Submitted to MBE, for consideration as a ***Letter***

Viral CpG deficiency provides no evidence that dogs were intermediate hosts for SARS-CoV-2

David D. Pollock[1,*], Todd A. Castoe[2,*], Blair W. Perry[2], Spyros Lytras[3], Kristen J. Wade[1], David L. Robertson[3], Edward C. Holmes[4], Maciej F. Boni[5], Sergei L. Kosakovsky Pond[6], Rhys Parry[7], Elizabeth J. Carlton[8], James L. N. Wood[9], Pleuni S. Pennings[10], and Richard A. Goldstein[11]

[*]joint primary authors

[1]Department of Biochemistry & Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, USA

[2]Department of Biology, 501 S. Nedderman Dr., University of Texas Arlington, Arlington, TX, USA

[3]MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow, UK.

[4]Marie Bashir Institute for Infectious Diseases & Biosecurity, School of Life & Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[5]Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA, USA.

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

[7]Australian Infectious Disease Research Centre, School of Biological Sciences, The University of Queensland, Brisbane, QLD 4072, Australia

[8]Department of Environmental and Occupational Health, Colorado School of Public Health, University of Colorado, Anschutz, Aurora, CO, USA

[9]Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge UK

[10]Department of Biology, San Francisco State University, San Francisco, CA, USA

[11]Division of Infection & Immunity, University College London, London, UK.

Correspondence and requests for materials should be addressed to David Pollock (email: David.Pollock@CUAnschutz.edu).

Abstract

Due to the scope and impact of the COVID-19 pandemic there exists a strong desire to understand where the SARS-CoV-2 virus came from and how it jumped species boundaries to humans. Molecular evolutionary analyses can trace viral origins by establishing relatedness and divergence times of viruses and identifying past selective pressures. However, we must uphold rigorous standards of inference and interpretation on this topic because of the ramifications of being wrong. Here, we dispute the conclusions of Xia (2020) that dogs are a likely intermediate host of a SARS-CoV-2 ancestor. We highlight major flaws in Xia's inference process and his analysis of CpG deficiencies, and conclude that there is no direct evidence for the role of dogs as intermediate hosts. Bats and pangolins currently have the greatest support as ancestral hosts of SARS-CoV-2, with the strong caveat that sampling of wildlife species for coronaviruses has been limited.

Introduction

The COVID-19 pandemic began following a cross-species transmission event of the causative virus, SARS-CoV-2, sometime in late 2019 (Li et al., 2020; Lu et al., 2020; Gorbalenya et al., 2020; Zhou P. et al., 2020). As the scientific community works to understand the origins, biology, impacts, and treatment strategies for this virus, it is key that we avoid over interpretation of findings and speculation not well supported by available evidence. Otherwise, we risk diversion of time and resources from following more plausible and scientifically justified leads. Accordingly, there is a heightened urgency for the scientific community to diligently survey and critically evaluate new research findings before they are accepted as sound or actionable knowledge.

Understanding the pre-human origins of SARS-CoV-2 is important because it may provide insight into how and why it was able to jump into human populations, in turn better defining the risks of future pandemics. Molecular evolutionary studies have an important role to play in inferring the origins of the virus because they can confirm the relatedness of viruses, shed light on evolutionary time-scales, and potentially identify past selective pressures that allowed the virus to successfully infect and replicate in human hosts. A recent study by Xia (2020) used patterns of CpG deficiency in SARS-CoV-2 and related coronaviruses, and a series of compounding assumptions, to promote "the importance of monitoring SARS-like coronaviruses in feral dogs". His conclusions rest upon the observation that values of CpG deficiency in SARS-CoV-2 (genus *Betacoronavirus*) resemble those observed in distantly related canine alphacoronaviruses that constitute a separate genus within the *Coronaviridae*. Here, we conduct a critical re-evaluation of the conclusions of Xia (2020), highlight key flaws in his underlying logic, and illustrate why his conclusion that dogs are likely intermediate hosts of SARS-CoV-2 is unjustified based on available data. We re-analyze viral CpG deficiency data to incorporate key pangolin viral genomes that were available but omitted from Xia's study. These data further undermine the key inferences and conclusions of Xia (2020).

Clarifying the uncertainty in SARS-CoV-2 origins

To date, the closest known relative of SARS-CoV-2 across its genome as a whole is the RaTG13 virus that was isolated from a horseshoe bat, the established reservoir of the earlier SARS coronaviruses that

emerged in 2002-2003 (Zhou H. et al., 2020). Interestingly, RmYN02, isolated from another horseshoe bat, is more closely related to SARS-CoV-2 in the long replicase 1a reading frame (orf1ab; Zhou P. et al., 2020). The next closest relative of SARS-CoV-2, pangolin-2020, was isolated from pangolins illegally smuggled into Guangdong province, China (Lam et al., 2020; Xiao et al., 2020). Thus, until a closer relative is identified, bats, followed by pangolins, are the most likely source of the originating or reservoir host species for SARS-CoV-2. However, all these viruses are divergent enough from SARS-CoV-2 on an evolutionary time-scale that their role is uncertain (Boni et al., 2020).

A potentially informative feature of the cluster of bat and pangolin coronaviruses similar to SARS-CoV-2 is a region of the Spike protein. This is a key viral feature that binds to the ACE2 receptor in SARS-CoV-2 to enter host cells, and shows strong signs of multiple past recombination events. The Spike binding regions of the pangolin-2020 coronavirus, and that of the 2017 pangolin coronavirus sequence, are more similar to SARS-CoV-2 than that of RaTG13. This suggests that there were multiple recombination events between ancestral viruses related to the bat RaTG13, RmYN02, pangolin-2020, and SARS-CoV-2 lineages (Boni et al., 2020). These findings suggest that such inter-viral recombination events occur commonly among coronaviruses in nature (Zhou H. et al., 2020). Further, there was likely a recombination event in the past involving the variable loop region of the bat RaTG13 virus, although current sampling is insufficient to determine what the parental and offspring sequences were in this recombination event (Boni et al., 2020). For these recombination events to have occurred, divergent viruses must have co-infected the same host. While bats are the only group known to host both ancestral forms of SARS-CoV-2, the two recent host-jumping events indicate that other organisms are also possible candidate hosts. The timing of these events is informed by the extent of divergence among these sequences and the viral mutation rate. Estimated divergence dates between SARS-CoV-2 and RaTG13, suggest that the coronavirus lineage that gave rise to SARS-CoV-2 circulated unnoticed for decades in bats or other intermediate hosts prior to infecting humans (Boni et al. 2020; Nielsen et al., 2020).

Genomic nucleotide content is not good evidence to implicate viral hosts

A well-known feature of most RNA viruses is that they tend to have lower levels of CpG dinucleotides than expected based on the relative frequencies of C and G nucleotides independently (Cheng, 2013;

Jenkins et al., 2001; Karlin et al., 1994; Rima and McFerran, 1997). The SARS-CoV-2 viral genome is more depleted in CpGs than many related coronaviruses (Fig. 1), a trait shared with distantly related alphacoronaviruses in dogs. Based primarily on this observation, Xia (2020) concluded that canines are a likely intermediate (pre-human) host for SARS-CoV-2. The idea is founded on the assumption that CpG levels in SARS-CoV-2 and dog alphacoronavirus are notably low, requiring an unusual environment to evolve, and that the gastro-intestinal tract of dogs is the singular prime candidate to provide that environment. However, the basis of this argument is undermined by the observation that the most closely related sequences from bats and pangolins, several of which were omitted from Xia's (2020) analysis, are also highly depleted in CpGs (Fig. 1 and Supplementary Table S1). In addition, many other RNA viruses are far more depleted in CpGs than is SARS-CoV-2, including pestiviruses that also happen to be found in the pangolin (Gao et al., 2020; Fig. 1). Hence, CpG depletion is not a unique feature of dog viruses or SARS-CoV-2.

Many factors can influence the genomic composition of viruses, including random genetic drift, recombination, and underlying stochastic mutational bias, as well as natural selection (Dunham et al. 2009; Jenkins et al., 2001; Theys et al., 2018). Normally in molecular evolutionary analyses, we assume mutation and drift as the null model, and inference of natural selection, adaptation, and recombination need to be demonstrated by obtaining strong evidence in their favor. Xia (2020), however, provided no compelling evidence for natural selection. It is reasonable to think that natural selection can play a role in viral CpG levels because viral CpG is a target for mammalian defense systems and viruses are likely to evolve to evade such host defense mechanisms. Nevertheless, the evolutionary reasons for low GC content are still debated in even exceptionally well-studied systems with unquestioned animal origins (e.g. HIV-1; Alinejad-Rokny et al., 2016; Antzin-Anduetza et al., 2017; Wasson et al., 2017). As Xia (2020) points out, the mammalian zinc finger antiviral protein (ZAP) binds to CpG dinucleotides in viral RNA genomes and inhibits viral replication and mediates viral degradation (Ficarelli et al., 2020; Ficarelli et al., 2019; Meagher et al., 2019; Takata et al., 2017). Additionally, mammalian APOBEC3G is known to modify viral RNA, deaminating C to U (Sharma et al., 2016; Sharma et al., 2015; Sharma et al., 2019). Notably, bats show unusual and extensive adaptation of APOBEC3G, potentially driving their anti-viral response and perhaps correlating with low CpG content in SARS-like coronaviruses in bats (Jebb et al., 2020). At any point in time, natural selection affecting CpG content may be in a rough

balance with mutation and drift, but differences in CpG content among species could be caused by strengthening or weakening of any of these factors. An altered host environment could induce more extensive targeting of CpGs and positive selection for their removal, or an altered viral life history could lead to stronger selection on viral protein function, including CpGs, and stronger selection for their retention. We can speculate that sequence context-dependency, such as that shown for GATC motifs (Henaut et al., 1996), may also play a role. Likewise, relaxed selection could influence CpG levels in either direction. Further, it has been shown that the genomic dinucleotide composition of RNA viruses is a poor-predictor of host species, suggesting that there is minimal host-specific impact on CpG suppression (Di Giallonardo et al., 2017). For these reasons, gross similarities in CpG depletion characteristics are unreliable for inferring their shared causative nature.

In summary, CpG depletion levels are known to be diverse among RNA viruses broadly, CpG levels are also depleted in non-canine viruses closely related to SARS-CoV-2, evidence that natural selection drove the CpG depletion in SARS-CoV-2 ancestors is lacking, and there are a variety of competing mechanisms for genomes to become relatively depleted in CpG over evolutionary time. Despite this, Xia (2020) speculated that low viral genomic CpG levels in SARS-CoV-2 required evolutionary time in a previous host species and tissue that more actively selected for CpG depletion than do bats. Because low CpG levels, similar to those in SARS-CoV-2, were observed in alphacoronaviruses that infect dog digestive tracts, he then concluded: *"… canine tissue infected by the canine coronavirus may provide a cellular environment selecting against CpG*", and "*This suggests the importance of monitoring SARS-like coronaviruses in feral dogs in the fight against SARS-CoV-2.*" However, there is no evidence for the logical premise of Xia's argument, considering that all mammals have digestive tracts. Additionally, a recent inoculation study found that while other domesticated mammalian hosts are highly susceptible to SARS-CoV-2, canines exhibited low susceptibility, and no traces of viral RNA were detectable in any dog organs (Shi et al., 2020). Further, it is notable that based on a study modeling ACE2 binding affinity with the Spike protein from SARS-CoV-2, it seems highly unlikely that dogs played an important role in the recent evolution of SARS-Cov-2 (Damas et al., 2020). These findings cast further doubt on the relevance of dogs as hosts of viruses related to SARS-CoV-2. Hence, there is no reason to conclude that dogs or dog digestive tracts are special in this respect.

Further analysis indicating that viral CpG depletion levels don't implicate dogs

We reanalyzed the "SARS-related" subset of the data shown in Fig. 1 from Xia (2020), but also including seven betacoronaviruses from pangolins and a bat (RmYN02), four additional dog alphacoronaviruses, and two additional non-coronaviruses (pestiviruses) from pangolins, using the same indices (I_{CpG} – a measure of genomic CpG deficiency, and genomic GC content; Fig. 1). The names of all viruses used in our analysis, along with estimated GC content and I_{CpG} estimates, are provided in Supplementary Table S1). Multiple bat and pangolin betacoronaviruses have low I_{CpG} comparable to SARS-CoV-2, and the other pangolin viruses have even lower I_{CpG}. This non-exhaustive sample is sufficient to refute the claim by Xia (2020) that "no betacoronaviruses from their natural hosts have the genomic I_{CpG} and GC% combination close to SARS-CoV-2 and BatCoV RaTG13". Notably, dog alphacoronaviruses are also not exceptional in terms of CpG deficiency. Furthermore, while humans and dogs have ZAP, which Xia (2020) hypothesizes targets and selects for CpG depletion, our analyses suggest ZAP is highly conserved in mammalian genomes. In particular, bat and pangolin genomes also appear to contain functional ZAP (Supplementary Table S2). APOBEC3G may also be conserved across mammals, but the results are less clear, as similarity to human APOBEC3G is low in other mammals; however, human APOBEC3G is more similar to genes in bats and the pangolin than in dogs (Supplementary Material Table S3). These results are relevant because they mean that bats and pangolins, the most likely pre-human hosts at present, have equal mechanistic potential to select against viral CpG content as dogs. While there is no evidence that SARS-CoV-2 has a low CpG content due to the action or evasion of these mechanisms (or if such a process is responsible for any CpG patterns in any organisms), the distribution of these proteins provides no prior mechanistic basis to exclude bats and pangolins as either reservoirs or intermediate hosts, and provides no evidence to specifically implicate dogs.

In addition to being unsupported by positive evidence, Xia's (2020) hypothesis for dogs as intermediate hosts of ancestral viruses giving rise to SARS-CoV-2 requires an unlikely history of cross-species viral transmission (see Fig. 2 for potential hypotheses) for which there is no evidence. Specifically, this hypothesis minimally requires: 1) an ancestral SARS virus in bats (the main reservoir for SARS-lineage viruses) was passed to dogs, which drove depletion of viral CpGs, 2) dogs passed this virus back to an unknown host or hosts that passed it to bats and pangolins (which gave rise to Pangolin 2020, bat

RmYN02, and bat RaTG13 observed coronaviruses), 3) and descendant lineages of this virus were passed to humans via an unknown host (Fig. 2). In addition to this primary hypothesis, Xia's manuscript and subsequent online comments further imply dogs were a more recent host of SARS-CoV-2, and thus the need for monitoring "in feral dogs" (Fig. 2). A simpler alternative to this improbable transmission hypothesis is that bats transferred this virus directly to humans or through a yet undetermined host (Fig. 2). In our view, it is a problem that potential wild animal hosts have not yet been well sampled. While it may be worthwhile to test dog samples as part of broader efforts to sample diverse potential hosts, a narrow focus on dogs is unjustified by existing evidence.

In summary, the proposition of Xia (2020) that dogs are a likely pre-human host for SARS-CoV-2 is not justified by available evidence. Xia (2020) did not demonstrate that the low CpG frequency in the SARS-CoV-2 genome was driven by a unique selective environment in dog digestive tracts. The SARS-CoV-2 is also less virulent than other human betacoronaviruses (SARS-CoV-1 and MERS-CoV; Chen, 2020; Munster et al., 2020), contradicting his assertion that CpG-deficient viruses are more virulent. Furthermore, closely related betacoronaviruses from bats and pangolins have CpG-deficiencies similar to SARS-CoV-2. Dogs are not more plausible than most other potential host species, and based on current data, far less plausible than bats or pangolins. Still, we are missing ~20-70 years of the recent evolutionary history of the lineage leading to SARS-CoV-2, and we must broadly survey a wide range of wild and domestic species to uncover the origin of SARS-like coronaviruses.

References

Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, and Ebrahimi D. 2016. Source of CpG depletion in the HIV-1 genome. *Mol Biol Evol* 33(12):3205–3212.

Anderson, KG, Rambaut A, Lipkin WI, Holmes EC, and Garry RF. 2020. The proximal origin of SARSCoV-2. *Nat Med* 26:450-452.

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

Boni, MF, Lemey P, Jiang X, Tsan-Yuk Lam T, Perry BW, Castoe TA, Rambout A, and Robertson DL 2020. Evolutionary origins of the SARC-COV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *bioRxiv* doi.org/10.1101/2020.03.30.015008.

Chen J. 2020. Pathogenicity and transmissibility of 2019-nCoV—A quick overview and comparison with other emerging viruses. *Microbes and Infection* 22(2):69-71.

Cheng X, Virk N, Chen W, Ji S, Ji S, Sun Y, and Wu X. 2013. CpG Usage in RNA viruses: data and hypotheses. *PLoS One* 8(9):e74109.

Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M, Hiller M, Koepfli K-P, Pfenning AR, Zho, H, et al. 2020. Broad host range of SARS-CoV-2 predicted by comparative structural analysis of ACE2 in vertebrates. *bioRxiv* doi.org/10.1101/2020.04.16.045302.

Di Giallonardo F, Schlub T E, Shi M, and Holmes EC. 2017. Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *J Virology* 91(8) https://doi.org/10.1128/jvi.02381-16.

Dunham, E J, Dugan VG, Kaser EK, Perkins SE, Brown IH, Holmes EC, and Taubenberger JK. 2009. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J Virol* 83:5485-5494.

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* 5(4):536–544.

Gao WH, Lin XD, Chen YM, Xie CG, Tan ZZ, Zhou JJ, Chen S, Holmes EC, and Zhang YZ. 2020. Newly identified viral genomes in pangolins with fatal disease. *Virus Evolution* 6(1):veaa020. doi.org/10.1093/ve/veaa020.

Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler W, Jermiin LS, Skirmuntt EC, Katzourakis, A, et al. 2020. Six new reference-quality bat genomes illuminate the molecular basis and evolution of 1 bat adaptations. *bioRxiv* doi.org/10.1101/836874.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, et al. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N Engl J Med* 382:1199-1207.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395:565-574.

Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SJD, Schulz R, and 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, and Swanson CM. 2019. KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. *eLife* 8.

Hénaut, A, Rouxel T, Gleizes A, Moszer I, and Danchin A. 1996. Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J Mol Biol* 257(3):574-585

256 Jenkins, GM, Pagel M, Gould EA, de Zanotto PM, and Holmes EC. 2001. Evolution of base composition
257 and codon usage bias in the genus Flavivirus. *J Mol Evol* 52:383-390.

258 Karlin S, Doerfler W, and Cardon LR. 1994. Why is CpG suppressed in the genomes of virtually all small
259 eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* 68:2889-2897.

260 Lam TT, Shum HHM, Zhu HC, Tong Y-G, Ni X-B, Liao YS, Wei W, and Cheung WYM, Li WJ, Li LF, et al.
261 2020. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*.
262 doi.org/10.1038/s41586-020-2169-0.

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

267 Munster VJ, Koopmans M, van Doremalen N, van Riel D, and de Wit E. 2020. A novel coronavirus
268 emerging in china — key questions for impact assessment. *N Engl J Med* 382:692-694.

269 Nielsen R, Wang H, and Pipes L. 2020. Synonymous mutations and the molecular evolution of SARS-
270 Cov-2 origins. *bioRxiv* https://doi.org/10.1101/2020.04.20.052019.

271 Rima BK, and McFerran NV. 1997 Dinucleotide and stop codon frequencies in single-stranded RNA
272 viruses. *J Gen Virol* 78:2859-2870.

273 Sharma S, Patnaik SK, Taggart RT, and Baysal BE. 2016. The double-domain cytidine deaminase
274 APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci Rep* 6:39100-39100.

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

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

281 Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, et al. 2020. Susceptibility of ferrets, cats, dogs, and
282 other domesticated animals to SARS–coronavirus 2. *Science* 7015:eabb7015.
283 https://doi.org/10.1126/science.abb7015.

284 Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, and Bieniasz PD. 2017. CG
285 dinucleotide suppression enables antiviral defense targeting non-self RNA. *Nature*
286 550(7674):124-127.

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

290 Wasson MK, Borkakoti J, Kumar A, et al. 2017. The CpG dinucleotide content of the HIV-1 envelope
291 gene may predict disease progression. *Sci Rep* 7:8162.

292 Xia A. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol*
293 *Biol Evol* doi: 10.1093/molbev/masa095.

Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ, Li N, Guo Y, Li X, Shen X, et al. 2020. Isolation and characterization of 2019-nCoV-like coronavirus from malayan pangolins. *bioRxiv* 10.1101/2020.02.17.951335v1.

Zhang YZ, and Holmes EC. 2020. Commentary: A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* 181, doi.org/10.1016/j.cell.2020.03.035.

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, D., Liu, D., Yang, J., Holmes, E.C., et al. 2020. A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19. BioRxiv. doi.org/10.1101/2020.03.02.974139.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* doi.org/10.1038/s41586-020-2012-7.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgements

DDP, TAC, and ELC are funded by NIH (1R01AI134673); DDP is also funded by NIH GM083127; TAC is also funded by National Science Foundation (DEB-1655571, IOS-1655735); ECH is funded by an ARC Australian Laureate Fellowship (FL170100022); SL and DLR are funded by the MRC (MC_UU_12014/12); JW is supported by The Alborada Trust.

Author contributions

D.D.P. and T.A.C wrote the manuscript; B.W.P. and S.L. analyzed data; all authors contributed concerns about the dog CpG paper and to editing the final manuscript.

Competing interests

The authors declare no competing interests.

Figures

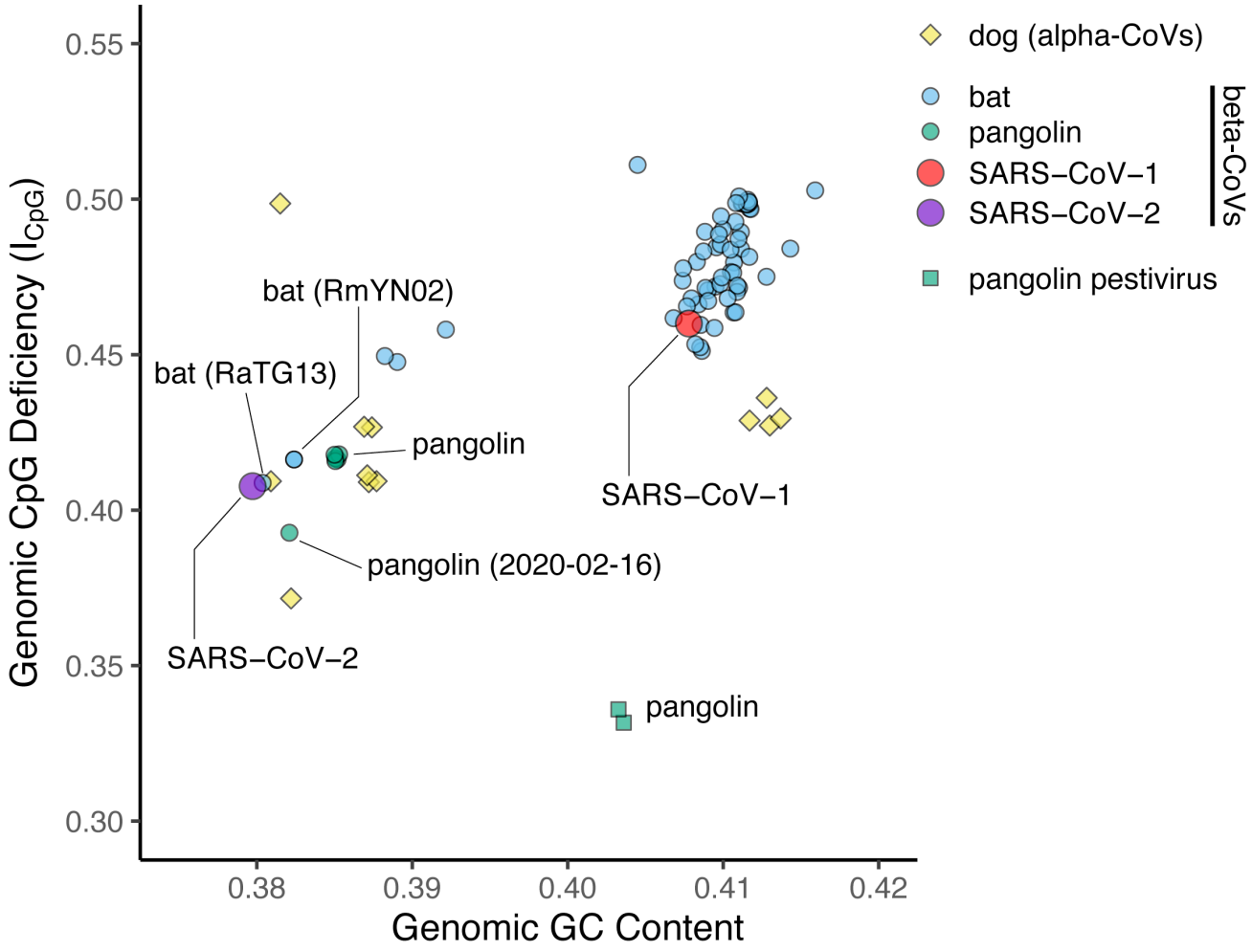

Figure 1. Coronavirus genomic CpG deficiency (I_{CpG}) versus viral genomic GC content for select betacoronaviruses (beta-CoVs), and dog alphacoronaviruses (alpha-CoVs). Pangolin pestiviruses are also shown to illustrate variation in I_{CpG} in a single host.

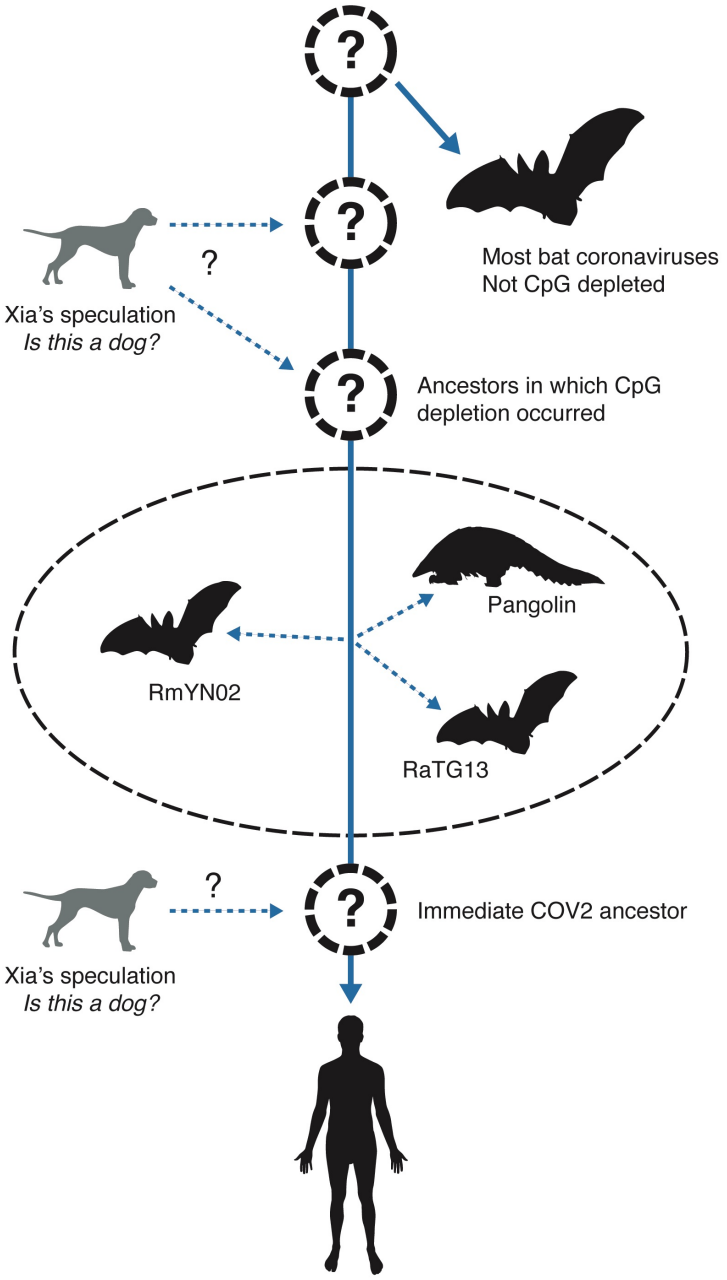

Figure 2. Prevailing origin and transmission hypotheses supported by recent literature. The organisms in black outline are host sources of viral sequences closely related to SARS-CoV-2. The dashed circles represent hosts carrying viruses on the ancestral lineage leading to SARS-Cov-2, with the large question marks indicating that despite the recurrence of bats as hosts of related viruses, the ancestral hosts are uncertain. Two ancestral hosts are indicated during the time of CpG depletion because this is a much

longer timespan, and there could plausibly have been multiple hosts from divergent species during this time. Dogs are represented by grey outlines because no viruses closely related to SARS-CoV-2 have been discovered in dogs. Question mark labeled dashed arrows represent Xia's (2020) dual speculations, that dogs may have been hosts during the process of CpG depletion and during recent ancestral SARS-CoV-2 evolution.

Supplementary Material Online.

Supplementary Table S1. Viral sequences used in Figure 1, along with I_{CpG} and GC content.

Supplementary Table S2. Blast results comparing human ZAP proteins to homologous annotated genes in bat, dog, and pangolin genomes.

Supplementary Table S3. Blast results comparing human APOBEC3G proteins to homologous annotated genes in bat, dog, and pangolin genomes.