Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense

Xuhua Xia[1,2]

University of Ottawa

Department of Biology, University of Ottawa, Ottawa, Canada K1N 6N5 Email:

xxia@uottawa.ca

Ottawa Institute of Systems Biology, Ottawa, Canada K1H 8M5

ABSTRACT

Wild mammalian species, including bats, constitute the natural reservoir of Betacoronavirus (including SARS, MERS, and the deadly SARS-CoV-2). Different hosts or host tissues provide different cellular environments, especially different antiviral and RNA modification activities that can alter RNA modification signatures observed in the viral RNA genome. The zinc finger antiviral protein (ZAP) binds specifically to CpG dinucleotides and recruits other proteins to degrade a variety of viral RNA genomes. Many mammalian RNA viruses have evolved CpG deficiency. Increasing CpG dinucleotides in these low-CpG viral genomes in the presence of ZAP consistently leads to decreased viral replication and virulence. Because ZAP exhibits tissue-specific expression, viruses infecting different tissues are expected to have different CpG signatures, suggesting a means to identify viral tissue-switching events. I show that SARS-CoV-2 has the most extreme CpG deficiency in all known Betacoronavirus genomes. This suggests that SARS-CoV-2 may have evolved in a new host (or new host tissue) with high ZAP expression. A survey of CpG deficiency in viral genomes identified a virulent canine coronavirus (Alphacoronavirus) as possessing the most extreme CpG deficiency, comparable to that observed in SARS-CoV-2. This suggests that the canine tissue infected by the canine coronavirus may provide a cellular environment strongly selecting against CpG. Thus, viral surveys focused on decreasing CpG in viral RNA genomes may provide important clues about the selective environments and viral defenses in the original hosts.

Keywords: SARS-CoV-2; viral evolution; canine intestine; zinc finger antiviral protein; COVID-19

INTRODUCTION

Coronaviruses (CoV) evolve in mammalian hosts and carry genomic signatures of their host-specific environment, especially the host-specific antiviral and RNA modification activities. Many pathogenic single-stranded RNA viruses, including coronaviruses, exhibit strong CpG deficiency (Atkinson *et al.*, 2014; Greenbaum *et al.*, 2008; Greenbaum *et al.*, 2009; Takata *et al.*, 2017; Yap *et al.*, 2003). Two mammalian enzymes are inferred to contribute to the observed CpG deficiency. The zinc finger antiviral protein (ZAP, known as ZC3HAV1 in mammals or hZAP in human), a key component in mammalian interferon-mediated immune response, binds specifically to CpG dinucleotides in viral RNA genomes via its RNA-binding domain (Meagher *et al.*, 2019). ZAP inhibits viral replication and mediates viral genome degradation (Ficarelli *et al.*, 2020; Ficarelli *et al.*, 2019; Meagher *et al.*, 2019; Takata *et al.*, 2017). ZAP has two isoforms (ZAP-L and ZAP-S); both participate in initiating antiviral activities but only ZAP-S mediates the return to homeostasis after the antiviral response (Schwerk *et al.*, 2019). ZAP acts against not only retroviruses such as HIV-1 (Ficarelli *et al.*, 2020; Ficarelli *et al.*, 2019), but also Echovirus 7 (Odon *et al.*, 2019) and Zika virus (Trus *et al.*, 2019), both being positive-sense single-stranded RNA viruses like coronaviruses. In particular, selection against CpG in viral RNA disappears in ZAP-deficient cells (Takata *et al.*, 2017), suggesting that ZAP may be the only cellular agent targeting CpG in viral RNA genomes.

Experimental evidence is consistent with the interpretation that CpG deficiency in RNA viruses has evolved in response to these cytoplasmic CpG-specific antiviral activities. During natural evolution of HIV-1 within individual patients, viral fitness decreased with increasing CpG dinucleotides (Theys *et al.*, 2018). Experimental increase of CpG dinucleotides in CpG-deficient viral genomes consistently leads to strong decrease in viral replication and virulence (Antzin-Anduetza *et al.*, 2017; Burns *et al.*, 2009; Fros *et al.*, 2017; Trus *et al.*, 2019; Tulloch *et al.*, 2014; Wasson *et al.*, 2017), prompting the proposal of vaccine-development strategies involving increasing CpG to attenuate pathogenic RNA viruses (Burns *et al.*, 2009; Ficarelli *et al.*, 2020; Trus *et al.*, 2019; Tulloch *et al.*, 2014).

Another antiviral enzyme is APOBEC3G, found in innate immune cells. APOBEC3G was originally thought specific to single-stranded DNA such as reverse-transcribed HIV-1, but is now known to modify a variety of RNA viruses, deaminating C to U (Sharma *et al.*, 2016; Sharma *et al.*, 2015; Sharma *et al.*, 2019). This would be effective against RNA viruses if the deaminated sites are functionally important. APOBEC3G co-purifies with highly edited mRNA substrates (Sharma *et al.*, 2016) and therefore could act on coronavirus genomes which are positive-sense single-stranded RNA. While APOBEC3G is not strongly CpG-specific, it could contribute to CpG deficiency when coupled with ZAP-mediated antiviral activities targeting CpG. Modification of CpG to UpG in non-functional regions could reduce viral susceptibility to CpG-mediated attack by ZAP relative to viruses with unmodified CpG dinucleotides.

Both ZAP and APOBEC3G exhibit tissue-specific expression patterns in human (Fagerberg *et al.*, 2014). Both are expressed in lungs, but ZAP is the most highly expressed where lymphocytes are the most abundant (bone marrow, lymph node, appendix, and spleen), whereas APOBEC3G is the most highly expressed in lymph node, spleen, and testis (Fagerberg *et al.*, 2014). A severely CpG deficient virus may indicate an evolutionary history in ZAP-abundant tissues, such as strongly CpG-deficient HIV-1 infecting host T cells in lymph organs where ZAP is abundant (Fagerberg *et al.*, 2014). The presence of such viruses indicates that they have found ways to evade ZAP-mediated cellular antiviral defense.

The differential expression of ZAP and APOBEC3G in different host or host tissues is expected to leave different genomic signatures on viral RNA genomes. We may use the conventional index of CpG deficiency (Cardon *et al.*, 1994; Karlin *et al.*, 1997) implemented in DAMBE (Xia, 2018):

$$I_{CpG} = \frac{P_{CpG}}{P_C P_G} \qquad (1)$$

The index is expected to be 1 with no deficiency or excess, smaller than 1 if deficient and greater than 1 if excess. The 1252 Betacoronavirus (BetaCoV) full-length genomes deposited in GenBank (of which 1127 are unique), have mean±SE value of 0.516±0.0017 for I_{CpG}, which is significantly ($p < 0.0001$) smaller than their null expectation of 1.

If a coronavirus infects a different host tissue with different ZAP abundance, then its RNA genome will experience different selection pressure against its CpG. This difference in cellular antiviral activity would result in differences in I_{CpG} during viral genomic evolution. In contrast, a coronavirus infecting a specific host tissue for a long time would experience the same cellular antiviral and RNA modification environment and is consequently expected to have similar and stable I_{CpG}.

RESULTS

SARS-CoV-2 and its most closely related known relative (BatCoV RaTG13) have the lowest I_{CpG} among its close relatives, both being outliers in a plot of viral genomic I_{CpG} versus GC% (Fig. 1). Three groups of BetaCoV most closely related to SARS-CoV-2 are represented in Fig 1. Group 1 consists of genomes of human CoV-HKU1 which is found only in human (Dominguez *et al.*, 2012) and circulates among human populations without any dependence on other mammalian species as intermediate or reservoir species. Group 2 includes Betacoronavirus 1 genomes with two types of hosts: 1) ungulates (with bovine and equine coronavirus as well as porcine hemagglutinating encephalomyelitis virus), and 2) human, with CoV-OC43 being a recent derivative of bovine coronaviruses (Hulswit *et al.*, 2019). Group 3 are all SARS-related coronaviruses from three types of hosts: 1) *Rhinolophus* bats which serve as a natural reservoir of SARS-related coronaviruses (Li *et al.*, 2005; Wu *et al.*, 2016a; Wu *et al.*, 2016b) and the new SARS-CoV-2 (Zhou *et al.*, 2020), 2) civets (from which coronavirus genomes with 99.6% identity to SARS virus genomes were identified) (Shi and Hu, 2008), and 3) human patients infected by SARS-CoV-2. Fig. 1 shows that genomic GC% and I_{CpG} can differ among different viral lineages in the same host, or among different hosts for the same viral lineage.

The most striking pattern in Fig. 1 is an isolated but dramatic shift in the lineage leading to BatCoV RaTG13 which was reported (Zhou *et al.*, 2020) to be sampled from a bat (*Rhinolophus affinis*) in Yunnan Province in 2013 but only sequenced by Wuhan Institute of Virology after the outbreak of SARS-CoV-2 infection in late 2019. This bat CoV genome is the closest phylogenetic relative of SARS-CoV-2 (Zhou *et al.*, 2020), sharing 96% sequence similarity. Many studies have shown an association between decreased CpG (low I_{CpG}) in viral RNA genomes and increased virulence, not only in HIV evolving within individual patients (Theys *et al.*, 2018), but also in experimentally CpG dinucleotide-enriched viral genomes (Antzin-Anduetza *et al.*, 2017; Burns *et al.*, 2009; Fros *et al.*, 2017; Trus *et al.*, 2019; Tulloch *et al.*, 2014; Wasson *et al.*, 2017). The association between decreased CpG and increased virulence in RNA viruses is mainly due to interferon-induced ZAP protein which binds to CpG dinucleotides in viral RNA genomes by its RNA-binding domain (Meagher *et al.*, 2019), inhibits viral replication and facilitates viral genome degradation (Ficarelli *et al.*, 2020; Ficarelli *et al.*, 2019; Meagher *et al.*, 2019; Takata *et al.*, 2017). Thus, a decreased I_{CpG} in a viral pathogen suggests an increased threat to public health, but an increased I_{CpG} decreases the threat because such viral pathogens, with increased I_{CpG} and reduced virulence, would be akin to natural vaccines. Many viral researchers have in fact proposed vaccine development by increasing CpG in viral RNA genomes (Burns *et al.*, 2009; Ficarelli *et al.*, 2020; Trus *et al.*, 2019; Tulloch *et al.*, 2014).

In this context, it is unfortunate that BatCoV RaTG13 was not sequenced in 2013, otherwise the downshifting in I_{CpG} might have served as a warning due to two highly significant implications. First, the virus likely evolved in a tissue with high ZAP expression which favors viral genomes with a low I_{CpG}. Second and more importantly, survival of the virus indicates that it has successfully evaded ZAP-mediated antiviral defense. In other words, the virus has become stealthy and dangerous to humans.

The I_{CpG} value for BatCoV RaTG13 genome is 0.40875, much lower than I_{CpG} values observed in all other betacoronavirus genomes sampled from bat species in the genus *Rhinolophus*. There are 56 betacoronavirus genomes sampled from *Rhinolophus* bats inhabiting south and southeastern Asia (but mostly from central and southern China). Nature had essentially inoculated betacoronaviruses into various *Rhinolophus* lineages and allowed genomic evolution to happen (Supplemental Fig. 1). While genomic I_{CpG} values have fluctuated in different viral lineages, only BatCoV RaTG13 has been observed to possess an extraordinarily low I_{CpG} (Supplemental Fig. 1). This suggests that the ancestor of BatCoV RaTG13 and SARS-CoV-2 may have evolved in a mammalian tissue with high expression of ZAP and emerged with an unusually low-I_{CpG}. This mammalian tissue likely is not in *Rhinolophus* bats because low I_{CpG} has not been observed in other BatCoV lineages (Supplemental Fig. 1). Identifying a virus with comparably low I_{CpG} would suggest candidate host species possessing tissues with cellular environments that select strongly against CpG in viral genomes.

Among all Betacoronaviruses available in GenBank on Feb. 3, 2020, there are 1127 unique genomes of which 927 genomes have explicit host designations (Supplemental File Betacoronavirus_CpG.xlsx). Surprisingly, no available Betacoronavirus genome from diverse natural hosts has a genomic I_{CpG} and GC% combination close to that observed in SARS-CoV-2 and BatCoV RaTG13 (Fig. 2). BetaCoV lineages parasitizing *Rhinolophus* bats overall have relatively low I_{CpG} values (Fig. 2).

BetaCoV infecting dromedary camels offers a weak hint that camel digestive system may select more strongly against CpG in viral genomes than camel respiratory system. Camel coronaviruses form two clusters. One cluster overlaps completely with MERS viruses (Fig. 2) that infect mammalian respiratory system (Fehr and Perlman, 2015; Li, 2016). The other cluster includes camel coronavirus HKU23 strains positioned close to bovine CoV (grouped under "Ungulate_CoV" in Figs. 1 and 2), both belonging to Embecovirus and infecting mainly mammalian digestive system but also respiratory systems (Athanassious *et al.*, 1994; Chae *et al.*, 2019; Fulton *et al.*, 2015; Ribeiro *et al.*, 2016; Symes *et al.*, 2018). Those viruses infecting camel digestive system have lower genomic I_{CpG} and GC% than those infecting camel respiratory system (Fig. 2).

To search for a mammalian host with the potential to select viral lineages with low I_{CpG} values, I expanded the search to include all complete alphacoronavirus (AlphaCoV) genomes (Supplemental File Alphacoronavirus_CpG.xlsx). All complete AlphaCoV genomes (>27000 nt) with explicit host information are plotted in I_{CpG} and GC content in Fig. 3. Five points are worth highlighting. First, only genomes from canine coronaviruses (CCoVs), which had caused a highly contagious intestinal disease worldwide in dogs (Le Poder, 2011; Pratelli, 2006), have genomic I_{CpG} and GC% values similar to those observed in SARS-CoV-2 and BatCoV RaTG13 (Fig. 3A). The genome (accession KP981644) is from the most virulent pantropic CCoV invading multiple canine organs (Buonavoglia *et al.*, 2006; Decaro *et al.*, 2007; Zappulli *et al.*, 2008). It belongs to a clade with the lowest observed I_{CpG} values (Fig. 3B).

Second, canids, like camels, also have coronaviruses infecting their respiratory system (canine respiratory coronavirus or CRCoV belonging to BetaCoV). There are two genomes sequenced for CRCoV (accessions JX860640 and KX432213). Their genomic I_{CpG} values are 0.4756 and 0.4684, respectively, substantially higher than those for CCoVs infecting the digestive system (Fig. 3A). Thus, similar to the pattern observed in coronaviruses infecting camels, CCoVs infecting canine digestive system have I_{CpG} much lower than CRCoVs infecting canine respiratory system.

Third, none of the available AlphaCoV genomes from bats or other mammalian host species possess genomic I_{CpG} and GC% values similar to those observed in SARS-CoV-2 and BatCoV RaTG13 (Fig. 3). Thus, although AlphaCoV infects a diverse array of bat lineages, these bat tissues do not seem to generate AlphaCoV strains with low I_{CpG} values comparable to SARS-CoV-2 and BatCoV RaTG13.

Fourth, I want to highlight one data point involving a CCoV genome represented as a green dot in Fig. 3 (highlighted by a green arrow in Fig. 3A, genome accession KC175339). The CCoV has a genomic GC% of 38.17% and I_{CpG} of 0.4986, much higher than the rest. The virus was originally isolated from a dog but had been propagated extensively in cell culture before being sequenced (Dr. Gary R. Whittaker, pers. comm.). Viruses are propagated in cells that expresses the right cellular receptor for viral entry, but do not mount an immune response to kill the virus or get killed by the virus (Banerjee *et al.*, 2019; Benfield and Saif, 1990). The consequent relaxation of selection against the virus (and against CpG in the CCoV genome) in cell culture would allow CpG in the viral RNA genome to rebound through mutation, which would explain the increased I_{CpG} (KC175339 in the phylogeny in Fig. 3B). This process of regaining CpG is reminiscent of CpG-specific methylation in *Mycoplasma* species where CpG was regained when some lineages lost CpG-specific methyltransferases, with a fast-evolving lineage (*M. pneumoniae*) regaining CpG faster than a slow-evolving lineage (*M. genitalium*) (Xia, 2003). This rapid change in I_{CpG} with environmental change as shown in Fig. 3B has two important implications. First, it suggests the feasibility of tracking certain host-switching or tissue-switching events (which would be impossible if it takes hundreds of years for a virus to change I_{CpG}). Second, many experimental studies (Burns *et al.*, 2009; Ficarelli *et al.*, 2020; Odon *et al.*, 2019; Trus *et al.*, 2019; Tulloch *et al.*, 2014) have demonstrated attenuated virulence in RNA viruses with increasing CpG in the viral RNA genome, and suggested this as an efficient means of vaccine development. The observed increase in I_{CpG} in the CCoV genome through cell culture propagation shows a simple way of increasing CpG by simply propagating the virus without selection against CpG dinucleotides in the viral genome.

Fifth, the cellular receptor for SARS-CoV-2 entry into the cell is ACE2 (angiotensin I converting enzyme 2) (Zhou *et al.*, 2020). ACE2 is pervasively expressed in human digestive system, at the highest levels in small intestine and duodenum (Fig. 3C), with relatively low expression in lung (Fagerberg *et al.*, 2014). This suggests that mammalian digestive system is likely to be infected by coronaviruses. This is consistent with the interpretation that the low I_{CpG} in SARS-CoV-2 was acquired by the ancestor of SARS-CoV-2 evolving in mammalian digestive system. The interpretation is further corroborated by a recent report that a high proportion of COVID-19 patients also suffer from digestive discomfort (Pan *et al.*, 2020). In fact, 48.5% presented with digestive symptoms as their chief complaint.

Humans are the only other host species observed to produce coronavirus genomes with low genomic I_{CpG} values, as shown by the cluster of human Alphacoronavirus NL63 genomes (Fig. 3). This virus mainly infects the respiratory system, but also causes digestive problems in 33% of the patients reporting respiratory problems (Vabret *et al.*, 2005). In a comprehensive study of the first 12 COVID-19 patients in US (Midgley and The COVID-19 Investigation Team, 2020), one patient reported diarrhea as the initial symptom before developing fever and cough (Midgley and The COVID-19 Investigation Team, 2020). Stool samples from 7 out of 10 patients tested positive for SARS-CoV-2, including 3 patients with diarrhea (Midgley and The COVID-19 Investigation Team, 2020), corroborating a previous report of SARS-CoV-2 detection in stool (Holshue *et al.*, 2020). In particular, live SARS-CoV-2 virus was isolated from stool of a COVID-19 patient (Zhang *et al.*, 2020). In this context, it is significant that BatCoV RaTG13, as documented in its genomic sequence in GenBank (MN996532), was isolated from a fecal swab. These observations are consistent with the hypothesis that SARS-CoV-2 has evolved in mammalian intestine or tissues associated with intestine.

Figs. 1 and 2 do not include all betacoronaviruses from all hosts, so betacoronaviruses from other mammalian species may also possess low I_{CpG} values. One example is viruses isolated from pangolins. Nine SARS-CoV-2-like genomes have recently been isolated and sequenced from pangolin and deposited in GISAID database (gisaid.org). The one with the highest sequence coverage (GISAID ID: EPI_ISL_410721) has an I_{CpG} value of 0.3929, close to the extreme low end of I_{CpG} values observed among available SARS-CoV-2 genomes. Thus, SARS-CoV-2, BatCoV RaTG13 and those from pangolin may either have a common ancestor with a low I_{CpG}, or have convergently evolved low I_{CpG} values.

Other than ZAP and ABOBEC3G, the enigmatic Dnmt2 (Okano, Xie, Li 1998b; Okano, Xie, Li 1998a; Dong et al. 2001), originally thought to be a DNA methyltransferase (Dnmt), may also contribute to viral RNA modification. However, Dnmt2 appears to methylate only small RNA (Jeltsch *et al.*, 2017). For this reason it may not be important in shaping I_{CpG} in large coronavirus RNA genomes, although it has been observed to relocate from the nucleus to cytoplasmic stress granules (Schaefer, Lyko 2010a; Schaefer, Lyko 2010b; Dev et al. 2017) where it may participate in the methylation of mRNA (Dev et al. 2017).

These observations allow formation of a hypothesis for the origin and initial transmission of SARS-CoV-2. First, the ancestor of SARS-CoV-2 and BatCoV RaTG13 infected the intestine of a mammalian species (e.g., canids or human ingesting bat meat). Second, the presumably strong selection against CpG in the viral RNA genome in canid intestine resulted in rapid evolution of the virus, with many CpG→UpG mutations leading to reduced genomic I_{CpG} and GC%. The licking of anal regions in canids during mating and other circumstances facilitate viral transmission from the digestive system to the respiratory system. Finally, the reduced viral genomic I_{CpG} allowed the virus to evade human ZAP-mediated immune response and became a severe human pathogen. This suggests the importance of monitoring SARS-like coronaviruses in feral dogs in the fight against SARS-CoV-2.

While the specific origins of SARS-CoV-2 are of vital interest in the current world health environment, this study more broadly suggests that important evidence of viral evolution can be revealed by consideration of the interaction of host defense with viral genomes, including selective pressure exerted by host tissues on viral genome composition.

Acknowledgements

This study is supported by Discovery Grant from Natural Science and Engineering Research Council (NSERC, RGPIN/2018-03878) of Canada. I am particularly grateful to K. Katoh and his colleagues for excellent comments and critical references, and to G. R. Whittacker for information on canine coronavirus (ACCN KC175339). I have also benefitted from discussion with D. Gray, X. Jiang, J. Mennigen, G. Wang, C.-I. Wu, J. Yang and W. Zhai. Five anonymous reviewers contributed significantly to the improvement of the manuscript. Heather Rowe corrected many grammatical errors.

References

Antzin-Anduetza I, Mahiet C, Granger LA, Odendall C, Swanson CM. 2017. Increasing the CpG dinucleotide abundance in the HIV-1 genomic RNA inhibits viral replication. Retrovirology 14(1):49.

Athanassious R, Marsolais G, Assaf R, Dea S, Descôteaux JP, Dulude S, Montpetit C. 1994. Detection of bovine coronavirus and type A rotavirus in neonatal calf diarrhea and winter dysentery of cattle in Quebec: evaluation of three diagnostic methods. Can Vet J 35(3):163-169.

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

Banerjee A, Falzarano D, Rapin N, Lew J, Misra V. 2019. Interferon Regulatory Factor 3-Mediated Signaling Limits Middle-East Respiratory Syndrome (MERS) Coronavirus Propagation in Cells from an Insectivorous Bat. Viruses 11(2):152.

Benfield DA, Saif LJ. 1990. Cell culture propagation of a coronavirus isolated from cows with winter dysentery. J. Clin. Microbiol. 28(6):1454-1457.

Buonavoglia C, Decaro N, Martella V, Elia G, Campolo M, Desario C, Castagnaro M, Tempesta M. 2006. Canine coronavirus highly pathogenic for dogs. Emerg. Infect. Dis. 12(3):492-494.

Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol 83(19):9957-69.

Cardon LR, Burge C, Clayton DA, Karlin S. 1994. Pervasive CpG suppression in animal mitochondrial genomes. Proc. Natl. Acad. Sci. USA. 91:3799-3803.

Chae J-B, Park J, Jung S-H, Kang J-H, Chae J-S, Choi K-S. 2019. Acute phase response in bovine coronavirus positive post-weaned calves with diarrhea. Acta Vet. Scand. 61(1):36-36.

Decaro N, Martella V, Elia G, Campolo M, Desario C, Cirone F, Tempesta M, Buonavoglia C. 2007. Molecular characterisation of the virulent canine coronavirus CB/05 strain. Virus Res. 125(1):54-60.

Dominguez SR, Sims GE, Wentworth DE, Halpin RA, Robinson CC, Town CD, Holmes KV. 2012. Genomic analysis of 16 Colorado human NL63 coronaviruses identifies a new genotype, high sequence diversity in the N-terminal domain of the spike gene and evidence of recombination. The Journal of general virology 93(Pt 11):2387-2398.

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K et al. . 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular & cellular proteomics : MCP 13(2):397-406.

Fehr AR, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods in molecular biology (Clifton, N.J.) 1282:1-23.

Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SJD, Schulz R, Swanson CM. 2020. CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. J Virol 94(6).

Ficarelli M, Wilson H, Pedro Galao R, Mazzon M, Antzin-Anduetza I, Marsh M, Neil SJ, Swanson CM. 2019. KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. eLife 8.

Fros JJ, Dietrich I, Alshaikhahmed K, Passchier TC, Evans DJ, Simmonds P. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. eLife 6.

Fulton RW, Herd HR, Sorensen NJ, Confer AW, Ritchey JW, Ridpath JF, Burge LJ. 2015. Enteric disease in postweaned beef calves associated with Bovine coronavirus clade 2. J Vet Diagn Invest 27(1):97-101.

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4(6):e1000079.

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

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

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A et al. . 2020. First Case of 2019 Novel Coronavirus in the United States. The New England journal of medicine:10.1056/NEJMoa2001191.

Hulswit RJG, Lang Y, Bakkers MJG, Li W, Li Z, Schouten A, Ophorst B, van Kuppeveld FJM, Boons G-J, Bosch B-J et al. . 2019. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. Proc. Natl. Acad. Sci. U S A 116(7):2681-2690.

Jeltsch A, Ehrenhofer-Murray A, Jurkowski TP, Lyko F, Reuter G, Ankri S, Nellen W, Schaefer M, Helm M. 2017. Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. RNA Biol 14(9):1108-1123.

Karlin S, Mrazek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179(12):3899-913.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537:39-64.

Le Poder S. 2011. Feline and canine coronaviruses: common genetic and pathobiological features. Adv Virol 2011:609465-609465.

Li F. 2016. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual review of virology 3(1):237-261.

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H et al. . 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310(5748):676-679.

Meagher JL, Takata M, Gonçalves-Carneiro D, Keane SC, Rebendenne A, Ong H, Orr VK, MacDonald MR, Stuckey JA, Bieniasz PD et al. . 2019. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. Proc. Natl. Acad. Sci. U S A 116(48):24303-24309.

Midgley CM, The COVID-19 Investigation Team. 2020. First 12 patients with coronavirus disease 2019 (COVID-19) in the United States. medRxiv:doi: https://doi.org/10.1101/2020.03.09.20032896.

Odon V, Fros JJ, 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 Res 47(15):8061-8083.

Pan L, Mi Mu, Ren HG, Yang P, Sun Y, Wang R, Yan J, Li P, Hu B, Jin Y et al. . 2020. Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study. Am. J. Gastroenterol. (in press).

Pratelli A. 2006. Genetic evolution of canine coronavirus and recent advances in prophylaxis. Vet. Res. 37(2):191-200.

Ribeiro J, Lorenzetti E, Alfieri AF, Alfieri AA. 2016. Molecular detection of bovine coronavirus in a diarrhea outbreak in pasture-feeding Nellore steers in southern Brazil. Trop. Anim. Health Prod. 48(3):649-653.

Schwerk J, Soveg FW, Ryan AP, Thomas KR, Hatfield LD, Ozarkar S, Forero A, Kell AM, Roby JA, So L et al. . 2019. RNA-binding protein isoforms ZAP-S and ZAP-L have distinct antiviral and immune resolution functions. Nature immunology 20(12):1610-1620.

Sharma S, Patnaik SK, Taggart RT, Baysal BE. 2016. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Scientific reports 6:39100-39100.

Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, Baysal BE. 2015. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nature communications 6:6881-6881.

Sharma S, Wang J, Alqassim E, Portwood S, Cortes Gomez E, Maguire O, Basse PH, Wang ES, Segal BH, Baysal BE. 2019. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol 20(1):37-37.

Shi Z, Hu Z. 2008. A review of studies on animal reservoirs of the SARS coronavirus. Virus Res. 133(1):74-87.

Symes SJ, Allen JL, Mansell PD, Woodward KL, Bailey KE, Gilkerson JR, Browning GF. 2018. First detection of bovine noroviruses and detection of bovine coronavirus in Australian dairy cattle. Aust. Vet. J. 96(6):203-208.

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

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

Trus I, Udenze D, Berube N, Wheler C, Martel MJ, Gerdts V, Karniychuk U. 2019. CpG-Recoding in Zika Virus Genome Causes Host-Age-Dependent Attenuation of Infection With Protection Against Lethal Heterologous Challenge in Mice. Frontiers in immunology 10:3077.

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

Vabret A, Mourez T, Dina J, van der Hoek L, Gouarin S, Petitjean J, Brouard J, Freymuth F. 2005. Human coronavirus NL63, France. Emerg. Infect. Dis. 11(8):1225-1229.

Wasson MK, Borkakoti J, Kumar A, Biswas B, Vivekanandan P. 2017. The CpG dinucleotide content of the HIV-1 envelope gene may predict disease progression. Scientific reports 7(1):8162.

Wu Z, Yang L, Ren X, He G, Zhang J, Yang J, Qian Z, Dong J, Sun L, Zhu Y et al. . 2016a. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. The ISME journal 10(3):609-620.

Wu Z, Yang L, Ren X, Zhang J, Yang F, Zhang S, Jin Q. 2016b. ORF8-Related Genetic Evidence for Chinese Horseshoe Bats as the Source of Human Severe Acute Respiratory Syndrome Coronavirus. The Journal of infectious diseases 213(4):579-583.

Xia X. 2003. DNA methylation and mycoplasma genomes. J. Mol. Evol. 57:S21-S28.

Xia X. 2018. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. Mol Biol Evol 35:1550–1552.

Yap YL, Zhang XW, Danchin A. 2003. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinformatics 4(1):43.

Zappulli V, Caliari D, Cavicchioli L, Tinelli A, Castagnaro M. 2008. Systemic fatal type II coronavirus infection in a dog: pathological findings and immunohistochemistry. Res. Vet. Sci. 84(2):278-282.

Zhang Y, Chen C, Zhu S, Shu C, Wang D, Song J. 2020. Isolation of 2019-nCoV from a Stool Specimen for use under a CC0 license. China CDC Wkly [Internet] 2(8):123-124.

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L et al. . 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature:10.1038/s41586-020-2012-7.

Supporting Online Material

Betacoronavirus_CpG.xlsx with Fig. 1 and Fig. 2 (and the compiled data for generating them).

Alphacoronavirus_CpG.xlsx with Fig. 3 (and the compiled data for generating them).

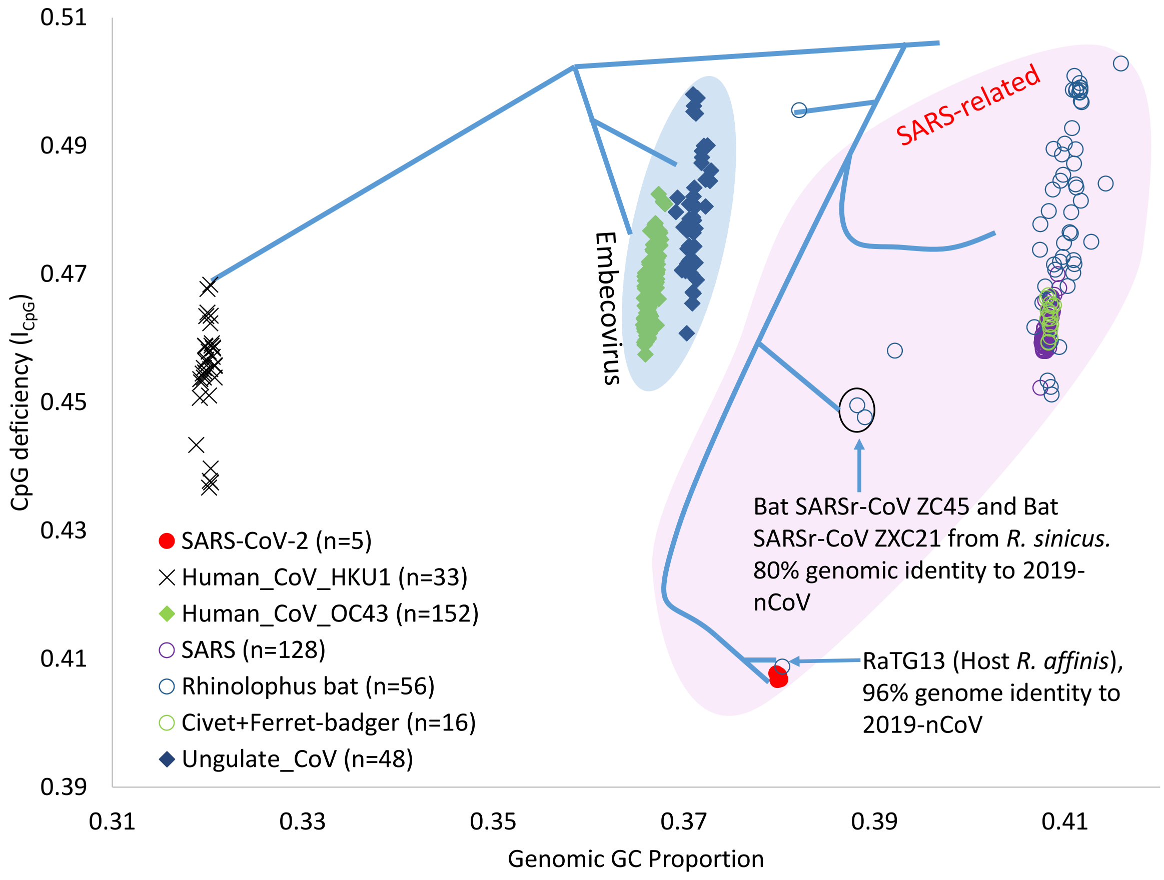

Fig. 1. Different host species of Betacoronavirus (BetaCoV) have different combinations of viral genomic GC% and $I_{CpG}(= P_{CpG}/P_C P_G)$. SARS-CoV-2 and BatCoV RaTG13 are clear outliers with extraordinarily small I_{CpG} values indicative of a host tissue with strong selection against CpG in the viral genome. The first three legends are viral taxonomic names. SARS: SARS CoVs; Rhinolophus bats: BetaCoVs isolated from bats in the genus *Rhinolophus;* Civet+Ferret-badger: BetaCoVs isolated from civet and ferret badger; Ungulate_CoV: BetaCoVs isolated from ungulates (including bovine and equine coronavirus as well as porcine hemagglutinating encephalomyelitis virus). Human coronavirus HKU1 and Human coronavirus OC43 are two members of the viral species Betacoronavirus 1.

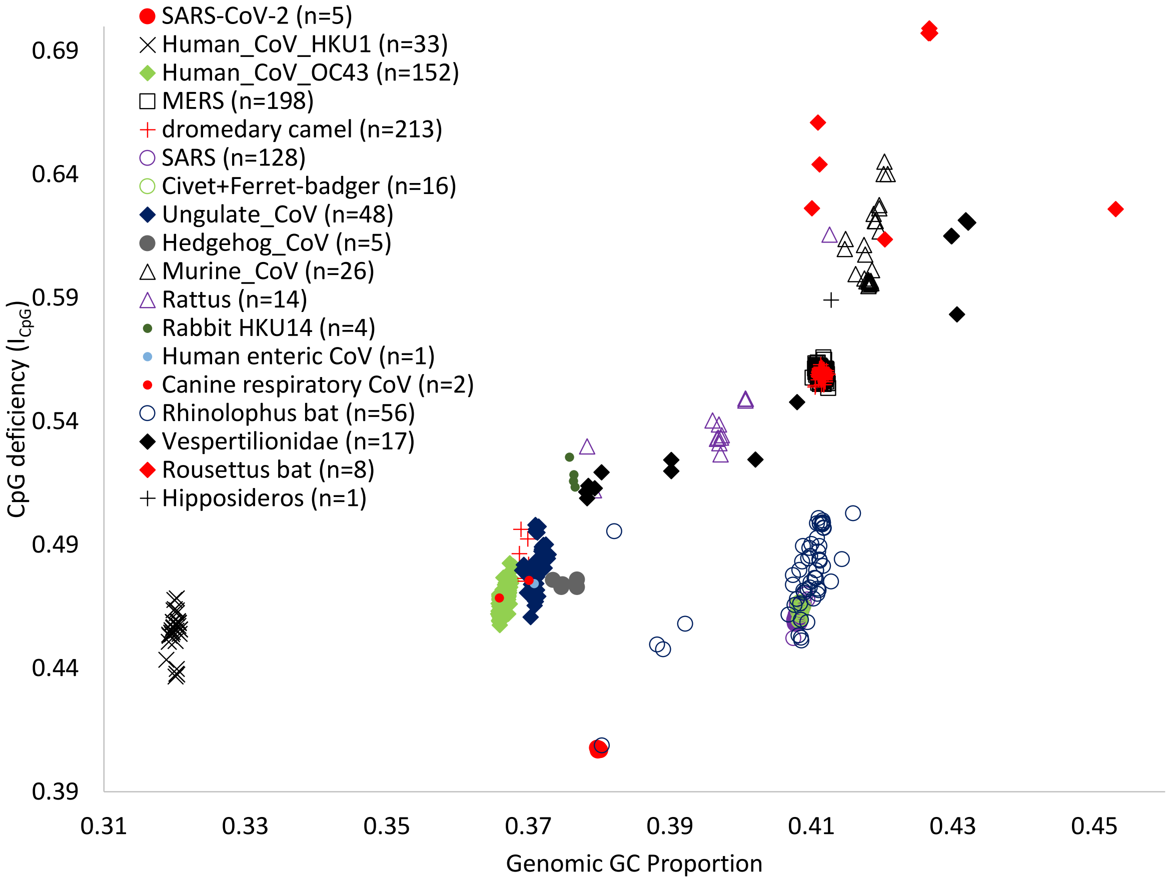

Fig. 2. Genomic GC proportion and I_{CpG} for all known betacoronaviruses (BetaCoV), with a complete genome (≥ 27000 nt) and host information. No Betacoronaviruses from their natural hosts have the genomic I_{CpG} and GC% combination close to SARS-CoV-2 and BatCoV RaTG13. New legends not explained in Fig. 1 are: MERS: MERS CoV; dromedary camel: BetaCoVs isolated from dromedary camels; Hedgehog CoV, Murine CoV, Rattus: BetaCoVs isolated from hedgehog, mouse, and rats; Rabbit HKU14: BetaCoV HKU14 strains isolated from rabbit; Human enteric CoV and Canine respiratory CoV are taxonomic names; Rhinolophus bat, Vespertilionidae, Rousettus bat, Hipposideros: BetaCoV isolated from bats in the genus *Rhinolophus*, in the family Vespertilionidae, in the genus *Rousettus*, and in the genus *Hipposideros*, respectively.

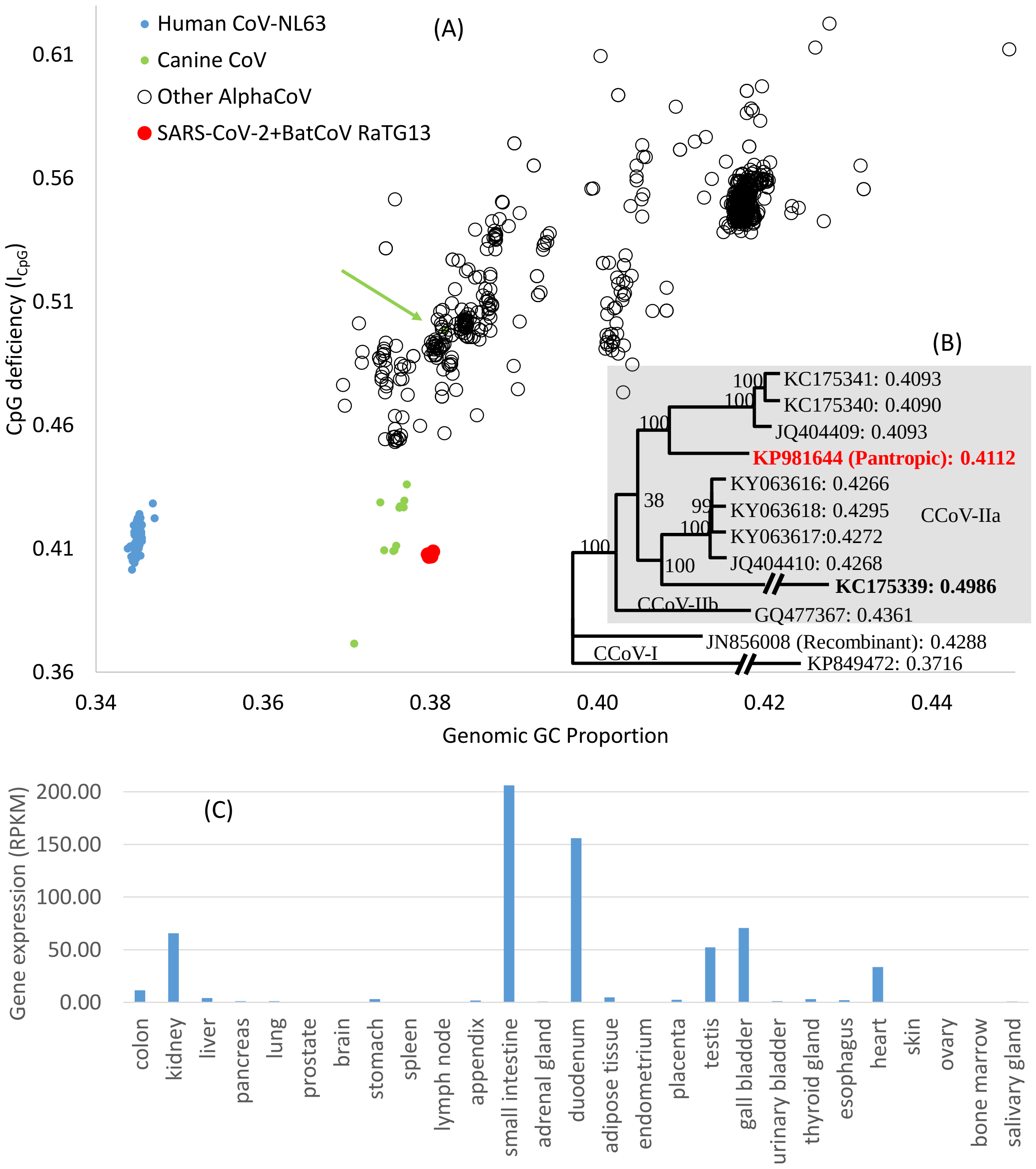

Fig. 3. SARS-CoV-2 may have evolved in mammalian digestive tract. (A) Genomic GC% and I_{CpG} for all alphacoronaviruses with a complete genome (≥ 27000 nt) and host information. Only canine coronavirus (CCoV, intestinal pathogen) genomes have GC% and I_{CpG} combination close to SARS-CoV-2+BatCoV RaTG13. The green arrow points to CCoV (accession KC175339 in the phylogeny) that had been propagated extensively in cell culture before sequencing. (B) Phylogeny from the alignment of all sequenced CCoV genomes, with leaf name in the format of (ACCN: I_{CpG}). Genomes were aligned with MAFFT (Katoh *et al.*, 2009) with the FFT-NS-2 option (more accurate than default). PhyML(Guindon and Gascuel, 2003) with the GTR+Γ substitution model and "best" option

were used to search for the best tree. (C) Tissue-specific gene expression of ACE2, with data from Fagerberg et al (2014).