out cross-validation, predictive accuracy on unseen coronaviruses consistently reached ~73%, indicating evolutionary signal in spike proteins to be just as informative as whole genome sequences. However, different composition biases were informative in each case. Applying optimised random forest models to classify human sequences of MERS-CoV and SARS-CoV revealed evolutionary signatures consistent with their recognised intermediate hosts (camelids, carnivores), while human sequences of SARS-CoV-2 were predicted as having bat hosts (suborder Yinpterochiroptera), supporting bats as the suspected origins of the current pandemic. In addition to phylogeny, variation in genome composition can act as an informative approach to predict emerging virus traits as soon as sequences are available. More widely, this work demonstrates the potential in combining genetic resources with machine learning algorithms to address long-standing challenges in emerging infectious diseases.

Background

The ongoing COVID-19 pandemic remains a significant public health emergency. Since the first identified cases in China in December 2019, this outbreak of respiratory disease has developed into a global crisis, with over 43 million cases worldwide to date (WHO, 2020). The causative virus was termed 'severe acute respiratory syndrome-related coronavirus 2' (SARS-CoV-2) (Gorbalenya et al., 2020) and is a previously unknown betacoronavirus that likely emerged through zoonotic transmission from contact with non-human animals (Andersen et al., 2020; Zhou et al., 2020). However, the precise origins of the current pandemic remain inconclusive at present (Zhang and Holmes, 2020).

Two other betacoronaviruses have zoonotically emerged to cause significant human epidemics. Severe acute respiratory syndrome-related coronavirus (SARS-CoV) emerged in China in 2002 via an intermediate host of masked palm civets (*Paguma larvata*) in live animal markets (Guan et al., 2003; Song et al., 2005), and Middle East respiratory syndrome related coronavirus (MERS-CoV) emerged in Saudi Arabia in 2012 via an intermediate host of dromedary camels (Alagaili et al., 2014; Chu et al., 2014), with considerable evidence that both originated in bats (Anthony et al., 2017; Cui et al., 2007; Hu et al., 2015; Lau et al., 2005). Four additional coronaviruses are known to be endemic within humans, causing mild common cold-like illness (*Alphacoronavirus:* Human coronaviruses 229E and NL63; *Betacoronavirus:* Human coronaviruses HKU1 and OC43).

All viruses in the family *Coronaviridae* feature similar structural proteins, including a spike glycoprotein on the outer viral surface. This protein attaches to host cell receptors and subsequently initiates cell entry. While SARS-CoV-2 shows high genetic similarity to bat coronaviruses, particularly bat coronavirus RaTG13 (matching 96% sequence identity) (Zhou et al., 2020), its spike protein instead exhibits differences among key amino acid residues of the receptor binding domain (RBD) (Andersen et al., 2020; Wan et al., 2020), the region which directly interacts with host receptors. Based on this region, SARS-CoV-2 is predicted through structural (Wan et al., 2020; Wrapp et al., 2020) and in vitro experimental models (Hoffmann et al., 2020; Letko and Munster, 2020) to have highly efficient binding to the human angiotensin-converting enzyme 2 (ACE2) receptor, a feature that has likely contributed to its efficient human-to-human transmissibility. As the key molecular determinant of host range (Graham and Baric, 2010), adaptation of coronavirus spike proteins therefore represents a key opportunity to further understand their host range constraints.

Beyond selection acting at specific loci, viral adaptation can also manifest through broad-scale genomic signatures. Viruses exhibit biased genome composition, for example, in non-uniform use of synonymous codons (Jenkins and Holmes, 2003). Furthermore, coevolution within different hosts may indirectly lead to selection for particular compositions, as reported for nucleotide and dinucleotide usage within avian and human influenzaviruses (Greenbaum et al., 2008; Rabadan et al., 2006) and codon pair usage of arboviruses within their insect vectors and mammalian hosts (Shen et al., 2015). The *Coronaviridae* are no exception - different coronaviruses (including SARS-CoV-2) vary in their genome composition, with particularly complex patterns of codon usage in spike protein coding sequences (Dilucca et al., 2020; Gu et al., 2020), which could potentially contain important evolutionary signal regarding host origin.

Machine learning has recently gained substantial attention as a methodology in comparative modelling of emerging diseases. These methods are capable of decomposing signal in high-dimensional genomic information (a limitation of regression frameworks) without the need for sequence alignment. Genomic machine learning analyses have demonstrated the ability to not only classify viruses from recurring viral genome motifs (Randhawa et al., 2020), but also classify their broad host origins (Babayan et al., 2018; Bartoszewicz et al., 2020; Young et al., 2020; Zhang et al., 2019). Specifically considering coronaviruses, support vector machines and random forests have been trained on various genomic features to predict host group, including nucleotide and dinucleotide biases (Tang et al., 2015), amino acid composition (Qiang et al., 2020) or sequence k-mers (Li and Sun, 2018). However, previous model predictions are mostly concentrated upon bats or humans, and few analyses explicitly address the spike protein (but see Li and Sun, 2018). The exact potential of genome composition to predict host origin therefore remains unclear.

We aimed to use machine learning to understand how the complex genomic signatures of coronaviruses might predict their hosts and determine the importance of such signature in the spike protein. Specifically, we trained random forest models on compositional biases for spike protein and whole genome nucleotide sequences and compared their performance. A limitation of these approaches is that model predictions can be strongly influenced by viral sampling biases and reflect virus lineage rather than host (Di Giallonardo et al., 2017). Therefore, we undersample sequences to create balanced training data and control for similarity between related sequences during hold-one-out validation. We demonstrate the use of machine learning as a reliable method to estimate host origins of future novel coronaviruses in humans and livestock.

Methods

Data extraction and processing

Spike protein or whole genome sequence data for coronaviruses were identified within GenBank, using search terms

'txid###[Organism:noexp] AND (spike[Title] OR "S gene"[Title] OR "S protein"[Title] OR "S glycoprotein"[Title] OR "S1 gene"[Title] OR "S1 protein"[Title] OR "S1 glycoprotein"[Title] OR peplomer[Title] OR peplomeric[Title] OR peplomers[Title] OR "complete genome"[Title]) NOT (patent[Title] OR vaccine OR artificial OR construct OR recombinant[Title])'

where successive searches were conducted replacing ### with taxonomic identifiers for each species and unranked sub-species belonging to the family *Coronaviridae* within the NCBI taxonomy database (Federhen, 2012) (n = 1585 taxonomic ids total). Matching sequences were then extracted and further filtered to exclude incomplete or truncated sequences based on a) metadata labels and b) length restrictions, discarding any spike protein sequences < 2 kilobases (kb) and any whole genome sequences outside a range of 20kb – 32kb. We accepted both spike protein coding sequences within whole genome sequences and standalone complete spike protein sequences, excluding those only covering individual S1 or S2 subunits. All sequence data searching, filtering and extraction was conducted with R package \`rentrez\` v1.2.2 (Winter, 2017; see also Brierley, 2020).

Host classification

For each spike protein or whole genome sequence, host names were extracted from the host organism metadata field before being resolved to the standard NCBI taxonomy using the R package \`taxizedb\` v0.1.9.93 (Chamberlain and Arendsee, 2020; see also Brierley, 2020). Host names were automatically resolved to the highest taxonomic resolution possible and any unmatched host names were resolved manually, discarding sequences with missing/unresolvable names.

We then constructed a new variable broadly describing host category of each sequence, defined at various taxonomic levels: human (species *Homo sapiens*), camelid (family *Camelidae*), swine (family *Suidae*), carnivore (order *Carnivora*), rodent (order *Rodentia*), and bird (class *Aves*). Following a previous analysis (Babayan et al., 2018), we included two categories to represent bats (order *Chiroptera*): suborder Yinpterochiroptera (families *Craseonycteridae, Hipposideridae, Megadermatidae, Pteropodidae, Rhinolophidae,* and *Rhinopomatidae*) and suborder Yangochiroptera (all other families), based on their evolutionary divergence (Tsagkogeorga et al., 2013) and differences in ecology and hosted viruses (Moratelli et al., 2015; Young and Olival, 2016). Sequences not conforming to any of the above host categories were excluded from further analysis.

Genomic feature calculation

We then calculated several features describing genome composition biases of each spike protein and whole genome coding sequence at nucleotide, dinucleotide or codon level. Firstly, nucleotide biases were calculated as simple proportion of A, C, G or U content. Dinucleotide biases were calculated as the ratio of observed dinucleotide frequency to expected based on nucleotide frequency, following (Babayan et al., 2018):

$$\frac{\frac{d_{xy}}{D}}{\left(\frac{n_x}{N}\cdot\frac{n_y}{N}\right)}$$

where d_{xy} denotes frequency of dinucleotide xy, n_x and n_y denote frequency of individual nucleotides x and y, and D denotes total dinucleotides and N total nucleotides for length of the given sequence. Biases were calculated separately for each dinucleotide at each position within codon reading frames (i.e. at positions 1-2, 2-3 or 3-1) as dinucleotides spanning adjacent codons may be subject to more extreme biases (Kunec and Osterrieder, 2016; Tulloch et al., 2014). Finally, Relative Synonymous Codon Usage (RSCU) was also calculated for each codon including stop codons, following (Sharp and Li, 1987):

$$\frac{c_{ij}}{\frac{1}{n_i}\cdot\sum_j^{n_i} c_{ij}}$$

where n_i denotes number of codons synonymous for amino acid i and c_{ij} denotes frequency of j^{th} codon encoding for such amino acid. In total, this gave 116 genomic features for use in predictive models. The Effective Number of codons (ENc) (Wright, 1990) was also calculated for each sequence as a summative metric of magnitude of codon bias. All calculations for whole genomes considered nucleotide sequences as-is rather than as-read; sequence strings duplicated by frameshifting among ORF1a and ORF1b replicase protein were discarded to avoid disproportionate weighting in modelling analyses.

Machine learning analysis

To quantify the potential for genome composition biases to predict coronavirus host category, we used random forests, an ensemble machine learning approach that aggregates over a large number of individual classification tree models (Breiman, 2001). Such predictive modelling methods are often sensitive to class imbalances in outcome variables (He and Garcia, 2009), as well as training data composition. As several viruses (e.g. Porcine epidemic diarrhea virus) and host categories (e.g. human) appeared highly over-represented, we therefore conducted stratified random undersampling prior to modelling, retaining a maximum of 20 sequences per host category per virus (Supplementary Figure S1).

Zoonotic or epizootic coronavirus sequences from their novel host (i.e. SARS-CoV, SARS-CoV-2, MERS-CoV in humans, totaling m = 9 taxonomic identifiers, full list in Supplementary Data Files; swine acute diarrhea syndrome coronavirus in swine) were held out from model training as their evolutionary signature in genome composition is much more likely to instead reflect the original or donor host, having experienced comparatively little coevolution within the novel host following cross-species transmission. We also excluded human enteric coronavirus, the zoonotic potential of which remains unclear.

All random forests were constructed using 1000 trees and implemented using R package `ranger`, v0.12.1 (Wright and Ziegler, 2017). Model parameters were optimized within an inner loop of 10-fold cross-validation (Supplementary Figure S1), retaining the parameter set yielding the highest prediction accuracy. This approach was repeated in an outer loop through hold-one-out validation applied to coronaviruses (Supplementary Figure S1), i.e., rather exclude a single sequence in each instance, we exclude all sequences from a single virus to control for the similarity of genome composition within-virus. This allows host predictions for novel viruses based on values of compositional bias, rather than indirectly predicting host by the proxy of viral identity.

Model performance was then assessed by applying each random forest to its respective held-out coronavirus sequences as a test set. Probabilities of host categories were obtained by dropping sequences down each individual tree model within the random forest and averaging host category prevalence of resulting terminal nodes (see Malley et al., 2012)). To investigate explanatory relationships between genomic biases and hosts of coronaviruses, variable importance (calculated as relative mean decrease in Gini impurity) and partial dependence (calculated as marginal probability of each host category) associated with each genomic feature were averaged across all random forests. Host category predictions were also generated for each zoonotic coronavirus sequence by averaging predicted probabilities across all random forests to investigate model utility in application to a newly-identified human epidemic coronavirus.

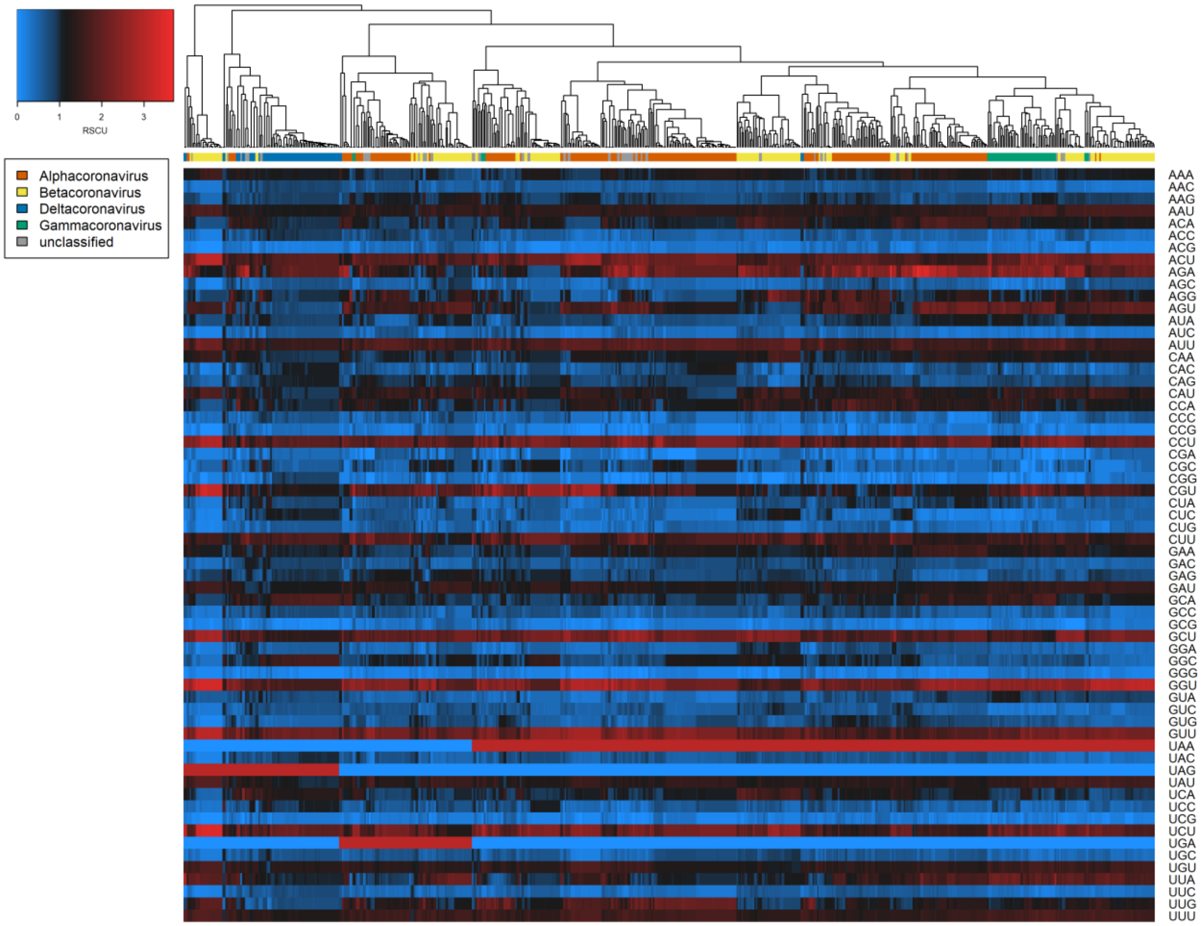
All data processing and modelling were conducted within R v4.0.0 (R Development Core Team, 2020).

Results

Genome composition across the *Coronaviridae*

In total, we identified n = 4960 nucleotide sequences for coronavirus spike proteins and n = 2987 whole genome sequences that met inclusion criteria. These were undersampled to n = 650 spike protein sequences from m = 225 coronaviruses and n = 511 whole genome sequences from m = 187 coronaviruses for use in further analysis (Supplementary Figure S1, Supplementary Table S1), spanning 40 identified host genera and 58 identified host species (Supplementary Data Files 1 & 2).

Broadly consistent genome composition biases were observed across the diversity of all coronavirus sequences examined. Specifically, ACU, AGA, GGU and GUU codons were over-represented among both coronavirus spike protein sequences and whole genomes (Figure 1). Hierarchical clustering based on RSCU values suggested codon usage was less distinct between genera within spike protein sequences (Figure 1A) than within whole genomes (Figure 1B). However, clear separation of deltacoronaviruses was observed for both cases, as these appeared to have less extreme biases in codon usage. This was confirmed by ENc calculation; deltacoronaviruses had higher ENc values than other coronavirus genera (Table 1). Considering dinucleotide biases, compositional bias was typically more extreme for dinucleotides spanning adjacent codons, i.e. position 3-1 (Supplementary Figure S2), and the characteristic coronavirus CpG suppression was also observed.



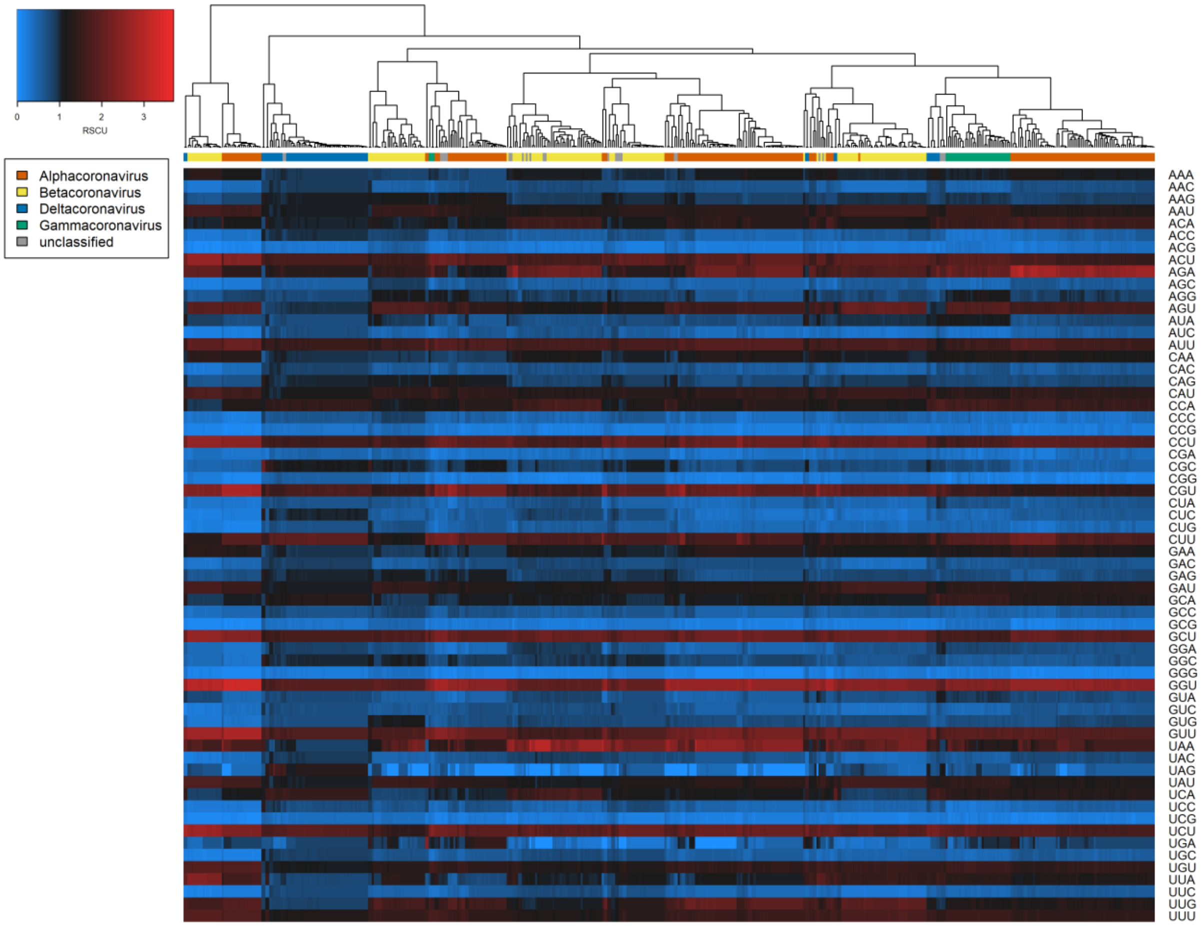


Figure 1. Codon biases (RSCU) across coronavirus genome sequences examined.
Heatmaps of coronavirus codon usage bias (RSCU) associated with each codon in each A) spike protein sequence (n = 650) and B) whole genome sequence (n = 511). Main colour scale denotes RSCU value, a null value of 1 (black) indicating no difference in codon usage from expectation, with blue and red representing under- or over-representation respectively. Dendrogram colour bar denotes taxonomic genus.

Table 1. ENc values across coronavirus genera.
Effective Number of Codons (ENc) for coronaviruses, stratified by genus and genome sequence type. ENc values are calculated as grand means, i.e. mean ENc was calculated per coronavirus by averaging sequences before means of means were calculated per genus by averaging coronaviruses. SD denotes standard deviation.

	Spike protein sequences		**Whole genome sequences**	
Genus	**Mean**	**SD**	**Mean**	**SD**
Alphacoronavirus	48.65	3.12	45.02	3.00
Betacoronavirus	47.77	5.04	47.03	4.79
Gammacoronavirus	46.36	1.83	46.11	0.74
Deltacoronavirus	54.79	3.43	51.75	3.22

(unclassified)	47.89	3.41	48.56	3.04
total	48.77	4.36	46.81	4.22

Host predictions of random forest models

Random forest models trained on nucleotide, dinucleotide and codon bias features of spike protein sequences predicted coronavirus hosts with 73% accuracy during hold-one-out cross-validation (Table 2). Genome composition of spike proteins appeared just as informative as whole genomes despite being much smaller in sequence length, as both models achieved very similar performance in all diagnostic measures (Table 2).

Table 2. Predictive performance of random forest models.
Model diagnostics describing overall performance when applied to predict host category of held-out coronaviruses. CI denotes confidence interval, Kappa denotes Cohen's Kappa statistic, mAUC denotes multiclass area-under-curve statistic. F1micro and F1macro denote F1 scores calculated using micro-averaging (performance on each host category weighted proportionally) and macro-averaging (performance on each host category weighted equally), respectively.

Predictor features	**Accuracy (95% CI)**	**Kappa**	**mAUC**	**$F1_{micro}$**	**$F1_{macro}$**
Spike protein	0.735 (0.700, 0.769)	0.696	0.898	0.848	0.757
Whole genome	0.728 (0.687, 0.766)	0.688	0.902	0.843	0.758

Patterns of host-specific predictive performance were evident during hold-one-out cross-validation. Random forests trained on both spike protein and whole genome sequence compositional features most easily distinguished bird, carnivore and rodent host categories (Figure 2, see also Supplementary Tables S2 & S3). Less powerful predictive performance was obtained for livestock (i.e., swine and camelid) host categories with these sequences often predicted as having bat (suborder Yangochiroptera) hosts, including all MERS-CoV sequences sampled from camels.

Human host origins appeared particularly difficult to characterise, with model-predicted hosts appearing more uncertain using spike protein features than whole genome features (Figure 2); while human coronaviruses HKU1 and NL63 were more confidently correctly classified based on whole genomes, human coronaviruses OC43 and 229E were more confidently misclassified as having camelid or Yinpterochiroptera hosts. The reciprocal also only occurred using whole genomes, i.e. several camel coronaviruses were predicted to have human hosts.

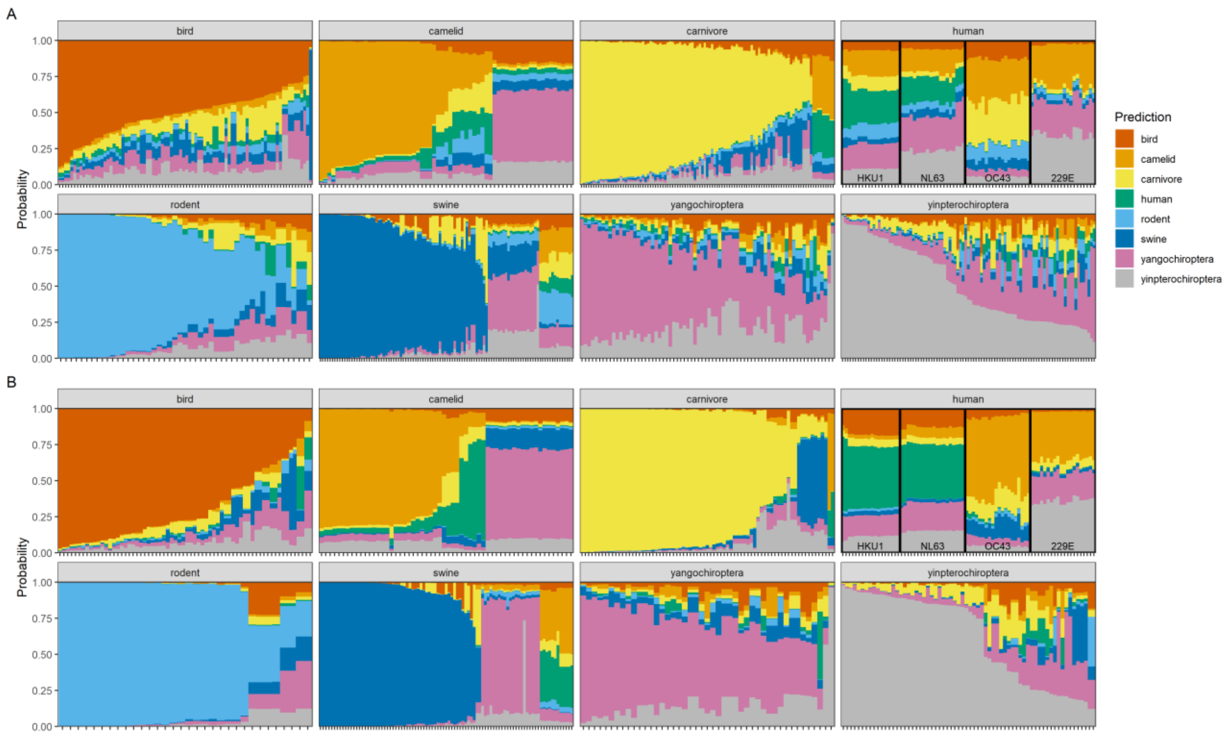


Figure 2. Random forest host predictions based on coronavirus genome composition.
Stacked bar plots of predicted probabilities of each host category for coronavirus sequences. Predictions were obtained from ensemble random forest models trained on A) spike protein and B) whole genome composition features. Panels depict sequences from each metadata-derived host category and colour coding denotes model-predicted host category. Stacks represent individual coronavirus sequences, ordered from largest to smallest probability of the correct host, i.e. greater panel area matching the correct host category indicates better overall model performance. Non-zoonotic coronavirus sequences originating from humans (human coronaviruses HKU1, NL63, OC43, 229E) are labelled for clarity.

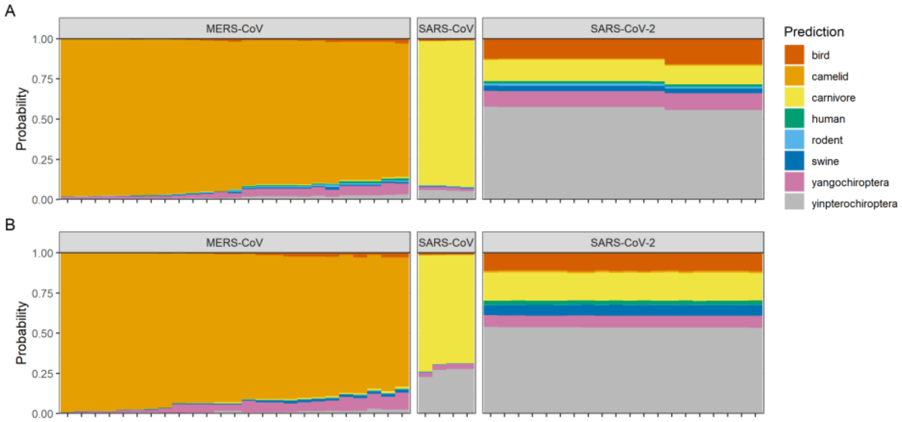


Figure 3. Random forest predictions based on zoonotic coronavirus genome composition.

Stacked bar plots of predicted probabilities of each host category for zoonotic coronavirus sequences sampled from humans. Predictions were obtained from ensemble random forest models trained on A) spike protein and B) whole genome composition features. Colour coding denotes model-predicted host category. Stacks represent individual coronavirus sequences.

We then applied random forests to those sequences of zoonotic viruses sampled from humans and excluded from model training: SARS-CoV, SARS-CoV-2, and MERS-CoV (Supplementary Figure S1, Supplementary Data Files 3 & 4). As these viruses have experienced little coevolution following zoonotic spillover, their genome composition signal likely gives indications about their ultimate or proximate animal host origins. MERS-CoV was overwhelmingly predicted to have camelid hosts (Figure 3) and SARS-CoV was predicted with less certainty as having carnivore hosts, consistent with the respective known intermediate hosts of camels and palm civets (order *Carnivora*). Contrastingly, SARS-CoV-2 was predicted mostly strongly to have a bat (suborder Yinpterochiroptera) host. Host predictions for zoonotic viruses were consistent between models using spike protein and whole genome features (Figure 3).

Variable importance of random forest models

The most informative genomic features towards predicting coronavirus hosts were a mixture of dinucleotide and codon biases (Figure 4), with dinucleotide biases appearing slightly more informative for spike protein sequences and codon biases appearing slightly more informative for whole genome sequence. However, predictive power of individual genomic features did not hold between spike protein and whole genome sequences; only weak correlation was observed between ranked variable importance from both analyses (Spearman's rank, ρ = 0.191, p = 0.042; Figure 4, see also Supplementary Figures S4 & S5). Partial dependence plots suggested the strongest individual discriminating feature to be GG dinucleotides at positions 1-2; an overrepresentation of this dinucleotide within the spike protein sequence clearly distinguished bird hosts from mammalian hosts (Supplementary Figure S4), consistent with the greatest predictive performance for bird coronaviruses (Supplementary Table S2).

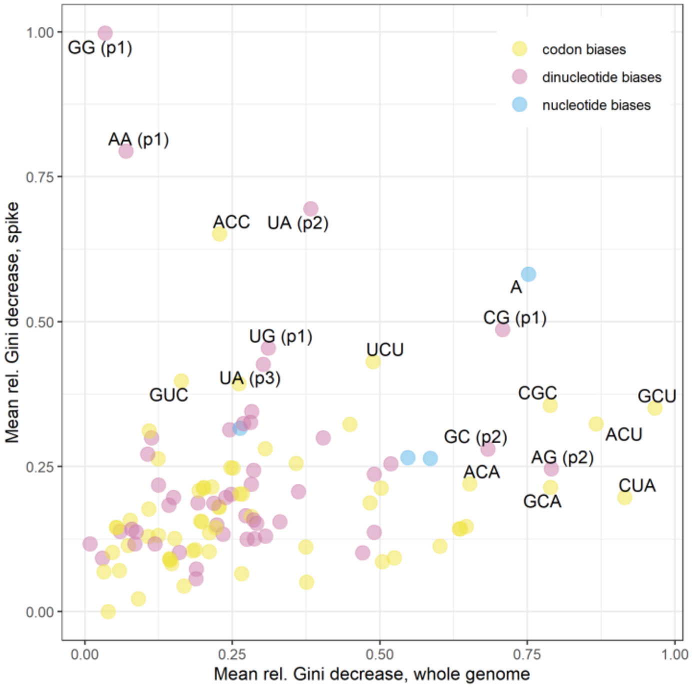


Figure 4. Variable importance of genomic features.

Variable importance of genome composition features in ensemble random forest models predicting coronavirus host category from whole genome sequences (x axis) and spike protein sequences (y axis), with labelling of top ten most informative features from both analyses. Points denote mean values of relative decrease in Gini impurity associated with each feature across A) m = 225 and B) m = 187 random forests during hold-one-out validation. Colour key denotes genomic feature type.

Discussion

We observe biased dinucleotide and nucleotide usage across the family *Coronaviridae*, and demonstrate that these genome composition biases contain sufficient evolutionary signal such that they can predict animal host origin. We show that training random forests on these features of spike proteins is equally as informative as using whole genome sequences in predicting hosts of novel (i.e., unseen) coronaviruses, with bird, carnivore and rodent viruses having the highest prediction accuracy. When applied to human coronavirus sequences from previous epidemics (SARS-CoV, MERS-CoV), random forest model predictions consistently represented the intermediate hosts. In the case of SARS-CoV-2, where the exact transmission pathway remains unknown, models predicted sequences to have a bat host (suborder Yinpterochiroptera).

Variability in genome composition

Among our dataset of 225 coronaviruses, we observed A- and U-ending codons to be overrepresented and G- and C- ending codons to be underrepresented (Figure 1), a commonly noted trait in other studies (Dilucca et al., 2020; Tort et al., 2020). Elsewhere, CpG dinucleotide bias has been proposed as a specific determinant of host (and tissue) range of coronaviruses (Xia, 2020), on the basis that CpG dinucleotides are targeted by zinc finger antiviral proteins and their suppression is therefore linked to immune evasion. We observed consistent CpG suppression (Supplementary Figure S2) and CG dinucleotides (positions 1-2) ranked 6th and 8th in feature importance for spike proteins and whole genomes (Figure 4), respectively.

However, spike proteins display different patterns of codon usage from other viral proteins (Gu et al., 2020), reflected in the lack of correlation between genome composition feature importance in random forests trained on spike proteins versus whole genomes (Figure 3, see also Supplementary Figures S4 & S5). This indicates spike proteins contain complex and distinct evolutionary signatures indicative of host-virus coevolution, supporting our approach in using many features beyond single dinucleotides (Pollock et al., 2020).

Model predictions of human coronaviruses

During model validation, human hosts appeared more challenging to correctly predict than other host types. Although the endemic human coronaviruses are common, they are also thought to have their ultimate evolutionary origins within non-human animals (Cui et al., 2007), which may explain this difficulty. In particular, human coronaviruses NL63 and HKU1 were more consistently predicted as having human hosts than human coronaviruses OC43 and 229E, especially when using whole genome sequences (Figure 2). Human coronavirus NL63 is estimated to have a more ancient common ancestor with bat coronaviruses than 229E (Huynh et al., 2012; Pfefferle et al., 2009), implying longer coevolutionary history within human hosts has resulted in a more consistently identifiable genomic signature.

Although several mutations of SARS-CoV-2 are becoming fixed in the population, e.g. D614G in the spike protein (Grubaugh et al., 2020), the virus has experienced only weak purifying selection (MacLean et al., 2020) and sequences remain extremely similar over the course of the pandemic. As such, our approach cannot identify host adaptation "in real-time"; rather, we examine variation generated over much longer macroevolutionary histories.

Instead, we would expect viruses that have transmitted cross-species more recently to retain the genome composition signature of their previous hosts, having experienced little coevolution within the novel host. Applying our finalised models to zoonotic human virus sequences may therefore give an indication of their proximate or ultimate animal host origin (Figure 3).

Human sequences of SARS-CoV were predicted to have a carnivore host, consistent with the known intermediate host of palm civets (*Paguma larvata)*. Much previous work has shown human and civet SARS-CoV sequences to have high similarity, with adaptive mutations concentrated within the spike protein (specifically, the receptor binding domain) (Graham and Baric, 2010; Guan et al., 2003; Song et al., 2005), which may explain the stronger prediction of carnivore hosts when using spike protein sequences. Similarly, human sequences of MERS-CoV were strongly predicted as having camelid hosts, consistent with the intermediate host of dromedary camels. The detection of evolutionary signatures corresponding to these intermediate hosts implies that these coronaviruses circulated in those hosts for sufficient time for coevolution to shape genome composition before cross-species transmission to humans. For MERS-CoV, camel infections have been recognised as far back as at least the 1980s (Müller et al., 2014; Sabir et al., 2016).

The origins of SARS-CoV-2 have been heavily speculated upon since its discovery, though there remains no compelling evidence towards the animal source of the first human infections. Our random forest models trained on genome composition of both spike protein and whole genome sequences predicted SARS-CoV-2 as having a bat host (suborder Yinpterochiroptera). Alignment-based and phylogenetic approaches showed the most closely related virus to be bat coronavirus RaTG13, a virus sampled from a horseshoe bat (*Rhinolophus affinis*) belonging to this suborder (Zhou et al., 2020), and more widely, the *Rhinolophidae* family are the most likely ancestral hosts of the *Sarbecovirus* genus (Latinne et al., 2020).

While our model predictions support bats as the ultimate origin of SARS-CoV-2, the involvement of intermediate hosts remains unclear. Although the Malayan pangolin (*Manis javanica*) was proposed early in the pandemic (Liu et al., 2020; Xiao et al., 2020), recent analyses have argued there is absence of evidence for this (Boni et al., 2020; Zhan et al., 2020). The methods used here are unable to identify intermediate hosts without sufficient sequence availability, and lack of such from pangolins (order *Pholidota*) preclude us from directly testing this hypothesis. However, selection analyses indicate SARS-CoV-2 could reasonably have exhibited efficient human infectivity and human-to-human transmissibility following direct transmission from bats (MacLean et al., 2020; Zhan et al., 2020), i.e., without strict need for prolonged selection within an intermediate host.

Future directions

A natural comparison to these methods is phylogenetic analyses, which can estimate traits such as host type from reconstructing viral ancestry based on sequence similarity. There is challenging confounding between molecular characteristics and sequence similarity, i.e., variation in genome composition may actually be predictive of viral lineage rather than host type (Di Giallonardo et al., 2017), essentially acting as a proxy for phylogenetic relatedness. To separate these signals, phylogeny needs to be considered in model construction (Young et al., 2020); by using a cross-validation procedure holding out individual coronaviruses rather than individual sequences, we attempt to distinguish genomic signatures arising from convergent evolution within specific hosts, rather than from viral similarity. A more generalised scope of study across multiple viral families would allow holdout of entire families during cross-validation (Young et al., 2020), further removing any phylogenetic proxy effects.

Additional challenges are created by the unavoidable, systematic gaps in sampling coverage. For example, disproportionate sampling to identify viruses in wildlife similar to those already known to affect humans or livestock may introduce bias to predictive models. Although we address this by undersampling, our model predictions, particularly for zoonotic coronaviruses, are likely influenced by the range of known viruses with available sequence data. These issues highlight the need for a careful choice of training dataset in modelling studies, but also for wider sampling and surveillance of coronaviruses among the wild virome (Carroll et al., 2018), especially considering their high public health risks.

Although we focus on compositional features, predictive approaches using other sequence properties may improve more mechanistic understanding of host range. For example, amino acid composition and physiochemical similarity between contiguous amino acid residues can predict human origin of coronavirus spike sequences (Qiang et al., 2020). Similarly, Young et al. (2020) have recently demonstrated the use of multiple types of genomic features in combination to predict infected hosts and found physicochemical classification of amino acid k-mers to achieve similar predictive power to nucleotide k-mers. More widely, hydrophobic and hydrophilic composition of host receptors shows some predictive signal towards virus sharing (Bae and Son, 2011), hydropathy being of mechanistic importance during virus-receptor binding, e.g. for murine coronavirus (Thackray et al., 2005). These properties could be used as additional features and improve machine learning model-derived predictions.

This emphasises an additional key question for future modelling studies distinct from host origin - whether genomic traits can predict the zoonotic potential of newly discovered animal coronaviruses. As this is strongly determined by molecular mechanisms of virus-host receptor interactions (Graham and Baric, 2010), these predictions may be best inferred by model frameworks combining genomic features of both spike proteins and host receptors.

Conclusion

By training machine learning models on genome composition across the *Coronaviridae*, we demonstrate a detectable evolutionary signature predictive of host type rooted in a region of the genome that is key to host shifts. Our random forest predictions add to the growing evidence COVID-19 ultimately originated within bats, though further work is needed to understand the potential for intermediate hosts. Characterising spike proteins (and by extension, their interaction with host receptors) may provide a fruitful path to further understanding zoonosis risk among coronaviruses.

Acknowledgements

We thank Cillian Courtney, Maya Wardeh, and Matthew Baylis for helpful support and commentary on this work. LB acknowledges funding from a Medical Research Council Skills Development Fellowship award, grant number MR/T027355/1.

References

Alagaili, A.N., Briese, T., Mishra, N., Kapoor, V., Sameroff, S.C., Wit, E. de, Munster, V.J., Hensley, L.E., Zalmout, I.S., Kapoor, A., Epstein, J.H., Karesh, W.B., Daszak, P., Mohammed, O.B., Lipkin, W.I., 2014. Middle East Respiratory Syndrome Coronavirus Infection in Dromedary Camels in Saudi Arabia. mBio 5. https://doi.org/10.1128/mBio.00884-14

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26, 450–452. https://doi.org/10.1038/s41591-020-0820-9

Anthony, S.J., Gilardi, K., Menachery, V.D., Goldstein, T., Ssebide, B., Mbabazi, R., Navarrete-Macias, I., Liang, E., Wells, H., Hicks, A., Petrosov, A., Byarugaba, D.K., Debbink, K., Dinnon, K.H., Scobey, T., Randell, S.H., Yount, B.L., Cranfield, M., Johnson, C.K., Baric, R.S., Lipkin, W.I., Mazet, J. a. K., 2017. Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. mBio 8. https://doi.org/10.1128/mBio.00373-17

Babayan, S.A., Orton, R.J., Streicker, D.G., 2018. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science 362, 577–580. https://doi.org/10.1126/science.aap9072

Bae, S.-E., Son, H.S., 2011. Classification of viral zoonosis through receptor pattern analysis. BMC Bioinformatics 12, 96. https://doi.org/10.1186/1471-2105-12-96

Bartoszewicz, J.M., Seidel, A., Renard, B.Y., 2020. Interpretable detection of novel human viruses from genome sequencing data. bioRxiv 2020.01.29.925354. https://doi.org/10.1101/2020.01.29.925354

Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.-Y., Perry, B.W., Castoe, T.A., Rambaut, A., Robertson, D.L., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 1–10. https://doi.org/10.1038/s41564-020-0771-4

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brierley, L., 2020. Using Open-access Tools (taxize, rentrez) to Find Coronaviruses, Their Genetic Sequences, and Their Hosts. ROpenSci Blog. URL https://ropensci.org/blog/2020/11/10/coronaviruses-and-hosts/

Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O., Mazet, J.A.K., 2018. The Global Virome Project. Science 359, 872–874. https://doi.org/10.1126/science.aap7463

Chamberlain, S., Arendsee, Z., 2020. taxizedb: Tools for Working with “Taxonomic” Databases.

Chu, D.K.W., Poon, L.L.M., Gomaa, M.M., Shehata, M.M., Perera, R.A.P.M., Abu Zeid, D., El Rifay, A.S., Siu, L.Y., Guan, Y., Webby, R.J., Ali, M.A., Peiris, M., Kayali, G., 2014. MERS coronaviruses in dromedary camels, Egypt. Emerg. Infect. Dis. 20, 1049–1053. https://doi.org/10.3201/eid2006.140299

Cui, J., Han, N., Streicker, D., Li, G., Tang, X., Shi, Z., Hu, Z., Zhao, G., Fontanet, A., Guan, Y., 2007. Evolutionary relationships between bat coronaviruses and their hosts. Emerg. Infect. Dis. 13, 1526–1532.

Di Giallonardo, F., Schlub, T.E., Shi, M., Holmes, E.C., 2017. Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. J. Virol. 91. https://doi.org/10.1128/JVI.02381-16

Dilucca, M., Forcelloni, S., Pavlopoulou, A., Georgakilas, A.G., Giansanti, A., 2020. Codon usage and evolutionary rates of the 2019-nCoV genes. bioRxiv 2020.03.25.006569. https://doi.org/10.1101/2020.03.25.006569

Federhen, S., 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40, D136–D143. https://doi.org/10.1093/nar/gkr1178

Gorbalenya, A.E., Baker, S.C., Baric, R.S., Groot, R.J. de, Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., Penzar, D., Perlman, S., Poon, L.L.M., Samborskiy, D., Sidorov, I.A., Sola, I., Ziebuhr, J., 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. 5, 536–544. https://doi.org/10.1038/s41564-020-0695-z

Graham, R.L., Baric, R.S., 2010. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. J. Virol. 84, 3134–3146. https://doi.org/10.1128/JVI.01394-09

Greenbaum, B.D., Levine, A.J., Bhanot, G., Rabadan, R., 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4, e1000079. https://doi.org/10.1371/journal.ppat.1000079

Grubaugh, N.D., Hanage, W.P., Rasmussen, A.L., 2020. Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear. Cell 182, 794–795. https://doi.org/10.1016/j.cell.2020.06.040

Gu, H., Chu, D.K.W., Peiris, J.S.M., Poon, L.L.M., 2020. Multivariate Analyses of Codon Usage of SARS-CoV-2 and other betacoronaviruses. bioRxiv 2020.02.15.950568. https://doi.org/10.1101/2020.02.15.950568

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., others, 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276–278.

He, H., Garcia, E.A., 2009. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hoffmann, M., Kleine-Weber, H., Krüger, N., Müller, M., Drosten, C., Pöhlmann, S., 2020. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. bioRxiv 2020.01.31.929042. https://doi.org/10.1101/2020.01.31.929042

Hu, B., Ge, X., Wang, L.-F., Shi, Z., 2015. Bat origin of human coronaviruses. Virol. J. 12, 221. https://doi.org/10.1186/s12985-015-0422-1

Huynh, J., Li, S., Yount, B., Smith, A., Sturges, L., Olsen, J.C., Nagel, J., Johnson, J.B., Agnihothram, S., Gates, J.E., Frieman, M.B., Baric, R.S., Donaldson, E.F., 2012. Evidence Supporting a Zoonotic Origin of Human Coronavirus Strain NL63. J. Virol. 86, 12816–12825. https://doi.org/10.1128/JVI.00906-12

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1–7. https://doi.org/10.1016/S0168-1702(02)00309-X

Kunec, D., Osterrieder, N., 2016. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. Cell Rep. 14, 55–67. https://doi.org/10.1016/j.celrep.2015.12.011

Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., Li, B., Zhang, W., Wang, L.-F., Shi, Z.-L., Daszak, P., 2020. Origin and cross-species transmission of bat coronaviruses in China. Nat. Commun. 11, 4235. https://doi.org/10.1038/s41467-020-17687-3

Lau, S.K.P., Woo, P.C.Y., Li, K.S.M., Huang, Y., Tsoi, H.-W., Wong, B.H.L., Wong, S.S.Y., Leung, S.-Y., Chan, K.-H., Yuen, K.-Y., 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. U. S. A. 102, 14040–14045. https://doi.org/10.1073/pnas.0506735102

Letko, M., Munster, V., 2020. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. bioRxiv 2020.01.22.915660. https://doi.org/10.1101/2020.01.22.915660

Li, H., Sun, F., 2018. Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. Sci. Rep. 8, 10032. https://doi.org/10.1038/s41598-018-28308-x

Liu, P., Jiang, J.-Z., Wan, X.-F., Hua, Y., Li, L., Zhou, J., Wang, X., Hou, F., Chen, Jing, Zou, J., Chen, Jinping, 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLOS Pathog. 16, e1008421. https://doi.org/10.1371/journal.ppat.1008421

MacLean, O.A., Lytras, S., Weaver, S., Singer, J.B., Boni, M.F., Lemey, P., Kosakovsky Pond, S.L., Robertson, D.L., 2020. Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. bioRxiv. https://doi.org/10.1101/2020.05.28.122366

Malley, J.D., Kruppa, J., Dasgupta, A., Malley, K.G., Ziegler, A., 2012. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. Methods Inf. Med. 51, 74–81. https://doi.org/10.3414/ME00-01-0052

Moratelli, R., Calisher, C.H., Moratelli, R., Calisher, C.H., 2015. Bats and zoonotic viruses: can we confidently link bats with emerging deadly viruses? Mem. Inst. Oswaldo Cruz 110, 1–22. https://doi.org/10.1590/0074-02760150048

Müller, M.A., Corman, V.M., Jores, J., Meyer, B., Younan, M., Liljander, A., Bosch, B.-J., Lattwein, E., Hilali, M., Musa, B.E., Bornstein, S., Drosten, C., 2014. MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983-1997. Emerg. Infect. Dis. 20, 2093–2095. https://doi.org/10.3201/eid2012.141026

Pfefferle, S., Oppong, S., Drexler, J.F., Gloza-Rausch, F., Ipsen, A., Seebens, A., Müller, M.A., Annan, A., Vallo, P., Adu-Sarkodie, Y., Kruppa, T.F., Drosten, C., 2009. Distant Relatives of Severe Acute Respiratory Syndrome Coronavirus and Close Relatives of Human Coronavirus 229E in Bats, Ghana. Emerg. Infect. Dis. 15, 1377–1384. https://doi.org/10.3201/eid1509.090224

Pollock, D.D., Castoe, T.A., Perry, B.W., Lytras, S., Wade, K.J., Robertson, D.L., Holmes, E.C., Boni, M.F., Kosakovsky Pond, S.L., Parry, R., Carlton, E.J., Wood, J.L.N., Pennings, P.S., Goldstein, R.A., 2020. Viral CpG Deficiency Provides No Evidence That Dogs Were Intermediate Hosts for SARS-CoV-2. Mol. Biol. Evol. 37, 2706–2710. https://doi.org/10.1093/molbev/msaa178

Qiang, X.-L., Xu, P., Fang, G., Liu, W.-B., Kou, Z., 2020. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect. Dis. Poverty 9, 33. https://doi.org/10.1186/s40249-020-00649-8

R Development Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project. org.

Rabadan, R., Levine, A.J., Robins, H., 2006. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J. Virol. 80, 11887–11891. https://doi.org/10.1128/JVI.01414-06

Randhawa, G.S., Soltysiak, M.P.M., Roz, H.E., Souza, C.P.E. de, Hill, K.A., Kari, L., 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. bioRxiv 2020.02.03.932350. https://doi.org/10.1101/2020.02.03.932350

Sabir, J.S.M., Lam, T.T.-Y., Ahmed, M.M.M., Li, L., Shen, Y., Abo-Aba, S.E.M., Qureshi, M.I., Abu-Zeid, M., Zhang, Y., Khiyami, M.A., Alharbi, N.S., Hajrah, N.H., Sabir, M.J., Mutwakil, M.H.Z., Kabli, S.A., Alsulaimany, F.A.S., Obaid, A.Y., Zhou, B., Smith, D.K., Holmes, E.C., Zhu, H., Guan, Y., 2016. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science 351, 81–84. https://doi.org/10.1126/science.aac8608

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Shen, S.H., Stauft, C.B., Gorbatsevych, O., Song, Y., Ward, C.B., Yurovsky, A., Mueller, S., Futcher, B., Wimmer, E., 2015. Large-scale recoding of an arbovirus genome to rebalance its insect versus mammalian preference. Proc. Natl. Acad. Sci. 112, 4749–4754. https://doi.org/10.1073/pnas.1502864112

Song, H.-D., Tu, C.-C., Zhang, G.-W., Wang, S.-Y., Zheng, K., Lei, L.-C., Chen, Q.-X., Gao, Y.-W., Zhou, H.-Q., Xiang, H., Zheng, H.-J., Chern, S.-W.W., Cheng, F., Pan, C.-M., Xuan, H., Chen, S.-J., Luo, H.-M., Zhou, D.-H., Liu, Y.-F., He, J.-F., Qin, P.-Z., Li, L.-H., Ren, Y.-Q., Liang, W.-J., Yu, Y.-D., Anderson, L., Wang, M., Xu, R.-H., Wu, X.-W., Zheng, H.-Y., Chen, J.-D., Liang, G., Gao, Y., Liao, M., Fang, L., Jiang, L.-Y., Li, H., Chen, F., Di, B., He, L.-J., Lin, J.-Y., Tong, S., Kong, X., Du, L., Hao, P., Tang, H., Bernini, A., Yu, X.-J., Spiga, O., Guo, Z.-M., Pan, H.-Y., He, W.-Z., Manuguerra, J.-C., Fontanet, A., Danchin, A., Niccolai, N., Li, Y.-X., Wu, C.-I., Zhao, G.-P., 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. U. S. A. 102, 2430–2435. https://doi.org/10.1073/pnas.0409608102

Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., Xia, X.-Q., 2015. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci. Rep. 5, 1–8. https://doi.org/10.1038/srep17155

Thackray, L.B., Turner, B.C., Holmes, K.V., 2005. Substitutions of conserved amino acids in the receptor-binding domain of the spike glycoprotein affect utilization of murine CEACAM1a by the murine coronavirus MHV-A59. Virology 334, 98–110. https://doi.org/10.1016/j.virol.2005.01.016

Tort, F.L., Castells, M., Cristina, J., 2020. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. Virus Res. https://doi.org/10.1016/j.virusres.2020.197976

Tsagkogeorga, G., Parker, J., Stupka, E., Cotton, J.A., Rossiter, S.J., 2013. Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats. Curr. Biol. 23, 2262–2267. https://doi.org/10.1016/j.cub.2013.09.014

Tulloch, F., Atkinson, N.J., Evans, D.J., Ryan, M.D., Simmonds, P., 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife 3, e04531. https://doi.org/10.7554/eLife.04531

Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. J. Virol. https://doi.org/10.1128/JVI.00127-20

WHO, 2020. Coronavirus disease (COVID-19) Weekly Epidemiological Update - 11. WHO, Geneva. [WWW Document]. URL https://www.who.int/docs/default-source/coronaviruse/situation-reports/weekly-epi-update-11.pdf (accessed 10.28.20).

Winter, D.J., 2017. rentrez: An R package for the NCBI eUtils API. R J. 9, 520–526.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.-L., Abiona, O., Graham, B.S., McLellan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. https://doi.org/10.1126/science.abb2507

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01

Xia, X., 2020. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. Mol. Biol. Evol. 37, 2699–2705. https://doi.org/10.1093/molbev/msaa094

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.-J., Li, N., Guo, Y., Li, X., Shen, X., Zhang, Zhipeng, Shu, F., Huang, W., Li, Y., Zhang, Ziding, Chen, R.-A., Wu, Y.-J., Peng, S.-M., Huang, M., Xie, W.-J., Cai, Q.-H., Hou, F.-H., Chen, W., Xiao, L., Shen, Y., 2020. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature 583, 286–289. https://doi.org/10.1038/s41586-020-2313-x

Young, C.C.W., Olival, K.J., 2016. Optimizing viral discovery in bats. PLoS ONE 11, e0149237. https://doi.org/10.1371/journal.pone.0149237

Young, F., Rogers, S., Robertson, D.L., 2020. Predicting host taxonomic information from viral genomes: A comparison of feature representations. PLOS Comput. Biol. 16, e1007894. https://doi.org/10.1371/journal.pcbi.1007894

Zhan, S.H., Deverman, B.E., Chan, Y.A., 2020. SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? bioRxiv 2020.05.01.073262. https://doi.org/10.1101/2020.05.01.073262

Zhang, Y.-Z., Holmes, E.C., 2020. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell 181, 223–227. https://doi.org/10.1016/j.cell.2020.03.035

Zhang, Z., Cai, Z., Tan, Z., Lu, C., Jiang, T., Zhang, G., Peng, Y., 2019. Rapid identification of human-infecting viruses. Transbound. Emerg. Dis. 66, 2517–2522. https://doi.org/10.1111/tbed.13314

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273. https://doi.org/10.1038/s41586-020-2012-7