Introductory remarks

Alexander Crits-Christoph, Karthik Gangavarapu, Jonathan E. Pekar, Niema Moshiri, Reema Singh, Joshua I. Levy, Stephen A. Goldstein, Marc A. Suchard, Saskia Popescu, David L. Robertson, Philippe Lemey, Joel O. Wertheim, Robert F. Garry, Angela L. Rasmussen, Kristian G. Andersen, Edward C. Holmes, Andrew Rambaut, Michael Worobey, Florence Débarre

In an attempt to ensure transparency and engagement towards global partnership, we would like to provide further context regarding the timeline of work and efforts to collaborate on the research presented in the report below.

On 4 March 2023 (dates in UTC), we discovered accessions posted publicly on the GISAID database corresponding to sequences from environmental samples collected at the Huanan Seafood Wholesale Market, Wuhan. On 9 March, we realized that those accessions were associated with raw metagenomic sequence read data files. We further recognised that it was the data underlying the preprint posted on Research Square by Gao *et al.* at the Chinese Center for Disease Control and Prevention (CCDC) on 25 February 2022 (DOI: 10.21203/rs.3.rs-1370392). The metadata on GISAID indicated these sequencing data had been uploaded in June 2022, however, they evidently had not been released at that time. We downloaded the public data to search for genetic sequences from non-human animals, which the CCDC did not identify in their February 2022 preprint. The preprint also posited that all SARS-CoV-2-positive samples in the market were the result of human infections, claiming that the market was a site of amplification of an already widespread epidemic. We and others therefore had urgently requested release of the data. The potential for analysis of samples for animal DNA had also been recommended in the mission report of the World Health Organization (WHO)-convened global study of origins of SARS-CoV-2: China Part, released March 2021[1].

Once the data were identified on GISAID, it became possible to test the veracity of these claims. We found information that was critical to understanding the nature of the origins of the human infections at the Huanan market, as this was the early epicenter of SARS-CoV-2 spread and was likely where spillover occurred and sustained human-to-human transmission was established.

Our analysis of these data found that genetic evidence of multiple animal species was present in locations of the market where SARS-CoV-2 positive environmental samples had been collected. This includes raccoon dogs, which are susceptible to SARS-CoV-2 infection and shed sufficient virus to transmit to other species. However, this also included

[1] https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part

other mammalian species that require consideration as possible intermediate hosts of SARS-CoV-2. Although live mammals had previously been observed at Huanan market in late 2019, their exact locations were not conclusively known, and some of the animal species we identify in the report below were not included in the list of live or dead animals tested at the Huanan market, as reported in the 2021 WHO-China joint report on the origin of the COVID-19 pandemic. Our results show that they were present. In some cases, the amount of animal genetic material was greater than the amount of human genetic material, consistent with the presence of SARS-CoV-2 in these samples being due to animal infections.

We contacted an author of the Gao *et al.* preprint on 9 March 2023 to inquire about the data, and were told that we could conduct an independent analysis. On 10 March we advised the same author that we had discovered the presence of animal genetic material in the samples. On 11 March 2023, we discovered that the data had been made unavailable (at the request of the submitter according to a statement on GISAID). On the same day we contacted both the corresponding author of the preprint as well as the author who had contributed the raw data to GISAID and asked if they would like to collaborate with us on analyses of these data. On 13 March 2023, those of us who had either downloaded the data, or associated metadata, or contacted the corresponding author of the preprint, received emails from the GISAID Secretariat admonishing us to comply with the GISAID terms of use[2], or in some cases falsely accusing us of having breached the GISAID terms of use. We are well aware of these terms of use, have not breached them, and have no intention of breaching them.

We informed WHO of our preliminary findings on 11 March 2023. On 12 March 2023, some of us met with WHO and some members of SAGO (the WHO-convened Scientific Advisory Group for the Origins of Novel pathogens) to discuss our observations. On 14 March 2023, the WHO convened a meeting with SAGO where some of us and representatives from CCDC presented our respective results. We cannot comment on the CCDC team's findings, as those are theirs to share, but some findings from our analyses have already been shared in the media and in public statements by the WHO[3]. This meeting constituted one of several efforts to establish a collaborative relationship with our colleagues at CCDC to share data and findings as rapidly as possible.

We acknowledge that these circumstances are unusual. We are proponents of open data sharing, and ensuring that data from our analyses are broadly accessible in public repositories is our standard practice. Although our colleagues at the CCDC have stated their intention to share these raw sequence data to support the publication currently undergoing review, they remain inaccessible through GISAID at the time of writing. There is

[2] https://gisaid.org/terms-of-use/

[3] https://www.who.int/news/item/18-03-2023-sago-statement-on-newly-released-sars-cov-2-metageno@mics-data-from-china-cdc-on-gisaid

no clear timeline for data availability, nor any indication of when data may become available if the manuscript is not recommended for publication after peer review. We have also encouraged our Chinese colleagues to seek to immediately share as a preprint their manuscript. At the time of writing, we are not aware that that has happened.

The GISAID terms of use do not preclude the public discussion of data as long as the data generators are acknowledged and best efforts have been made to collaborate with the contributors. CCDC has thus far declined to collaborate on this. We respect our CCDC colleagues' right to be first to publish a manuscript on their own data and do not plan to submit a paper that would compete with their manuscript currently undergoing review. We note, however, that by providing a data generator with the ability to embargo data (for nearly 8 months – the CCDC's data is recorded as having been uploaded on 2 June 2022), GISAID has deviated from its stated mission to overcome "disincentive hurdles and restrictions, which discourage or prevented sharing of virological data prior to formal publication"[4]. Samples from the Huanan Market were collected in January and February 2020 and, given their importance to understanding the origin of the pandemic, we feel this is an unreasonable amount of time to have passed.

[4] https://gisaid.org/about-us/mission/

Genetic evidence of susceptible wildlife in SARS-CoV-2 positive samples at the Huanan Wholesale Seafood Market, Wuhan

Analysis and interpretation of data released by the Chinese Center for Disease Control

Alexander Crits-Christoph, Karthik Gangavarapu, Jonathan E. Pekar, Niema Moshiri, Reema Singh, Joshua I. Levy, Stephen A. Goldstein, Marc A. Suchard, Saskia Popescu, David L. Robertson, Philippe Lemey, Joel O. Wertheim, Robert F. Garry, Angela L. Rasmussen, Kristian G. Andersen*, Edward C. Holmes, Andrew Rambaut, Michael Worobey*, Florence Débarre*

(See Appendix C for affiliations)

* To whom correspondence should be addressed: worobey@arizona.edu, andersen@scripps.edu, florence.debarre@sorbonne-universite.fr

20-Mar-2023

Some of the information contained in this report was initially communicated to the World Health Organization (WHO) on 11 March 2023 (PST), and presented to the WHO's Scientific Advisory Group on the Origin of Novel Pathogens (SAGO) on 14 March 2023. This report is not intended for publication in a journal.

Key Points

- Using metagenomic sequencing data publicly available on GISAID, we provide evidence for the co-occurrence of SARS-CoV-2 and the genetic material of susceptible wildlife in environmental samples from the Huanan market during the start of the COVID-19 pandemic.
  This finding corroborates reports of putative intermediate animal hosts for SARS-CoV-2 being sold live in the market in late 2019 and adds to the body of evidence identifying the Huanan market as the spillover location of SARS-CoV-2 and the epicenter of the COVID-19 pandemic.

- We assembled complete or near complete mitochondrial genomes (DNA) of five wildlife species from SARS-CoV-2-positive environmental samples.
  This provides new leads to investigate the sources of wild-caught or farmed mammals sold live at the Huanan market when the pandemic began.

- We performed *de novo sequence* assembly to obtain sequences from susceptible mammals, including raccoon dogs, from samples that were reported positive for SARS-CoV-2. We identified sequences corresponding to chromosomal DNA and/or RNA.
  The identification of genomic DNA and/or RNA affirms key findings based on mitochondrial DNA.

- The data used to generate this report were made inaccessible on 11 March 2023 (UTC) from the GISAID database. As of 20 March 2023, these data are still not available. In addition, more metagenomic data related to the Huanan market exist but have not yet been shared publicly. Sharing these data is essential to further decipher the sequence of events that likely led to the emergence of SARS-CoV-2 at the Huanan Wholesale Seafood Market.

Overview

A preponderance of COVID-19 cases with the earliest known symptom onsets have been directly or indirectly linked to the Huanan Wholesale Seafood Market, located in the Jianghan district of the city of Wuhan, China[1–3]. Various wild and farmed wildlife species susceptible to infection and capable of transmitting SARS-CoV-2 are documented to have been sold at this market in late 2019[2,4]. Widespread environmental sampling on surfaces and wastewater throughout the market from 1 January 2020 into February 2020 revealed that SARS-CoV-2 was predominantly present in the south-western section of the market[2,5,6], where live mammals were being sold and many human COVID-19 patients with the earliest-known onset of symptoms worked[2].

On or before 9 March 2023, raw sequence data from these market samples, obtained via metatranscriptomic sequencing, were made public through the GISAID SARS-CoV-2 sequence repository, shared by the scientific team responsible for the environmental sampling at the Huanan market. A scientific team lead by the Chinese CDC released a preprint based on these metagenomic data in February 2022, reporting the presence of human nucleic acids in the samples positive for SARS-CoV-2; they did not identify which, if any, other animal species' nucleic acids were present in the samples[6].

We conducted independent analyses of these data and identified genetic material from a variety of wildlife species in the sequenced samples. We find that the RNA and/or DNA sequence reads from susceptible animals are at the highest frequency in wildlife stalls in the southwest corner of the market. This quadrant of the market is where most SARS-CoV-2 environmental RNA was detected, and live mammals were being sold[2]. For multiple samples, mitochondrial nucleic acids from susceptible, or potentially susceptible, animals, including raccoon dogs, were significantly more abundant than human mitochondrial nucleic acids (**Figure 1**).

In some samples, we found that animal mitochondrial DNA (mtDNA) sequences were sufficiently abundant that we were able to assemble near-complete mitochondrial genomes (**Figure 2**). We reconstructed complete or mostly complete mitochondrial genomes for the common raccoon dog (*Nyctereutes procyonoides*), Malayan porcupine (*Hystrix brachyura*), Amur hedgehog (*Erinaceus amurensis*), masked palm civet (*Paguma larvata*), and hoary bamboo rat (*Rhizomys pruinosus*) from wildlife stalls positive for SARS-CoV-2. These sequencing data can aid in tracing the sources of these animals upstream of the market for future investigations of the origins of SARS-CoV-2 within the wildlife trade.

We performed *de novo* assembly using sequencing reads from a wildlife stall where five SARS-CoV-2 positive environmental samples had been collected and previously identified in Worobey *et al.*[2]. These sequences were assigned to their host species of origin using available whole genomes from a subset of the species (including human)

detected in the market. In one sample, we found that these contigs matched both coding transcripts and genomic DNA from the raccoon dog genome at 100% or >99% identity, indicating the presence of both RNA and DNA from this species at that wildlife stall in the market. This sample yielded no contigs that matched the human genome at >99% identity, demonstrating that the RNA/DNA in this sample was largely contributed by other animals (**Figure 3**).

The co-occurrence of SARS-CoV-2 virus and susceptible animal RNA/DNA in the same samples, from a specific section of the Huanan market, and often at greater abundance than human genetic material, identifies these species, particularly the common raccoon dog, as the most likely conduits for the emergence of SARS-CoV-2 in late 2019.

Analysis

The results presented here are based on the analysis of raw sequence data downloaded from GISAID on 9-10 March 2023 (Accessions EPI_ISL_13052298 to EPI_ISL_13052346; see Appendix B for acknowledgments of data generators). They were assigned via their corresponding sample IDs to their market location and sample metadata using the supplementary data from refs.[2,6] (**Table S1-2**). All corresponding samples contained SARS-CoV-2 RNA or had been previously shown to be positive for SARS-CoV-2, either by PCR or sequencing[6]. Importantly, numerous other environmental samples from the market tested negative for SARS-CoV-2, but no metagenomic data have yet been shared from them.

We identified the presence of mitochondrial DNA (mtDNA) sequences of multiple mammalian species either known or predicted to be susceptible to SARS-CoV-2 in environmental samples positive for SARS-CoV-2, including some animals that were not mentioned in the WHO report nor in the preprint by Gao *et al*.[6] (**Table S3-4**). Notably, these samples were from stalls known to have sold live mammals, in the section of the market with the highest density of reported SARS-CoV-2 positive samples[2] (**Figure 1**). These animals include raccoon dogs (*Nyctereutes procyonoides*; present in six samples, and at extremely high abundance in one cart sample), Siberian weasels (*Mustela sibirica*), Amur hedgehog (*Erinaceus amurensis*), and hoary bamboo rat (*Rhizomys pruinosis*). Other animal mtDNA sequences found in these samples include Malayan porcupine (in a sample from a metal cage), dog (*Canis lupus familiaris*, from a feather/hair removal machine), and Himalayan marmot (*Marmota himalayana*, on the carts). Masked palm civet (*Paguma larvata*) mtDNA was found at lower abundances on a sample from one cart.

We detected raccoon dog mtDNA-related sequences in six samples from two stalls in the southwest corner of the market, both previously identified as selling wildlife products. Raccoon dog genetic material was abundant on a SARS-CoV-2 positive sample from a cart, and much more so than human genetic material from the same sample. This finding supports photographic evidence[7] of the presence of live raccoon dogs in this area. The mitochondrial DNA of other mammals experimentally

demonstrated or predicted to be susceptible to SARS-CoV-2 infection was also exclusively identified in SARS-CoV-2-positive samples from the southwest corner of the Huanan market. This includes mtDNA sequences from fox (*Vulpes spp.*), hoary bamboo rat (*R. pruinosis*), and masked palm civet (*P. larvata*).

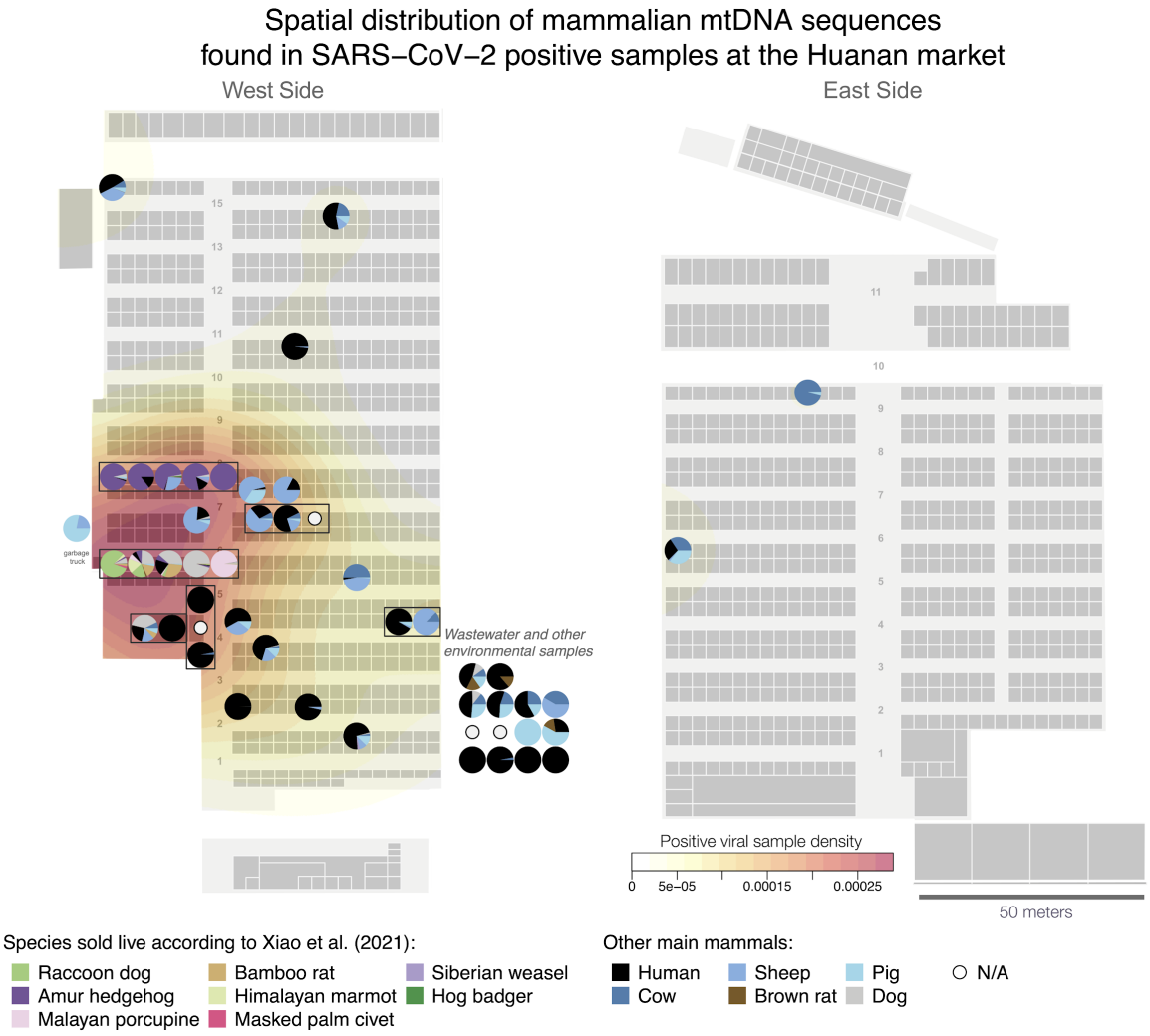

Figure 1: Spatial distribution of mammalian mtDNA sequences found in SARS-CoV-2 positive samples at the Huanan market. Distributions of the main mammal species (susceptible to SARS-CoV-2, and/or abundant reads at the market), by mtDNA sequence read count. These species are the raccoon dog (*Nyctereutes procyonoides*), Amur hedgehog (*Erinaceus amurensis*), Malayan porcupine (*Hystrix brachyura*), hoary bamboo rat (*Rhizomys pruinosus*), Himalayan marmot (*Marmota himalayana*), masked palm civet (*Paguma larvata*), Siberian weasel (*Mustela sibirica*), hog badger (*Arctonyx collaris*), human (*Homo sapiens*), cow (*Bos taurus*), sheep (*Ovis orientalis*), brown rat (*Rattus norvegicus*), pig (*Sus scrofa*), dog (*Canis lupus*). The dark contours group samples from the same stall. Samples in the middle of the map do not

have a clearly recorded location (they include sewage samples). “N/A” means that the sample had no detectable mammalian mtDNA reads that passed mapping cutoffs described in the Methods. The map also shows the density of SARS-CoV-2-positive environmental samples in the market, which was highest in the southwest corner of the western section of the market (see Figure 4 of Worobey et al. 2022 and Methods therein; input data set updated for **Figure 1**, here).

We found SARS-CoV-2-positive samples in which human mtDNA sequences were most abundant scattered throughout the rest of the market, indicating that in other samples the virus had been shed by infected persons. Other samples were dominated by non-human animal genetic material corresponding to the type of stall where they were collected, such as cattle at a stall selling beef product, and fish in a sample collected on a fish packaging surface. Hence, the most abundant animal in the sequencing data of a particular sample is not necessarily the source of the virus in that sample.

In samples where genetic material was sufficiently abundant, it was possible to assemble partial or near-complete consensus mitochondrial genomes (raccoon dog, Amur hedgehog, Malayan porcupine, hoary bamboo rat, and masked palm civet) and infer phylogenetic trees (**Figure 2)**. This analysis confirmed the presence of these animals at the wildlife stalls and taxonomic assignment strongly supported phylogenetic clustering with the mitochondrial reference genomes of these species. The closest reported relative of the complete raccoon dog mitochondrial genome sequence was a *Nyctereutes procyonoides* individual collected in the wild in China[8]. Further, the mitochondrial genome was distinct from that of *N. p. ussuriensis*, the subspecies commonly raised in northern China for fur[9]. These mitochondrial genome sequences can be helpful in identification of particular subspecies—and perhaps even local populations of wild or farmed animals—and thereby potentially aid in tracing the origins of susceptible hosts in the wildlife trade that were present at the Huanan market at the time the pandemic began.

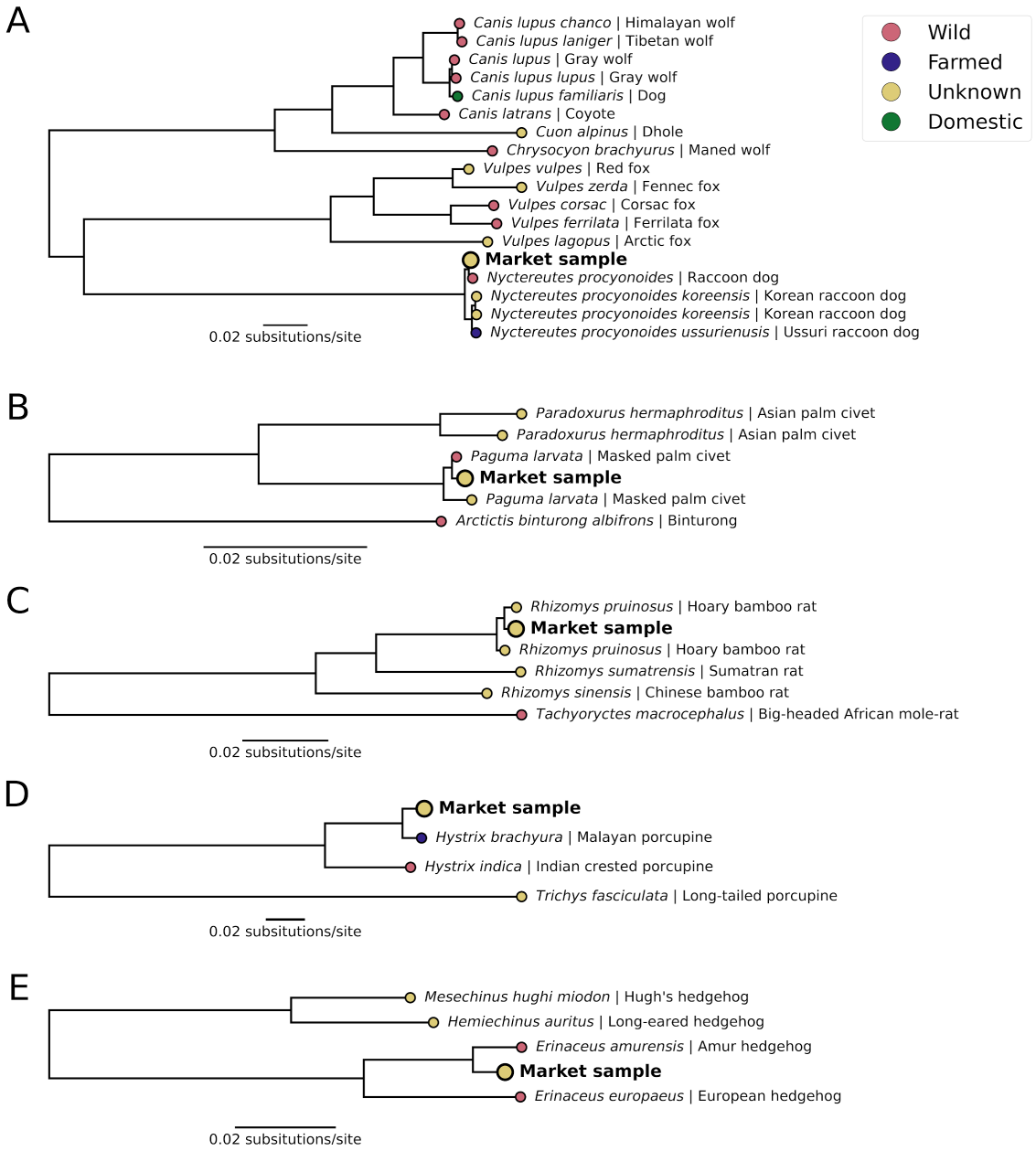

Figure 2: Maximum likelihood phylogenies of near-complete consensus mtDNA genomes of market animals from environmental samples and close relatives. A. Raccoon dog mitochondrial genome (99.4% complete) assembled from sample Q61 ("cart 1"), B. Masked palm civet mitochondrial genome (68.4% complete) assembled from sample Q61 ("cart 1"), C. Hoary bamboo rat mitochondrial genome (88.3% complete) assembled from a merge of samples Q61 ("cart 1"), Q64 ("cart 2"), and Q68 ("Ground surface"), D. Malayan porcupine mitochondrial genome (97.2% complete) assembled from sample Q70 (“Metal cage in inner room”), and E. Amur hedgehog mitochondrial genome (94.0% complete) assembled from sample W-8-25-D2. Tip colors denote whether a reference species was wild, farmed, domestic, or unknown. All nodes shown had support values >0.95.

We found other taxa that were detected with a lower prevalence of mtDNA reads. However, species level classification of rare species is challenging, especially without the exact reference species in the RefSeq mitochondrial sequence data. For example, a very small number of reads from one sample are classified as the fox genus *Urocyon*; however, we believe that these sequence reads are likely from another canid. Reads from the rodent genus, *Rattus*, were also found throughout the market samples at low abundances, likely from feral rats.

We performed *de novo* sequence assembly of the samples from a wildlife stall previously identified as containing the largest number of SARS-CoV-2-positive environmental swabs from animal-related surfaces[2]. Assembled contigs were aligned using BLAST, reporting the best hit with >99% identity against a custom database of reference genomes from wildlife species reported to be at Huanan market, as well as dog, cat, cow, sheep, goat, mouse, rat, and human (**Table S5**). In samples collected at this stall, we detected sequences that aligned with 100% identity to multiple species that are either confirmed or potentially susceptible to SARS-CoV-2, including raccoon dog, rabbit, dog, and red fox.

We restricted contig alignments to those at least 300 nucleotides in length. The cage in sample Q70 contained 816 contigs with 100% identity to Malayan porcupine, although it is unknown if this animal is susceptible to SARS-CoV-2. The cart in sample Q61 contained 1252 contigs with 100% identity to the raccoon dog genome, and 0 sequences with unambiguous and 100% identity to the human genome (**Figure 3**). 43.0% of all reads in this sample mapped to the *N. procyonoides* reference genome using BWA-MEM. To account for diversity that may not be captured in the reference genome, we also examined sequences with 99% identity to the corresponding reference genome. We identified 4622 raccoon dog sequences and 0 human sequences that met this criterion (**Figure 3**). These sequences were clearly distinguishable from other canids such as dog and red fox, which were also detected in samples from this cart. Translating open reading frames in the raccoon dog transcripts and searching for human orthologs, we identified numerous genic contigs (**Table S6**), including sequences for both genes that are often constitutively expressed, as well as several genes with tissue-specific transcription such as mucin 16 and olfactory receptor 2AG1, which could be indicative of potential nasal excretions if transcribed. Accurately estimating the respective proportions of RNA and DNA in these samples remains an open question for future work.

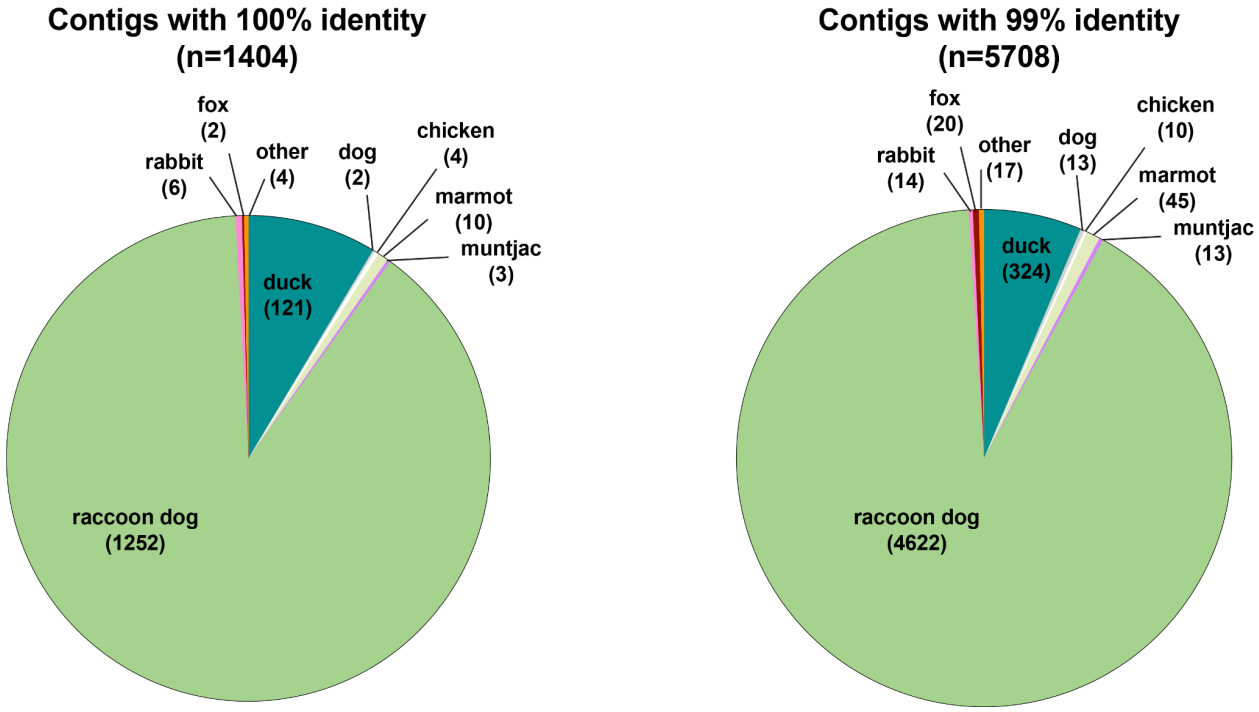

Figure 3: Mapping of sequence contigs assembled by Trinity[10] to reference host genomes from Cart 1 at a stall in the southwestern corner of Huanan market. We quantified contig alignments longer than 300 nt in sample Q61 with 100% sequence identity (left) and greater than 99% sequence identity (right) to identify the abundance of alignments to avian and mammalian genomes: *Anas platyrynchos* (duck), *Canis lupus familiaris* (dog), *Gallus gallus* (chicken), *Marmota himalayana* (Himalayan marmot), *Muntiacus reevesi* (muntjac), *Nyctereutes procyonoides* (raccoon dog), *Oryctolagus cuniculus* (rabbit), and *Vulpes vulpes* (red fox). Alignments to "other" genomes include *Bos taurus* (cow), *Erinaceus europaeus* (European hedgehog), *Meles meles* (European badger), *Neovison vison* (mink), *Ovis aries* (sheep), *Rattus rattus* (black rat), and *Sus scrofa* (pig/wild boar).

Susceptibility of Identified Species

Raccoon dogs are of particular note because they are susceptible to productive SARS-CoV-2 infection, shed high titers of virus, and can transmit to uninfected susceptible animals[11]. However, the environmental samples also contained additional sequences corresponding to other animal species known to be susceptible to SARS-CoV-2 infection. These include red foxes, rabbits, cats, and dogs[12–16]. In addition, several species were identified as potentially susceptible based on known susceptibility of closely related species at the genus level, such as Siberian weasels. This list of confirmed susceptible species present at the market is not exhaustive: SARS-CoV-2 has broad host tropism and may also be able to cause productive

infections in the other wildlife species detected at the market that have not yet been tested. Animals such as hog badgers (*Arctonyx spp.*) and masked palm civets (*Paguma larvata*) have also been hypothesized to be susceptible based on *in vitro* studies.

Conclusion and Recommendations

Our analysis of metagenomic sequence data provides genomic evidence of the presence of SARS-CoV-2 susceptible live animals at the Huanan market, Wuhan, before it was closed on 1 January 2020. Importantly, the genomic data confirming the presence of wildlife species occurred in the same market stalls where some of these animals had been documented to be sold. These samples were positive for SARS-CoV-2 and often had a lower abundance of human genetic material than other animal genetic material, in some cases including animals known to be susceptible to SARS-CoV-2. That these genomic traces are present in samples in which SARS-CoV-2 was detected identifies these wildlife species as potential intermediate hosts between the *Rhinolophus spp.* (horseshoe bat) animal reservoir and humans. Human genetic material was abundant and predominant in many samples with no evidence of non-human mammal DNA, indicating that the SARS-CoV-2 virus at these sites was likely shed by humans.

Although we cannot identify the intermediate animal host species from these data, a plausible explanation for the co-occurence of the genetic material of SARS-CoV-2 and susceptible animals is that a subset of these animals were infected. Combined with the previously published observation of the strong association of the earliest reported COVID-19 cases with the west side of the market, and the clustering of SARS-CoV-2-containing environmental samples near the wildlife stalls[2], this provides further support for the hypothesis that wildlife were the source of the first human SARS-CoV-2 infections. Once the initial spillovers had taken place, the market likely became a place of widespread human-to-human transmission.

Our analysis is based on partial data, corresponding only to a subset of environmental samples from the Huanan market in which SARS-CoV-2 had been detected. Other sequencing data exist for at least some of these samples, obtained with different sequencing technology (Illumina NextSeq 550 instead of DNBSEQ-T7 in the data we analyzed)[6]. Additional data also exist for the remaining SARS-CoV-2 positive samples, as well as for other environmental samples from the Huanan market that were negative for SARS-CoV-2. These data have not yet been shared, precluding further analysis. Given that sampling was likely to be uneven across the market, more detailed spatial analyses of the distributions of animals and SARS-CoV-2 require the raw sequencing files from all samples, and the accompanying metadata. All of these missing data could provide valuable information on the timeline of events at the Huanan market and the provenance of the virus.

It would be tempting to pursue a correlational epidemiological analysis between the presence of animal genetic material and SARS-CoV-2 positivity using data from market

samples that tested negative for the virus, when they become available. However, this approach is complicated by several factors. First, virus throughout the market may have been shed both by humans and by one or more species of other animals. Second, viral introductions may have been from a particular supply source or subspecies of a species at the market, not the entire species. Third, samples were obtained over a period of weeks, and viral and host genetic material would have degraded over time. Finally, the presence of RNA and/or DNA may reflect different signals of host recency. Any epidemiological analysis would need to take these considerations into account.

Our findings reinforce the need for further analyses to address transmission between and from possible intermediate host species, and provide potential clues to the upstream events that led to the emergence of SARS-CoV-2 into humans. Notably, observing the exact moment of spillover between an animal and a human has never been achieved for a pandemic pathogen. Nonetheless, our findings contribute to and underscore the large body of evidence supporting a natural origin of SARS-CoV-2.

Data accumulated since the beginning of the COVID-19 pandemic point clearly towards a zoonotic origin of SARS-CoV-2.

1. A preponderance of the earliest hospitalized COVID-19 patients were linked to a single location[3], one of just four locations in a city of 11 million people where plausible intermediate hosts of SARS-CoV-2 were sold live. There are hundreds or thousands of other sites that would have been equally or more likely than this wildlife market to have the first-detected and largest cluster of early cases had the outbreak there not been associated with wildlife sales[2].
2. The locations of early, severe COVID-19 cases without a clear epidemiological link to the Huanan market nevertheless were so centered on and close to the Huanan market that it is clear that community transmission of SARS-CoV-2 began in this local area and only later expanded across Wuhan, and the rest of the world[2]. Importantly, this includes those infected with lineage A viruses (and not just lineage B), indicating that early community spread of both early lineages of the virus radiated out from the market[2].
3. The locations of SARS-CoV-2-positive environmental samples in the Huanan market were close to or exactly where susceptible live mammals were sold[2,6].
4. The early genetic diversity of SARS-CoV-2 suggests multiple spillovers[17], and both early lineages, “A” and “B”, were directly observed at the market[6] and geographically associated with the market’s location in a way not expected by chance[2].
5. Live susceptible animals such as raccoon dogs had been reported to be on sale in this market[4], including during the month (November 2019) when the first human SARS-CoV-2 infection is estimated to have occurred[17,18].

6. Evidence collected and generated by China CDC and analyzed here shows that genetic material of such potential intermediate hosts was detected in SARS-CoV-2-positive environmental samples.

These arguments stand in stark contrast to the absence of evidence for any other SARS-CoV-2 emergence route.

Our results provide leads for further investigation of the upstream events that likely led to the presence of infected animals at the Huanan market and its role as the epicenter of the pandemic. Further studies on the origin of SARS-CoV-2 should include investigation of all the supply chains of the stalls identified here as selling wildlife where SARS-CoV-2 was detected, as well as population genetic studies of wildlife farms supplying the market and of wild populations in the vicinity of Wuhan and beyond. However, as the events in question occurred over three years ago, the window of opportunity for these investigations is closing.

Acknowledgements

We gratefully acknowledge all data contributors for generating these metagenomic sequences and metadata on which this research is based and sharing them via the GISAID Initiative. We thank those scientists who collected the samples, sequenced them and shared them via GISAID. Specifically, this analysis is based on accessions EPI_ISL_13052298 – EPI_ISL_13052346, by William J. Liu, Peipei Liu, Wenwen Lei, Zhiyuan Jia, Xiaozhou He, Linlin Liu, Weifeng Shi, Yun Tan, Shumei Zou, Xiang Zhao, Gary Wong, Ji Wang, Feng Wang, Gang Wang, Kun Qin, Rongbao Gao, Jie Zhang, Min Li, Wenling Xiao, Yuanyuan Guo, Ziqian Xu, Yingze Zhao, Jingdong Song, Jing Zhang, Wei Zhen, Wenting Zhou, Beiwei Ye, Juan Song, Mengjie Yang, Weimin Zhou, Yuhai Bi, Kun Cai, Dayan Wang, Wenjie Tan, Jun Han, Wenbo Xu, George F. Gao, Guizhen Wu. We thank Marion Koopmans and Stuart Neil for constructive comments.

Declaration of Interest

J.O.W. receives funding from the Centers for Disease Control and Prevention (CDC) through contracts to his institution unrelated to this research. M.A.S. receives contracts from the US Food & Drug Administration, US Department of Veterans Affairs and Janssen Research & Development, all outside the scope of this work. R.F.G. is a cofounder of Zalgen Labs, a biotechnology company developing countermeasures for emerging viruses. M.W., A.L.R., J.E.P., A.R., M.A.S., E.C.H., S.A.G., J.O.W., and K.G.A. have received consulting fees and/or provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic.

Funding

J.I.L. acknowledges support from the NIH (grant 5T32AI007244-38). S.A.G. acknowledges support from the NIH (grant F32AI152341). J.E.P. acknowledges support

from the NIH (grant T15LM011271) and UC San Diego Merkin Fellowship. J.O.W. acknowledges support from NIH (grants AI135992). D.L.R. acknowledges support from the Medical Research Council (grants MC_UU_12014/12 and MR/V01157X/1). M.A.S., P.L., and A.R. acknowledge support from the Wellcome Trust (collaborators award 206298/Z/17/Z – ARTIC network), the European Research Council (grant no. 725422 – ReservoirDOCS), and the NIH (grant R01AI153044). A.L.R. and R.S. are supported by the Canadian Institutes of Health Research as part of the Coronavirus Variants Rapid Response Network (CoVaRR-Net; CIHR FRN#175622) and acknowledge that VIDO receives operational funding from the Canada Foundation for Innovation – Major Science Initiatives Fund and from the Government of Saskatchewan through Innovation Saskatchewan and the Ministry of Agriculture. R.F.G. acknowledges support from the NIH (grants R01AI132223, R01AI132244, U19AI142790, U54CA260581, U54HG007480, and OT2HL158260), the Coalition for Epidemic Preparedness Innovation, the Wellcome Trust Foundation, Gilead Sciences, and the European and Developing Countries Clinical Trials Partnership Programme. E.C.H. is funded by an NHMRC Investigator grant (GNT2017197) and by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China. K.G.A. acknowledges support from the NIH (grants U19AI135995, U01AI151812, and UL1TR002550). This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services (contract no. 75N93021C00015 to M.W.) F.D. received funding from the MODCOV19 platform of the National Institute of Mathematical Sciences and their Interactions (Insmi, CNRS; 2022).

References

1. Tan W, Zhao X, Ma X, et al. A novel Coronavirus genome identified in a cluster of pneumonia cases - Wuhan, China 2019-2020. China CDC Wkly 2020;2(4):61–2.

2. Worobey M, Levy JI, Malpica Serrano L, et al. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science 2022;377(6609):951–9.

3. Worobey M. Dissecting the early COVID-19 cases in Wuhan. Science 2021;374(6572):1202–4.

4. Xiao X, Newman C, Buesching CD, Macdonald DW, Zhou Z-M. Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. Sci Rep 2021;11(1):11898.

5. Organization WH, Others. WHO-convened global study of origins of SARS-CoV-2: China Part. 2021;Available from: https://apo.org.au/node/311637

6. Gao G, Liu W, Liu P, et al. Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market [Internet]. Research Square. 2022;Available from:

https://www.researchsquare.com/article/rs-1370392/latest.pdf

7. Zhang Y-Z, Holmes EC. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell 2020;181(2):223–7.

8. Zhang H, Chen L. The complete mitochondrial genome of the raccoon dog. Mitochondrial DNA [Internet] 2010 [cited 2023 Mar 20];21(3-4). Available from: https://pubmed.ncbi.nlm.nih.gov/20795776/

9. Horecka B, Jakubczak A, Ślaska B, Jeżewska-Witkowska G. Raccoon dog (Nyctereutes procyonoides) phylogeography including the Polish population: local and global aspects. The European Zoological Journal 2022;89(1):641–52.

10. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29(7):644–52.

11. Freuling CM, Breithaupt A, Müller T, et al. Susceptibility of Raccoon Dogs for Experimental SARS-CoV-2 Infection. Emerg Infect Dis 2020;26(12):2982–5.

12. Porter SM, Hartwig AE, Bielefeldt-Ohmann H, Bosco-Lauth AM, Jeffrey Root J. Susceptibility of Wild Canids to SARS-CoV-2 - Volume 28, Number 9—September 2022 - Emerging Infectious Diseases journal - CDC. [cited 2023 Mar 18];Available from: https://wwwnc.cdc.gov/eid/article/28/9/pdfs/22-0223.pdf

13. Mykytyn AZ, Lamers MM, Okba NMA, et al. Susceptibility of rabbits to SARS-CoV-2. Emerg Microbes Infect 2021;10(1):1.

14. Shi J, Wen Z, Zhong G, et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. Science 2020;368(6494):1016.

15. Sit THC, Brackman CJ, Ip SM, et al. Canine SARS-CoV-2 infection. Nature 2020;586(7831):776.

16. Halfmann PJ, Hatta M, Chiba S, et al. Transmission of SARS-CoV-2 in Domestic Cats. N Engl J Med 2020;383(6):592.

17. Pekar JE, Magee A, Parker E, et al. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Science 2022;eabp8337.

18. Jijón S, Czuppon P, Blanquart F, Débarre F. Using early detection data to estimate the date of emergence of an epidemic outbreak [Internet]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.01.09.23284284v1

19. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754–60.

20. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet] 2021 [cited 2023 Mar 20];10(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931819/

21. GitHub - wwood/CoverM: Read coverage calculator for metagenomics [Internet]. GitHub. [cited 2023 Mar 20];Available from: https://github.com/wwood/CoverM

22. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30(14):3059.

23. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 2020;37(5):1530.

24. GitHub - rambaut/figtree: Automatically exported from code.google.com/p/figtree [Internet]. GitHub. [cited 2023 Mar 20];Available from: https://github.com/rambaut/figtree

25. GitHub - evogytis/baltic: baltic - backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualisation. Development is back on the evogytis/baltic branch (i.e. here) [Internet]. GitHub. [cited 2023 Mar 20];Available from: https://github.com/evogytis/baltic

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet] 1990 [cited 2023 Mar 20];215(3). Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

27. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 2019;20(1):1–14.

Appendix A – Methods

Metagenomic sequencing data were downloaded from GISAID (Accessions EPI_ISL_13052298 – EPI_ISL_13052346; see Appendix B for acknowledgments of data generators) and assigned via their corresponding sample IDs to their market location and sample metadata using the supplementary data from[2,6] (**Table S1-2**). Samples were either 43 bp single end reads or 2x100 bp paired-end reads, although the exact experimental procedure used to generate these data is uncertain. The set of full length Metazoa mitochondrial genomes available in the NCBI Refseq database were downloaded and de-replicated using Mash distances of 93% average nucleotide identity, resulting in a data set of 10,117 dereplicated mtDNA genomes. Reads from each sample were mapped as single reads using BWA-MEM[19] to the mtDNA data set. Reads were filtered from the resulting mappings, requiring a minimum mapping quality score of 20, with samtools[20]. Read counts, the number of bases covered, and the breadth of coverage (%) for each mitochondrial genome in each sample were calculated with coverM[21], filtering out reads smaller than 40 bp (--min-read-aligned-length 40), excluding reads mapping to the first and last 100 bp of contigs (--contig-end-exclusion 100), and retaining reads that mapped with at least 95% identity to the reference (--min-read-percent-identity 0.95). A minimum of 400 covered bases (which corresponds to ~10 43-bp reads) was required to mark a species as positive in a given sequencing run of a sample. To identify the best representative for each mtDNA genome cluster, after mapping to the set of mtDNA genomes dereplicated at 93% identity, reads were remapped to all mtDNA genomes within each detected mammalian genome cluster. The genome within each cluster that recruited the highest percentage of 100% identity reads was then chosen as the representative sequence for that cluster, and a final dereplicated database including these representatives was mapped and filtered to in an identical manner to that described above. This iterative process helped select the genome representative in the database most similar to that in the data and only affected a minority of species.

Mitochondrial consensus genomes of four species were called from the sample(s) in which those species were found to be most abundant. The most abundant sequences of reads that uniquely mapped to the mtDNA of *Nyctereutes procyonoides* (Raccoon dog), *Erinaceus amurensis* (Amur hedgehog), *Hystrix brachyura* (Malayan porcupine), and *Rhizomys pruinosus* (hoary bamboo rat) were determined using a custom Python script consensus_from_bam.py, available on GitHub, taking the consensus base mapped to each position and filling reference positions covered by 0 mapped reads with 'N'. The raccoon dog mtDNA genome could also be also assembled *de novo* with SPAdes in one nearly complete contig, and *de novo* assemblies of the other species' mtDNA were fragmented due to their lower genome coverage.

For each reconstructed consensus mtDNA genome, we used nucleotide BLAST to find the most closely related 10-17 publicly available genomes. The mtDNA genomes were then aligned with MAFFT[22], using the --globalpair and --maxiterations 1000 flags. We

used IQTREE-2[23], specifying a general time reversible (GTR) substitution model with four discretized-gamma rate categories and fixed empirical nucleotide frequencies, to infer maximum likelihood trees and calculated support at internal nodes with an approximate Bayes test (the --abayes flag). The trees were subsequently midpoint rooted and plotted in the FigTree[24] and Baltic[25] visualization software. For the civet, bamboo rat, porcupine, and hedgehog phylogenies, we extracted subtrees with the 3–5 most closely phylogenetically related mtDNA genomes.

We generated *de novo* transcriptome assemblies for each sample using Trinity[10] (version 2.15.1; with parameters --seqType fq --max_memory 50G --SS_lib_type F --CPU 6 --no_normalize_reads --single).

The *de novo* assembled transcripts of each sample were searched against the NCBI whole genome shotgun (WGS) database (database downloaded on 14 March 2023 https://ftp.ncbi.nlm.nih.gov/blast/db/) using BLASTn[26] (v 2.10.1+). The specific parameters used are: -outfmt '6 qseqid sseqid pident evalue score bitscore length qstart qend sstart send stitle' -max_target_seqs 2.

BLAST (version 2.13.0+.) searches were also performed against an in-house database of genome sequences assemblies from *Myocastor coypus* (GCA_004027025.1), *Marmota himalayana* (GCA_005280165.1), *Hystrix brachyura* (GCA_016801275.1), *Muntiacus reevesi* (GCA_020226045.1), *Sciurus vulgaris* (GCA_902686455.2), *Nyctereutes procyonoides* (GCA_905146905.1), *Homo sapiens* (GRCh38.p14), *Mus musculus* (GRCm39), *Canis lupus familiaris* (NC_006583.4), *Sus scrofa* (NC_010443.5), *Erinaceus europaeus* (EriEur2.0), *Capra hircus* (ARS1.2), *Bos taurus* (ARS-UCD1.3), *Vulpes vulpes* (VulVul2.2), *Oryctolagus cuniculus* (UM_NZW_1.0), *Rattus rattus* (Rrattus_CSIRO_v1), *Anas platyrhynchos* (ZJU1.0), *Gallus gallus* (bGalGal1), *Ovis aries* (ARS-UI_Ramb_v2.0), *Felis catus* (F.catus_Fca126_mat1.0), *Neogale vison* (ASM_NN_V1), *Meles meles* (GCF_922984935.1), *Paguma larvata* (NC_029403.1) and *Arctonyx collaris* (NC_020645.1). The genome assemblies were downloaded from NCBI WGS genome assembly database. The BLASTn search parameters used were: -outfmt '6 qseqid sseqid pident evalue score bitscore length qstart qend sstart send stitle' -max_target_seqs 2. The output files were filtered to exclude hits with alignment length of less than 100. The unique best hit transcript sequences were annotated using Transdecoder and the final annotated proteins were used for ortholog mapping against the human proteome (GRCh38.p14; downloaded from NCBI) using OrthoFinder[27] (version 2.5.4) at default parameters.

Analysis scripts and supplementary tables are available on Github https://github.com/sars-cov-2-origins/huanan-environmental.

Appendix B – Data Acknowledgments

Accession ID	Authors	Originating/Submitting Laboratory	Notes
EPI_ISL_13052297	Peipei Liu, Wenting Zhou, Wei Zhen, Xiaozhou He, Yu Lan, William J. Liu, George F. Gao, Guizhen Wu	National Institute for Viral Disease Control and Prevention, China CDC Address: No. 155 Changbai Road, Changping District, Beijing 102206 China	No data in this accession
EPI_ISL_13052298			Unavailable as of 2023-03-16
EPI_ISL_13052299			Unavailable as of 2023-03-16
EPI_ISL_13052300			Unavailable as of 2023-03-16
EPI_ISL_13052301			Unavailable as of 2023-03-16
EPI_ISL_13052302			Unavailable as of 2023-03-16
EPI_ISL_13052303			Unavailable as of 2023-03-16
EPI_ISL_13052304			Unavailable as of 2023-03-16
EPI_ISL_13052305			Unavailable as of 2023-03-16
EPI_ISL_13052306			Unavailable as of 2023-03-16
EPI_ISL_13052307			Unavailable as of 2023-03-16
EPI_ISL_13052308			Unavailable as of 2023-03-16
EPI_ISL_13052309			Unavailable as of 2023-03-16
EPI_ISL_13052310			Unavailable as of 2023-03-16
EPI_ISL_13052311			Unavailable as of 2023-03-16
EPI_ISL_13052312			Unavailable as of 2023-03-16
EPI_ISL_13052313			Unavailable as of 2023-03-16
EPI_ISL_13052314			Unavailable as of 2023-03-16
EPI_ISL_13052315			Unavailable as of 2023-03-16
EPI_ISL_13052316			Unavailable as of 2023-03-16
EPI_ISL_13052317			Unavailable as of 2023-03-16
EPI_ISL_13052318			Unavailable as of 2023-03-16
EPI_ISL_13052319			Unavailable as of 2023-03-16
EPI_ISL_13052320			Unavailable as of 2023-03-16
EPI_ISL_13052321			Unavailable as of 2023-03-16
EPI_ISL_13052322			Unavailable as of 2023-03-16
EPI_ISL_13052323			Unavailable as of 2023-03-16
EPI_ISL_13052324			Unavailable as of 2023-03-16
EPI_ISL_13052325			Unavailable as of 2023-03-16
EPI_ISL_13052326			Unavailable as of 2023-03-16
EPI_ISL_13052327			Unavailable as of 2023-03-16
EPI_ISL_13052328			Unavailable as of 2023-03-16
EPI_ISL_13052329			Unavailable as of 2023-03-16
EPI_ISL_13052330			Unavailable as of 2023-03-16
EPI_ISL_13052331			Unavailable as of 2023-03-16
EPI_ISL_13052332			Unavailable as of 2023-03-16
EPI_ISL_13052333			Unavailable as of 2023-03-16
EPI_ISL_13052334			Unavailable as of 2023-03-16
EPI_ISL_13052335			Unavailable as of 2023-03-16
EPI_ISL_13052336			Unavailable as of 2023-03-16
EPI_ISL_13052337			Unavailable as of 2023-03-16
EPI_ISL_13052338			Unavailable as of 2023-03-16
EPI_ISL_13052339			Unavailable as of 2023-03-16
EPI_ISL_13052340			Unavailable as of 2023-03-16
EPI_ISL_13052341			Unavailable as of 2023-03-16
EPI_ISL_13052342			Unavailable as of 2023-03-16
EPI_ISL_13052343			Unavailable as of 2023-03-16
EPI_ISL_13052344			Unavailable as of 2023-03-16
EPI_ISL_13052345			Unavailable as of 2023-03-16
EPI_ISL_13052346			Unavailable as of 2023-03-16

Appendix C – Affiliations

KGA: Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

FD: Institut d'Écologie et des Sciences de l'Environnement (IEES-Paris, UMR 7618), CNRS, Sorbonne Université, UPEC, IRD, INRAE, Paris, France. ORCID: 0000-0003-2497-833X

KG: Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90024, USA.

RFG: Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA 70112, USA.
Zalgen Labs, Frederick, MD 21703, USA.
Global Virus Network (GVN), Baltimore, MD 21201, USA.

SAG: Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA.

ECH: Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

PL: Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium.

JIL: Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

NM: Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA, USA.

JEP: Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA.

SP: George Mason University, Schar School of Public Policy and Government, Department of Biodefense, Arlington, VA 22201, USA.

AR: Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK.

ALR: Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon, SK, Canada. ORCID: 0000-0001-9462-3169

DLR: MRC-University of Glasgow Center for Virus Research, Glasgow, G61 1QH, UK. ORCID: 0000-0001-6338-0221

RS: Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon, SK, Canada.

MAS: Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA 90024, USA.

JOW: Department of Medicine, University of California San Diego, La Jolla, CA, USA.

MW: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA.