bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans

Oscar A. MacLean[1,#], Spyros Lytras[1,#], Joshua B. Singer[1], Steven Weaver[2], Sergei L. Kosakovsky Pond[2,*], David L. Robertson[1,*]

[1]MRC-University of Glasgow Centre for Virus Research, Scotland, UK.

[2]Temple University, Institute for Genomics and Evolutionary Medicine, Philadelphia, USA.

[#]Joint first authors.

[*]To whom correspondence should be addressed: david.l.robertson@glasgow.ac.uk, spond@temple.edu

Abstract

RNA viruses are proficient at switching to novel host species due to their fast mutation rates. Implicit in this assumption is the need to evolve adaptations in the new host species to exploit their cells efficiently. However, SARS-CoV-2 has required no significant adaptation to humans since the pandemic began, with no observed selective sweeps to date. Here we contrast the role of positive selection and recombination in the *Sarbecoviruses* in horseshoe bats to SARS-CoV-2 evolution in humans. While methods can detect some evidence for positive selection in SARS-CoV-2, we demonstrate these are mostly due to recombination and sequencing artefacts. Purifying selection is also substantially weaker in SARS-CoV-2 than in the related bat *Sarbecoviruses*. In comparison, our results show evidence for positive, specifically episodic selection, acting on the bat virus lineage SARS-CoV-2 emerged from. This signature of selection can also be observed among synonymous substitutions, for example, linked to ancestral CpG depletion on this bat lineage. We show the bat virus RmYN02 has recombinant CpG content in Spike pointing to coinfection and evolution in bats without involvement of other species. Our results suggest the non-human progenitor of SARS-CoV-2 was capable of human-human transmission as a consequence of its natural evolution in bats.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Main text

In December 2019, a novel coronavirus emerged in the city of Wuhan, China, causing coronavirus disease-2019 (COVID-19) characterised by respiratory or gastrointestinal viral symptoms, and in severe cases, additionally, acute respiratory distress syndrome, cardiovascular dysfunction, thrombosis and other symptoms[1]. Evolutionary analysis placed this new human virus in the same subgenus of *Betacoronavirus*, the *Sarbecoviruses* (Figure 1A), that SARS emerged from[2], and it was named SARS-CoV-2[3] – the seventh known human-infecting member of the *Coronaviridae*. The initial outbreak of human cases of the virus was connected to the Huanan Seafood Wholesale Market in Wuhan[4], and while related viruses have been found in horseshoe bats[5] and pangolins[6], their divergence represents decades of evolution[7] leaving the direct origin of the pandemic unknown. In addition to the importance of understanding the route from animals to humans, key questions for assessing future risk of emergence are: what is the extent of evolution required to permit a bat virus to transmit to humans, and what subsequent evolution needs to occur for efficient transmission once the virus is established within the human population?

While both the first SARS virus outbreak in 2002/2003, causing approximately 8,000 infections, and a re-emergence in late 2003, causing four infections, were linked to Himalayan palm civets and raccoon dogs in marketplaces in Guangdong province[8,9], it became clear that these animals were conduits for spillover to humans and not true viral reservoirs[10]. Extensive surveillance work subsequently identified related viruses circulating in horseshoe bats in China some of which can replicate readily in human cells[11,12]. The bat viruses most closely related to SARS-CoV-1 (red variants, Figure 1A), can use human ACE2, while the addition of protease is required for the more divergent bat viruses tested (grey lineages, Figure 1A)[13]. Collectively these results demonstrate that, unlike other RNA viruses that usually acquire adaptations after switching to a new host species, the *Sarbecoviruses* – which already transmit frequently among bat species[14] – can utilise this generalist property, facilitating successful infection of non-bat species, including humans. We test this hypothesis by investigating the extent of positive selection (a measure of molecular adaptation) in the virus circulating in humans since the COVID-19 outbreak began, and contrast this to historic selection acting on the related bat viruses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

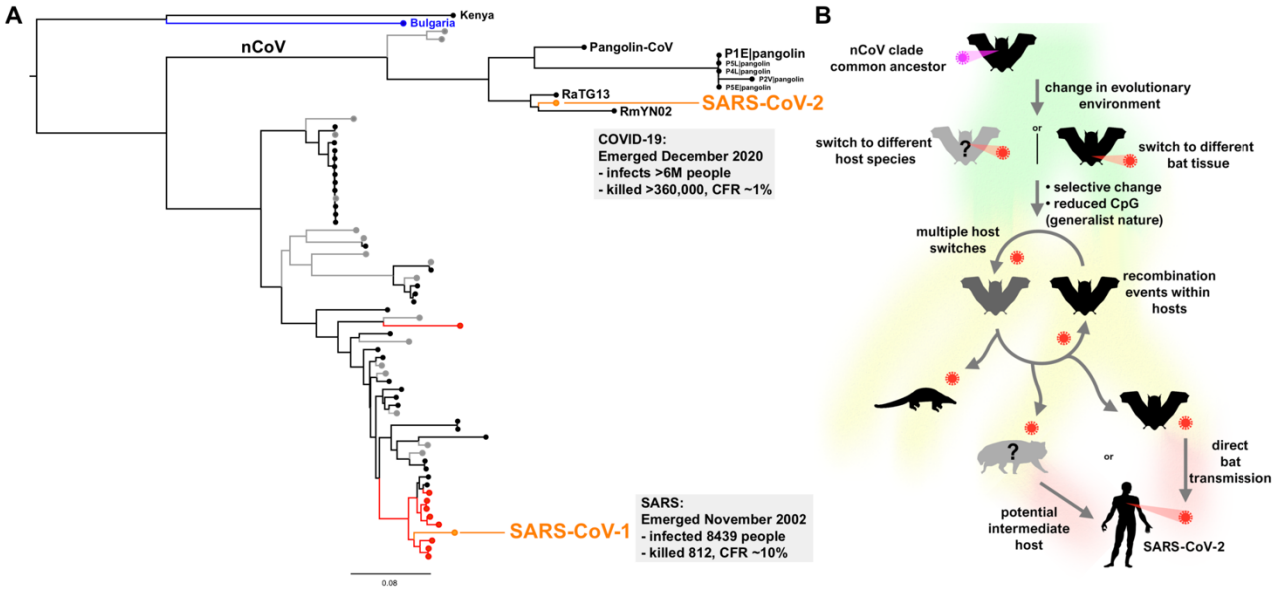

Figure 1. ***(A)*** *Phylogenetic tree (from RAxML[49]) showing the relationship of SARS-CoV-1 & -2 (orange) to related bat and pangolin Sarbecoviruses. Grey, red and blue variants are coloured according to Letko et al, 2020 who showed experimentally some viruses are able to use human ACE2 (red), while some require the addition of protease (grey)[13]; the red outlier is a known recombinant. Black indicates not tested by Letko et al, 2020, while the virus in blue, sampled in Bulgaria, could not been induced to infect human cells. Case fatality rate (CFR) from Verity et al.[15]. The scale bar is in expected nucleotide substitutions per site.* ***(B)*** *Schematic of our proposed evolutionary history of the nCoV lineage leading to SARS-CoV-2.*

The significance of mutation

There is intense interest in the mutations emerging in the SARS-CoV-2 pandemic[16–18]. This is important as, for example, Spike amino acid replacements could reduce the efficacy of vaccines targeting epitopes overlapping these mutations, replacements in proteases and polymerases could result in acquired drug resistance, and other mutations could change the biology of the virus, e.g., enhancing its transmissibility or severity[19]. Although the vast majority of mutations and any associated amino acid replacements are expected to be 'neutral'[20,21], changes with functional significance to the virus will eventually arise, as they have in most other viral epidemics and pandemics. Many nonsynonymous mutations which cause amino acid replacements are expected to be deleterious to the virus. These are likely to be removed from the population through the action of purifying selection, as viruses which possess these mutations transmit less frequently. One way to begin to understand the functional impact of mutations is to characterise the selective regime they are under. Mutations which are under positive selection are of particular interest as they are most likely to possess some functional significance. However, mutations that might provide a strong selective advantage for a virus, for example a 1% increase in growth rates during early infection, will not necessarily have any

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

observable impact on virulence or transmission rates particularly when so many hosts are susceptible as is the case for SARS-CoV-2. This is because many other factors determine individual transmission rates and disease severity, and intraspecific variation in these factors may dominate over variance across viruses.

Evidence of relatively weak purifying selection in SARS-CoV-2. We first analyse selection acting on the encoded amino acids in 15537 genome sequences, a sample of the SARS-CoV-2 variants circulating in humans (also see Supplementary text 2). Purifying selection would be expected to act more strongly on nonsynonymous sites, supported by the estimate that nonsynonymous sites evolve at only 4% of the speed of synonymous sites in the wider *Sarbecovirus* phylogeny[7]. We compared the relative frequencies of nonsynonymous and synonymous mutations in the pandemic data and found that, after adjusting for the greater number of nonsynonymous sites, there are fewer nonsynonymous mutations than synonymous across at all frequency intervals (Figure 2). This depletion of nonsynonymous mutations across all frequencies indicates that selection is filtering out nonsynonymous mutations before they are observed by sequencing of the viral population. The vast majority of observed mutations occur at low frequency, with only ~10% of mutations observed in more than six of the 15537 sequences (Figure 2). Nonsynonymous mutations appear to spread into the highest frequency categories less frequently, suggesting that selection is additionally suppressing amino acid replacements observed in the population, i.e., purifying selection is the overwhelming signal in the pandemic.

As observed in other virus outbreaks mutation rate estimates of related coronaviruses appear to decline with increasing sampling time[7,22]. This is because mildly deleterious mutations in viral outbreaks fail to persist over longer time periods, being gradually purged by purifying selection[23,24]. This can be observed in the relatively suppressed frequency of nonsynonymous mutations in the SARS-CoV-2 outbreak (Figure 2). In addition, as worldwide suppression strategies reduce the effective reproductive number (Rt) towards or below 1, it is likely that many of the putatively deleterious segregating mutations, which are driving the initially elevated nonsynonymous substitution rate, will be purged by purifying selection[25], decreasing the observed dN/dS values further.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

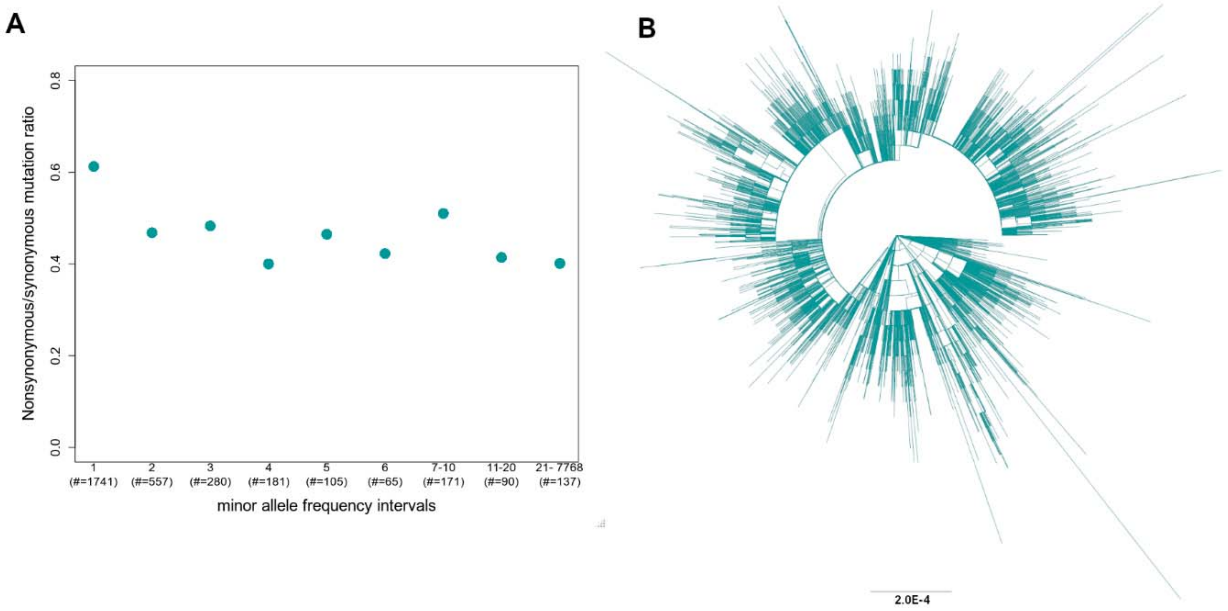

***Figure 2**. **(A)** The ratio of synonymous and nonsynonymous mutations at each allele frequency interval, scaled by the relative number of nonsynonymous to synonymous sites (2.76; estimated using the PAML4 1x4 model; Yang 2007). Sites with fewer than 95% of samples sequenced were excluded from the analysis. 15537 sequences were included in the analysis; data as of May 14th, 2020. **(B)** Phylogenetic tree (from FastTree[26]) of these SARS-CoV-2 concatenated open reading frame sequences used in the analysis.*

Little evidence for positive selection in SARS-CoV-2. Next, we performed selection analysis on 396 SARS-CoV-2 sequences from mid-March, a sufficient number of variants to capture the emergence of SARS-CoV-2 and any early associated adaptations. This analysis using the FUBAR method[27] from the HyPhy package[28] yielded sparse evidence of positively selected sites in the pandemic data (Supplementary table 1). Interestingly, we are able to attribute most of the candidate sites to either artefactual lab recombination, or potential hypermutation (see Supplementary figures 1, 2 and 3). Finding evidence of recombinant sequences has several important consequences. First, it violates the assumption that a single phylogenetic tree describes the evolutionary history of the sample sequences, which may have implications for molecular epidemiology inferences. Second, genomic searches for 'beneficial' mutations in the pandemic must incorporate the possibility that multiple origins of mutations may in fact just be recombination events, whether real or artificial sequencing errors. Failure to control for this possibility will lead to high false positive rates, and the mistaken inference of adaptation[29].

Genuine detectable recombination events would require natural co-infections by genetically distinct viruses. Secondary infection is much more likely to occur early on in the initial infection period when competition is lower. Later during infection the first virus will likely have colonised a

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

greater proportion of susceptible tissues, causing competitive suppression of the second infection, or have initiated an inhibitory immune response. Therefore, it is likely that recombination rates scale non-linearly with prevalence and mostly occur only in very high prevalence areas. It will be important to monitor for their occurrence as they could provide additional mechanisms for novel genotypes to be generated once sufficient diversity exists in the human population, or in the event of emergence of a third SARS-like coronavirus.

Importantly, as the SARS-CoV-2 sequence data continues to accumulate (recently surpassing 30,000 genomes in GISAID) the dominant evolutionary signal in the data is one of purifying selection (see, http://hyphy.org/covid/).

What about in bats? Positive selection in *Sarbecoviruses*. Coronaviruses are known to frequently recombine in their bat hosts, with the Spike open reading frame (ORF) being an apparent hotspot for this process, which might have adaptive implications for the viruses in the context of immune evasion[10,30–33]. We therefore separately tested each ORF of 69 *Sarbecoviruses* including SARS-CoV-2, SARS-CoV-1 and their close relatives (Supplementary table 2), and further separated the two longest ORFs, Orf1ab and Spike, into five putatively non-recombinant regions each, based on Boni et al. (2020)[7]. We define as the 'nCoV lineage' the set of viruses closest to SARS-CoV-2 in the phylogeny (Figure 1A). These vary across genomic regions according to the recombination patterns observed. The viruses that are present in every definition of the lineage in this analysis are the following: SARS-CoV-2, RaTG13, Pangolin-CoV and the pangolin-infecting viral cluster: P2V, P5L, P1E, P5E and P4L. Genomic sites are generally subject to conservation in this nCoV lineage, with 8184/9744 (84%) of codon sites conserved at the amino-acid level, and 4274 (43.7%) sites, of which 3388 were variable at the nucleotide level, showing evidence of purifying selection in this lineage (using the FEL method[34]).

We sought evidence of episodic diversifying selection on the nCoV lineage using BUSTED[S][35], coupled with a hidden Markov model (HMM) with three rate categories to describe site-specific synonymous rate variation (SRV) and allow spatial autocorrelation in these rates[36]. Non-recombinant regions of Orf1ab, Spike and ORF N show evidence of episodic diversifying positive selection on the nCoV lineage (Figure 3). This finding is consistent with evidence of positive selection operating on Orf1ab in MERS[37], and Spike and N being essential for antigenic

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

recognition. Eighty five individual sites were inferred to evolve subject to episodic diversifying selection in the nCoV lineage (Figure 3) using the MEME[38] method.

We next looked for branch-specific evidence of selection in these flagged regions, using the aBSREL method[39]. Our analysis found that diversifying selection left its imprints primarily at the deepest branches of the nCoV lineage, with no evidence of selection in the terminal branch leading to SARS-CoV-2 (Figure 3). This is consistent with the non-human progenitor of SARS-CoV-2 requiring little or no novel adaptation to successfully infect humans. Still, no model can detect all signatures of historic genomic adaptation, and mutations which may enable SARS-CoV-2 to infect humans could have arisen by genetic drift in the reservoir host.

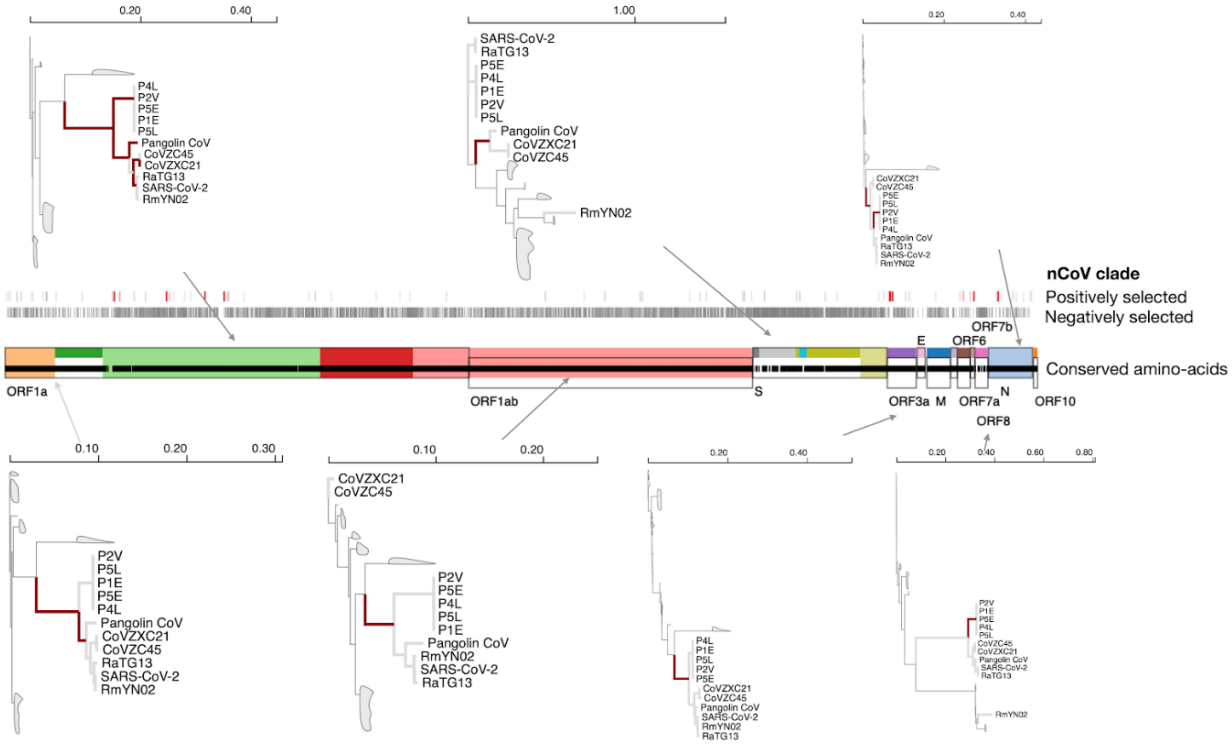

***Figure 3**. Schematic of the non-recombinant ORF regions used for the nCoV lineage selection analyses. Regions with significant evidence of positive selection based on BUSTED[S] filled in with wide colour bars. Phylogenies with highlighted branches are presented for regions with branch-specific evidence of selection in aBSREL. Non-nCoV lineages are collapsed for clarity and greyed out and branch lengths are shown as estimated under the aBSREL mixture model. Three categories of individual sites (conserved, negatively selected, positively selected) are shown as tracks in or above the schematic. For positively selected sites, colouring reflects the fraction of the branches in the nCoV lineage inferred to be under selection at the site (gray: smaller, red: larger).*

The BUSTED[S] method also partitioned synonymous rate variation into three rate classes across the sites. The majority of regions showed large, in some cases more than 20-fold, differences between rate classes, with all three classes representing a substantial proportion of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sites for most regions (Supplementary figure 4), with varying degrees of spatial autocorrelation. This suggests that strong purifying selection is acting on some synonymous sites (e.g., conserved motifs or RNA features), and some synonymous mutations in the SARS-CoV-2 genome may not be selectively neutral or occur at sites that are hypervariable. Some synonymous rate variation may also be attributed to the 5' and 3' context-specific mutation rate variation observed in SARS-CoV-2[22].

Patterns of CpG depletion in the nCoV lineage. Genome composition measures, such as dinucleotide representation and codon usage can also be an informative tool for characterising the host history of a virus[40]. Various host antiviral mechanisms accelerate the depletion of CpG dinucleotides in virus genomes. This is thought to be primarily mediated either through selective pressures by a CpG-targeting mechanism involving the Zinc finger Antiviral Protein (ZAP)[41] or C to U hypermutation by APOBEC3 cytidine deaminases[42]. These forces are likely to vary across different tissues within a host and across different mammalian hosts. Thus, a smaller or greater level of CpG depletion in particular viral lineages may be indicative of a switch in the evolutionary environment of that lineage or its ancestors. Care must be taken to not over-interpret these results and conjure unsupported narratives (Pollock et al. 2020, under review).

We examined the CpG representation in Orf1ab of the *Sarbecoviruses* using the corrected Synonymous Dinucleotide Usage (SDUc) measure, controlling for amino acid abundance and single nucleotide composition bias in the sequences[43]. We find a downward shift in CpG depletion levels at the base of the nCoV lineage, in comparison to the rest of the phylogeny (Figure 4a-d). This may indicate a change of evolutionary environment, e.g., host or tissue preference, since CpG depletion following host switches has been observed in other human infecting RNA viruses, such as Influenza B[40]. Given the decades of divergence between SARS-CoV-2 and the most closely related bat viruses[7] could this evolution have occurred in a different host species?

Zhou et al. (2020)[44] report a novel bat-infecting *Sarbecovirus* sample, RmYN02, which possesses the highest sequence similarity to SARS-CoV-2 of known *Sarbecoviruses* for most of its genome. Yet, part of the RmYN02 Spike ORF is recombinant and is placed in the non-nCoV clade of the *Sarbecovirus* phylogeny. This viral sequence offers an opportunity to test if the recombination scenario was consistent with the lineage-specific CpG depletion patterns. A sliding window of CpG relative dinucleotide abundance (RDA)[45] shows that CpG levels of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SARS-CoV-2 and RmYN02 only differ at the recombinant region (Figure 4e). This is further demonstrated when contrasting the SDUc values of the nCoV and non-nCoV parts of RmYN02 Spike to those of SARS-CoV-2 (Figure 4f). The finding that RmYN02 is a recombinant between the high and low CpG lineages suggests that viruses from both lineages are co-infecting the same bat species. The CpG depletion is therefore probably not being driven by unique selection or mutational pressures across lineages, but instead by a lineage-specific effect, such as functional polymerase differences.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

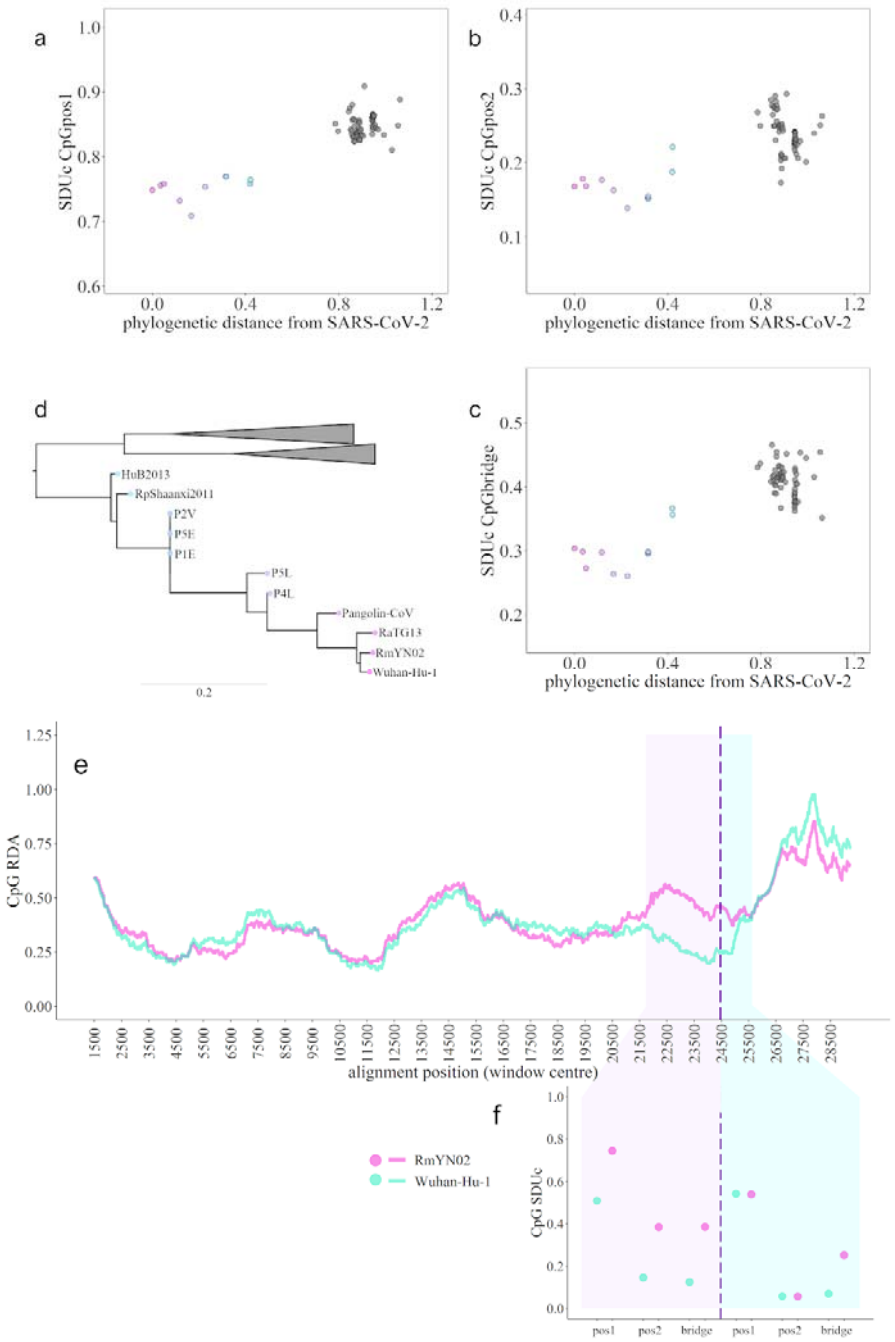

Figure 4. *Corrected synonymous dinucleotide usage (SDUc) values for the Orf1ab of each Sarbecovirus, frame positions 1 (a), 2 (b) and bridge (c) plotted against patristic distance from SARS-CoV-2. The tip colours of the phylogeny (d) correspond to the SDUc data points. The non-nCoV part of the phylogeny has been excised for clarity. (e) 3kb sliding window plot of relative dinucleotide abundance (RDA) across the whole-genome alignment of Wuhan-Hu-1 (turquoise) and RmYN02 (magenta). Shaded regions depict the Spike ORF region in the alignment. The dashed line indicates the inferred RmYN02 Spike recombination breakpoint, splitting the shaded region into non-nCoV (pink) and nCoV (blue). (f) SDUc values calculated for each frame position of the two RmYN02 Spike non-recombinant regions and the corresponding Wuhan-Hu-1 regions.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conclusion. The evidence of positive selection on the nCoV lineage SARS-CoV-2 emerged from – coupled with the change in CpG composition in this lineage and evidence primarily of purifying selection in human circulating SARS-CoV-2 – indicates that the significant SARS-CoV-2 evolution occurred prior to spillover into humans. The immediate 'success' of this bat virus in several species following cross-species transmission supports the hypothesis that this is a viral lineage with a relatively generalist nature (Figure 1B). We suggest that the early ancestors of the nCoV lineage developed this generalist phenotype through a change in its evolutionary environment (host switch, or tissue tropism) probably in bat species, allowing for multiple spillover events to pangolins, and now humans, and potentially other wild mammals we have yet to sample. While no evidence yet points to an intermediate host playing anything more than a conduit role in the SARS-CoV-2 transmission to humans, there are still large gaps in our knowledge of its recent non-human origins, as the closest bat viruses are relatively divergent in time[7]. Recombination between a proximal ancestor of RmYN02 and a non-nCoV *Sarbecovirus* is another indication that all the viruses in this subgenus co-circulate in bat reservoirs and occasionally transmit to other mammals causing infection. In terms of controlling viral emergence, we must dramatically ramp up surveillance at the human-animal interface. Serological studies of communities in China that come into contact with bats indicate that incidental and 'dead-end' spillover of SARS-like viruses into humans do take place[46,47]. Due to the high diversity of *Sarbecoviruses* and the generalist nature of these coronaviruses, a future emergence is likely and could be sufficiently divergent to evade either natural or vaccine-acquired immunity, as demonstrated for SARS-CoV-1 versus SARS-CoV-2[48]. While gradual 'antigenic drift' could become an issue as the new virus diverges in the human population, it will be important to monitor for abrupt 'antigenic shifts', e.g. those facilitated by recombination with divergent *Sarbecoviruses* in the context of spillover events. Such events appear routine in bat species and like it or not we are now part of the host range of these viruses.

Methods

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SARS-CoV-2 GISAID sequence filtering. To reduce the impact of sequencing errors on selection analysis the data from GISAID was filtered by excluding all sequences which meet any of the following criteria: any sequence of length less than 29000 nucleotides or greater than 35000; any sequence with a non-human host, e.g., bat, pangolin; sequences from environmental samples; any sequences marked with a warning flag, as having quality issues on GISAID; any sequence with more than 30 unique (across the whole dataset) single nucleotide mutations relative to the SARS-CoV-2 reference sequence; and any sequence which has a frameshifting deletion or insertion relative to the SARS-CoV-2 reference sequence.

SARS-CoV-2 positive selection. To search for signatures of positive selection in the phylogenetic tree of the current SARS-CoV-2 outbreak we ran the Bayesian FUBAR software from the HyPhy package[27,28]. This software searches for evidence of positive selection by estimating phylogeny-wide ratios of nonsynonymous (dN) and synonymous (dN) substitutions rates for each site in the alignment. It estimates a posterior probability that each site is under positive selection across the phylogeny (dN/dS >1), with a posterior probability>0.9 used as the threshold for significance, as suggested by the authors. FUBAR was run using an alignment exported from CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/) on the 16th March 2020, using a tree generated in RAxML[49] under the GTR+Γ model.

SARS-CoV-2 recombination. As recombination is known to confound FUBAR, and other methods in the HyPhy package, the maximum likelihood recombination detection software GARD[29] was used to test for recombination before using FUBAR. This software searches for recombination by introducing potential breakpoints and optimising tree topologies either side of the new breakpoint. If the Akaike information criterion (AIC)[50] is improved by the optimisations with breakpoints in, this provides significant evidence of recombination. If significant evidence of recombination is found, the method can then generate multiple non-recombinant partitions in the sequence alignment for use in downstream analyses. However, if the samples are highly related, as in the SARS-CoV-2 dataset, this phylogeny-based approach is limited in power as each recombination event introduces a large number of additional number of parameters, substantially penalising the AIC[50]. To detect recombination with more power for closely related samples, we also used the pairwise homoplasy index[51], which tests for excessive homoplasies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

However, this method cannot tell if homoplasies are due to recombination or convergent evolution through parallel adaptation due to shared selection pressures.

***Sarbecoviruses* alignment and recombination.** To avoid the confounding effects of recombination, we have analysed each open reading frame (ORF) separately, and divided the Orf1ab and Spike ORFs into putative non-recombinant regions, based on the seven major recombination breakpoints presented in Boni et al. (2020)[7]. This produces five non-recombinant regions for Orf1ab (regions A to E) and five regions for Spike (regions A to D, and the variable loop - region VL). The protein sequences of the non-recombinant regions SARS-CoV-2, SARS-CoV-1 and 67 closely related viruses with non-human hosts (bats and pangolins) were aligned using MAFFT version 7 (L-INS-i)[52]. Subsequent manual corrections were made on the protein alignments and pal2nal[53] was used to convert them to codon alignments. Phylogenies for each codon alignment were inferred using RAxML with a GTR+Γ model[49].

***Sarbecovirus* selection analysis.** We used an array of selection detection methods to examine whether the lineage leading to SARS-CoV-2 has experienced episodes of diversifying positive selection. Each non-recombinant region was examined separately. We separated each region's phylogeny into an nCoV and non-nCoV/SARS-CoV-1 lineage. The nCoV lineage includes SARS-CoV-2 and the viruses that it is phylogenetically most closely related to. These are the bat-infecting CoVZC45, CoVZXC21, RmYN02 and RaTG13, and the pangolin-infecting Pangolin-CoV and P2V, P5L, P1E, P5E, P4L cluster. Note, some recombinant regions of the first three viruses do not belong to the nCoV lineage.

We tested for evidence of episodic diversifying selection on the internal branches of the nCoV lineage using BUSTED[S], accounting for synonymous rate variation (SRV) as described in Wisotsky et al. (2020)[35]. We developed an extension to BUSTED[S], that included a hidden Markov model (HMM) with three rate categories to describe site-specific synonymous rate variation (SRV)[36]. This HMM allows explicit incorporation of autocorrelation in synonymous rates across codons. This autocorrelation would be expected if selection or mutation rate variation were spatially localised within ORFs. The rate switching parameter between adjacent codons of the HMM describes the extent of autocorrelation, with values under 1/N (N = number of rate classes) suggestive of autocorrelation. Standard HMM techniques (e.g. the Viterbi path) applied to these models can reveal where the switches between different rate types occur, thereby

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

partitioning the sequence into regions of weaker or stronger constraint on synonymous substitutions.

aBSREL method[39] was used on all branches of the nCoV lineage to determine which specific branches drive the inference of selection. Finally, we examined which specific codon sites are under negative selection on average over the nCOV lineage using FEL[34], and under pervasive or episodic diversifying positive selection on the nCoV lineage using MEME[38]. P-values of ≤0.05 for the likelihood ratio tests, specific to each method, were taken as evidence of statistical significance. All selection analyses were performed in the HyPhy software package v.2.5.14[28].

CpG depletion. To quantify over/under representation of CpG dinucleotides in the *Sarbecovirus* genomes we developed a modified version of the Synonymous Dinucleotide Usage (SDU) metric[43], which now accounts for biased base composition. The original SDU metric compares the observed proportion of synonymous CpG, *o*, for each pair of frame positions, *h*, in a coding sequence to that expected under equal synonymous codon usage, *e*, for each amino acid (or amino acid pair), *i*, that can have CpG containing codons (or codon pairs). The SDU metric is the mean of these ratios weighted by the number of informative amino acids (or pairs), *n*, in the sequence (Equation 1).

To incorporate the biased, and variable base composition of SARS-CoV-2 and other *Sarbecoviruses*[22], here we have estimated expected codon usage based on each virus's whole-genome nucleotide composition. We term this new metric the corrected Synonymous Dinucleotide Usage (SDUc). We use observed base frequencies from each virus to generate the corrected null expectation of the metric, *e'*, instead of assuming equal usage, (Equation 1). The expected proportion, *e'*, for every amino acid / amino acid pair was estimated by randomly simulating codons based on the whole-genome single nucleotide proportions of each virus. This *e'* was then used for all SDUc calculations of the corresponding virus.

As this metric is susceptible to error when used for short coding sequences, we applied SDUc on the longest ORF, Orf1ab, of all the viruses. To estimate the extent of phylogenetic independence between synonymous sites across SDUc datapoints, we measured the pairwise synonymous divergence (Ks) between viruses. Pairwise Ks values were calculated using the seqinr R package[54] which utilises the codon model of Li (1993)[55], demonstrating the partial but

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not complete independence within the two lineages. The Ks median and maximum is 0.54 and 0.89 within the nCoV lineage, and 0.34 and 1.09 respectively within the non-nCoV lineage.

$$SDUc_{CpG,h} = \frac{\sum_{i=1}^{k} n_i \times \frac{o_{i,h}}{e'_{i,h}}}{N}$$

(Equation 1; N = total number of amino acids)

Spike recombination analysis. To determine the recombination breakpoint on the Spike ORF of the RmYN02 virus we used the RDP5 method suite[56], implementing seven methods: RDP, GENECONV, Chimaera, MaxChi, BootScan, SiScan and 3seq. We first performed the analysis on the whole-genome alignment of the *Sarbecoviruses* and then determined the relevant breakpoint within the Spike ORF by rerunning the method on the Spike-only alignment. The accepted breakpoint (position 24058 in the RmYN02 genome) was consistently called by six out of the seven tested methods (RDP, GENECONV, Maxchi, Chimaeara, SiSscan, 3seq).

Acknowledgements

We would like to thank all the authors who have kindly deposited and shared genome data on GISAID. A table with genome sequence acknowledgments can be found in the supplementary material. We thank Joseph Hughes for thankful comments on the manuscript. DR and SL are funded by the MRC (MC_UU_1201412). SLKP and SW are supported in part by the NIH (R01 AI134384 (NIH/NIAID)) and the NSF (award 2027196).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).

2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

3. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* vol. 5 536–544 (2020).

4. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).

5. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

6. Lam, T. T. Y. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2169-0.

7. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *bioRxiv* 2020.03.30.015008 (2020) doi:10.1101/2020.03.30.015008.

8. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. *Science (80-.).* **302**, 276–278 (2003).

9. Song, H. D. *et al.* Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2430–2435 (2005).

10. Graham, R. L. & Baric, R. S. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* **84**, 3134–3146 (2010).

11. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

12. Ge, X. Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

13. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Lin, X. D. *et al.* Extensive diversity of coronaviruses in bats from China. *Virology* **507**, 1–10 (2017).

15. Verity, R. *et al.* Articles Estimates of the severity of coronavirus disease 2019: a model-based analysis. (2020) doi:10.1016/S1473-3099(20)30243-7.

16. Grubaugh, N. D., Petrone, M. E. & Holmes, E. C. We shouldn't worry when a virus mutates during disease outbreaks. *Nature Microbiology* vol. 5 529–530 (2020).

17. Maclean, O. A., Orton, R. J., Singer, J. B. & Robertson, D. L. No evidence for distinct types in the evolution of SARS-CoV-2. doi:10.1093/ve/veaa034.

18. Van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. doi:10.1101/2020.05.21.108506.

19. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020.04.29.069054 (2020) doi:10.1101/2020.04.29.069054.

20. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

21. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).

22. Simmonds, P. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. *bioRxiv* 2020.05.01.072330 (2020) doi:10.1101/2020.05.01.072330.

23. Pybus, O. G. *et al.* Phylogenetic Evidence for Deleterious Mutation Load in RNA Viruses and Its Contribution to Viral Evolution. doi:10.1093/molbev/msm001.

24. Duchêne, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B Biol. Sci.* **281**, 20140732 (2014).

25. Ewens, W. J. The probability of survival of a new mutant in a fluctuating environment. *Heredity (Edinb).* **22**, 438–443 (1967).

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

27. Murrell, B. *et al.* FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.* (2013) doi:10.1093/molbev/mst030.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

28. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. doi:10.1093/molbev/msz197.

29. Kosakovsky, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. BIOINFORMATICS APPLICATIONS NOTE GARD: a genetic algorithm for recombination detection. **22**, 3096–3098 (2006).

30. Dudas, G. & Rambaut, A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. doi:10.1093/ve/vev023.

31. Anthony, S. J. *et al.* Further evidence for bats as the evolutionary source of middle east respiratory syndrome coronavirus. *MBio* **8**, (2017).

32. Wong, M. C., Cregeen, S. J. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* **2013**, 2020.02.07.939207 (2020).

33. Li, X. *et al.* Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. *bioRxiv* 2020.03.20.000885 (2020) doi:10.1101/2020.03.20.000885.

34. Kosakovsky, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. doi:10.1093/molbev/msi105.

35. Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D. & Muse, S. V. Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. doi:10.1093/molbev/mst012.

36. Felsenstein, J. & Churchill, G. A. *A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution*. https://academic.oup.com/mbe/article-abstract/13/1/93/1055515.

37. Forni, D. *et al.* Extensive Positive Selection Drives the Evolution of Nonstructural Proteins in Lineage C Betacoronaviruses. *J. Virol.* **90**, 3627–3639 (2016).

38. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).

39. Smith, M. D. *et al.* Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. doi:10.1093/molbev/msv022.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40. Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadan, R. Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLoS Pathog* **4**, 1000079 (2008).

41. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124–127 (2017).

42. Bishop, K. N., Holmes, R. K., Sheehy, A. M. & Malim, M. H. APOBEC-mediated editing of viral RNA. *Science (80-.).* **305**, 645 (2004).

43. Lytras, S. & Hughes, J. Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses. *Viruses* **12**, 462 (2020).

44. Zhou, H., Chen, X., Hughes, A. C., Bi, Y. & Shi, W. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* (2020) doi:10.1016/j.cub.2020.05.023.

45. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* vol. 11 283–290 (1995).

46. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virologica Sinica* vol. 33 104–107 (2018).

47. Li, H. *et al.* Human-animal interactions and bat coronavirus spillover potential among rural residents in Southern China. *Biosaf. Heal.* **1**, 84–90 (2019).

48. Anderson, D. E. *et al.* Lack of cross-neutralization by SARS patient sera towards SARS-CoV-2. *Emerg. Microbes Infect.* **9**, 900–902 (2020).

49. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma. Appl.* **30**, 1312–1313 (2014).

50. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in 199–213 (Springer, New York, NY, 1998). doi:10.1007/978-1-4612-1694-0_15.

51. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).

52. Katoh, K. & Standley, D. M. Article Fast Track MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. doi:10.1093/molbev/mst010.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.122366; this version posted May 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

53. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. doi:10.1093/nar/gkl315.

54. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. in 207–232 (Springer, Berlin, Heidelberg, 2007). doi:10.1007/978-3-540-35306-5_10.

55. Li, W. H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* vol. 36 96–99 (1993).

56. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. doi:10.1093/ve/vev003.