Imperfect data can still provide insights: SARS-CoV-2 and animals in metagenomic samples from the Huanan market

F. Débarre

Institute of Ecology and Environmental Sciences, CNRS UMR 7618, Sorbonne Université, UPEC, IRD, INRAE, Paris, France
https://orcid.org/0000-0003-2497-833X
Corresponding author: florence.debarre@normalesup.org

Abstract

While the exact context of the emergence of SARS-CoV-2 remains uncertain, data accumulated since 2020 now provide a more precise picture of Wuhan's Huanan Seafood Wholesale Market, to which the earliest clusters of human cases of Covid-19 were linked. After the market closed on January 1st 2020, teams from the Chinese Center for Disease Control and Prevention collected environmental samples, and sequenced them. Metagenomic sequencing data from these samples were shared in early 2023. These data confirmed that non-human animals susceptible to SARS-CoV-2 were present in the market before it closed, but also that these animals were located in the side of the market with most human cases, and in a corner with relatively more SARS-CoV-2-positive environmental samples. The environmental samples were however collected after abundant human-to-human transmission had taken place in the market, precluding the identification of a potential non-human animal host. In spite of this acknowledged limitation, Jesse Bloom recently investigated associations between SARS-CoV-2 and non-human animals, concluding that his analyses could not indicate whether animals were infected by SARS-CoV-2. Here I explain why such analyses cannot identify infected animals, and I rebut the suggestion that such analyses had been encouraged. I show that Bloom's investigation ignores the temporal and spatial structure of the data, which led to incorrect interpretations. Finally, I show that criteria put forward to question the role of raccoon dogs as source of environmental SARS-CoV-2 would also exclude humans. There needs to be clear distinctions between scientific studies and news articles written about them, even for topics with broad media interest.

1. Introduction

The question of the origin of the Covid-19 pandemic is inherently scientific, even though its broader political context and implications cannot be ignored (Gostin and Gronvall, 2023). Uncertainty remains regarding the exact conditions of SARS-CoV-2's emergence and spillover to humans. The level of available details and the amount of data on the early days of Covid-19 are however remarkable compared to other emerging diseases, and in particular to SARS (Xu *et al.*, 2004; Keusch *et al.*, 2022).

In early 2020, a team from the Chinese Center for Disease Control and Prevention (CCDC) collected environmental samples in Wuhan's Huanan Seafood Wholesale Market (hereafter "Huanan market"), to which cases of a new pneumonia were linked (Chen *et al.*, 2020; Worobey, 2021). The market was sampled multiple times over several weeks, with various sampling strategies (World Health Organization, 2021). Results from the first two sampling trips, which took place on January 1st and 12th 2020, were communicated to the press at

the end of January 2020[1,2], and were detailed in a report, which was not public at the time[3] (Wu, 2020). When these first results were reported, the CCDC team had collected 585 samples, of which 33 had been found positive for the new coronavirus. The positive samples were in their vast majority (31/33) from the west side of the market, where wildlife was sold. The potential role of wildlife sold at the market was explicitly mentioned by CCDC (Wu, 2020; Tan *et al.*, 2020).

The existence of metagenomic sequencing data related to the Huanan market environmental samples was mentioned in the report of the 2021 Joint World Health Organization (WHO)–China mission (World Health Organization, 2021, p.97). However, only results on SARS-CoV-2 detection were reported, and raw data were not publicly shared.

The first results on the metagenomic content of the market samples were provided in a preprint by Liu et al. in early 2022 (Liu *et al.*, 2022). The corresponding raw data were not shared either at the time. The preprint did not describe in detail the contents of the environmental samples; humans were the only species named. A figure (Figure 4A in Liu *et al.* (2022)) however indicated that other species besides humans had been detected in the data. In the absence of labels on the figure, of any description in the preprint's main text, and of raw data, the identities of these species were unknown to the readers. In addition, the figure itself was criticized both for its form and its interpretation. It displayed Pearson correlation as a function of Spearman correlation, i.e. a correlation of correlations, and did not include error bars, therefore giving a false impression of certainty. The figure was used to conclude that humans were the source of SARS-CoV-2 in the Huanan market, a conclusion that could not be drawn from the analysis performed.

Multiple researchers publicly regretted the absence of raw data, and some, led by Jesse Bloom, even organized a petition to ask for the publication of these data[4].

On January 30th 2023, a subset of the raw sequencing data became public on GISAID[5]. The submission date displayed on GISAID was however in June 2022. The new data therefore appeared down in the database, and they apparently remained unseen for weeks. I first came across the data by chance on March 4th 2023, while looking for related information in the database. In the table of results on GISAID's website, the records were reported as being 3-nucleotide long; the FASTA field only contained "`AAA`". Puzzled, and initially erroneously thinking that the newly found accessions were just placeholders awaiting data, I notified over the course of the next few days some colleagues of my discovery. These colleagues included Alex Crits-Christoph (on March 4th) and Jesse Bloom (on March 9th). After a brief email exchange with Jesse Bloom, I noticed the presence of FASTQ buttons on GISAID, indicating that raw data were in fact available. I communicated this new information to Alex Crits-Christoph and Jesse Bloom, together with the accession numbers of the data. Of these two, only Alex Crits-Christoph showed immediate interest in the data, downloaded them and started analyzing them.

The dataset shared on GISAID at the time corresponded to 50 samples from the Huanan market, and data were available for 49 of them (Crits-Christoph *et al.*, 2023a). The dataset only contained samples that had been tested positive for SARS-CoV-2.

Because previous work (Worobey *et al.*, 2022a) had identified a specific market stall with an unusual number of samples reported as positive for SARS-CoV-2, the initial analysis focused on the five available samples from this stall. Worobey *et al.* (2022a) had made the direct prediction that SARS-CoV-2 susceptible wildlife may be present and detectable within

[1]References that are not scientific articles or reports are presented as footnotes.

[2]e.g., `http://www.xinhuanet.com/english/2020-01/27/c_138735677.htm`, `https://news.cgtn.com/news/2020-01-27/Experts-confirm-Wuhan-seafood-market-was-source-of-novel-coronavirus--NAHPUtsPgA/index.html`

[3]The report was leaked later in 2020; `https://www.epochtimes.com/gb/20/5/31/n12150755.htm`

[4]Twitter discussion, February 27th 2022; `https://x.com/jbloom_lab/status/1497743165226303489`

[5]The date range was delimited by comparing metadata downloads from various time points, and it was confirmed by GISAID to a journalist; `https://www.vanityfair.com/news/2023/06/raccoon-dog-george-gao-covid-origins`

the sequencing samples from this stall. The very first result we obtained with the data downloaded from GISAID was the detection of raccoon dog mitochondrial DNA (mtDNA) in the data from sample Q61. This species, shown to be susceptible to SARS-CoV-2 and able to transmit it (Freuling *et al.*, 2020), was not named by Liu *et al.* (2022), and was not listed as a species present in the market in the 2021 WHO-China report (World Health Organization, 2021). A study published in the summer 2021 had identified them as animals sold in Wuhan markets (Xiao *et al.*, 2021), but the results combined multiple markets and all the 31 months during which the study had been conducted, leading to the speculation that the animals may have been absent in the late Fall 2019. Finding genetic traces of these animals was therefore an important result. This initial result was later confirmed: raccoon dog reads were abundant across wildlife stall samples, and particularly in wildlife stalls positive for SARS-CoV-2 than any other mammal (Crits-Christoph *et al.*, 2023b).

In the early hours of March 11th 2023, the data could not be accessed on GISAID anymore (Crits-Christoph *et al.*, 2023a, foreword), and would remain inaccessible until March 29th 2023.

The discovery of the data and further initial results were communicated to the WHO's Scientific Advisory Group for the Origins of Novel Pathogens (SAGO). Shortly after an online meeting with SAGO on March 14th, a journalist contacted one member of our group[6], thereby revealing to us that the news had leaked to the press. This first journalist only knew of generic points communicated to SAGO, and not of the context of the discovery of the data. At that time, we had only generated some figures that were presented to the committee, but our analyses were not finalized, and we did not yet have a written report to share. We were not ready, nor willing, to have our results publicized. Another journalist aware of the news however went ahead, managing to convince some of us to be interviewed. Importantly it was not even clear to some of those who were interviewed that the news would be published before our report was ready. The scoop was published on March 16th with a flashy title misrepresenting our results[7]. I had explicitly and at multiple times declined to be interviewed by this journalist, and yet I ended up being quoted in a revised version of the article — the quote was taken, without consent, from an email asking for a correction that I sent two days after the article first came out. This illustrates that scientists have no control over what journalists write, and it is therefore important to differentiate a scientific article from what journalists may write about it.

We published a report a few days later (Crits-Christoph *et al.*, 2023a), accompanied by a foreword detailing the peculiar context in which the work was done. In it, we explained why the results could not be reproduced, as data were no longer available. The report focused on the description of animals susceptible to SARS-CoV-2 in the market environmental sequences, including but not limited to raccoon dogs. This report contained a paragraph (pp.13-14) preemptively listing issues that would be met if attempting to conduct a correlation analysis if the whole dataset ever became available.

Nine days later, Liu et al. published an updated version of their study on ChinaXiv (Liu *et al.*, 2023a), and made their whole dataset available on public repositories (NGDC and SRA). There were now sequence data available for 172 samples (159 from the market, 3 from related warehouses, 10 from sewage in the area). This new dataset notably now included samples tested negative for SARS-CoV-2. In their updated preprint, which was soon accepted in Nature, Liu *et al.* (2023b) briefly described animal species detectable in the data. The Figure 4A of their preprint was replaced after peer review by supplementary tables of tests of co-presence (Supplementary Table 9) and quantitative comparisons of animal contents in PCR-positive vs. PCR-negative samples (Supplementary Table 7).

[6] Later publicly confirmed by a tweet; `https://archive.ph/2MDB9`

[7] Katherine J. Wu, *The Strongest Evidence Yet That an Animal Started the Pandemic*; `https://www.theatlantic.com/science/archive/2023/03/covid-origins-research-raccoon-dogs-wuhan-market-lab-leak/673390/`

Finally, in April 2023, Bloom (2023a) published a preprint analyzing the sequence data from Liu *et al.* (2023b), eventually published in Virus Evolution. In the article, Bloom (2023b) reproduced versions of Liu *et al.* (2022)'s Figure 4A, and concluded that the analysis did not allow for the identification of infected animals. While I agree with the conclusion — and more generally, with the fact that it is not possible to prove with these data that animals were infected, I would like to point out several important issues with Bloom's article.

2. Results

2.1 Questioning the motivation for doing an analysis that we knew would fail

There are statistical errors, technical limitations, experimental imbalances, and conceptual failings behind the correlation analyses as described in Bloom (2023b), rendering them ultimately meaningless. Several conceptual issues were noted *a priori* (Crits-Christoph *et al.*, 2023a), and were further described in our reply to Jesse Bloom's preprint on bioRxiv[8], and in a recent preprint (Crits-Christoph *et al.*, 2023b).

It is not possible to prove that non-human animals were infected because the Huanan market samples were collected too late. This inescapable limitation was already identified by Jesse Bloom in 2022[9], and clearly stated in Bloom (2023b). (For simplicity, in the rest of the text, "animals" means non-human animals). There were a lot of human cases in the market, and most SARS-CoV-2 detected in the market are definitely of human origin. Given that it is impossible, with these data, to distinguish between viruses shed by humans and viruses shed by animals, any animal signal at the market scale was covered by human signal before the samples were collected. A correlation at the scale of the whole market cannot therefore identify infected animals, if there were any.

Even if the samples had been collected in the very early days of SARS-CoV-2's presence in the Huanan market, a correlation analysis could have failed to identify the source species. Such an analysis indeed implicitly assumes that most animals of a species were infected. Different stalls selling the same species did not necessarily have the same suppliers. This was for instance the case for bamboo rats, which came from origins as different as Hubei, Guangxi, Hunan, Yunnan provinces (World Health Organization, 2021, annexes pp.189–191). If only a fraction of animals of a given species were infected, a correlation analysis may fail to identify the species as the source. We observe this phenomenon in the market data with another virus, H3N2 influenza virus (Crits-Christoph *et al.*, 2023b). This human virus is present in a single stall in the market, probably because only a limited number of humans had this flu virus in Fall 2019 in the market. A market-wide correlation with human reads is not significant, because many other humans in the market were not infected (log proportion of total reads, cor = 0.076 (95% CI: -0.081, 0.23), $p = 0.34$; data: Crits-Christoph *et al.* (2023b)). This shows that even if the samples had been collected much earlier, before the many human infections that followed, an animal source of SARS-CoV-2 would still probably not have been identifiable with a correlation analysis.

Then why conduct a market-wide correlation analysis in the first place, when it is acknowledged that the samples were collected comparatively too late for such an analysis to be insightful? A motivation is given by Bloom (2023b): The absence of labels in Liu *et al.* (2022)'s Figure 4A had been widely noted (Bloom (2023b) cited three news articles by Jon Cohen to back up the claim), but neither Crits-Christoph *et al.* (2023a) nor Liu *et al.* (2023b) had "addressed this omission"; Bloom (2023b) study is here "[t]o remedy this omission".

This is a misrepresentation of the reasons why the absence of labels was noted. None of the scientists quoted by Jon Cohen, nor even Jon Cohen himself (personal communications and contemporary statements[10]), meant that it would be worth reproducing Figure 4A. The

[8] Reply posted on bioRxiv; `http://disq.us/p/2u34bgh`.

[9] As noted publicly in a tweet; `https://x.com/jbloom_lab/status/1497627222206664706`.

[10] Twitter thread by Andrew Rambaut, 2022-02-26; `https://archive.ph/tuhgM`

figure was only used as evidence that species were undescribed in Liu *et al.* (2022). This omission was addressed by Crits-Christoph *et al.* (2023a) and Liu *et al.* (2023b), who described and quantified animals reads detectable in the metagenomic sequence data. There is just no good rationale for conducting a market-wide univariate correlation analysis to identify a potential infected non-human animal with these data.

In his article, Bloom (2023b) emphasized that his results were inconsistent with "media articles that emphasized the co-mingling of raccoon dog and viral material (Wu 2023; Mueller 2023)"[11]. It seems that countering these media articles was an important motivation for Bloom's study. However, the meaning of co-mingling in these articles is mischaracterized by Bloom: it was not meant at the scale of the whole market, but in a specific stall. Both Wu's and Mueller's articles emphasized the fact that raccoon dog genetic material and SARS-CoV-2 had been found together in a sample (Q61). A quote in Mueller's article clarified that the data "can't prove definitely there was an infected animal at that stall", and Mueller specified that "even if a raccoon dog had been infected, it would not be clear that the animal had spread the virus to people" – i.e., Bloom's conclusions. Wu used the word "co-mingled" in "Finding the genetic material of virus and mammal so closely co-mingled—enough to be extracted out of a single swab—isn't perfect proof", and this sentence was clearly again about sample Q61. Neither of these articles was about market-wide co-presence, so it is incorrect to suggest that a market-wide analysis contradicts Mueller's and Wu's articles.

Bloom's study itself was misinterpreted by commenters and misrepresented in the media, notably as evidence that the raccoon dogs were not infected (instead of as absence of evidence that they were). This and the confusion between our report and news articles that preceded it led to particularly vitriolic attacks on our own work (e.g., "Why Does Bad Science on Covid's Origin Get Hyped?" in a New York Times newsletter[12]).

Even if there had been good reasons to conduct such a correlation analysis at the scale of the whole market, there are substantial issues with the way it is done in Bloom (2023b).

2.2 Temporal and spatial structures in the data cannot be omitted

The unbalanced sequencing design in the data shared by Liu *et al.* (2023b) precludes any market-wide analysis. The Huanan market was sampled multiple times; all samples were tested by PCR for SARS-CoV-2, and only a fraction were subjected to metagenomic sequencing. The fraction and characteristics of the samples that were sequenced varied over time. The first collection trip, on January 1st 2020, focused on stalls with human cases, and only samples positive by PCR were then sequenced. The second collection trip, on January 12th 2020, focused on wildlife stalls; an equal number of samples were collected from each stall, and all were sequenced irrespective of PCR positivity. For later trips, only a fraction of samples were sequenced, both PCR positive and PCR negative. Later trips also included samples from outside the market (including warehouses related to the market; sewage around the market). The proportion of SARS-CoV-2 positive samples among the sequenced samples therefore widely varied over time, as did the composition of animal species (wildlife stalls had different species than other stalls in the market). It is therefore incorrect to present correlation analyses mixing up data from all sampling trips, as is done in most of Bloom (2023b)'s article. Mixing up these data leads to instances of Simpson's paradox: some correlations substantially change once the time structure is taken into account (Crits-Christoph *et al.*, 2023b, Figure S4).

One subset of the data can be considered as balanced. All samples collected on January 12th 2020 were sequenced, irrespective of SARS-CoV-2 PCR positivity; ten samples were collected in each of the seven stalls identified as selling wildlife. Because these samples

[11] Wu's article is referenced in footnote 7; Benjamin Mueller, *New Data Links Pandemic's Origins to Raccoon Dogs at Wuhan Market* `https://www.nytimes.com/2023/03/16/science/covid-wuhan-market-raccoon-dogs-lab-leak.html`

[12] David Wallace-Wells; `https://www.nytimes.com/2023/05/03/opinion/covid-origin-science.html`.

Virus	Host	Dates	Dataset	Result
SARS-CoV-2	Human	Jan 12	ACC	cor = 0.2 (95% CI: -0.036, 0.42), $p = 0.095$
			JB	cor = 0.22 (95% CI: -0.017, 0.43), $p = 0.068$
SARS-CoV-2	Raccoon dog	Jan 12	ACC	cor = -0.042 (95% CI: -0.27, 0.19), $p = 0.73$
			JB	cor = -0.00055 (95% CI: -0.24, 0.23), $p = 1$
SARS-CoV-2	Human	Jan 01	ACC	cor = 0.18 (95% CI: -0.23, 0.54), $p = 0.38$
			JB	cor = 0.3 (95% CI: -0.1, 0.62), $p = 0.14$
			JB2	cor = 0.33 (95% CI: -0.09, 0.66), $p = 0.12$
SARS-CoV-2	Spotted bass	Jan 01	ACC	cor = -0.016 (95% CI: -0.41, 0.38), $p = 0.94$
	Largemouth bass		JB	cor = 0.18 (95% CI: -0.23, 0.54), $p = 0.39$
			JB2	cor = 0.41 (95% CI: 0.0024, 0.71), $p = 0.049$
H3N2	Human	all	ACC	cor = 0.076 (95% CI: -0.081, 0.23), $p = 0.34$

Table 1: Pearson correlations between viruses and potential hosts, computed on log proportions of total reads. Zero reads are converted to the half minimum possible proportion. ACC: Crits-Christoph *et al.* (2023b); JB: Bloom (2023b); JB2: Bloom (2023b) without samples F13 and F54.

were collected late, however, SARS-CoV-2 reads are rare: six samples are positive, with five in the same stall ("stall A" in Crits-Christoph *et al.* (2023b)). There are 1, 2, 2, 5, and 7 reads in the positive samples from stall A, and 5 reads from the positive samples from stall B; 0 reads in the 64 other samples collected on January 12th 2020. It does not make much sense to compute a correlation with such data. However, if we do it, humans are not significantly associated with SARS-CoV-2 (log proportions of total reads; cor = 0.22 (95% CI: -0.017, 0.43), $p = 0.068$; data: Bloom (2023b) from inside the market; see Table 1). Raccoon dogs are not significantly associated either (cor = -0.00055 (95% CI: -0.24, 0.23), $p = 1$). A correlation analysis on this subset of data therefore does not exclude raccoon dogs as potential source, and does not identify humans as potential source.

Consideration of the spatial and temporal structure of the samples helps confirm their positive or negative status. In his article, Bloom (2023b) emphasized the low number of reads in sample Q61 ("1 of 200,000,000 reads"), and stressed the lack of significant difference between 1 read in Q61 vs. 0 reads in E-10-29-2. This result was widely interpreted as Q61 being a false positive. Bloom's analysis however neglected the context in which the samples were collected. When taking into account the spatial and temporal structure of the data, Q61 is clearly positive. The maximum number of SARS-CoV-2 reads in the other samples collected on January 12th 2020 is 7, i.e. also very low. Out of the six positive samples collected on that date, five including Q61 come from the very same stall, and three of them were positive by PCR (Crits-Christoph *et al.*, 2023b). Finally, Q61 was also clearly identified as positive in January 2020 by China CDC (Wu, 2020). Even though it only contained one SARS-CoV-2 read, Q61 is positive. Likewise, sample E-10-29-2 is clearly negative: no sample from inside the market was tested positive by PCR or sequencing on the date it was collected (February 20th 2020) nor after, and no other sample in the stall and adjacent locations was ever tested positive either. Even though a chi-squared test on the proportions of reads in Q61 vs. E-10-29-2 is non significant, their spatial and temporal contexts confirm the results.

2.3 Criteria excluding raccoon dogs as hosts also exclude humans

Bloom (2023b) argued that only one sample (Q61) had a substantial proportion of raccoon dog reads and any presence of SARS-CoV-2, "substantial" being defined as at least 20% of chordate reads. In the previous section, I explained why this sample is clearly SARS-CoV-2-positive in spite of its very low number of reads; here I focus on the choice of criteria to

call a proportion of reads of a given species "substantial". After these criteria had been pointed out, Jesse Bloom argued that the threshold had been chosen to produce a table of reasonable length; yet the argument remained in the abstract. Twenty percent of chordate reads is a very stringent threshold. If we apply the same threshold to human reads on Bloom's data, then among the 33 samples from inside the market with SARS-CoV-2 reads, only eight have at least 20% of chordate reads belonging to *Homo sapiens*. Among these six samples, five are from stalls with human cases on the map of the Joint China-WHO mission (World Health Organization, 2021), two are in stalls with human cases on another map produced by CCDC (The BMJ, 2021), and one is next to stalls with human cases on this other map. Twenty-five other samples from inside the market contain SARS-CoV-2 reads but less than 20% human reads among chordates reads. Yet SARS-CoV-2 in at least one of these is almost certainly of human origin: sample B5 comes from a stall with two to three cases, SARS-CoV-2 was so abundant in the sample that a full sequence could be assembled, which matches the reference sequence. In spite of this clear association with a human case, the percentage of human reads in this sample is 0.12%.

Market-wide correlations do not identify humans as hosts either. In the text of the article, Bloom highlights the negative correlation between raccoon dog reads and SARS-CoV-2 reads. As already explained, correlations on all samples are plagued by the fact that the sampling and sequencing strategies changed over time, and Simpson's paradox is at play. Focusing on a homogeneous subset of the data, collected on January 12th, I also already pointed out that correlations were not statistically significant for raccoon dogs nor for humans. We can additionally consider the January 1st data: this first trip focused on stalls with human cases; only samples positive by PCR were sequenced. Even on this subset of data for which humans were the likely source of potentially all SARS-CoV-2 viruses, the correlation is not significant for humans (cor = 0.3 (95% CI: -0.1, 0.62), $p = 0.14$; data: Bloom (2023b)). Finally, Bloom (2023b) highlighted a high positive correlation between SARS-CoV-2 and a fish species, largemouth bass. This non-sensical result by itself illustrated that market-wide correlations cannot identify hosts. We however need to note that the result was largely driven by the exclusion of two samples, F13 and F54. Once these metagenomic samples are added back, the correlation is much weaker and non-significant (see Table 1).

3. Discussion

There are major scientific issues with the original results presented in Bloom (2023b). The market-wide correlations on all samples from all locations and at all dates lump together data that cannot be analyzed together, because the samples were collected with different purposes and sequenced for different reasons. The correlation values are therefore uninterpretable, and many of these values are affected by an instance of Simpson's paradox: they change once structure in the data is taken into account. Many of the values put forward are actually not statistically significant. The criteria that are applied to non-human animals are so stringent that they would exclude humans as a source of SARS-CoV-2 in many samples. Finally, the discussion of the SARS-CoV-2-positive or -negative nature of some samples ignores the spatial and temporal structure of the data.

There is also a major issue with the presentation of the motivation for the study. Contrary to what is written in Bloom (2023b), and as previously highlighted in a reply to the first version of the preprint of this work[13], it was not suggested that it would be valuable to reproduce Liu *et al.* (2022)'s Figure 4A. It was obvious before doing it that such an approach would not identify an animal host – not to mention the odd choice of plotting a correlation of correlations. In addition, Bloom (2023b) does not contradict news reports on "co-mingling" of SARS-CoV-2 and raccoon dogs, because these articles were focused on one stall, which was clearly positive for SARS-CoV-2, while Bloom's study is at the scale of

[13] See footnote 8.

the whole market. In the news articles criticized by Bloom, none of the interviewed collaborators of Crits-Christoph *et al.* (2023a) mentioned co-presence of SARS-CoV-2 and raccoon dogs at the scale of the whole market.

The metagenomic sequence data from the Huanan market are observational: by nature, they cannot prove any hypothesis. They could however disprove hypotheses if they were shown to be incompatible. Before the contents of these data were revealed, falsifiable predictions had been made:

i) If the origin is at the market, then lineage A should be in the market. Early SARS-CoV-2 sequences were grouped into two lineages, A and B, separated by two characteristic mutations. Initially however, lineage A had not been associated with the market. Absence of evidence is not evidence of absence, so this absence of detection did not disprove a market origin, but it was nevertheless seen as a serious limitation for a single Huanan market-origin scenario (Zhang *et al.*, 2020; Bloom, 2021). Worobey *et al.* (2022b) had predicted that, given the location of the two earliest publicly known lineage A cases, lineage A had to be in the market. Liu *et al.* (2022) revealed that lineage A was indeed in the market, found in sample A20, which independently validated the prediction.

ii) If the virus comes from infected animals, then there should be human infections and positive environmental samples near stalls selling animals. Most cases are from the West side of the market, and it was recognized as early as January 2020 that this was where wildlife stalls were. The details published by Liu *et al.* (2023b) on the numbers of samples collected in each stall, and the associated PCR and sequencing results, allowed the identification of a positivity hotspot in the market, controlling for sampling effort (Crits-Christoph *et al.*, 2023b). This result rebutted the claim that more positive samples had been found in the south-west corner of the West side because more samples had been collected there.

iii) If the virus comes from infected animals, then there should exist stalls in which the genetic materials of SARS-CoV-2 and of susceptible animals are detectable together. This prediction was also confirmed by the metagenomic sequence data from the Huanan market. Strikingly, the stall (stall A) in which susceptible animals and SARS-CoV-2 are detected together happens to be next to the positivity hotspot.

Predictions ii) and iii) could have been disproved by the Huanan market data. Positive samples or cases could have been predominantly in the East side of the market. Susceptible animals could have been totally disjoint from SARS-CoV-2, as is the case a few streets up stall A, where wildlife was sold but no SARS-CoV-2 was detected, even though the stalls were heavily sampled (Liu *et al.*, 2023b; Crits-Christoph *et al.*, 2023b).

An animal origin does not imply that all SARS-CoV-2 in the market was shed by (non-human) animals, nor that all humans were infected by animals. Bloom (2023b) cited an article by Courtier-Orgogozo and De Ribera (2022) as an alternative interpretation of Liu *et al.* (2022)'s data. This article proposes that the toilets or a mahjong room, both in the South-West corner of the West side of the market, were locations where human-to-human transmission took place. It is important to note that these suggestions are not incompatible with an animal origin — quite the contrary actually. Most SARS-CoV-2 detected in the market was shed by humans, and most humans were infected by other humans. A fraction of these infections can have taken place in places where people gathered, without precluding in any way the fact that a couple of people may have been initially infected by animals. The news article[14] at the origin of the suggestion that toilets may have played a role in transmission is actually even a good argument for an origin linked to the wildlife trade in the market, when the quote is read in full (emphasis added): "Looking back, Ms. [W.] thinks she might have been infected via the *toilet she shared with the wild meat sellers* and others on the market's west side.".

Proving that non-human animals were infected would not prove that SARS-CoV-2 first entered the Huanan market inside of non-human animals. It could still be argued, as is

[14] Jeremy Page, Wenxin Fan and Natasha Khan, *How It All Started: China's Early Coronavirus Missteps*; `https://www.wsj.com/articles/how-it-all-started-chinas-early-coronavirus-missteps-11583508932`

done in Liu *et al.* (2023b) and Bloom (2023b) that the animals could have been infected by humans. Likewise, it can be argued that the detection of SARS-CoV-2 on a cage could be from a human coughing on the cage, and not from an animal inside of it. At some point however, making sense of all the available data – including a feature not discussed yet, the distribution of early cases around the market (Worobey *et al.*, 2022a) – becomes unreasonably difficult under a lab leak scenario. It then becomes necessary to invoke deliberately hidden data, as is done in Bloom (2023b) with the reference to Hvistendahl & Mueller's news article[15]. This reference does not only amount to suggesting that the Chinese government is withholding data, but also that Chinese scientists are hiding or manipulating such data in their own papers.

To conclude, the metagenomic data from the Huanan market cannot by nature prove that a non-human animal was infected, and it was known before even conducting it that a market-wide correlation analysis on these data would not prove it either. Crits-Christoph *et al.* (2023a) did not pretend it would be the case. In addition, Bloom (2023b)'s analysis had substantial scientific issues. The article listed "lab leak" as a keyword, even though the Huanan market data do not directly inform this other hypothesis. Whether or not keeping the discussion alive was the intention of the article, this was its outcome in the media. Maybe such aims are better served by Op-Eds, rather than by conducting a study doomed to fail to argue that nothing can be concluded from a dataset. The metagenomic sequencing data from Huanan market samples are among the most precious and insightful datasets related to the origin of the pandemic shared by Chinese researchers. Even though they were shared late, we can be grateful that these data exist and are now publicly available.

4. Methods

I downloaded analysis results by Bloom (2023b) (`https://github.com/jbloom/Huanan_market_samples`) and by Crits-Christoph *et al.* (2023b) (`https://github.com/sars-cov-2-origins/huanan-market-environment`). I excluded samples from outside the market: sewage from surrounding areas, and warehouses, because SARS-CoV-2 in such samples can have been deposited after the closure of the Huanan market. I also excluded samples corresponding to amplicon-based SARS-CoV-2 whole genome sequecing. I kept all samples from the Huanan market and subjected to metagenomic sequencing.

The `cor.test` function in R was used to compute correlations, confidence intervals, p values.

R scripts to reproduce the results are available at `https://github.com/flodebarre/Huanan-env_Bloom-reply`.

5. Acknowledgements

I thank Alex Crits-Christoph and Zach Hensel for discussions and comments.

References

Bloom JD. 2021. Recovery of Deleted Deep Sequencing Data Sheds More Light on the Early Wuhan SARS-CoV-2 Epidemic. Molecular Biology and Evolution. 38:5211–5224. `https://academic.oup.com/mbe/article/38/12/5211/6353034`.

Bloom JD. 2023a. Association between SARS-CoV-2 and metagenomic content of samples from the Huanan Seafood Market. Preprint. Microbiology. `https://www.biorxiv.org/content/10.1101/2023.04.25.538336v1`.

[15] Mara Hvistendahl and Benjamin Mueller, *Chinese Censorship Is Quietly Rewriting the Covid-19 Story*; `https://www.nytimes.com/2023/04/23/world/europe/chinese-censorship-covid.html`

Bloom JD. 2023b. Association between SARS-CoV-2 and metagenomic content of samples from the Huanan Seafood Market. Virus Evolution. 9:vead050. https://academic.oup.com/ve/article/9/2/vead050/7249794.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y *et al*. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. The Lancet. 395:507–513. https://www.sciencedirect.com/science/article/pii/S0140673620302117.

Courtier-Orgogozo V, De Ribera FA. 2022. SARS-CoV-2 infection at the Huanan seafood market. Environmental Research. 214:113702. https://www.sciencedirect.com/science/article/pii/S0013935122010295.

Crits-Christoph A, Gangavarapu K, Pekar JE, Moshiri N, Singh R, Levy JI, Goldstein SA, Suchard MA, Popescu S, Robertson DL *et al*. 2023a. 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. Technical report. Zenodo. https://zenodo.org/record/7754299.

Crits-Christoph A, Levy JI, Pekar JE, Goldstein SA, Singh R, Hensel Z, Gangavarapu K, Rogers MB, Moshiri N, Garry RF *et al*. 2023b. Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic. Preprint. Genomics. https://www.biorxiv.org/content/10.1101/2023.09.13.557637v1.full.

Freuling CM, Breithaupt A, Müller T, Sehl J, Balkema-Buschmann A, Rissmann M, Klein A, Wylezich C, Höper D, Wernike K *et al*. 2020. Susceptibility of Raccoon Dogs for Experimental SARS-CoV-2 Infection. Emerging Infectious Diseases. 26:2982–2985. http://wwwnc.cdc.gov/eid/article/26/12/20-3733_article.htm.

Gostin LO, Gronvall GK. 2023. The Origins of Covid-19 — Why It Matters (and Why It Doesn't). New England Journal of Medicine. 388:2305–2308. http://www.nejm.org/doi/10.1056/NEJMp2305081.

Keusch GT, Amuasi JH, Anderson DE, Daszak P, Eckerle I, Field H, Koopmans M, Lam SK, Das Neves CG, Peiris M *et al*. 2022. Pandemic origins and a One Health approach to preparedness and prevention: Solutions based on SARS-CoV-2 and other RNA viruses. Proceedings of the National Academy of Sciences. 119:e2202871119. https://pnas.org/doi/10.1073/pnas.2202871119.

Liu W, Liu P, Lei W, Jia Z, He X, Liu LL, Shi W, Tan Y, Zou S, Zhao X *et al*. 2022. Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market. https://www.researchsquare.com/article/rs-1370392/v1.

Liu WJ, Liu P, Lei W, Jia Z, He X, Shi W, Tan Y, Zou S, Wong G, Wang J *et al*. 2023a. Surveillance of SARS-CoV-2 at the Huanan seafood market. http://chinaxiv.org/abs/202303.10351.

Liu WJ, Liu P, Lei W, Jia Z, He X, Shi W, Tan Y, Zou S, Wong G, Wang J *et al*. 2023b. Surveillance of SARS-CoV-2 at the Huanan Seafood Market. Nature. https://www.nature.com/articles/s41586-023-06043-2.

Tan W, Zhao X, Ma X, Wang W, Niu P, Xu W, F. Gao G, Wu G, MHC Key Laboratory of Biosafety, National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China, Center for Biosafety Mega-Science, Chinese Academy of Sciences, Beijing, China. 2020. A Novel Coronavirus Genome Identified in a Cluster of Pneumonia Cases — Wuhan, China 2019-2020. China CDC Weekly. 2:61–62. http://weekly.chinacdc.cn/en/article/doi/10.46234/ccdcw2020.017.

The BMJ. 2021. Origins of covid. https://youtu.be/eLSv4Iwk_jM?feature=shared&t=1893.

World Health Organization. 2021. *WHO-convened Global Study of Origins of SARS-CoV-2: China Part: Joint WHO-China Study, 14 January-10 February 2021 : Joint Report*. WHO. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

Worobey M. 2021. Dissecting the early COVID-19 cases in Wuhan. Science. 374:1202–1204. https://www.science.org/doi/10.1126/science.abm4454.

Worobey M, Levy JI, Serrano LM, Crits-Christoph A, Pekar JE, Goldstein SA, Rasmussen AL, Kraemer MUG, Newman C, Koopmans MPG *et al*. 2022a. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science. p. abp8715. https://www.science.org/doi/10.1126/science.abp8715.

Worobey M, Levy JI, Serrano LMM, Crits-Christoph A, Pekar JE, Goldstein SA, Rasmussen AL, Kraemer MUG, Newman C, Koopmans MPG *et al*. 2022b. The Huanan market was the epicenter of SARS-CoV-2 emergence. https://zenodo.org/record/6299600.

Wu G. 2020. Report from the Chinese Center for Disease Control and Prevention's Office of Virus Control and Prevention, Regarding the Results of Environmental Sample Testing for the Novel Coronavirus Epidemic in Wuhan. Technical Report 53. Chinese Center for Disease Control and Prevention. https://archive.ph/xPBFD.

Xiao X, Newman C, Buesching CD, Macdonald DW, Zhou ZM. 2021. Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. Scientific Reports. 11:1–7. https://www.nature.com/articles/s41598-021-91470-2.

Xu RH, He JF, Evans MR, Peng GW, Field HE, Yu DW, Lee CK, Luo HM, Lin WS, Lin P *et al*. 2004. Epidemiologic Clues to SARS Origin in China. Emerging Infectious Diseases. 10:1030–1037. http://wwwnc.cdc.gov/eid/article/10/6/03-0852_article.htm.

Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z, Jia X, Wu M, Shi B, Xu S *et al*. 2020. Viral and host factors related to the clinical outcome of COVID-19. Nature. 583:437–440. https://www.nature.com/articles/s41586-020-2355-0.