Intra-genome variability in the dinucleotide composition of SARS-CoV-2

Citation for published version:
Digard, P., Lee, H-M., Sharp, C., Grey, F. & Gaunt, E. 2020, 'Intra-genome variability in the dinucleotide composition of SARS-CoV-2', Virus Evolution. https://doi.org/10.1093/ve/veaa057

Digital Object Identifier (DOI):
10.1093/ve/veaa057

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Virus Evolution

Publisher Rights Statement:
©The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Intra-genome variability in the dinucleotide composition of SARS-CoV-2

Paul Digard ¹, Hui Min Lee ¹, Colin Sharp ¹, Finn Grey ¹, Eleanor Gaunt ¹,*

¹The Roslin Institute, The University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK

*To whom correspondence should be addressed. Tel: 0044 131 6519 152; Email: Elly.Gaunt@ed.ac.uk
Intra-genome variability in the dinucleotide composition of SARS-CoV-2

ABSTRACT

CpG dinucleotides are under-represented in the genomes of single stranded RNA viruses, and SARS-CoV-2 is no exception to this. Artificial modification of CpG frequency is a valid approach for live attenuated vaccine development; if this is to be applied to SARS-CoV-2, we must first understand the role CpG motifs play in regulating SARS-CoV-2 replication. Accordingly, the CpG composition of the SARS-CoV-2 genome was characterised. CpG suppression amongst coronaviruses does not differ between virus genera, but does vary with host species and primary replication site (a proxy for tissue tropism), supporting the hypothesis that viral CpG content may influence cross-species transmission. Although SARS-CoV-2 exhibits overall strong CpG suppression, this varies considerably across the genome, and the Envelope (E) open reading frame (ORF) and ORF10 demonstrate an absence of CpG suppression. Across the Coronaviridae, E genes display remarkably high variation in CpG composition, with those of SARS and SARS-CoV-2 having much higher CpG content than other coronaviruses isolated from humans. This is an ancestrally-derived trait reflecting their bat origins. Conservation of CpG motifs in these regions suggests that they have a functionality which over-rides the need to suppress CpG; an observation relevant to future strategies towards a rationally attenuated SARS-CoV-2 vaccine.

INTRODUCTION

CpG dinucleotides are under-represented in the DNA genomes of vertebrates (Cooper and Krawczak, 1989; Simmonds et al., 2013). Cytosines in the CpG conformation may become methylated, and this methylation is used as a mechanism for transcriptional regulation (Medvedeva et al., 2014). Methylated cytosines have a propensity to undergo spontaneous deamination (and so conversion to a thymine). Over evolutionary time, this has reduced the frequency of CpGs in vertebrate genomes (Cooper and Krawczak, 1989). However, loss of CpGs in promoter regions would affect transcriptional regulation, and so CpGs are locally retained, resulting in functionally important ‘CpG islands’ found in around half of all vertebrate promoter regions (Deaton and Bird, 2011).

Single strand RNA (ssRNA) viruses infecting vertebrate hosts reflect the CpG dinucleotide composition of their host in a type of mimicry (Simmonds et al., 2013). It was hypothesised that this is because vertebrates have evolved a CpG sensor which flags transcripts with aberrant CpG frequencies (Atkinson et al., 2014; Gaunt et al., 2016). This idea was strengthened by the discovery that the cellular protein Zinc-finger Antiviral Protein (ZAP) binds CpG motifs on viral RNA and directs them for degradation (Takata et al., 2017), and further supported by observations that CpGs can be synonymously introduced into a viral genome to the detriment of virus replication without negatively
impacting transcriptional or translational efficiency (Gaunt et al., 2016; Tulloch et al., 2014). Current understanding is therefore that ssRNA viruses mimic the CpG composition of their host at least in part to subvert detection by ZAP. ssRNA viruses also under-represent the UpA dinucleotide, but to a far more modest extent (Simmonds et al., 2013), and the reasons behind UpA suppression are less well understood. A consequence of dinucleotide bias is that certain codon pairs are under-represented (Kunec and Osterrieder, 2016; Tulloch et al., 2014) (so, for example, codon pairs of the conformation NNC-GNN are among the most rarely seen codon pairs in vertebrates (Tats et al., 2008)). Whether the two phenomena of CpG suppression and codon pair bias (CPB) are discrete remains controversial (Futcher et al., 2015; Groenke et al., 2020; Kunec and Osterrieder, 2016).

The Coronaviridae have a generally low genomic cytosine content (Berkhout and van Hemert, 2015), but as with other ssRNA viruses, nonetheless still under-represent CpG dinucleotides to a frequency below that predicted from individual base frequencies of cytosine and guanine (Woo et al., 2007). The Coronavirus family comprises four genera – the alpha, beta, gamma and delta-coronaviruses. Human-infecting coronaviruses (HCoVs) have been identified belonging to the alpha and beta genera (Hu et al., 2015). Alphacoronaviruses infecting humans include HCoV-229E and the more recently discovered HCoV-NL63 (van der Hoek et al., 2004). Betacoronaviruses include HCoV-OC43, HCoV-HKU1 (Woo et al., 2005), severe acute respiratory syndrome (SARS)-CoV (Rota et al., 2003), Middle East respiratory syndrome (MERS)-CoV (Zaki et al., 2012) and the recently emerged SARS-CoV-2 (Lu et al., 2020; Zhu et al., 2020). Prior to the emergence of SARS-CoV-2, SARS-CoV had the strongest CpG suppression across human-infecting coronaviruses (Woo et al., 2007). The reason(s) for this are uncertain, but loss of CpG from a virus genome upon zoonotic transfer into the human host has previously been reported for influenza A virus (Greenbaum et al., 2008), potentially indicating an advantage of reduced CpG content for infection of the human respiratory tract. All human-infecting coronaviruses are thought to be derived from ancestral bat viruses, though intermediate hosts may have facilitated zoonotic passage in some cases (Banerjee et al., 2019).

During replication, coronaviruses synthesise transcriptionally active negative sense sub-genomic RNAs which are of varying length. Sub-genomic RNAs are synthesised by the viral polymerase copying the genome up to a 5' leader sequence (Liao and Lai, 1994) which is repeated upstream of most open reading frames (ORFs) in the coronavirus genome (such repeats are referred to as transcription regulation sequences (TRSS)); this complementarity allows viral polymerase jumping from the 5' leader sequence to directly upstream of ORFs preceded by a TRS (Sawicki and Sawicki, 1998). The negative sense sub-genomic RNAs serve as efficient templates for production of mRNAs (Sawicki et al., 2007). Generally, only the first ORF of a sub-genomic mRNA is translated (Perlman and Netland, 2009), although leaky ribosomal scanning has been reported as a means for accessing alternative ORFs for several coronaviruses including SARS-CoV (Schaecher et al., 2007).

SARS-CoV-2 was recently reported to have a CpG composition lower than other members of the betacoronavirus genus, comparable to certain canine alphacoronaviruses; an observation used to draw inferences over its origin and/or epizootic potential (Xia, 2020). Here we show that
coronaviruses have a broad range of CpG composition which is partially host and tissue tropism dependent, and that there is no difference in CpG content across coronavirus genera. There is however a striking disparity in CpG composition between SARS-CoV-2 ORFs, with the Envelope (E) protein ORF and ORF10 over-representing CpG dramatically. E ORF and ORF10 also have higher UpA dinucleotide composition and lower CPB scores than other ORFs. E ORF displays CpG suppression in all human-infecting viruses except SARS-CoV and SARS-CoV-2, suggesting a potential correlation between CpG presentation and disease severity in human-infecting coronaviruses.

MATERIALS AND METHODS

Sequences.

For a comparison of GC content versus CpG ratio, all SARS-CoV-2 complete genome sequences of high coverage (as defined on the GISAID website) were downloaded from GISAID (www.gisaid.org) on 26 March 2020 (1163 sequences in total) and aligned against the SARS-COV-2 reference sequence (Accession number NC_045512) using Simmonics software (Simmonds, 2012) SSE v1.4 (pre-release download kindly provided by Prof. Peter Simmonds, Oxford University). All sequences represented human isolates except for one sequence of bat origin (hCoV-19/bat/Yunnan/RaTG13/2013; EPI_ISL_402131) and one sequence from a pangolin (hCoV-19/pangolin/Guangdong/1/2019; EPI_ISL_410721). All complete genome sequences of all coronaviruses were downloaded from NCBI on the 16 April 2020 (3407 sequences in total). Sequences were then aligned and sequences less than 10% divergent at the nucleotide level, identified using the ‘identify similar/identical sequences’ function in SSE v1.4 were removed from the dataset. Sequences were annotated into animal groups and genera based on their description in the NCBI database. The trimmed dataset (Table S1) included 215 complete genome coronavirus sequences. Individual groups were made for sequences originating from the following hosts: bat (n = 109), avian (35), camelid (3), canine (7), feline (9), human (7), mustelids (5), rodents (8), swine (13), ungulates (10) and ‘other’ (which included bottle-nosed dolphin (2), hedgehog (2), rabbit (2), beluga whale (1), civet (1) and pangolin (1)). Groups were loosely defined based on taxonomic orders, with some exceptions made to examine our specific research questions. Bats are of the order Chiroptera; multiple avian orders were grouped together (Galliformes, Anseriformes, Passeriformes, Gruiformes, Columbiformes and Pelicaniformes); even toed (Artiodactyla) and odd toed (Perissodactyla) ungulate orders were grouped, with camelids analysed separately due to their association with MERS-CoV (Azhar et al., 2014); Canidae (canine) and Pantherinae (feline) sequences of the Carnivora order were analysed separately, as canines have previously been suggested as an intermediate host species for SARS-CoV-2 (Xia, 2020) and cat infections with SARS-CoV-2 have been reported (Shi et al., 2020); humans were the only representatives from the Primate order; all remaining Carnivora, with the exception of a single civet sequence, belonged to the Mustelidae (mustelids); rodents belong to the Rodentia order; and swine belong to the Artiodactyla order; whales are also Artiodactyla but swine

Digard et. al.  https://mc.manuscriptcentral.com/vevolu
were considered separately due to considerable interest in porcine coronaviruses (Vlasova et al., 2020). Sequences were also annotated for genus by reference to the NCBI description (203 of the 215 sequences were assigned to a genus), and for primary replication site by literature reference (refer to Table S1). Replication site annotations were based on the sample type from which a coronavirus sequence was obtained – ‘enteric’ for faecal/ gastrointestinal samples, ‘respiratory’ for nasal, oropharyngeal and other respiratory samples; ‘multiple’ if samples from multiple systems tested positive, ‘other’ if the sample was collected from a site not falling into the enteric or respiratory categories (e.g. brain), or ‘unknown’ if a sample type could not be determined. If only one sampling route was tested and returned a positive result, the sequence was categorised in accordance with the sole sampling route. The sequence datasets used in this paper are summarised in Fig. 1.

**Analyses of dinucleotide content.**

CpG and UpA composition of complete genomes or of individual ORFs were calculated using the composition scan in SSE v1.4. CpG frequencies were measured as observed: expected (O:E) ratios, using the formula $f(CpG) / f(C)*f(G)$. Individual ORFs were identified using a combination of ORF finder ([https://www.ncbi.nlm.nih.gov/orffinder/](https://www.ncbi.nlm.nih.gov/orffinder/)), visual inspection of nucleotide alignments in SSE v1.4, comparison with previous literature and information available from nextstrain.org. Sliding window analyses were performed on the 1163 aligned SARS-CoV-2 sequences and the related bat and pangolin sequences by performing composition scans in SSE v1.4 for 100 nucleotide genomic regions, at 25 nucleotide iterations. For the SARS-CoV-2 sequences, mean CpG O:E ratios for each window were calculated. CPB (Gutman and Hatfield, 1989) scores across the SARS-CoV-2 ORFeome were calculated using the SSE v1.4 composition scan function. Individual ORFs were concatenated with a separating ‘NNN’ codon for analysis, and secondary overlapping ORFs were not included due to coding constraints imposed in these regions.

To examine the extent of CpG retention in E ORF, the same analyses were performed with an additional correction for amino acid composition (Corr_CpG dataset produced by SSE v1.4).

**Codon usage analysis.**

To examine the use of rare codons, codon adaptation index values were calculated ([https://www.biologicscorp.com/tools/CAICalculator](https://www.biologicscorp.com/tools/CAICalculator)).

**Phylogenetic analyses.**

Of the 215 divergent sequences included in the analysis, E ORF could be identified in 178 by homology with E ORFs previously annotated in NCBI. Of these 178 E ORFs, 7 were sequences isolated from humans and 96 were from bats; these sequences were selected for analysis. E ORFs were aligned in MEGA X (Kumar et al., 2018) using the Clustal method. Phylogenetic reconstruction was performed using an unrooted maximum likelihood tree, with gamma distributed variation in rates between branches and 100 bootstraps (also in MEGA X).
RESULTS

CpG suppression within coronavirus genomes varies between host species and tissue tropism but not between genera.

The genomic CpG composition of all complete genome coronavirus sequences ($n = 3407$; downloaded and further processed as described in the methods section and Fig. 1) were calculated using observed: expected (O:E) ratios, with any value below 1 indicating CpGs are under-represented relative to the genomic content of cytosine and guanine bases. A substantial range in GC content (from ~ 0.32 – 0.47) was seen across the Coronaviridae, and as expected, all viruses exhibited some degree of CpG suppression, with CpG O:E ratios ranging from 0.37 to 0.74 (Fig. 2A). To investigate the root of this variation, the coronavirus sequence dataset was refined to remove sequences with more than 90% nucleotide identity to reduce sampling biases (so, for example, SARS-CoV sequences of human origin were stripped from over 1000 representative sequences to just one). The CpG compositions of the remaining 215 sequences (Table S1) were compared between coronavirus genera (alpha, beta, gamma and delta). For the 215 representative sequences, a genus could be assigned for 203. No differences in CpG composition between coronavirus genera were apparent, although the gamma genus exhibited a tighter range (Fig. 2B). Next, we examined whether differences in CpG composition between viruses isolated from different hosts explained the range in CpG composition across the Coronaviridae. For the 215 representative sequences, a host could be assigned to 210. Coronavirus sequences were divided into host groups, and groups with at least three divergent sequences were compared; this included bat, avian, camelid, canine, feline, human, mustelid, rodent, swine and ungulate viruses. Variation in CpG composition between coronaviruses detected in different host species was evident across and between groups, with coronaviruses detected in canine and human species having lower CpG content and rodent and bat coronaviruses having the highest (Fig. 2C). All frequency ranges overlapped however, indicating viral CpG frequency alone seems to be a poor predictor of virus origin, contradicting the recent suggestion of a canine origin of SARS-CoV-2 (Xia, 2020). Where sequences in a host group representative of both alpha and betacoronaviruses were available (which was the case for bat, camelid, canine, human, rodent and swine viruses), these sequences were split by genus and compared to determine whether coronavirus genera influenced coronavirus CpG frequencies in a host species-specific manner. By this method, the lack of difference in CpG composition of coronaviruses of different genera was maintained (Fig. 2D).

To test the hypothesis that coronavirus CpG content varies according to tissue tropism (Xia, 2020), we classified the viruses according to their primary site of replication, where this was known or could be inferred from the sampling route. Samples were split into five categories – ‘respiratory’, ‘enteric’, ‘multiple’, ‘other’, or ‘unknown’. Altogether, 206 of the 215 sequences were classifiable (detailed in Table S1), with 9 sequences categorised as ‘unknown’ and excluded from further analyses. By this admittedly inexact approach, viruses infecting the respiratory tract had a lower mean CpG composition than viruses with enteric tropism (Fig. 2E). However, the spread of respiratory virus CpG frequencies was contained entirely within the range exhibited by enteric viruses. Furthermore, 124
sequences were assigned to the enteric group, and only 22 to the respiratory group. Of these 146 sequences, bat viruses accounted for 80, all of which were assigned to the enteric group (despite reasonable sampling of respiratory tract in bats) and this cohort of viruses maintained almost the full spread of CpG frequencies (Fig. 2E, Table S1). Thus, while coronavirus CpG frequency may show some correlation with replication site, the dataset available does not permit strong conclusions to be drawn or predictions about zoonotic potential to be made.

**Heterogeneities in the dinucleotide composition of SARS-CoV-2.**

By our methods for calculating CpG O:E ratios, SARS-CoV-2 has a genomic CpG ratio of 0.408 (representing the mean of 1163 complete genome sequences). This is similar to the value calculated previously for a much smaller sample \( (n = 5) \) of SARS-CoV-2 sequences (Xia, 2020). As this previous study noted, this is at the bottom end of the range of genomic CpG O:E ratios for betacoronaviruses and for coronaviruses detected in humans (Figs 2B, C and D). However, as noted above, vertebrate DNA genomes contain localised islands of higher CpG content (Deaton and Bird, 2011). To determine if similar heterogeneity in CpG frequency was evident in the SARS-CoV-2 genome, the composition of individual ORFs was examined. Overall, most ORFs had CpG O:E ratios which were comparable to the genomic CpG ratio. However, two ORFs in particular, E ORF and ORF10, had CpG ratios higher than 1, indicating an absence of CpG suppression in those regions (Fig. 3A). These two ORFs also did not suppress the UpA dinucleotide, in contrast with other SARS-CoV-2 ORFs (Fig. 3B).

Due to the difficulties in distinguishing between dinucleotide bias and CPB, CPB scores were also calculated for each ORF and plotted against CpG composition (Fig. 3C). CPB scores provide an indication of whether the codon pairs encoded in each ORF are congruous with usage in vertebrate genomes. A score below 0 indicates use of codon pairs that are disfavoured in host ORFs. An approximately linear relationship between CpG O:E ratio and CPB score for each SARS-CoV-2 ORF was apparent \( (R^2 = 0.80) \). E ORF and ORF10 both had negative CPB scores, indicating that they use under-represented codon pairs and in keeping with the observation that both ORFs over-represent CpG and UpA dinucleotides.

To examine the precise location of the CpG hotspots, a sliding window analysis of CpG content across the 3’ end of the SARS-CoV-2 genome (averaged over 1163 complete genome sequences) as well as the closely related bat and pangolin sequences was performed. As expected, marked increases in CpG O:E ratio were observed concomitant with the genomic regions associated with E ORF and ORF10 (Fig. 3D). The E ORF and ORF10 regions associated with high CpG composition were maintained across the bat, pangolin and human sequences, indicating that since the bat sample was collected in 2013, the higher CpG frequency in this region has not been negatively selected. While the increase in CpG presentation was apparent across the entire E ORF, starting at the 3’ end of ORF3 and ending at the beginning of the M gene, the CpG spike in ORF10 was more narrowly associated with the putative coding region. Additionally, a CpG spike between the 3'-'end of ORF8 and the 5'-'end of the N gene was evident. The 5'-'end of the N ORF also contains the overlapping ORF9b
gene, which when considered alone, has a CpG O:E ratio approaching 1 (Fig. 3A), and is the ORF with the third-highest CpG O:E ratio after E ORF and ORF10. The usual coding plasticity afforded to nucleotides in the third position of a codon is nullified when overlapping reading frames are present, and so the CpG spike at this gene boundary is not surprising. Thus, although the SARS-CoV-2 genome exhibits high CpG suppression overall, there are local heterogeneities associated with individual ORFs, most notably E.

On the origins of the high CpG content of E ORF of SARS-CoV-2.

To determine whether the high CpG content of E ORF is evolutionarily conserved (ORF10 is poorly conserved and only encoded by a subset of SARS-like coronaviruses, so it was not analysed), attempts to identify the E ORF by nucleotide alignment for the set of 215 coronavirus sequences was undertaken, compared with E ORFs already annotated in NCBI. Of the 215 sequences, E ORF was identifiable in 178, with the remaining sequences too divergent to be confident of gene assignment. CpG composition for E ORF for the 178 sequences was measured and plotted according to host. Amino acid conservation within the short ORF of E could bias levels of CpG; for example, amino acids encoded by codons containing C and G in combinations other than CpG could be disproportionately represented. To account for this possibility, CpG O:E ratios were corrected for amino acid composition across this region (Fig. 4A). A diverse distribution of CpG content was evident in viruses from every host group except ungulates, with bats in particular displaying a notable range from total suppression to overrepresentation. Otherwise, most viruses from most species still maintained some level of CpG suppression in E ORF. The exceptions with high CpG O:E ratios in E ORF were avian coronaviruses and notably, SARS-CoV and SARS-CoV-2. In contrast, other human-infecting coronaviruses (HCoV-229E, HCoV-HKU1, HCoV-NL63 and HCoV-OC43) all strongly under-represented CpG in E ORF, while MERS-CoV E ORF had an intermediate CpG O:E ratio of 0.6. To confirm E ORF over-represented CpG relative to the rest of the genome in SARS-CoV and SARS-CoV-2, ratios for E ORF: genomic CpG O:E were calculated (Fig. 4B). In non-bat non-avian host genomes, E ORF usually displayed CpG suppression in line with or stronger than that seen at the genome level, whereas SARS-CoV and SARS-CoV-2 starkly contrasted with this, displaying far less CpG suppression in this region. This could be linked with their recent emergence from bat reservoirs, as genome composition is more likely to be optimised for replication in that host, and the CpG composition of E ORF for both SARS-CoV-2 and SARS-CoV falls within the E ORF CpG heterogeneity apparent across bat-derived sequences.

As another check of whether differential codon usage might explain the CpG composition disparity in E ORF, we calculated codon adaptation index (CAI) scores baselined against the human transcriptome. While SARS-CoV and SARS-CoV-2 E ORFs had CAI scores that were lower than those for other coronavirus E ORFs, the differences were small and did not explain the large differences in CpG ratios (Table 1).

To investigate the evolutionary history of E ORF CpG composition in the human-infecting coronaviruses, a phylogenetic reconstruction of all 7 human coronavirus and 96 bat coronavirus E
genes was performed to determine whether CpG ratios in this region were ancestrally derived. As expected (Cotten et al., 2013; Lu et al., 2020), the human viruses were interspersed among the bat viruses, reflective of their independent emergence events (Fig. 4C). The CpG compositions of the human coronavirus E ORFs, although diverse, were similar to the CpG compositions of their phylogenetically proximal bat relatives, demonstrating that CpG composition in E ORF is an ancestrally derived trait selected prior to emergence in the human population.

**DISCUSSION**

We have examined the CpG O:E ratios of all the currently available complete genome sequences of coronaviruses and uncovered a noteworthy diversity. Generally, the CpG O:E ratio of coronavirus genomes from a single host species varied considerably. For bats, which serve as a coronavirus reservoir (Banerjee et al., 2019) and which had the largest number of representative sequences, the CpG O:E range was from 0.41 to 0.70, demonstrating the genome plasticity of coronaviruses and indicating that their evolution is not overtly restricted by a requirement to minimise CpG composition in the natural reservoir. The antiviral CpG-detector protein, ZAP (Takata et al., 2017), has been identified as a target for several viral proteins including the 3C protease of enterovirus 71 (Xie et al., 2018) and NS1 of influenza A virus (Tang et al., 2017) – two viruses with overall low CpG content (Atkinson et al., 2014; Gaunt et al., 2016). This highlights the importance of CpG as a pathogen-associated molecular pattern (PAMP), and so this diversity in CpG expression within the **Coronaviridae** is striking. If coronaviruses also produce a protein with anti-ZAP activity, it is possible that this has variable efficacy between strains, explaining the ability of coronaviruses to fluctuate CpG composition considerably. Alternatively (or in addition), this may be host driven; we show that average CpG suppression varies with host species (Fig. 2C) and, as previously suggested (Xia, 2020), this may be linked with ZAP expression levels. We have demonstrated that CpG variation is not related to viral taxonomic grouping (Fig. 2B) but we did find an association between viral CpG composition and primary replication site, with respiratory coronaviruses having a lower CpG composition than enteric ones (Fig. 2E). This is the opposite of what has been previously suggested (Xia, 2020), though this proposal was not supported by any comprehensive investigation. Nevertheless, our meta-analysis was subject to the sampling preferences of many labs who have performed surveillance for coronaviruses and many of the tissue tropism assignments we made have not been verified by experimental infections. Another limitation of this analysis is that only sequences of greater than 10% divergence were included, and while this overcomes some sampling bias we cannot assume that datapoints are independent (which is why statistical comparisons are not included). Notably, tissue tropism can be defined by much smaller divergences; for example, a deletion in the spike protein of transmissible gastroenteritis virus (a porcine coronavirus) altered the tropism of the virus from enteric to respiratory, while nucleotide identity was preserved at 96% (Cox et al., 1990; Rasschaert et al., 2017).
1990). Further study on tissue tropisms of coronaviruses, as well as tissue expression profiles and antiviral activities of ZAP are needed to validate these analyses.

Loss of CpG motifs during adaptation to the human host has been previously described for influenza A virus (Greenbaum et al., 2008), highlighting the importance of CpG composition for host adaptation. For SARS-CoV-2, we determined a genomic CpG O:E ratio of 0.408, which is similar to the human genome CpG O:E ratio of 0.2-0.4 (McClelland and Ivarie, 1982; Sved and Bird, 1990; Tomso and Bell, 2003). Mimicry of the CpG composition of the host by ssRNA viruses is considered a mechanism to subvert detection by the innate immune response (Simmonds et al., 2013; Takata et al., 2017) and speculatively this may indicate that SARS-CoV-2 was genetically predisposed to make a host switch into humans. Similarly, the genomic CPB score of 0.048 indicates that SARS-CoV-2 uses codon pairs which are preferentially utilised in the human ORFeome, which may mean that the virus was well suited for translational efficiency in humans at its time of emergence.

In coding regions which do not have overlapping ORFs, there is no requirement at the coding level for CpG motifs to be retained (Kanaya et al., 2001). E ORF and ORF10 are not known to be in overlapping reading frames; conversely, ORF9b overlaps with the ORF for nucleocapsid (N). Some CpG retention in this region is therefore inevitable and may explain the high CpG composition of ORF9b. This nevertheless leaves open the question of why CpG motifs are retained in the E ORF and ORF10 regions (if this is not an ancestrally derived evolutionary hangover; as CpGs have not been lost from these regions between 2013 and now (Fig. 3D), this seems unlikely). CpG motifs may serve various non-exclusive purposes, including providing secondary structure (Rima and McFerran, 1997), intentionally stimulating ZAP activity (by analogy with multiple viruses intentionally triggering NF-kB (Hiscott et al., 2001)), or providing m5c methylation sites (Dev et al., 2017; Khoddami and Cairns, 2013; Squires et al., 2012).

It is also possible that CpG enrichment serves as a strategy for regulating translation. Conceivably, the high CpG content at the 5’ end of the E ORF transcript destines this for degradation via ZAP or CBP-associated mechanisms (Groenke et al., 2020; Guo et al., 2007) more rapidly than other viral transcripts. This could be intentional, or an evolutionarily accepted trade-off to preserve a higher importance role for CpGs. Alternatively, E ORF and ORF10 proteins may only be required late during infection (parallels with which can be drawn from the differential temporal expression and translational efficiencies of transcripts of the coronavirus mouse hepatitis virus strain A59 (Irigoyen et al., 2016)), by which time an as-yet unidentified inhibitor of ZAP (or other CpG/CBP sensor(s)) may render CpG suppression unnecessary, as suggested for human cytomegalovirus (Lin et al., 2020).

ORF9b and ORF10 do not have their own TRSs and so whether or how these open reading frames are accessed is currently controversial; nevertheless, peptides from both have been identified by mass spectrometry from SARS-CoV-2 infected cells (Davidson et al., 2020). The ORF9b AUG transcription initiation site, which has a strong Kozak context (Kozak, 1986), is the first AUG after and 10 nucleotides downstream of the initiation site for N ORF (which displays moderate Kozak context). It is therefore credible to think that ORF9b is accessed via leaky ribosomal scanning - a well
characterised method for accessing alternative ORFs used by coronaviruses and other viruses (Chenik et al., 1995; Firth and Atkins, 2010; Irigoyen et al., 2016; Lin and Lo, 1992; O'Connor and Brian, 2000; Ryabova et al., 2006; Schneider et al., 1997; Senanayake and Brian, 1997; Wise et al., 2011). There is a lack of evidence that ORF10 is accessed via production of its own subgenomic RNA (Kim et al., 2020); possibly, this ORF is accessed via leaky scanning from the leader immediately preceding the N ORF. However, visual inspection of the SARS-CoV-2 genome indicated that the AUG encoding ORF10 is 24 AUGs downstream from the one initiating N ORF, making this hypothesis speculative at best. Whether the anomalous CpG composition of ORF10 is somehow involved in priming its transcription remains to be determined.

The transcript encoding E ORF incorporates an additional ~3.4kb of RNA and ORF10, if accessed from the transcript produced from the TRS upstream of N ORF, is present on a transcript of approximately 1.6kb in length. Whether the described CpG enriched regions are relevant as PAMPs in these contexts is currently unclear from what is known about ZAP recognition of CpG motifs. It is also worth noting that the body TRS sequence ahead of the E gene is relatively weak in SARS-CoV-2, as it is in SARS-CoV (Marra et al., 2003), suggesting that this subgenomic mRNA may be of relatively low abundance. Of the SARS-CoV-2 transcripts which use a canonical TRS for synthesis, the donor site upstream of E ranked seventh when comparing sequencing read frequency across this site (behind reads spanning the TRS sites upstream of N, spike, ORF7a, ORF7b, ORF3a, ORF8 and M ORF respectively) in Vero cells infected at a low MOI for 24 hours, indicating that E ORF is of lower abundance than most other transcripts (Kim et al., 2020). It is therefore possible that E ORF is of sufficiently low abundance for a high CpG frequency to be physiologically inconsequential. Similar logic can be applied to ORF10, which is just 117 nucleotides in length.

Synonymous addition of CpGs into a virus genome has been suggested as a potential novel approach to vaccine development by us and others (Atkinson et al., 2014; Burns et al., 2009; Gaunt et al., 2016; Moratorio et al., 2017). Here we explore the evolutionary space occupied by coronaviruses in the context of their CpG composition and find that SARS-CoV-2 has a low CpG composition in comparison with other coronaviruses, but with CpG ‘hotspots’ in genomically disparate regions. This highlights the potential for large scale recoding of the SARS-CoV-2 genome by introduction of CpGs into multiple regions of the virus genome as a mechanism for generation of an attenuated live vaccine. Introduction of CpG into multiple sites could also be used to subvert the potential of the virus to revert to virulence through recombination. A challenge of live attenuated vaccine manufacture is to enable sufficient production of a vaccine virus that has a replication defect. Introduction of CpGs into specific regions of the virus genome under normal circumstances can be expected to cause a viral replication defect. However, if genome regions such as conserved secondary structures and overlapping reading frames are preserved, the detrimental effects of CpG addition may be circumvented by growing virus in a ZAP-knockout system (Ficarelli et al., 2019; Odon et al., 2019), thus allowing the generation of high titre replication-defective vaccine virus stocks.
ACKNOWLEDGEMENTS

We would like to thank Prof. Peter Simmonds (Oxford University) for providing a pre-release version of SSE v1.4. We are grateful to Dr James Glover (the Roslin Institute) and two anonymous reviewers for comments on the manuscript. Figure 1 was created using BioRender. This work was supported by Biotechnology and Biosciences Research Council Roslin Institute Strategic Program Grant funding (no. BB/P013740/1 to PD and FG) and Wellcome Trust/ Royal Society Fellowship (211222/Z/18/Z to EG).
REFERENCES

Atkinson, N., Witteveldt, J., Evans, D., Simmonds, P., 2014. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Research 42(7), 4527-4545.

Azhar, E.I., El-Kafrawy, S.A., Farraj, S.A., Hassan, A.M., Al-Saeed, M.S., Hashem, A.M., Madani, T.A., 2014. Evidence for Camel-to-Human Transmission of MERS Coronavirus. 370(26), 2499-2505.

Banerjee, A., Kulcsar K, Misra V, Frieman M, Mossman, K., 2019. Bats and Coronaviruses. Viruses 11(1), 41.

Berkhout, B., van Hemert, F., 2015. On the biased nucleotide composition of the human coronavirus RNA genome. Virus Research 202, 41-47.

Burns, C.C., Campagnoli, R., Shaw, J., Vincent, A., Jorba, J., Kew, O., 2009. Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. 83(19), 9957-9969.

Chenik, M., Chebli, K., Blondel, D., 1995. Translation initiation at alternate in-frame AUG codons in the rabies virus phosphoprotein mRNA is mediated by a ribosomal leaky scanning mechanism. 69(2), 707-712.

Cooper, D.N., Krawczak, M., 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Human Genetics 83(2), 181-188.

Cotten, M., Lam, T.T., Watson, S.J., Palser, A.L., Petrova, V., Grant, P., Pybus, O.G., Rambaut, A., Guan, Y., Pillay, D., Kellam, P., Nastouli, E., 2013. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. Emerg Infect Dis 19(5), 736-742B.

Cox, E., Hoooyberghs, J., Penzaert, M.B., 1990. Sites of replication of a porcine respiratory coronavirus related to transmissible gastroenteritis virus. Res Vet Sci 48(2), 165-169.

Davidson, A., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M., Heesom, K., Zambon, M., Ellis, J., Lewis, P., Hiscox, J., Matthews, D., 2020. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. bioRxiv.

Deaton, A.M., Bird, A., 2011. CpG islands and the regulation of transcription. 25(10), 1010-1022.

Dev, R.R., Gani, R., Singh, S.P., Mahalingam, S., Banerjee, S., Khosla, S., 2017. Cytosine methylation by DNMT2 facilitates stability and survival of HIV-1 RNA in the host cell during infection. Biochemical Journal 474(12), 2009-2026.

Ficarelli, M., Wilson, H., Pedro Galão, R., Mazzon, M., Antzin-Anduetza, I., Marsh, M., Neil, S.D., Swanson, C.M., 2019. KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. Elife 8, e46767.

Firth, A.E., Atkins, J.F., 2010. Candidates in Astroviruses, Seadornaviruses, Cytorhabdoviruses and Coronaviruses for +1 frame overlapping genes accessed by leaky scanning. Virology Journal 7(1), 17.

Futcher, B., Gorbateyvych, O., Shen, S.H., Stauff, C.B., Song, Y., Wang, B., Leatherwood, J., Gardin, J., Yurovsky, A., Mueller, S., Wimmer, E., 2015. Reply to Simmonds et al.: Codon pair and dinucleotide bias have not been functionally distinguished. 112(28), E3635-E3636.

Gaunt, E., Wise, H.M., Zhang, H., Lee, L.N., Atkinson, N.J., Nicol, M.Q., Highton, A.J., Kleinerman, P., Beard, P.M., Dutia, B.M., Digard, P., Simmonds, P., 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. Elife 5, e12735-e12735.

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 pathogens 4(6), e1000079-e1000079.
Groenke, N., Trimpert, J., Merz, S., Conradie, A.M., Wyler, E., Zhang, H., Hazapis, O.-G., Rausch, S., Landthaler, M., Osterrieder, N., Kunec, D., 2020. Mechanism of Virus Attenuation by Codon Pair Deoptimization. Cell Reports 31(4), 107586.

Guo, X., Ma, J., Sun, J., Gao, G., 2007. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. 104(1), 151-156.

Gutman, G.A., Hatfield, G.W., 1989. Nonrandom utilization of codon pairs in Escherichia coli. 86(10), 3699-3703.

Hiscott, J., Kwon, H., Génin, P., 2001. Hostile takeovers: viral appropriation of the NF-kB pathway. The Journal of Clinical Investigation 107(2), 143-151.

Hu, B., Ge, X., Wang, L.-F., Shi, Z., 2015. Bat origin of human coronaviruses. Virology Journal 12(1), 221.

Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y.W., Siddell, S.G., Brierley, I., 2016. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. PLoS pathogens 12(2), e1005473-e1005473.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T., 2001. Codon Usage and tRNA Genes in Eukaryotes: Correlation of Codon Usage Diversity with Translation Efficiency and with CG-Dinucleotide Usage as Assessed by Multivariate Analysis. Journal of Molecular Evolution 53(4), 290-298.

Khoddami, V., Cairns, B.R., 2013. Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nature Biotechnology 31(5), 458-464.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. 2020.2003.2012.988865.

Kozak, M., 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44(2), 283-292.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 35(6), 1547-1549.

Kunec, D., Osterrieder, N., 2016. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. Cell Reports 14(1), 55-67.

Liao, C.L., Lai, M.M., 1994. Requirement of the 5'-end genomic sequence as an upstream cis-acting element for coronavirus subgenomic mRNA transcription. 68(8), 4727-4737.

Lin, C.-G., Lo, S.J., 1992. Evidence for involvement of a ribosomal leaky scanning mechanism in the translation of the hepatitis B virus Pol gene from the viral pregenome RNA. Virology 188(1), 342-352.

Lin, Y.-T., Chiweshe, S., McCormick, D., Raper, A., Wickenhagen, A., DeFillipis, V., Gaunt, E., Simmonds, P., Wilson, S.J., Grey, F., 2020. Human cytomegalovirus evades ZAP detection by suppressing CpG dinucleotides in the major immediate early genes. 2020.2001.2007.897132.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 395(10224), 565-574.

Marra, M.A., Jones, S.J.M., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S.N., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Czub, M., Drebot, M., Fernandez, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipple, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C.,
Roper, R.L., 2003. The Genome Sequence of the SARS-Associated Coronavirus. 300(5624), 1399-1404.
McClelland, M., Ivarie, R., 1982. Asymmetrical distribution of CpG in an ‘average’ mammalian gene. Nucleic Acids Research 10(23), 7865-7877.
Medvedeva, Y.A., Khamis, A.M., Kulakovskiy, I.V., Ba-Alawi, W., Bhuyan, M.S.I., Kawaji, H., Lassmann, T., Harbers, M., Forrest, A.R.R., Bajic, V.B., The, F.c., 2014. Effects of cytosine methylation on transcription factor binding sites. BMC Genomics 15(1), 119.
Moratorio, G., Henningsson, R., Barbezange, C., Carrau, L., Bordería, A.V., Blanc, H., Beaucourt, S., Poirier, E.Z., Vallet, T., Boussier, J., Mounce, B.C., Fontes, M., Vignuzzi, M., 2017. Attenuation of RNA viruses by redirecting their evolution in sequence space. Nature Microbiology 2(8), 17088.
O'Connor, J.B., Brian, D.A., 2000. Downstream Ribosomal Entry for Translation of Coronavirus TGEV Gene 3b. Virology 269(1), 172-182.
Odon, V., Fros, J.J., Goonawardane, N., Dietrich, I., Ibrahim, A., Alshaikhahmed, K., Nguyen, D., Simmonds, P., 2019. The role of ZAP and OAS3/RNaseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. Nucleic Acids Research 47(15), 8061-8083.
Perlman, S., Netland, J., 2009. Coronaviruses post-SARS: update on replication and pathogenesis. Nature Reviews Microbiology 7(6), 439-450.
Rasschaert, D., Duarte, M., Laude, H., 1990. Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. 71(11), 2599-2607.
Rima, B.K., McFerran, N.V., 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. 78(11), 2859-2870.
Rota, P.A., Oberste, M.S., Monroe, S.S., Nix, W.A., Campagnoli, R., Icenogle, J.P., Peñaranda, S., Bankamp, B., Maher, K., Chen, M.-h., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J.L., Chen, Q., Wang, D., Erdman, D.D., Peret, T.C.T., Burns, C., Ksiazek, T.G., Rollin, P.E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Foucher, R., Günther, S., Osterhaus, A.D.M.E., Drosten, C., Pallansch, M.A., Anderson, L.J., Bellini, W.J., 2003. Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. 300(5624), 1394-1399.
Ryabova, L.A., Pooggin, M.M., Hohn, T., 2006. Translation reinitiation and leaky scanning in plant viruses. Virus Research 119(1), 52-62.
Sawicki, S.G., Sawicki, D.L., 1998. A New Model for Coronavirus Transcription. In: Enjuanes, L., Siddell, S.G., Spaan, W. (Eds.), Coronaviruses and Arteriviruses. Springer US, Boston, MA, pp. 215-219.
Sawicki, S.G., Sawicki, D.L., Siddell, S.G., 2007. A Contemporary View of Coronavirus Transcription. 81(1), 20-29.
Schaecher, S.R., Mackenzie, J.M., Pekosz, A., 2007. The ORF7b Protein of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) Is Expressed in Virus-Infected Cells and Incorporated into SARS-CoV Particles. 81(2), 718-731.
Schneider, P.A., Kim, R., Lipkin, W.I., 1997. Evidence for translation of the Borna disease virus G protein by leaky ribosomal scanning and ribosomal reinitiation. 71(7), 5614-5619.
Senanayake, S.D., Brian, D.A., 1997. Bovine coronavirus I protein synthesis follows ribosomal scanning on the bicistronic N mRNA. Virus Research 48(1), 101-105.
Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., Liu, R., He, X., Shuai, L., Sun, Z., Zhao, Y., Liu, P., Liang, L., Cui, P., Wang, J., Zhang, X., Guan, Y., Tan, W., Wu, G., Chen, H., Bu, Z., 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. eabb7015.
Simmonds, P., 2012. SSE: a nucleotide and amino acid sequence analysis platform. BMC Research Notes 5(1), 50.
Simmonds, P., Xia, W., Baillie, J.K., McKinnon, K., 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla—selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics 14(1), 610.

Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., Preiss, T., 2012. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Research 40(11), 5023-5033.

Sved, J., Bird, A., 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. 87(12), 4692-4696.

Takata, M.A., Gonçalves-Carneiro, D., Zang, T.M., Soll, S.J., York, A., Blanco-Melo, D., Bieniasz, P.D., 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550(7674), 124-127.

Tang, Q., Wang, X., Gao, G., 2017. The Short Form of the Zinc Finger Antiviral Protein Inhibits Influenza A Virus Protein Expression and Is Antagonized by the Virus-Encoded NS1. 91(2), e01909-01916.

Tats, A., Tenson, T., Remm, M., 2008. Preferred and avoided codon pairs in three domains of life. BMC Genomics 9(1), 463.

Tomso, D.J., Bell, D.A., 2003. Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands. Journal of Molecular Biology 327(2), 303-308.

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.

van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J.M., Wolthers, K.C., Wertheim-van Dillen, P.M.E., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nature Medicine 10(4), 368-373.

Vlasova, A.N., Wang, Q., Jung, K., Langel, S.N., Malik, Y.S., Saif, L.J., 2020. Porcine Coronaviruses. Emerging and Transboundary Animal Viruses, 79-110.

Wise, H.M., Barbezange, C., Jagger, B.W., Dalton, R.M., Gog, J.R., Curran, M.D., Taubenberger, J.K., Anderson, E.C., Digard, P., 2011. Overlapping signals for translational regulation and packaging of influenza A virus segment 2. Nucleic acids research 39(17), 7775-7790.

Woo, P.C.Y., Lau, S.K.P., Chu, C.-m., Chan, K.-h., Tsoi, H.-w., Huang, Y., Wong, B.H.L., Poon, R.W.S., Cai, J.J., Luk, W.-k., Poon, L.L.M., Wong, S.S.Y., Guan, Y., Peiris, J.S.M., Yuen, K.-y., 2005. Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. 79(2), 884-895.

Xia, X., 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Molecular Biology and Evolution.

Xie, L., Lu, B., Zheng, Z., Miao, Y., Liu, Y., Zhang, Y., Zheng, C., Ke, X., Hu, Q., Wang, H., 2018. The 3C protease of enterovirus A71 counteracts the activity of host zinc-finger antiviral protein (ZAP). 99(1), 73-85.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D.M.E., Fouchier, R.A.M., 2012. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. 367(19), 1814-1820.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine 382(8), 727-733.
Table 1. Comparison of the CpG dinucleotide composition of E ORF of coronaviruses that infect humans. CAI, codon adaptation index.

| Virus          | CpG frequency | GC content | CpG O:E ratio | CAI  |
|----------------|---------------|------------|---------------|------|
| SARS-CoV-2     | 11            | 0.387      | 1.262         | 0.58 |
| SARS-CoV       | 13            | 0.408      | 1.308         | 0.56 |
| HCoV-229E      | 1             | 0.338      | 0.175         | 0.63 |
| HCoV-NL63      | 3             | 0.349      | 0.221         | 0.66 |
| HCoV-HKU1      | 2             | 0.290      | 0.586         | 0.63 |
| HCoV-OC43      | 2             | 0.349      | 0.306         | 0.67 |
| MERS-CoV       | 6             | 0.402      | 0.642         | 0.61 |
FIGURE LEGENDS

FIGURE 1. Workflow for sequence processing. Two sequence datasets were used for analysis; all coronavirus complete genome sequences available on NCBI, and SARS-CoV-2 complete genome sequences available on the GISAID platform (left hand pink shaded boxes). The coronavirus complete genome sequences were cleaned by removal of sequences with 90% nucleotide identity or greater to remove epidemiologic biases, leaving 215 complete genome sequences (central yellow shaded box). These were then categorised by genera, host, and tissue tropism. The subset of 215 sequences were also aligned over the E ORF and grouped by host (blue shaded boxes). Each box firstly describes each dataset used, the number of sequences in that dataset is then indicated in italicized font, and the figure to which the dataset corresponds is indicated in bold font.

FIGURE 2. Comparison of the CpG ratios of complete genomes of coronaviruses. SARS-CoV is represented by a blue circle, SARS-CoV-2 and its related bat sequence RaTG13 by purple circles and MERS-CoV by a green circle throughout. A. GC content versus CpG ratio for all complete genome sequences of coronaviruses downloaded from Genbank (3407 sequences). The sequence dataset in (A) was then stripped to include only one representative from sequences with less than 10% nucleotide diversity to overcome epidemiologic biases (215 representative sequences), which were analysed in the subsequent sub-figures. B. Coronavirus genus against genomic CpG content. Other human-infecting coronaviruses (HCoV-229E, HCoV-NL63 (alphacoronaviruses) and HCoV-HKU1 and HCoV-OC43 (betacoronaviruses) are represented using orange circles. C. Vertebrate host of coronavirus against genomic CpG content. D. Vertebrate host of coronavirus, with further sub-division into coronavirus genus, against genomic CpG content. Alphacoronaviruses are denoted with filled circles and betacoronaviruses with open circles. E. Primary replication site against genomic CpG content by host. For a full breakdown of how these were assigned, please refer to Table S1.

FIGURE 3. Heterogeneities in the dinucleotide composition of the SARS-CoV-2 genome. A-C. Comparison of the dinucleotide and coding compositions of SARS-CoV-2 open reading frames (ORFs) for A. CpG observed: expected (O:E) ratios, B. UpA O:E ratios and C. Codon pair bias (CPB) scores. Average scores across the genome are indicated using open circles. D. Sliding window analysis of CpG content of SARS-CoV-2 (green line) and closely related bat (RaTG13; black line) and pangolin (purple line) isolates. The CpG O:E ratio of the 3’ end of the genome was measured in 100 nucleotide windows in 25 nucleotide increments. The mean of 1163 complete genome sequences is presented for SARS-CoV-2.

FIGURE 4. Evolutionary conservation of E ORF CpG content. MERS-CoV (green circle), SARS-CoV (blue circle) and SARS-CoV-2 and its bat sequence relative RaTG13 (purple circles) are indicated in all panels. A. CpG O:E ratios for E ORF for 178 coronavirus E ORFs are plotted by host. B. CpG O:E ratios for E ORF were divided by the genomic CpG O:E ratio for 178 coronavirus sequences and grouped by host. C. Phylogenetic reconstruction of E ORF of human and bat coronaviruses. Maximum composite likelihood tree (100 bootstraps) representing the seven human-infecting coronaviruses (HCoV-229E, HCoV-HKU1, HCoV-NL63, HCoV-OC43 are indicated by black circles) and 96 bat coronaviruses for which E ORF could be identified by alignment with the human coronaviruses. CpG O:E ratios for the E gene are indicated by large font numbers, and the sequences to which they relate are either bracketed or represented by triangles scaled to indicate the number of sequences they represent.

https://mc.manuscriptcentral.com/vevolu
Figure 1

338x190mm (96 x 96 DPI)
Figure 2

200x264mm (300 x 300 DPI)
Figure 3

338x190mm (96 x 96 DPI)
| Accession number | Description                                      | genus | Host      | GC content | Ratio_CpG |
|------------------|--------------------------------------------------|-------|-----------|------------|-----------|
| AC_000192.1      | Murine hepatitis virus strain bg                 | rodent| rodent    | 0.421      | 0.6453    |
| AF201929.1       | Murine hepatitis virus strain bg                 | rodent| rodent    | 0.414      | 0.61      |
| AY319651.1       | Avian infectious bronchitis vi g                 | avian | avian     | 0.382      | 0.478     |
| AY572034.1       | SARS coronavirus civet007, cb                    | misc  | human     | 0.408      | 0.4621    |
| AY595412.1       | SARS coronavirus LLJ-2004, cb                    | human | human     | 0.408      | 0.4602    |
| DQ011855.1       | Porcine hemagglutinating en b                    | swine | bat       | 0.373      | 0.4901    |
| DQ071615.1       | Bat SARS coronavirus Rp3, cc b                    | bat   | bat       | 0.409      | 0.4523    |
| DQ084200.1       | bat SARS coronavirus HKU3-:b                     | bat   | bat       | 0.411      | 0.4988    |
| DQ412042.1       | Bat SARS coronavirus Rf1, co b                   | bat   | bat       | 0.411      | 0.487     |
| DQ412043.1       | Bat SARS coronavirus Rm1, co b                   | bat   | bat       | 0.411      | 0.4893    |
| DQ648794.1       | Bat coronavirus (BtCoV/133/ b                    | bat   | bat       | 0.379      | 0.513     |
| DQ648856.1       | Bat coronavirus (BtCoV/273/ b                    | bat   | bat       | 0.412      | 0.4834    |
| DQ648857.1       | Bat coronavirus (BtCoV/279/ b                    | bat   | bat       | 0.411      | 0.4839    |
| DQ811785.1       | TGEV Miller M6, complete genome                  | swine | swine     | 0.377      | 0.4787    |
| DQ811788.1       | TGEV Purdue P115, complete genome               | swine | swine     | 0.377      | 0.4559    |
| DQ848678.1       | Feline coronavirus strain FCo a                  | feline| feline    | 0.384      | 0.5258    |
| EF065508.1       | Bat coronavirus HKU4-4, con b                    | bat   | bat       | 0.378      | 0.509     |
| EF065512.1       | Bat coronavirus HKU5-5, con b                    | bat   | bat       | 0.43       | 0.6151    |
| EF065516.1       | Bat coronavirus HKU9-4, con b                    | bat   | bat       | 0.416      | 0.6119    |
| EF424621.1       | Sable antelope coronavirus Lb                    | ungulate| ungulate | 0.371      | 0.4702    |
| EF424624.1       | Calf-giraffe coronavirus US/b                    | ungulate| ungulate | 0.37       | 0.4718    |
| EU111742.1       | Coronavirus SW1, complete genome                 | misc  | misc      | 0.393      | 0.5097    |
| EU420137.1       | Bat coronavirus 1B strain AF/c a                 | bat   | bat       | 0.385      | 0.4641    |
| FJ425185.1       | Waterbuck coronavirus US/Cb                      | ungulate| ungulate | 0.371      | 0.4715    |
| FJ425186.1       | Waterbuck coronavirus US/Cb                      | ungulate| ungulate | 0.37       | 0.4716    |

https://mc.manuscriptcentral.com/vevolu
| Accession   | Name                                      | Host     | Similarity 1 | Similarity 2 |
|------------|-------------------------------------------|----------|--------------|--------------|
| FJ425187.1 | White-tailed deer coronavirus              | ungulate | 0.371        | 0.4771       |
| FJ425190.1 | Sambar deer coronavirus US                  | ungulate | 0.37         | 0.4711       |
| FJ588686.1 | Bat SARS CoV Rs672/2006, ccb               | bat      | 0.411        | 0.4722       |
| FJ938054.1 | Feline coronavirus UU4, com a              | feline   | 0.386        | 0.5199       |
| FJ938058.1 | Feline coronavirus UU16, cor a             | feline   | 0.388        | 0.5146       |
| FJ938065.1 | Bovine respiratory coronavirus             | ungulate | 0.373        | 0.4699       |
| FJ938066.1 | Bovine respiratory coronavirus             | ungulate | 0.37         | 0.4744       |
| GQ152141.1 | Feline coronavirus strain FC               | feline   | 0.383        | 0.48         |
| GQ153548.1 | Bat SARS coronavirus HKU3-::b              | bat      | 0.411        | 0.4992       |
| GQ427176.1 | Turkey coronavirus strain TCi g             | avian    | 0.382        | 0.4954       |
| GQ477367.1 | Canine coronavirus strain CC a             | canine   | 0.377        | 0.4362       |
| HM211100.1 | Bat coronavirus HKU9-10-1, rb              | bat      | 0.426        | 0.6973       |
| HM211101.1 | Bat coronavirus HKU9-10-2, rb              | bat      | 0.411        | 0.6264       |
| HM245925.1 | Mink coronavirus strain WD1a               | mustelid | 0.375        | 0.5315       |
| HQ012369.1 | Feline coronavirus UU21, cor a             | feline   | 0.386        | 0.5061       |
| HQ850618.1 | Infectious bronchitis virus isc g          | avian    | 0.38         | 0.49         |
| JF705860.1 | Duck coronavirus isolate DK/ g             | avian    | 0.382        | 0.4652       |
| JN129834.1 | Human coronavirus OC43 str b               | human    | 0.367        | 0.4688       |
| JN856008.2 | Canine coronavirus strain A7 a             | canine   | 0.374        | 0.4289       |
| JN874562.1 | Rabbit coronavirus HKU14 st b              | misc     | 0.376        | 0.5255       |
| JQ404409.1 | Canine coronavirus strain 1-7a             | canine   | 0.374        | 0.4084       |
| Accession   | Description                                           | Species | Identity | Similarity |
|-------------|-------------------------------------------------------|---------|----------|------------|
| JQ989271.1  | Rousettus bat coronavirus HKU1                        | bat     | 0.385    | 0.4959     |
| JX993987.1  | Bat coronavirus Rp/Shaanxi2 b                         | bat     | 0.416    | 0.5028     |
| JX993988.1  | Bat coronavirus Cp/Yunnan2 b                          | bat     | 0.408    | 0.4832     |
| KC008600.1  | Infectious bronchitis virus, ccg                      | avian   | 0.38     | 0.4851     |
| KC175339.1  | Canine coronavirus strain 17 a                        | canine  | 0.381    | 0.4989     |
| KC461237.1  | Feline infectious peritonitis v a                      | feline  | 0.381    | 0.5012     |
| KC869678.4  | Coronavirus Neoromicia/PM b                           | bat     | 0.402    | 0.517      |
| KF530123.1  | Feline coronavirus strain Feli a                      | feline  | 0.387    | 0.5117     |
| KF569996.1  | Rhinolophus affinis coronavirus                        | bat     | 0.407    | 0.4617     |
| KF686346.1  | Human coronavirus HKU1 str b                          | human   | 0.321    | 0.4559     |
| KF793826.1  | Bottlenose dolphin coronavirus                        | misc    | 0.392    | 0.4991     |
| KF931628.1  | Infectious bronchitis virus isc g                      | avian   | 0.381    | 0.5056     |
| KJ473796.1  | BtMf-AlphaCoV/JX2012, comma                          | bat     | 0.382    | 0.4566     |
| KJ473798.1  | BtMf-AlphaCoV/HuB2013, cc a                           | bat     | 0.419    | 0.5888     |
| KJ473800.1  | BtMf-AlphaCoV/HeN2013, cc a                           | bat     | 0.42     | 0.5969     |
| KJ473810.1  | BtMs-AlphaCoV/GS2013, cor a                           | bat     | 0.386    | 0.5392     |
| KJ473811.1  | BtRf-BetaCoV/JL2012, compl b                         | bat     | 0.41     | 0.4903     |
| KJ473812.1  | BtRf-BetaCoV/HeB2013, com b                           | bat     | 0.41     | 0.4855     |
| KJ473813.1  | BtRf-BetaCoV/SX2013, comp b                           | bat     | 0.41     | 0.4887     |
| KJ473814.1  | Strand=AC*BtRs-BetaCoV/Hb                               | bat     | 0.414    | 0.4813     |
| KJ473815.1  | BtRs-BetaCoV/GX2013, com b                            | bat     | 0.409    | 0.4512     |
| KJ473816.1  | BtRs-BetaCoV/YN2013, comp b                           | bat     | 0.411    | 0.4715     |
| KJ473821.1  | BtVs-BetaCoV/SC2013, comp b                           | bat     | 0.43     | 0.5834     |
| KJ473822.1  | BtTp-BetaCoV/GX2012, com b                            | bat     | 0.378    | 0.5139     |
| KJ569769.1  | Porcine coronavirus HKU15 s d                         | swine   | 0.433    | 0.5599     |
| KM454473.1  | Duck coronavirus isolate DK/ g                        | avian   | 0.393    | 0.499      |
| KP143512.1  | Feline coronavirus isolate 26 a                       | feline  | 0.384    | 0.4992     |
| KP849472.1  | Alphacoronavirus 1 strain 23,a                        | canine  | 0.37     | 0.3717     |
| KP981644.1  | Canine coronavirus strain CB,a                        | canine  | 0.375    | 0.4113     |
| KR061459.1  | Swine enteric coronavirus str a                       | swine   | 0.381    | 0.4911     |
| Accession Number | Description                                           | Host | Similarity Score 1 | Similarity Score 2 |
|------------------|-------------------------------------------------------|------|---------------------|---------------------|
| KR270796.1       | Porcine respiratory coronavirus                        | swine| 0.371               | 0.5014              |
| KR822424.1       | European turkey coronavirus                            | avian| 0.383               | 0.5087              |
| KT029139.1       | Middle East respiratory syndrome                      | human| 0.411               | 0.5561              |
| KT253327.1       | Camel alphacoronavirus                                | camelid| 0.385               | 0.4957              |
| KU182964.1       | Bat coronavirus isolate JTM C                          | bat | 0.41                | 0.4944              |
| KU182965.1       | Bat coronavirus isolate JPDB: b                       | bat | 0.381               | 0.5341              |
| KU973692.1       | UNVERIFIED: SARS-related c                            | bat | 0.411               | 0.4814              |
| KX266757.1       | Infectious bronchitis virus is c                       | avian| 0.382               | 0.4804              |
| KX432213.1       | Canine respiratory coronavirus                         | canine| 0.364               | 0.4621              |
| KX442564.1       | Hypsugo bat coronavirus HKI b                         | bat | 0.422               | 0.5494              |
| KX512809.1       | Ferret enteric coronavirus str                         | mustelid| 0.389               | 0.5504              |
| KX580958.1       | Porcine epidemic diarrhea vi a                         | swine| 0.417               | 0.5425              |
| KX722530.1       | Feline coronavirus isolate Ca a                        | feline| 0.386               | 0.5101              |
| KY073745.1       | NL63-related bat coronavirus a                         | bat | 0.428               | 0.6224              |
| KY073746.1       | NL63-related bat coronavirus a                         | bat | 0.426               | 0.613               |
| KY073747.1       | 229E-related bat coronavirus a                         | bat | 0.379               | 0.4597              |
| KY073748.1       | 229E-related bat coronavirus a                         | bat | 0.388               | 0.4935              |
| KY352407.1       | Severe acute respiratory syn b                        | bat | 0.392               | 0.458               |
| KY406735.1       | Porcine respiratory coronavirus a                      | swine| 0.372               | 0.4855              |
| KY417145.1       | Bat SARS-like coronavirus iso b                        | bat | 0.413               | 0.4749              |
| KY419103.1       | Porcine hemagglutinating en b                         | swine| 0.371               | 0.4836              |
| KY419109.1       | Porcine hemagglutinating en b                         | swine| 0.372               | 0.49                |
| KY770850.1       | Bat coronavirus isolate Anlor n                       | bat | 0.421               | 0.5909              |
| KY770851.1       | Bat coronavirus isolate Anlor n                       | bat | 0.419               | 0.588               |
| KY770859.1       | Bat coronavirus isolate Anlor n                       | bat | 0.411               | 0.4636              |
| KY770860.1       | Bat coronavirus isolate Jiuwa n                       | bat | 0.41                | 0.4837              |
| KY799179.1       | Myotis lucifugus coronavirus a                         | bat | 0.406               | 0.5063              |
| KY805846.1       | Infectious bronchitis virus is c                       | avian| 0.383               | 0.486               |
| KY933090.1       | Avian coronavirus strain L11 g                         | avian| 0.379               | 0.47                |
| Accession   | Description                                      | Host     | Identity | AAI      | CCI      |
|-------------|--------------------------------------------------|----------|----------|----------|----------|
| KY938558.1  | Bat coronavirus strain 16BO1n                    | bat      | 0.41     | 0.4984   |
| KY994645.1  | Porcine hemagglutinating en b                    | swine    | 0.372    | 0.4846   |
| LC022792.1  | Porcine epidemic diarrhea vi a                   | swine    | 0.418    | 0.548    |
| LC061274.1  | Equine coronavirus RNA, con b                    | ungulate | 0.372    | 0.4893   |
| LC215871.1  | Ferret coronavirus genomic f a                   | mustelid | 0.391    | 0.546    |
| LC364344.1  | Pigeon coronavirus UAE-HKU d                     | avian    | 0.396    | 0.4904   |
| LC364346.1  | Quail coronavirus UAE-HKU3 d                     | avian    | 0.423    | 0.5493   |
| LC469308.1  | Bat coronavirus Vs-CoV-1 germ                     | bat      | 0.43     | 0.5914   |
| LC494172.1  | Bovine coronavirus TCG-6 RN b                    | bovine   | 0.393    | 0.5311   |
| LN610099.1  | Guinea fowl coronavirus GfC g                     | avian    | 0.382    | 0.4971   |
| MF113046.1  | Alphacoronavirus Mink/China/1/2016, a            | mustelid | 0.372    | 0.4898   |
| MF167434.1  | Porcine enteric alphacorona a                    | swine    | 0.393    | 0.5311   |
| MF370205.1  | Rhinolophus bat coronavirus b                    | Swine    | 0.395    | 0.5379   |
| MF618253.1  | Murine hepatitis virus strain b                  | rodent   | 0.42     | 0.6041   |
| MF924725.1  | Infectious bronchitis virus isc g                | avian    | 0.382    | 0.5075   |
| MG021452.1  | Middle East respiratory synd b                   | bat      | 0.408    | 0.548    |
| MG233398.1  | Infectious bronchitis virus isc g                | avian    | 0.383    | 0.4782   |
| MG428703.1  | Human coronavirus NL63 iso a                     | human    | 0.345    | 0.4042   |
| MG596802.1  | Middle East respiratory synd b                   | bat      | 0.39     | 0.5204   |
| MG738155.1  | Infectious bronchitis virus isc g                | avian    | 0.38     | 0.4917   |

https://mc.manuscriptcentral.com/vevolu
| Accession   | Description                                      | Host | Similarity 1 | Similarity 2 |
|------------|--------------------------------------------------|------|--------------|--------------|
| MG762674.1 | Rousettus bat coronavirus HI b                   | bat  | 0.421        | 0.6138       |
| MG772933.1 | Bat SARS-like coronavirus isolate bat            | bat  | 0.389        | 0.4477       |
| MG772934.1 | Bat SARS-like coronavirus isolate bat            | bat  | 0.388        | 0.4496       |
| MG812378.1 | Sparrow deltacoronavirus strain d                | avian| 0.435        | 0.5409       |
| MG916901.1 | Bat coronavirus BtCoV/Rh/YF n                    | bat  | 0.411        | 0.575        |
| MG916903.1 | Bat coronavirus BtCoV/Rh/YF n                    | bat  | 0.413        | 0.5597       |
| MG916904.1 | Bat coronavirus BtCoV/Rh/YF n                    | bat  | 0.39         | 0.4841       |
| MG987420.1 | Middle East respiratory syndrome b               | bat  | 0.423        | 0.5674       |
| MH002337.1 | Tylonycteris bat coronavirus b                    | bat  | 0.378        | 0.5101       |
| MH002338.1 | Tylonycteris bat coronavirus b                    | bat  | 0.378        | 0.5092       |
| MH002341.1 | Pipistrellus bat coronavirus b                    | bat  | 0.431        | 0.6159       |
| MH002342.1 | Pipistrellus bat coronavirus b                    | bat  | 0.431        | 0.6287       |
| MH021175.1 | Avian coronavirus strain D27 g                    | avian| 0.382        | 0.5237       |
| MH181793.1 | Infectious bronchitis virus strain g              | avian| 0.384        | 0.4906       |
| MH532440.1 | Quail deltacoronavirus strain d                  | avian| 0.423        | 0.5559       |
| MH687951.1 | Alphacoronavirus sp. strain Va                   | bat  | 0.404        | 0.549        |
| MH687953.1 | Alphacoronavirus sp. strain Va                   | bat  | 0.401        | 0.5257       |
| MH687960.1 | Alphacoronavirus sp. strain Va                   | bat  | 0.401        | 0.494        |
| MH687966.1 | Alphacoronavirus sp. strain Va                   | bat  | 0.402        | 0.4924       |
| MH687968.1 | Betacoronavirus sp. strain VZb                   | rodent| 0.379       | 0.5301       |
| MH687970.1 | Betacoronavirus sp. strain V2b                   | rodent| 0.379       | 0.5122       |
| MH938448.1 | Alphacoronavirus Bat-CoV/P.a                     | bat  | 0.424        | 0.548        |
| MH938449.1 | Alphacoronavirus Bat-CoV/P.a                     | bat  | 0.404        | 0.4843       |
| MH938450.1 | Alphacoronavirus Bat-CoV/P.a                     | bat  | 0.403        | 0.473        |
| MK211369.1 | Coronavirus BtSk-AlphaCoV/a                      | bat  | 0.404        | 0.517        |
| MK211371.1 | Coronavirus BtSk-AlphaCoV/a                      | bat  | 0.4          | 0.5208       |
| MK211373.1 | Coronavirus BtRs-AlphaCoV/a                      | bat  | 0.413        | 0.5763       |
| MK211374.1 | Coronavirus BtRI-BetaCoV/Scb                     | bat  | 0.409        | 0.4895       |
| MK211375.1 | Coronavirus BtRs-BetaCoV/Yb                      | bat  | 0.411        | 0.4765       |
| MK211376.1 | Coronavirus BtRs-BetaCoV/Yb                      | bat  | 0.409        | 0.4661       |
| MK211377.1 | Coronavirus BtRs-BetaCoV/Yb                      | bat  | 0.411        | 0.4701       |
| MK211379.1 | Coronavirus BtRt-BetaCoV/Gb                      | bat  | 0.38         | 0.4867       |
| MK357909.1 | Middle East respiratory syndrome b               | camelid| 0.412       | 0.5588       |
| Accession   | Description                                               | Host     | Identity 1 | Identity 2 |
|-------------|-----------------------------------------------------------|----------|------------|------------|
| MK359255.1  | Canada goose coronavirus strain                            | avian    | 0.384      | 0.4641     |
| MK423877.1  | Pheasant coronavirus strain                               | avian    | 0.382      | 0.4649     |
| MK472067.1  | Alphacoronavirus sp. isolate                               | bat      | 0.401      | 0.6095     |
| MK472069.1  | Alphacoronavirus sp. isolate                               | bat      | 0.423      | 0.5484     |
| MK472070.1  | Alphacoronavirus sp. isolate                               | bat      | 0.41       | 0.5892     |
| MK472071.1  | Alphacoronavirus sp. isolate                               | bat      | 0.427      | 0.5427     |
| MK492263.1  | Bat coronavirus strain                                     | bat      | 0.383      | 0.4848     |
| MK574043.1  | Infectious bronchitis virus strain                         | avian    | 0.379      | 0.4693     |
| MK581202.1  | Infectious bronchitis virus strain                         | avian    | 0.381      | 0.4834     |
| MK581206.1  | Infectious bronchitis virus strain                         | avian    | 0.382      | 0.4931     |
| MK679660.1  | Hedgehog coronavirus 1, complete genome                   | misc     | 0.373      | 0.4757     |
| MK720944.1  | Tylonycteris bat coronavirus                              | bat      | 0.37       | 0.4679     |
| MK720946.1  | Rhinolophus bat coronavirus                                 | bat      | 0.403      | 0.5248     |
| MK907287.1  | Erinaceus hedgheg coronav b                                | bat      | 0.408      | 0.5156     |
| MN065811.1  | Bat alphacoronavirus strain                               | bat      | 0.408      | 0.5156     |
| MN514964.1  | Dromedary camel coronavirus                               | camelid  | 0.369      | 0.488      |
| MN611517.1  | Rousettus aegyptiacus bat cc a                             | bat      | 0.391      | 0.4746     |
| MN611518.1  | Miniopterus pusillus bat cor a                             | bat      | 0.42       | 0.5836     |
| MN611519.1  | Tylonycteris pachypus bat co b                             | bat      | 0.38       | 0.5195     |
| MN611521.1  | Scotophilus kuhlii bat corona a                            | bat      | 0.403      | 0.4891     |
| MN611522.1  | Rhinolophus affinis bat cor a                              | bat      | 0.4        | 0.5559     |
| MN611523.1  | Hipposideros pomona bat co a                               | bat      | 0.383      | 0.4744     |
| MN611525.1  | Hipposideros pomona bat co a                               | bat      | 0.387      | 0.5203     |
| MN690611.1  | Bottlenose dolphin coronavirus                             | misc     | 0.393      | 0.4899     |
| MN996532.1  | Bat coronavirus RaTG13, complete genome                    | bat      | 0.381      | 0.4088     |
| MT039887.1  | Severe acute respiratory syn b                            | human    | 0.38       | 0.4077     |
| MT072864.1  | Pangolin coronavirus isolate b                             | misc     | 0.385      | 0.4178     |
| NC_002645.1 | Human coronavirus 229E, co a                               | human    | 0.383      | 0.4965     |
| NC_009019.1 | Bat coronavirus HKU4-1, con b                              | bat      | 0.378      | 0.5116     |
| NC_009021.1 | Bat coronavirus HKU9-1, con b                              | bat      | 0.411      | 0.6712     |
| NC_009657.1 | Scotophilus bat coronavirus                                 | bat      | 0.401      | 0.5093     |
| NC_009988.1 | Bat coronavirus HKU2, comp a                               | bat      | 0.393      | 0.5203     |
| NC_010438.1 | Bat coronavirus HKU8, comp a                               | bat      | 0.418      | 0.5959     |
| NC_011547.1 | Bulbul coronavirus HKU11-9:d                              | avian    | 0.387      | 0.4339     |

https://mc.manuscriptcentral.com/vevolu
| Accession  | Species                    | Host     | Similarity | Identity  |
|------------|----------------------------|----------|------------|-----------|
| NC_011549.1| Thrush coronavirus HKU12-6d | avian    | 0.38       | 0.4146    |
| NC_011550.1| Munia coronavirus HKU13-3!d | avian    | 0.425      | 0.5583    |
| NC_014470.1| Bat coronavirus BM48-31/Bcn | bat      | 0.404      | 0.5102    |
| NC_016991.1| White-eye coronavirus HKU1d  | avian    | 0.398      | 0.4538    |
| NC_016992.1| Sparrow coronavirus HKU17, d | avian    | 0.445      | 0.5764    |
| NC_016993.1| Magpie-robin coronavirus HKd | avian    | 0.467      | 0.6662    |
| NC_016994.1| Night-heron coronavirus HKLd | avian    | 0.381      | 0.4131    |
| NC_016995.1| Wigeon coronavirus HKU20, d | avian    | 0.393      | 0.5233    |
| NC_017083.1| Rabbit coronavirus HKU14, cb | misc     | 0.377      | 0.5185    |
| NC_022103.1| Bat coronavirus CDPHE15/USa | bat      | 0.408      | 0.5066    |
| NC_025217.1| Bat Hp-betacoronavirus/Zhejb | bat      | 0.413      | 0.5892    |
| NC_026011.1| Betacoronavirus HKU24 strai b | rodent   | 0.401      | 0.5491    |
| NC_028811.1| BtMr-AlphaCoV/SAX2011, coa | bat      | 0.41       | 0.5717    |
| NC_028814.1| BtRf-AlphaCoV/HuB2013, coa | bat      | 0.383      | 0.527     |
| NC_028824.1| BtRf-AlphaCoV/YN2012, coma | bat      | 0.378      | 0.488     |
| NC_028833.1| BtNv-AlphaCoV/SC2013, cora | bat      | 0.419      | 0.573     |
| NC_030292.1| Ferret coronavirus isolate FRa | mustelid | 0.39       | 0.5741    |
| NC_030886.1| Rousettus bat coronavirus isb | bat      | 0.453      | 0.6261    |
| NC_032107.1| NL63-related bat coronavirus a | bat      | 0.392      | 0.5634    |
| NC_032730.1| Lucheng Rn rat coronavirus ia | rodent   | 0.402      | 0.5938    |
| NC_034440.1| Bat coronavirus isolate PRED n | bat      | 0.412      | 0.5254    |
| NC_034972.1| Coronavirus AcCoV-JC34, cora | rodent   | 0.401      | 0.5256    |
| Ratio_UpA | Original Paper PMID | Sample | Replication Site |
|----------|---------------------|--------|-----------------|
| 0.9294   |                     |        | Neurotropic     |
|          |                     |        | Hepatotropic,   |
|          |                     |        | weakly          |
| 0.9313   |                     |        | neurotropic     |
| 0.9008   | 32214717            | Unknown| Unknown         |
|          |                     |        | Respiratory     |
| 0.7975   | 16485471            | Throat/rectal | Enteric |
| 0.7965   | 12730501            | Respiratory | Enteric |
|          | 17649356            | tonsils | neural |
| 0.7951   | 16195424            | Blood, faecal | Enteric |
| 0.8042   | 16169905            | Blood, nasoph | Enteric |
|          |                     | Blood, faeces, | and throat |
| 0.7957   | 16195424            | swab    | Enteric |
| 0.7836   | 16195424            | Blood, faecal | Enteric |
| 0.9593   | 16840328            | Oropharyngeal | Enteric |
| 0.7933   | 16840328            | Oropharyngeal | Enteric |
| 0.7847   | 16840328            | Oropharyngeal | Enteric |
| 0.813    | 1379786             | can’t access orig | Enteric |
| 0.8062   | 1379786             | can’t access orig | Enteric |
| 0.8457   | 17363313            | Jejunum/ liver t | Macrophage |
| 0.9595   | 17121802            | Alimentary/res | Enteric |
| 0.8265   | 17121802            | Alimentary/res | Enteric |
|          |                     | Alimentary/ | respiratory |
|          |                     | specimen    | Enteric |
| 1.0107   | 17121802            | Faeces     | Enteric |
| 0.9352   | 18842722            | Faeces     | Enteric |
| 0.9353   | 17344285            | Faeces     | Enteric |
| 0.9088   | 18353961            | Liver tissue | Hepatotropic |
| 0.8317   | 18420807            | Faeces     | Enteric |
| 0.9354   | 18842722            | Faeces     | Enteric |
| 0.9355   | 18842722            | Faeces     | Enteric |

https://mc.manuscriptcentral.com/vevolu
| Score | Code        | Sample Type          | Pathogen Type |
|-------|-------------|----------------------|---------------|
| 0.935 | 18842722    | Faeces               | Enteric       |
| 0.9343| 18842722    | Faeces               | Enteric       |
| 0.7812| 20016037    | Faeces               | Enteric       |
| 0.8452| Unknown     | Unknown              | Macrophage    |
| 0.8437| Unknown     | Unknown              | Unknown       |
| 0.9334| Unknown     | Unknown              | Enteric       |
| 0.936 | Unknown     | Unknown              | Respiratory?  |
| 0.8331| 23239278    | Pleural effusion     | Respiratory   |
|       |             | Respiratory/         | alimentary    |
|       |             | specimen             | Enteric       |
| 0.8072| 20071579    |                      |               |
| 0.9067| 20022075    | Intestines           | Enteric       |
| 0.7957| Unknown     | Unknown              | Unknown       |
| 1.0908| 20702646    | Alimentary/ res      | Enteric       |
| 1.0511| 20702646    | Alimentary/ res      | Enteric       |
| 0.8856| 21346029    | Faecal sample        | Enteric       |
| 0.8385| Unknown     | Faecal sample        | Enteric       |
| 0.8985| 22966194    | Unknown              | Respiratory   |
|       |             | Respiratory/         | nephrotropic/|
|       |             | enteric/             |               |
| 0.8888| 27164844    | Kidney/ liver sarhep|               |
| 0.9219| 21849456    | Nasopharyngea       | Respiratory   |
| 0.797 | 22609354    | Unknown              | Enteric       |
| 0.9237| 22398294    | Faeces               | Enteric       |
| 0.7811| Unknown     | Unknown              | Enteric       |
| Alimentary/ respiratory specimen | Enteric |
|----------------------------------|--------|
| 0.7991                           | 22933277 |
| 0.7824                           | 23739658 |
| 0.7906                           | 23739658 |
| 0.8851                           | Unknown |
| 0.8466                           | Unknown |
| 0.8469                           | 23996606 |
| 0.9157                           | 24050621 |
| 0.8353                           | Unknown |
| 0.797                            | 24719429 |
| 0.9595                           | 24394697 |
| 0.9045                           | 24227844 |

| Faeces                            | Enteric |
|-----------------------------------|--------|
| 0.8967                           | 25721384 |
| 0.8117                           | 26262818 |
| 0.8495                           | 26262818 |
| 0.8539                           | 26262818 |
| 0.8025                           | 26262818 |
| 0.8029                           | 26262818 |
| 0.7995                           | 26262818 |
| 0.801                            | 26262818 |
| 0.7778                           | 26262818 |
| 0.8007                           | 26262818 |
| 0.7888                           | 26262818 |
| 0.8991                           | 24960574 |
| 0.9615                           | 26262818 |
| 0.8719                           | 24744332 |
| 0.8579                           | 26053682 |
| 0.8446                           | 25667330 |
| 0.7629                           | 26221765 |

| Vaccine                          | Respiratory |
|----------------------------------|-------------|
| 0.7778                           | 25953186 |
| 0.7668                           | 26689738 |

| Pharyngeal/ anal sw Enteric/ Respiratory |
|-------------------------------------------|
| 0.8117                                   | 26262818 |
| 0.8495                                   | 26262818 |
| 0.8539                                   | 26262818 |
| 0.8025                                   | 26262818 |
| 0.8029                                   | 26262818 |
| 0.7995                                   | 26262818 |
| 0.801                                    | 26262818 |
| 0.7778                                   | 26262818 |
| 0.8007                                   | 26262818 |
| 0.7888                                   | 26262818 |
| 0.8991                                   | 24960574 |
| 0.9615                                   | 26262818 |
| 0.8719                                   | 24744332 |
| 0.8579                                   | 26053682 |
| 0.8446                                   | 25667330 |
| 0.7629                                   | 26221765 |

| Anal swab Enteric |
|-------------------|
| 0.9615            | 26262818 |

| Throat/ anal sw Enteric/ Respiratory |
|--------------------------------------|
| 0.8117                               | 26262818 |
| 0.8495                               | 26262818 |
| 0.8539                               | 26262818 |
| 0.8025                               | 26262818 |
| 0.8029                               | 26262818 |
| 0.7995                               | 26262818 |
| 0.801                                | 26262818 |
| 0.7778                               | 26262818 |
| 0.8007                               | 26262818 |
| 0.7888                               | 26262818 |
| 0.8991                               | 24960574 |
| 0.9615                               | 26262818 |
| 0.8719                               | 24744332 |
| 0.8579                               | 26053682 |
| 0.8446                               | 25667330 |
| 0.7629                               | 26221765 |
| 0.8719                               | 24744332 |

| Faecal/ intestinal Enteric |
|---------------------------|
| 0.7778                    | 25953186 |
| 0.7668                    | 26689738 |

| Faeces/ orophary Respiratory |
|------------------------------|
| 0.8579                       | 26053682 |

| Tissue lesion sa Macrophage |
|------------------------------|
| 0.8446                       | 25667330 |

| Lung sample Pantropic |
|-----------------------|
| 0.7778                | 25953186 |

| faecal/ intestinal Enteric |
|---------------------------|
| 0.8112                    | 26262818 |
| Score   | Code         | Sample Type                        | Virus Type      |
|---------|--------------|------------------------------------|-----------------|
| 0.8237  | 26250391     | unknown                           | respiratory     |
| 0.9117  | 26585962     | Duodenum sample                   | Enteric         |
| 0.8972  | 26272558     | Sputum                            | Respiratory     |
| 0.8121  | 27528677     | Nasal swab                        | Respiratory     |
| 0.808   | 26847648     | Respiratory/ intestinal            | Enteric         |
| 0.9571  | 26847648     | Respiratory/ intestinal            | Enteric         |
| 0.7993  | 28352124     | Urine/ faeces                      | Enteric         |
| 0.8859  | 27876864     | Unknown                            | Nephrotropic    |
| 0.9422  | 28506792     | Nasopharyngea                      | Respiratory     |
| 0.8696  | 29346682     | Alimentary sample                  | Enteric         |
| 0.8709  |              | Unknown                            | Enteric         |
| 0.8529  | 25217400     | intestinal or faecal               | respiratory     |
| 0.8553  | Unknown      | Lung sample                        | respiratory?    |
| 0.7791  | 28077633     | Faecal/ oral sw Enteric/ Respirator| Enteric         |
| 0.7805  | 28077633     | Faecal/ oral sw Enteric/ Respirator| Enteric         |
| 0.8027  | 28077633     | Faecal/ oral sw Enteric/ Respirator| Enteric         |
| 0.8182  | 28077633     | Faecal/ oral sw Enteric/ Respirator| Enteric         |
| 0.869   | 31296683     | Rectal swab                        | Enteric         |
| 0.8218  | 30850666     | Nasal swab                         | Respiratory     |
| 0.7895  | 29190287     | Faeces/ anal sw Enteric            | Enteric         |
| 0.925   | 28628449 ?   | nasal swab                         | respiratory     |
| 0.929   | 28628449 ?   | nasal swab                         | respiratory     |
| 0.8508  | 28384506     | Faeces                            | Enteric         |
| 0.8504  | 28384506     | Faeces                            | Enteric         |
| 0.7943  | 28384506     | Faeces                            | Enteric         |
| 0.7905  | 28384506     | Faeces                            | Enteric         |
|         |              | brain, intestines, liver, kidney,  |     |
|         |              | spleen tissue                      | Enteric         |
| 0.8757  | 28840816     |                                  | Respiratory     |
| 0.9156  | 28495648     | Trachea/ kidney                    | Nephrotropic    |
| 0.8999  | 28917538     | Vaccine                            | Respiratory     |

https://mc.manuscriptcentral.com/vevolu
| Score | ID       | Description                                      | Location                        |
|-------|----------|--------------------------------------------------|---------------------------------|
| 0.8121| 28725945 | Oral swab/ faeces Enteric/ Respiratory and brain  |
| 0.922 | 28956766 | Link not provided not stated; present enteric     |
| 0.8525| 26272566 | Faecal/ nasal sw Enteric                          |
| 0.8762| 28820366 | Faeces Enteric                                   |
| 0.9491| 29769348 | Faeces Enteric                                   |
| 0.911 | 29769348 | Faeces Enteric                                   |
| 0.9067| Unknown  | Unknown Unknown Unknown                           |
| 0.8879| 25712772 | Faeces Enteric                                   |
| 0.879 | Unknown  | Unknown Unknown                                  |
| 0.9126| 28654418 | Intestinal sample Enteric                         |
| 0.9126| 29102111 | Faeces/ intestin Enteric                           |
| 0.9172| 29010211 | Faeces/ intestin Enteric Hepatotropic, neurotropic|
| 0.9241|          |                                                 |
| 0.8976| 30463206 | Unknown Nephrotropic                             |
| 0.8718| 29669833 | Alimentary spec Enteric                           |
|       |          | Respiratoy/ nephrotopic/                         |
| 0.9002| 30159377 | Trachea/ kidney enteric                            |
| 0.8735| 29741740 | Nasopharyngea Respiratory                         |
| 0.9289| 29258555 | Pool of viscera Enteric                           |
|       |          | Respiratoy/ nephrotropic/                         |
| 0.9035| 29337625 | Unknown enteric                                   |
| Code   | Date     | Sample Type          | Full Pathogen Groups  |
|--------|----------|----------------------|-----------------------|
| 1.0278 | 29500692 | Faeces/ anal sw Enteric | Enteric/ respiratory/ |
| 0.8205 | 30209269 | Intestinal tissue Brain? | Enteric/ respiratory/ |
| 0.8278 | 30209269 | Intestinal tissue Enteric | Enteric/ respiratory/ |
| 0.8494 | 29872066 | Faeces Enteric | Enteric/ respiratory/ |
| 0.9324 | 31022925 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8491 | 31022925 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8746 | 31022925 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8702 | 29669833 | Alimentary spec Enteric | Enteric/ respiratory/ |
| 0.9612 | 29669833 | Alimentary spec Enteric | Enteric/ respiratory/ |
| 0.9668 | 29669833 | Alimentary spec Enteric | Enteric/ respiratory/ |
| 0.8293 | 29669833 | Alimentary spec Enteric | Enteric/ respiratory/ |
| 0.8412 | 29669833 | Alimentary spec Enteric | Enteric/ respiratory/ |
| 0.8989 | 30533915 | Tracheal sample Enteric | Enteric/ respiratory/ |
| 0.89 | 18792774 | Kidney tissue Nephrotropic | Enteric/ respiratory/ |
| 0.9222 | 30758768 | Intestines Enteric | Enteric/ respiratory/ |
| 0.8109 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8222 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8241 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.8152 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.935 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.9244 | 30568804 | Faeces Enteric | Enteric/ respiratory/ |
| 0.7797 | 30447246 | Faeces/ organ sw Enteric | Enteric/ respiratory/ |
| 0.789 | 30447246 | Faeces/ organ sw Enteric | Enteric/ respiratory/ |
| 0.7837 | 30447246 | Faeces/ organ sw Enteric | Enteric/ respiratory/ |
| 0.8235 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.8362 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.9198 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.806 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.7927 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.7966 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.7795 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.9945 | 31474969 | Anal swab Enteric | Enteric/ respiratory/ |
| 0.8987 | 30482895 | Nasal swab Respiratory | Enteric/ respiratory/ |
| Score  | Reference | Sample Type                  | Pathotype  |
|--------|-----------|------------------------------|------------|
| 0.9356 | 30976080  | Cloacal swab                 | Enteric    |
| 0.9018 | 31902509  | Tracheal/ kidney sample      | Respiratory|
| 0.8626 | 31847282  | Faeces/serum                 | Enteric    |
| 0.6957 | 31847282  | Faeces/serum                 | Enteric    |
| 0.8203 | 31847282  | Faeces/serum                 | Enteric    |
| 0.7551 | 31847282  | Faeces/serum                 | Enteric    |
| 0.9862 | 31418677  | Faeces/rectal swab          | Enteric    |
| 0.8982 | 31255833  | Unknown                      | Respiratory|
| 0.8943 | 31917362  | Unknown                      | Respiratory|
| 0.8931 | 31917362  | Unknown                      | Nephrotropic|
| 0.9539 | 31063092  | Faeces/enteric               | Enteric    |
| 0.8559 | 31067830  | Oral/ alimentary/ Enteric    | Respiratory|
| 0.8656 | 31067830  | Oral/ alimentary/ Enteric    | Respiratory|
| 0.9289 | 31653070  | Alimentary sample            | Enteric    |
| 0.9862 | 31503522  | Faeces                       | Enteric    |
| 0.935  | 31534035  | Nasal swab                   | Enteric/Respiratory |
| 0.795  | 31996413  | Rectal swab                  | Enteric    |
| 0.8452 | 31996413  | Rectal swab                  | Enteric    |
| 0.9655 | 31996413  | Rectal swab                  | Enteric    |
| 0.8333 | 31996413  | Rectal swab                  | Enteric    |
| 0.9445 | 31996413  | Rectal swab                  | Enteric    |
| 0.8001 | 31996413  | Rectal swab                  | Enteric    |
| 0.8472 | 31996413  | Rectal swab                  | Enteric    |
| 0.9081 | Unknown   | faeces                       | Enteric    |
| 0.833  | Unknown   | faecal swab                  | Enteric    |
| 0.8286 | Unknown   | Unknown                       | Respiratory|
|        |           |                              | Enteric/Respiratory |
| 0.818  | 32218527  | Intestine-lung               | Respiratory |
| 0.793  | 11369870 (seq paper) | Lab strain             | Respiratory |
| 0.9573 | 17121802  | Alimentary/ res              | Enteric    |
| 1.0833 | 17121802  | Alimentary/ res              | Enteric    |
| 0.8266 | 16840328  | Oropharyngeal/ Enteric       | Enteric    |
| 0.9069 | 17617433  | Respiratory/ ali             | Enteric    |
| 0.8529 | 18420807  | Faeces                       | Enteric    |
| 0.8867 | 18971277  | Throat/cloacal               | Enteric    |

https://mc.manuscriptcentral.com/vevolu
| Score  | Run   | Sample Type                        | Pathogen(s)          |
|--------|-------|-----------------------------------|----------------------|
| 0.8766 | 18971277 | Throat/ cloacal                        | Enteric              |
| 0.93   | 18971277 | Throat/ cloacal                        | Enteric              |
| 0.8546 | 20686038 | Faeces                              | Enteric              |
| 0.9197 | 22278237 | Rectal swab                         | Enteric              |
| 0.8237 | 22278237 | Rectal swab                         | Enteric              |
| 0.933  | 22278237 | Rectal swab                         | Enteric              |
| 0.9229 | 22278237 | Rectal swab                         | Enteric              |
| 1.0157 | 22278237 | Rectal swab                         | Enteric              |
| 1.0034 | 22278237 | Rectal swab                         | Enteric              |
| 0.9216 | 22398294 | Faeces                              | Enteric              |
| 0.8746 | Unknown   | Unknown                             | Unknown              |
| 0.8245 | 26262818 | Pharyngeal/ anal                   | Enteric/ Respiratory |
| 0.9142 | 25552712 | Alimentary sample                  | Enteric              |
| 0.8605 | 26262818 | Pharyngeal/ anal                   | Enteric/ Respiratory |
| 0.809  | 26262818 | Pharyngeal/ anal                   | Enteric/ Respiratory |
| 0.9463 | 26262818 | Pharyngeal/ anal                   | Enteric/ Respiratory |
| 0.834  | 26262818 | Pharyngeal/ anal                   | Enteric/ Respiratory |
| 0.8852 | 27283016 | Rectal swab                         | Enteric              |
| 0.9461 | 27676249 | Rectal swab                         | Enteric              |
| 0.7816 | 28077633 | Faecal/ oral sw                     | Enteric/ Respiratory |
| 0.896  | 25463600 | Faeces/ tissues                    | Enteric              |
| 0.9029 | 28377531 | Blood/ oral/ rect                   | Enteric              |
| 0.9127 | 28549438 | Intestinal sample                  | Enteric              |

https://mc.manuscriptcentral.com/vevolu
| Replication Site Grouped | PMID                          |
|-------------------------|-------------------------------|
| Other                   | 18137294; 8f                  |
| Multiple                | 11502093; 6:                  |
| Unknown                 |                               |
| Multiple                | 16485471                      |
| Respiratory             | 14990596; 1:                  |
| Multiple                | 16809333                      |
| Enteric                 | 16195424                      |
|                         | 16169905;                     |
|                         | 17121802;                     |
| Enteric                 | 16647731                      |
| Enteric                 | 16195424                      |
| Enteric                 | 16840328                      |
| Enteric                 | 16840328                      |
| Enteric                 | 17023013                      |
| Enteric                 | 17023013                      |
| Other                   | 17363313;                     |
|                         | 19254859                      |
| Enteric                 | 16647731;                     |
| Enteric                 | 23720729                      |
| Enteric                 | 23720729                      |
|                         | 17121802;                     |
|                         | 16647731;                     |
| Enteric                 | 29500692                      |
| Enteric                 |                               |
| Enteric                 | 17344285                      |
| Other                   | 18353961                      |
| Enteric                 | 8586714                       |
| Condition       | Count       |
|-----------------|-------------|
| Enteric         | 8586714     |
| Enteric         | 8586714     |
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Other           | 28675506    |
| Unknown         |             |

| Condition       | Count       |
|-----------------|-------------|
| Multiple        |             |
| Respiratory     |             |

| Condition       | Count       |
|-----------------|-------------|
| Respiratory     | 23239278    |
|                 | 20071579    |
|                 | 17121802    |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Unknown         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Unknown         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Multiple        |             |
| Respiratory     |             |

| Condition       | Count       |
|-----------------|-------------|
| Respiratory     | 22198411    |

| Condition       | Count       |
|-----------------|-------------|
| Multiple        | 27164844    |
| Respiratory     | 21849456    |

| Condition       | Count       |
|-----------------|-------------|
| Respiratory     | 23337903    |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         | 22609354    |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         | 18635322    |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         |             |

| Condition       | Count       |
|-----------------|-------------|
| Enteric         | 1259219     |

https://mc.manuscriptcentral.com/vevolu
| Disease       | Code       |
|--------------|------------|
| Enteric      | 22933277;  |
| Multiple     | 17121802   |
| Unknown      |            |
| Unknown      |            |
| Other        | 3038290    |
| Enteric      |            |
| Unknown      |            |
| Enteric      |            |
| Respiratory  | 15613317;  |
| Enteric      | 24227844   |
| Multiple     |            |
| Other        | 25667330   |
| Enteric      | 26221765   |
| Multiple     |            |
| Enteric      | 28634353   |
| Multiple     |            |
| Other        | 25667330   |
| Enteric      | 26221765   |
| Multiple     |            |
| Enteric      | 16704791;  |
| Multiple     | 17275120   |
| Enteric      |            |

https://mc.manuscriptcentral.com/vevolu
| Enteric       | 30146717 |
|--------------|---------|
| Respiratory  | 23075143; |
| Respiratory  | 26049252 |
| Enteric      | 26847648 |
| Enteric      |         |
| Enteric      |         |
| Multiple     | 27876864 |
| Respiratory  | 28506792 |
| Enteric      |         |
| Enteric      | 16499943 |
| Enteric      | 28284624 |
| Respiratory  |         |
| Enteric      |         |
| Respiratory  | 30850666 |
| Enteric      |         |
| Respiratory  | 30959064 |?
| Enteric      |         |
| Respiratory  | 30959064 |?
| Enteric      |         |
| Enteric      |         |
| Enteric      | 28840816 |
| Enteric      |         |
| Multiple     | 28495648; |
| Respiratory  | 30744668 |
| Respiratory  | 21331953 |

https://mc.manuscriptcentral.com/vevolu
| Type     | Accession Numbers |
|----------|-------------------|
| Enteric  | 23648375          |
| Enteric  | 28820366          |
| Enteric  | 29769348          |
| Enteric  | 29769348          |
| Unknown  | 25712772;         |
| Enteric  | 30842318          |
| Enteric  | 30103259          |
|          | 29102111;         |
|          | 29618817;         |
| Enteric  | 31554686          |
| Multiple | 6324031; 30(      |
| Multiple | 20153350          |
| Enteric  | 27928919          |
|          | 15034574;         |
|          | 17079323;         |
| Respiratory | 17944272       |
| Enteric  | 29337625;         |
| Multiple | 30173857          |
| Category   | Code     |
|------------|----------|
| Enteric    |          |
| Respiratory| 31902509 |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Respiratory| 31255833 |
| Respiratory| 31917362 |
| Other      | 31917362 |
| Enteric    | 31063092 |
| Multiple   |          |
| Enteric    |          |
| Enteric    |          |
| Multiple   |          |
| Enteric    |          |
| Multiple   | 24655427 |
| Enteric    |          |
| Enteric    |          |
| Respiratory| 32159775; 32235945 |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Respiratory| 32159775; 32235945 |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Enteric    |          |
| Respiratory| 11369870; 22926811; 16603522 |
| Enteric    | 16647731; 23720729 |
| Enteric    | 17121802; 29500692 |
| Enteric    | 16840328; 17617433; 17121802; |
| Enteric    | 16647731; |
| Enteric    | 27048154 |
| Enteric    | 18971277; |
| Enteric    | 22278237 |
Enteric 18971277;
22278237
Enteric 18971277;
22278237
Unknown Multiple
Enteric
Multiple
Multiple
Multiple
Multiple
Enteric
27283016
Enteric 29500692;
Multiple
Enteric
Enteric
28377531
Also referred to as MHV-4 or JHM SD. A lot of papers since 1940s suggest that JHM is a neurotropic virus. Mice infected with JHM developed encephalitis and demyelinating diseases. Many studies have shown MHV-2 is mainly hepatotropic and weakly neurotropic. Intracerebral inoculation of MHV-2 into mice showed viral replication in liver and brain. Mice infected with MHV-2 developed meningitis and hepatitis.

CoV was detected in both throat and rectal swab. Many studies showed that respiratory tract is the primary target of infection.

PHEV-WV572 first isolated from tonsils of two piglets with encephalomyelitis; virus has strong tropism for epithelial cells of URT and CNS; transmitted through nasal secretions. 3 of 30 bat faecal swabs were positive for RdRp and N gene, none detected in the throat swabs tested. RdRp was detected in the anal swab of 23 bats from the same species, but not in the nasopharyngeal swab. 3 samples were sequenced and they were 3 different HKU3 isolates; Refer to 5) EF065516.1 for more HKU3 isolates by the same group.

1 of 8 bats was positive for RdRp and N gene in faecal swab; not tested in throat swab. 1 of 8 bats was positive for RdRp and N gene in faecal swab, none detected in the throat swabs tested. 2 of 14 Tylonycteris pachypus anal swabs were positive for RdRp, none detected in throat swab. 4 of 41 Rhinolophus ferrumequinum anal swabs were positive for RdRp, none detected in throat swab. 1 of 38 Rhinolophus macrotis anal swabs were positive for RdRp, none detected in throat swab. TGEV isolates are enteric; there is a naturally occurring deletion variant, Purdue, which has respiratory tropism. TGEV isolates are enteric; there is a naturally occurring deletion variant, Purdue, which has respiratory tropism.

C1JE was identified as FIPV type I and it was isolated from a cat with a histopathologically confirmed diagnosis of FIP. The nucleotide sequences of the virus isolated from jejunum and liver are identical.; FIPV is known to target and replicate in macrophages.

Another HKU4 isolate by the same group. Refer to 69) NC_009019.1 and 5) EF065516.1; Another study by the same group which showed 55 of 216 alimentary samples were positive for RdRp, none detected in respiratory samples.

52 bats (~10% of total bats) were positive for CoV RdRp (bat CoV HKU2 (6), HKU3 (1), HKU8 (2), HKU9 (42), HKU10(1)) in RT-PCR. RdRp was found in all alimentary specimen, but it was also detected in one respiratory sample (HKU2 strain, also positive in alimentary specimen). Previous study by the same group also found CoV RdRp (bat CoV HKU2 (2), HKU3 (21), HKU4 (4), HKUS (4), HKU6 (1), HKU7 (1), HKU8 (4)) in 37 bat anal swabs (~12%), but not in nasopharyngeal swabs; Refer to 70) MG762674.1 for more info of HKU9.

Isolated from a sable antelope with diarrhea. Isolated from a giraffe with diarrhea. Inoculation of a BCoV-seronegative gnotobiotic calf with GiCoV-OH3 orally caused severe diarrhea. The beluga whale died from illness characterised by generalized pulmonary disease and terminal acute liver failure. CoV was detected in liver tissue.

Gnotobiotic calf-adapted strain of OH-WD358. Refer to 214) FJ425186.1. Isolated from a waterbuck with diarrhea. Viral antigen was detected in the respiratory tract and intestinal tissues from calves inoculated with WbCoV-OH-WD358, particularly high in colon tissues (minimal staining in respiratory tract tissues).
Isolated from a white-tailed deer with diarrhea. Viral antigen was detected in the respiratory tract and intestinal tissues from calves inoculated with WtDCoV-OH-WD470, particularly high in colon tissues (minimal staining in respiratory tract tissues)

Isolated from a sambar deer with diarrhea. Viral antigen was detected in the respiratory tract and intestinal tissues from calves inoculated with SDCoV-OH-WD388, particularly high in ileum and colon tissues (minimal staining in respiratory tract tissues)

HEK293 cell line stably express the ectodomain of S proteins derived from UU4 was established. By using recombinant S protein as probe, S protein IHC was performed on cat tissues with feline infectious peritonitis (FIP). UU4 S protein bound to macrophages in lymph node, liver, spleen and lung tissues, which was correlated with FCoV antigens IHC that stained macrophages in livers, spleens, kidneys, lymph nodes and leptomeninges, but not the intestinal tracts.

Previous study by the same group submitted the sequences showed that AH187 CoV RNA was detected in both nasal and rectal swab of infected calf with overt respiratory and enteric symptoms. The accession numbers for the isolate from nasal and rectal swab are different from this.

As its name suggests 
NTU156 was isolated from a kitten with FIP by the co-cultivation of pleural effusion with feline fcfw-4 cells 
RdRp was detected in respiratory specimens from 2 of 1337 bats and in alimentary specimens from 126 of 1337 bats (in general for this SARSr-Rh-BatCoV strain); Refer to 5) EF065516.1 for more isolates of HKU3 by the same group
Virus was isolated from a turkey with acute enteritis; Virus inoculation of 1-day-old pouls caused diarrhea by 48 hpi and lesions in the GI tract. Viral RNA was detected in the samples from GI tract.

Another HKU9 isolate by the same group of 5) EF065516.1. Refer to 5) EF065516.1 and 70) MG762674.3

Another HKU9 isolate by the same group of 5) EF065516.1. Refer to 5) EF065516.1 and 70) MG762674.2
Infected minks had diarrhea. RdRp and N genes were detected in faecal samples
Associated with FIP
Isolated from a chicken with respiratory problems

Isolated from ducks with mild respiratory and diarrhea symptoms. Necropsy also found kidney swelling and bleeding bursal and thymus organs
Isolated from a patient with upper respiratory tract infection; the most common symptoms of OC43 infection are fever, cough and upper respiratory tract infection

The authors stated that the dog, which the virus was isolated from, presented primarily with enteritis; In general, canine alphacoronavirus are known as canine enteric coronavirus (CCoV) whereas canine betacoronavirus are recognised as canine respiratory coronavirus (CRCoV). Another isolate of HKU14. Refer to 209) NC_017083.1
Neonatal dogs inoculated orally with CoV 1-71 showed enteritis and diarrhea. CoV was detected in faeces and intestines.
Only 3 bats were positive for this isolate of HKU10 and they were all detected only in alimentary samples; Another isolate of HKU10 (Hipposideros pomona bat HKU10) was identified in the same study where 36 positive alimentary samples and 3 positive respiratory samples (2 of which were also tested positive in alimentary samples) were detected; An isolate of HKU10 was also found in bat alimentary specimen in another study by the same group

WSU 79-1146 is known as FIPV

The first paper reported HKU1 isolated the virus from nasopharyngeal aspirate of a patient with pneumonia. Viral RNA was not detected in urine and faecal samples; Respiratory tract infections are usually observed in HKU1-positive patients

RdRp was found in 3 of 48 faecal samples, none detected in respiratory swab

VicS-del has 40 nucleotide deletion in 3' UTR compared to VicS-v (predominant subpopulation in VicS vaccine). Chickens inoculated with VicS-v by eye drop showed high viral concentrations in the trachea and low virus load was detected in the cecal tonsil (no virus found in kidney). Chickens inoculated with VicS-del showed minimal virus concentrations in trachea and cecal tonsil, none detected in kidney. Chickens inoculated with VicS-del and VicS-v showed similar symptoms, but less severe with VicS-del infection. Lesions were observed in trachea, but not in kidney

2 of 18 bats were positive for RdRp in anal swabs, but not in pharyngeal swabs

Related publication PMID - 26433221
Related publication PMID - 26433221
Related publication PMID - 26433221
Related publication PMID - 26433221
Related publication PMID - 26433221
Related publication PMID - 26433221
Related publication PMID - 26262818, 26433221
HKU15 strain IN2847; detected in samples from diarrhoeic pigs; also detected in respiratory samples (second paper- only NPAs and no GI samples collected)

Tissue lesion sample was derived from a kitten with FIP. IHC identified a large number of macrophages with abundant viral antigen (i.e. N protein) within the lesions

Isolated from the faecal specimen of a dog with diarrhea

Clinical symptoms of the sick dogs included fever, lethargy, inappetence, vomiting, hemorrhagic diarrhea, and neurologic signs. Necropsy showed abnormalities in abdominal cavity, lungs, liver, spleen, kidney and lymph node. CoV RNA was detected in faeces, intestine, lungs, spleen, liver, kidney, lymph node and brain. CoV antigen was found in lungs, kidneys, liver, spleen, gut, and lymph nodes

Paper vague; samples collected from pigs with diarrhoea, samples by EM with CoV particles partially sequenced, and representative (two) samples whole genome sequenced. One was PEDV and one was a recombinant of PEDV and TGEV
No access to paper but virus name is 'porcine respiratory virus strain OH7269' isolated from 42-day-old turkeys with poult enteritis complex. Viral antigen was detected in the intestinal tissues. The first report of MERS isolated the virus from sputum of a patient with pneumonia; in general, patients infected with MERS usually develop pneumonia.

1 of 30 bats was positive for RdRp in the intestinal specimen, but not in respiratory specimen. Shared 99% identity with SARSr-BatCoV Rf1; Refer to 38) DQ412042.1

5 of 95 bats were found to be RdRp-positive in intestinal specimens, none in respiratory specimens.

1-day-old chicks were challenged with 3575/08 intranasally. H&E staining showed abnormalities of tracheal and kidney tissues. Viral RNA was detected in lung and kidney samples.

Isolated from a dog with mild respiratory infection.

FECV (or FRECV) were shown to be associated with epizootic catarrhal enteritis which caused diarrhea, lethargy, anorexia and vomiting in ferrets. FECV-N gene was detected in faeces and saliva. FECV RNA and antigen were detected in the cytoplasm of enterocytes at the villi tips in the jejunum, not found in the large intestine, lymph nodes, spleen, esophagus, stomach, and parotid salivary glands.

PEDV highly virulent strain PC22A passaged 100 times in Vero cells; sample and replication site refer to original sample and paper.

Associated with FIP so should be enteric, but lung sample recorded on NCBI.

The lab had been isolating a lot of viruses from fresh intestines and faecal samples, but they did not detect this strain. This strain was isolated during their attempt to isolate other viruses from the nasal swab.

Nasal swabs collected from swine presenting with ILI; PMID link is presumptive based on author names.

Nasal swabs collected from swine presenting with ILI; PMID link is presumptive based on author names.

Detected CoV RdRp only in the intestinal tissues of 53 bats. In CoV N IHC, N antigen was detected in 5 lung samples which also showed positive viral RNA in the intestines.

Isolated from trachea and kidneys collected from a chicken with severe respiratory signs and diarrhea. Post mortem examination showed abnormalities in trachea and kidneys; Virus was detected in tracheal and oropharyngeal swab of 1-day-old chicks which were infected oculonasally.

Isolated from chickens with respiratory disease. Chickens infected with L1148 at low passage showed ciliostasis and lesions at the kidney. Virus at high passage (>80) lost its replication at kidney, but the virus was still detected in trachea (IHC) and oropharyngeal swab (RT-PCR).
This virus isolate was detected in oral swab. The authors collected a lot of oral swabs, but only a few of other samples and we don’t know if the other samples were from the same bat. Since the virus was isolated from oral swab, it might suggest the virus to be respiratory. Related publication PMID - 31076983

Porcine haemagglutinating encephalomyelitis virus; PHEV propagates via neural circuitry and replicates in nerve cells of CNS

Sequences are derived from an outbreak of diarrhoea in pigs, though sample collection type and replication site are not described; this is presumptive based on knowledge of PEDV

Infected horses had diarrhea. N gene was detected in faecal sample and negative in nasal swab. Authors suggest the virus to be enteric because the nucleotide sequence shared >90% identity with FRECV MSU-2. Refer to 120) KXS12809.1 for details of FRECV

1 of 18 pigeon samples tested were positive for HKU30 CoV

5 of 10 quail samples tested were positive for HKU30 CoV

Isolated from guinea fowl with peracute enteritis; H&E staining showed lesions in duodenum and intestines. Viral antigen was also found in duodenum and lower intestinal tract by IHC

Newborn piglets infected with GDS04 via oral feeding developed mild diarrhea at 1-4 dpi and showed severe watery diarrhea, vomiting and dehydration at 5-12 dpi. Tissues from two GDS04-infected piglets showed the virus detection in 2/2 hearts, 2/2 livers, 2/2 spleens, 2/2 kidneys, 2/2 stomachs, and 1/2 lungs. Lesion was only detected in intestinal tract, not any other organ.

Five 3-day-old piglets infected with SeACoV showed acute vomiting and watery diarrhea at 27-40 hpi; Another strain of SeACoV (SADS-CoV) was detected only in intestinal tissues of sick pigs (n=3); Infection of SeACoV in C57BL/6J mice showed the presence of N gene in intestinal, stomach and spleen tissues at 1 dpi, but replication of virus was detected only in 3 dpi spleen tissues (not intestines) by IHC staining of dsRNA, virus M and NSP3 (replication in mice might be different).

A lot of studies suggest dual tropism of A59. Mice infected with A59 showed hepatitis, mild encephalitis and demyelination.

Chickens were inoculated by the eye drop method with K2. Infected chickens showed respiratory problems and nephritis. Virus was successfully re-isolated from trachea, lung, cecal tonsil, kidney and bursa.

Chicken infected with IS-1494 showed depression, ruffled feathers, nasal discharge, coughing and mild watery diarrhea. Gross examination and H&E staining found abnormalities in trachea, lungs and kidney. Viral RNA was detected in trachea, lung, faeces, cecal tonsil, proventriculus, spleen and kidney

The first paper reported NL63 isolated the virus from nasopharyngeal aspirate of a patient with bronchiolitis and conjunctivitis. Patients infected with NL63 usually develop respiratory diseases.

Isolated from a chicken with swollen kidney; Chickens were inoculated with IBS130/2015 intraoculary. Infected chickens developed lesions in trachea, lungs and kidneys. Viral antigen was detected in trachea, kidney, lung, jejunum and cecal tonsil, but not in proventriculus. Viral RNA was also found in all the organs tested including proventriculus.
Heart, liver, spleen, lung, kidney, brain, and intestinal tissues were collected from 3 HKU9-infected bats. 2 bats showed HKU9 gene expression in kidney, heart, lung and intestinal tissues whereas the remaining bat showed the presence of HKU9 only detected in kidney, heart and intestinal tissues. All the tissues exhibited similar levels of HKU9 except one of the bats had about 10000-fold higher levels of HKU9 in intestinal tissues than the other tissues.; Refer to 5) EF065516.1 for more isolates of HKU9
3-day-old suckling BALB/c rats were intracerebrally inoculated with 20 μl of volume grinding supernatant of ZC45 intestinal tissue. Brain, lungs, intestine, and liver tissues were removed from infected rats on 14 dpi. Varying degrees of inflammation were observed in H&E staining of those tissues, but inflammation in brain tissues was the most evident. N protein antigen was shown in brain and lung tissues (highest level in brain) by western blotting. qRT-PCR detected the highest levels of virus in lung tissues.

Another HKU4 isolate. Refer to 69) NC_009019.1 and 5) EF065516.1

Another HKU4 isolate. Refer to 69) NC_009019.1 and 5) EF065516.1

Another HKU5 isolate. Refer to 75) EF065512.1 and 5) EF065516.1

Another HKU5 isolate. Refer to 75) EF065512.1 and 5) EF065516.1
Isolated from tracheal samples. The authors by including the upper part of the ceca (containing the cecal tonsils), this increased the isolation scores greatly.
Isolated from a chicken with nephritis, depression, and slight respiratory signs. Twenty-day-old chickens inoculated with HH06 exhibited depression, coughing, sneezing, dyspnea, and diarrhea. Necropsy of the infected chickens also showed abnormalities in kidney and respiratory tract.
Isolated from quails with acute enteritis. Short-lived diarrhea appeared in the parents and offspring developed diarrhea on day 2–4 of life

https://mc.manuscriptcentral.com/vevolu
Isolated from a pheasant with tracheitis and nephritis. Virus was detected in the oropharyngeal swab of chickens infected with this pheasant I0623/17 virus. No virus detected in cloacal swab. Viral replication was also found in the trachea and lung, but not kidney, of the infected chickens.

LHB/110617 was isolated from a chicken with respiratory disease. LHB/110615 (isolated by the same group), which has 99.7% sequence identity with LHB/110617, showed high virus titre in trachea sample (minimal virus was detected in kidney sample).

Isolated from a chicken with respiratory problems

Isolated from chickens with nephritis

CoV infection was associated with hedgehogs with green and yellow faeces.

There are other HKU23 strains isolated from faecal samples

229E-related

HKU4-related

This is the most closely related bat sequence to SARS-CoV-2; sample collected in 2013 (RaTG13)

In general, respiratory tract specimens showed the highest positive rates of SARS-CoV-2

In the same study, 6 more pangolins were tested positive and the virus was isolated from intestines (2), lungs (2), blood (1) and scale (1).

NC_002645.1 is a laboratory-adapted strain; Another 229E strain was isolated from the nasopharyngeal swab of a patient with fever and respiratory infection; 229E causes common cold and occasionally associated with more severe respiratory infections in children

Refer to 5) EF065516.1; Another study by the same group showed 29 of 99 alimentary samples were positive for RdRp, none detected in the respiratory samples

Another HKU9 isolate by the same group of 5) EF065516.1. Refer to 5) EF065516.1 and 70) MG762674.1

5 of 43 Scotophilus bats were positive for RdRp and they were all detected in anal samples

RdRp was detected in the alimentary specimen of 36 bats, but not in respiratory specimen;

Refer to 5) EF065516.1 for more isolates of HKU2 by the same group,

Refer to 5) EF065516.1; Smith et al. also detected HKU8 RNA in faecal samples

15 of 420 bulbuls tested were positive for HKU11 CoV (not sure whether it is from throat or cloacal swab); 10 more HKU11 strains were isolated from rectal swab in another study by the same group.
HKU12 was found in 4 of 365 thrushes tested (not sure whether it is from throat or cloacal swab); A HKU12 strain was isolated from rectal swab in another study by the same group.

HKU13 was detected in 2 of 127 munias tested (not sure whether it is from throat or cloacal swab); 6 more HKU13 strains were isolated from rectal swab in another study by the same group.

HKU16 was detected in 3 of 35 CoV-positive bird specimens.
HKU17 was detected in 7 of 35 CoV-positive bird specimens.
HKU18 was detected in 1 of 35 CoV-positive bird specimens.
HKU19 was detected in 5 of 35 CoV-positive bird specimens.
HKU20 was detected in 1 of 35 CoV-positive bird specimens.
HKU21 was detected in 1 of 35 CoV-positive bird specimens.

Related publication PMID - 26433221

Sequence analysis of the partial spike gene showed that this virus strain is more closely related to FRSCV (CoV detected in ferrets with systemic pyogranulomatous inflammation) than FRECV. Heart, liver, spleen, lung, kidney, brain, and intestinal tissues were collected from 3 GCCDC1-infected bats. The presence of GCCDC1 was only detected in intestinal tissues.

RdRp was detected in rectal swab, but not in oral swab and blood.
Table S1. Summary of the sequences included in analyses of coronavirus genome dinucleotide composition, broken down by species.

| Species | Sequences Included |
|---------|--------------------|
| SARS-CoV-2 | 10,000             |
| SARS-CoV  | 5,000              |
| MERS-CoV  | 2,500              |

Sequences were downloaded from NCBI and sequences less than 10% divergent at the nucleotide level were removed from the dataset.
tion, broken down by genus, host, and replication site. All complete coronavirus genomes were downloa
added from NCBI and sequences less than 10% divergent at the nucleotide level were removed from the c
dataset.
| EPI_ISL_414616, EPI_ISL_414617, EPI_ISL_414618, EPI_ISL_414620, EPI_ISL_414621, EPI_ISL_414622 | Servicio Microbiología, Hospital Clínico Universitario, Valencia | Sequencing and Bioinformatics Service and Molecular Epidemiology Research Group, PSIBio-Public Health Unit, Barcelona |
| EPI_ISL_414623 | Laboratoire de Virologie Institut de Virologie - INSERM U1190 Hôpitaux Universitaires de Strasbourg | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414624 | Centre Hospitalier Universitaire de Rouen Laboratoire de Virologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414625 | Centre Hospitalier Régional Universitaire de Nantes Laboratoire de Virologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414626 | Hôpital Robert Debré Laboratoire de Virologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414631, EPI_ISL_414632 | Hôpital René Dubois Laboratoire de Microbiologie - Bât A | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414633 | Centre Hospitalier Compagne Laboratoire de Biologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414634, EPI_ISL_414635, EPI_ISL_414636, EPI_ISL_414637, EPI_ISL_414638 | Centre Hospitalier Compagne Laboratoire de Biologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_414639 | Department of Virology and Immunology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland | Pathogen Sequencing Lab, National Institute for Biomedical Research (INRB) |
| EPI_ISL_414640, EPI_ISL_414641, EPI_ISL_414642, EPI_ISL_414643, EPI_ISL_414644 | Viral Recombination Lab, National Institute for Biomedical Research (INRB) | Andersen Lab, The Scripps Research Institute |
| EPI_ISL_414646 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_414667 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_414689, EPI_ISL_414690, EPI_ISL_414691, EPI_ISL_414692 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_414934, EPI_ISL_414936, EPI_ISL_414937, EPI_ISL_414938, EPI_ISL_414940, EPI_ISL_414941 | Shandong Provincial Center for Disease Control and Prevention, Weifang City, China | Beijing Institute of Microbiology and Epidemiology |
| EPI_ISL_415010 | Wales Specialist Virology Centre | Public Health Wales Microbiology Cardiff |
| EPI_ISL_415015 | Laboratório Central de Saúde Pública Professor Gonçalo Moniz - LACENBA | Instituto Oswaldo Cruz Fiocruz - Laboratory of Respiratory Viruses and Measles (LVRS) |
| EPI_ISL_415128 | Respiratory Virus Unit, Microbiology Services Colindale, Public Health England | Respiratory Virus Unit, Microbiology Services Colindale, Public Health England |
| EPI_ISL_415151 | Gorgias Clinical Microbiology Laboratories | MSH Pathogen Surveillance Program |
| EPI_ISL_415152 | Gorgias Memorial Institute for Health Studies | Gorgias Memorial Institute for Health Studies |
| EPI_ISL_415153, EPI_ISL_415154 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415155 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415156, EPI_ISL_415157, EPI_ISL_415158, EPI_ISL_415159 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415435 | Wales Specialist Virology Centre | Public Health Wales Microbiology Cardiff |
| EPI_ISL_415445 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |
| EPI_ISL_415455, EPI_ISL_415456, EPI_ISL_415457 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |
| EPI_ISL_415458, EPI_ISL_415459 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |

| EPI_ISL_415460 | Servicio Microbiología, Hospital Clínico Universitario, Valencia | Sequencing and Bioinformatics Service and Molecular Epidemiology Research Group, PSIBio-Public Health Unit, Barcelona |
| EPI_ISL_415462 | Centre Hospitalier Universitaire de Rouen Laboratoire de Virologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_415464 | Hôpital Robert Debré Laboratoire de Virologie | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris |
| EPI_ISL_415466 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_415468 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_415469 | State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China | The First Affiliated Hospital of Guangzhou Medical University & BGI-Shenzhen |
| EPI_ISL_415493, EPI_ISL_415496, EPI_ISL_415497, EPI_ISL_415513 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415512, EPI_ISL_415513, EPI_ISL_415514, EPI_ISL_415515, EPI_ISL_415516, EPI_ISL_415517 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415518, EPI_ISL_415519 | KU Leuven, Clinical and Epidemiological Virology | KU Leuven, Clinical and Epidemiological Virology |
| EPI_ISL_415535 | Wales Specialist Virology Centre | Public Health Wales Microbiology Cardiff |
| EPI_ISL_415544 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |
| EPI_ISL_415545, EPI_ISL_415546, EPI_ISL_415547 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |
| EPI_ISL_415548, EPI_ISL_415549 | Hôpitaux universitaires de Genève Laboratoire de Virologie | Hôpitaux universitaires de Genève Laboratoire de Virologie |
see above

EPI_ISL_417383
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Maddocks S, Kok J and Dwyer DE for the 2019-nCoV Study Group

EPI_ISL_417384
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Lam C, Dwyer DE, Rockett R and Eden J-S for the 2019-nCoV Study Group

EPI_ISL_417385
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Lam C, Dwyer DE, Rockett R and Eden J-S for the 2019-nCoV Study Group

EPI_ISL_417386
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Lam C, Dwyer DE, Rockett R and Eden J-S for the 2019-nCoV Study Group

EPI_ISL_417387
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C and Gray K for the 2019-nCoV Study Group

EPI_ISL_417388
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K and Timms V for the 2019-nCoV Study Group

EPI_ISL_417389
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V and Gall M for the 2019-nCoV Study Group

EPI_ISL_417390
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V and Gall M for the 2019-nCoV Study Group

EPI_ISL_417391
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V and Gall M for the 2019-nCoV Study Group

EPI_ISL_417392
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V and Gall M for the 2019-nCoV Study Group

EPI_ISL_417393
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V and Gall M for the 2019-nCoV Study Group

EPI_ISL_417394
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H and Holmes EC for the 2019-nCoV Study Group

EPI_ISL_417395
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H and Holmes EC for the 2019-nCoV Study Group

EPI_ISL_417396
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV and Sintchenko V for the 2019-nCoV Study Group

EPI_ISL_417397
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V and Chen SC for the 2019-nCoV Study Group

EPI_ISL_417398
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC and Maddocks S for the 2019-nCoV Study Group

EPI_ISL_417399
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S and Kok J for the 2019-nCoV Study Group

EPI_ISL_417400
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Rockett R, Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S and Kok J for the 2019-nCoV Study Group

EPI_ISL_417401
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Eden J-S, Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R and Eden J-S for the 2019-nCoV Study Group

EPI_ISL_417402
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Lam C, Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R and Eden J-S for the 2019-nCoV Study Group

EPI_ISL_417403
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Gray K, Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C and Gray K for the 2019-nCoV Study Group

EPI_ISL_417404
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Timms V, Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C and Gray K for the 2019-nCoV Study Group

EPI_ISL_417405
Centre for Infectious Diseases and Microbiology Public Health
NSW Health Pathology - Institute of Clinical Pathology and Medical Research; Westmead Hospital; University of Sydney
Gall M, Amott A, Sadsad R, Carter I, Rahman H, Holmes EC, O'Sullivan MV, Sintchenko V, Chen SC, Maddocks S, Kok J, Dwyer DE, Rockett R, Eden J-S, Lam C, Gray K and Timms V for the 2019-nCoV Study Group

https://mc.manuscriptcentral.com/vevolu
