# On some unpublished early SARS-CoV-2 sequences

*@Babarlelephant @Engineer2The @Drinkwater5Reed @Franciscodeasis*

June 2022

# Abstract

The origin of SARS-CoV-2 is still unknown: the chain of events that brought a virus whose close relatives are found in Rhinolophus bats in the Yunnan province and in Laos [18, 19] to the Huanan seafood market in Wuhan in early December 2019 remains to be elucidated. In particular, the non-market patients and the genetically more ancestral Lineage A remain mysterious.

A retrospective analysis identified 174 patients with onset in December [3a, 10], among them only 15 have been sequenced and published, often multiple times [3b, 11].

By collating as much data as possible on early cases – research papers, sequence databases, news articles, social media posts – we found some data on 65 patients with onset in December 2019 [Note 0]. Furthermore, we detected two patients who had been sequenced, but whose sequences were never uploaded to a public database and whose raw reads, although published [1], were not reanalyzed.

We also present some information on the first Beijing patient, who had an onset date of December 17, 2019 and was related to the Huanan market outbreak.

Using the collated information, significant progress has been made towards solving the discrepancies in the early sequences. A phylogeny of 19 early patients is presented, based on onset dates, as well as several tMRCA estimates – falling in late November.

# Main text

## Table of early sequences

Data was gathered from multiple sources and metadata on the sequences of Covid-19 cases with onset date in December 2019 are presented in Table 1 below.

As in [11] the patients' identifiers consist of their age and sex, followed by a number in case of duplication.

| Age&Sex | Onset | Admission / hospital | Sequence names |
|---|---|---|---|
| 21F | 2019-12-26 | 2019-12-28<br>Zhongnan then Jinyintan | WH03<br>WHU02 |

| | | | |
|---|---|---|---|
| 32M | 2019-12-19 | 2019-12-29<br>Jinyintan | WIV02<br>HBCDC-HB-02/2019<br>HBCDC-HB-04/2019 (low quality)<br>IVDC-HB-GX02<br>IVDC-HB-05 [Note 10]<br>WH19008<br>IME-WH03 [Note 9] |
| 39M1 | 2019-12-20 | 2019-12-25<br>Zhongnan then Jinyintan | WHU01 |
| 39M2 | 2019-12-27 | ?<br>PLA | WH04 |
| 40M1 | 2019-12-17 | 2019-12-27<br>Integrated [Note 8] then Jinyintan | WH19003 |
| 40M2 | 2019-12-22 | 2019-12-28<br>Integrated [Note 8] then Jinyintan | WIV06<br>WH19010 (low quality)<br>IME-WH02 [Note 9] |
| 41M1 | 2019-12-16 | 2019-12-22<br>Jiangxia then Central (Nanjing) then Jinyintan | IPBCAMS-WH-03 [Note 12]<br>Beijing Boao report leaked on Wechat (unpublished) |
| 41M2 | 2019-12-20 | 2019-12-26<br>Central (Houhu) then Jinyintan | Wuhan-Hu-1 |
| 41M3 | 2019-12-23 | 2019-12-29<br>Tongji then Jinyintan | WH19053 |
| 43M | unclear | unclear [Note 7] | WH02 |
| 44M | 2019-12-17 | 2019-12-24<br>Tongji then Jinyintan | WH01 |
| 49F | 2019-12-23 | 2019-12-27<br>Integrated [Note 8] then Jinyintan | WIV04<br>WH19001 (cultured as WH19005)<br>IVDC-HB-01<br>IPBCAMS-WH-02<br>HBCDC-HB-01/2019<br>HBCDC-HB-03/2019 |
| 51M | 2019-12-17<br>travel to Beijing the same day | 2019-12-27<br>Beijing 5th PLA | Beijing-01 |

| | | | |
|---|---|---|---|
| 52F | 2019-12-22 | 2019-12-29<br>Integrated [Note 8] then Jinyintan | WIV05<br>IME-WH05<br>WH19002<br>IPBCAMS-WH-04 [Note 12] |
| 53F | 2019-12-25 | 2019-12-31<br>likely Jinyintan | WH19016 |
| 56M | 2019-12-20 | 2019-12-30<br>Jinyintan | WIV07<br>IME-WH04<br>WH19012 (low quality) |
| 61M | 2019-12-20 | 2019-12-27<br>Puren then Jinyintan<br>(died 2020-01-09) | IVDC-HB-04<br>WH19004<br>IPBCAMS-WH-05 [Note 12] |
| 62M | 2019-12-12 | 2019-12-27 [Note 11]<br>Integrated [Note 8] then Jinyintan | IME-WH01 [Note 12] |
| 65M | 2019-12-13 | 2019-12-18<br>Central (Nanjing) then Tongji then Jinyintan<br>(may have died on 2020-01-30) | IPBCAMS-WH-01 [Note 16] |
| 57F | 2019-12-11 | 2019-12-18<br>Union | Huanan environmental sequence IVDC-HBF54 |
| ? | ? | no matched patient, no information | Huanan environmental sequences A20, A18, A2 from a single stall [16] |
| ? | ? | // | Environmental sequences F13 and B5 |

*Table 1 - matching between patients and early sequences*

- All the sequences are lineage B with the exception of 62M and 39M2 and the environmental sample A20.
- All the lineage B patients are linked to the Huanan market with the exception of 41M1, who lived 30km south of the market.
- Those linked to the market are workers, with the exception of 41M2, 41M3, 61M who are frequent shoppers, 21F who had close contact with a vendor, 51M who was a pharmacist near the entrance of the market, and 53F who was married to 61M.
- It is important to note that none of the workers were found to be linked with wildlife trade and game food, which are known to be present in the Huanan market in late 2019 [21, 8] although at a much smaller scale than in Guangdong markets.
- The Huanan market environmental sequences were collected on January 1.
- Several other low quality and unpublished environmental sequences [7,16] aren't included in the table.

The sequences WH19002 [Note 15] WH19003 WH19010 WH19012 WH19016 WH19053 were assembled from the raw reads provided in [1b] since they were not published in public repositories.

In the process of consolidating data, it was found that two early patients had been sequenced, but their sequences have not been shared publicly so far.

- 53F is the wife of a 61 year-old man who was the first patient to die on January 9 and who had a liver condition [2, 6]. This couple [Note 1] is mentioned in [4a] and among the 15 early patients previously known to be sequenced no-one was found to be a plausible match.[Note 2]
- Multiple lines of evidence allow us to match WH19003 with 40M1 [Note 4] a Huanan market seller who worked at the same stall as 32M [4a].

| | | |
|---|---|---|
| sample name | WH19016 (nCoV7) [1] | [1] |
| age & sex | 53 F | [1, 6] |
| onset | 2019-12-25 | [1a, 4, 5] |
| link to Huanan market | wife of 61M who was a frequent shopper at Huanan | [1, 4, 5, 6] |
| admission | Dec 31 | [4a] |
| hospital | likely Jinyintan | [6] |
| collection date | 2020-01-01 | [1c] |
| severity | mild pneumonia, recovered | [1a, 6a] |
| mutation (wrt Hu-1) | 3059 ambiguous sites | |

| | | |
|---|---|---|
| sample name | WH19003 (nCoV3) [1] | [1] |
| age & sex | 40 M | [1, 6] |
| onset | 2019-12-17 | [1a, 4, 5, 12a] |
| link to Huanan market | worker, same stall as 32M | [4a, 5] |
| admission | 2019-12-27 | [4a, 12a] |
| hospital | Integrated, then Jinyintan. | [12a, 22] |
| collection date | 2020-12-30 | [1c] |
| severity | severe pneumonia | [1a, 12a, 6a] |
| mutation (wrt Hu-1) | A24325G<br>3942 ambiguous sites | |

*Table 2 - main informations on two unpublished sequences*

# 53F was infected by 61M

Contact tracing and onset dates make it clear that 53F was infected by her husband 61M and this is consistent with the consensus genome of the two patients, both identical to Wuhan-Hu-1.

## Possible transmission from 40M1 to 32M

32M and 40M1 worked in the same stall [4a] and their consensus genomes are identical, with one mutation A24325G. 40M1 had symptom onset two days earlier than 32M and both were severe according to [12a]. It is thus possible that 40M1 infected his colleague. However, contamination of 40M1 by 32M or contamination of both by a third unsequenced patient cannot be excluded.

## A third unpublished patient

The same paper [1] may include a third unpublished sequence and patient. Our inference is that WH19053 corresponds to 41M3. [Note 5.] This inference is based on careful analysis of the available data, but may be reevaluated when additional clarifying data becomes available. [Note 13]

## The first case in Beijing

We found that the sequence Beijing-01, collected on January 3 and published on GISAID in August 2020, matches with a Chinese thesis [20] which reveals that the patient was a pharmacist near the entrance of the Huanan market, who had an onset of symptoms on December 17 and traveled to Beijing on the same day, were he was eventually hospitalized on December 27.

This patient, who had a pre-existing HIV and hepatitis B condition, was then diagnosed with novel coronavirus pneumonia combined with type I respiratory failure on January 4, 2020. This is surprising because officially the first case in Beijing was not announced before January 20, 2020.

## Missing data and confusion in the WHO report

The WHO report [3b] includes a very useful deduplication table. We discovered that six sequenced early patients are missing:

- 53F,
- 40M1,
- 41M3 (the WH19053 sequence should have been included in the table even if it happened to correspond instead to either 41M1 or 41M2),
- 51M, collection date on January 3,
- 49M2, collection date on January 5,
- 43M (lower quality genome, we reassembled a sequence with no mutation and 12045 N from [14]).

There are also several errors in the table:

- S01 corresponds to 41M1 – not linked to the market – and the ambiguity of his onset date (Dec 16 instead of Dec 8) has been discussed before. [Note 3] The WHO table gives IPBCAMS-WH-05 for the sequence whereas its Gisaid metadata corresponds to 61M.
- We assume that the Gisaid metadata is correct, so that IPBCAMS-WH-03 IPBCAMS-WH-04 IPBCAMS-WH-05 are misplaced in the WHO table.
- Based on the mutations there is a mix-up between IME-WH02 and IME-WH03, which correspond to WIV02 and WIV06 respectively.
- The onset date of 62M is disputed [12a] but it certainly can not be claimed to be the day of admission and diagnosis of pneumonia [24a, 22].

# Phylogeny of early patients

Our investigation yielded 37 exploitable sequences from 19 patients (Table 1) represented in the following tree

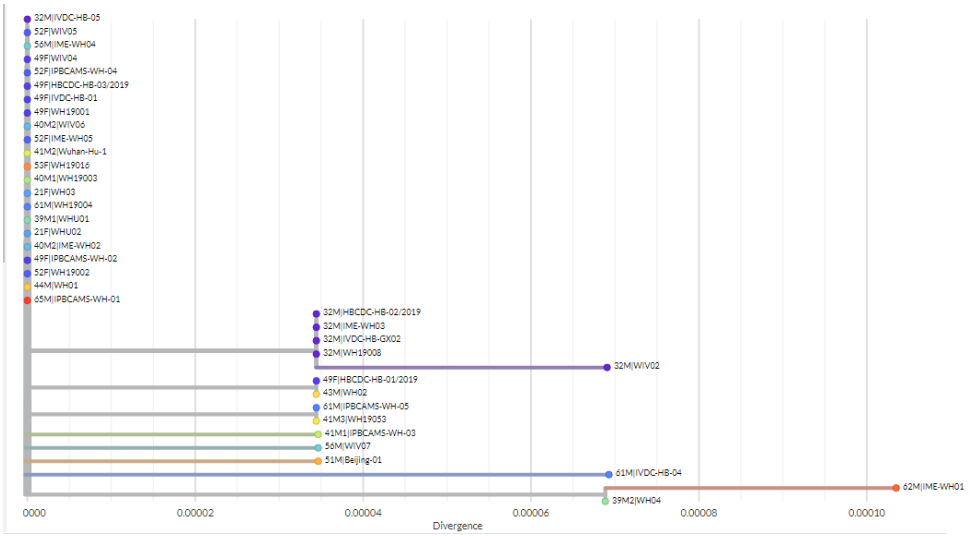



*Figure 1 - Tree of the 37 genomes from 19 patients link*

A consensus genome for each patient was then obtained by adding a mutation with respect to Wuhan-Hu-1 only if it appeared in at least half of the patient's genomes.

There is not enough temporal signal to reliably estimate the clockrate just from the early sequences. So all our timetrees used a fixed clockrate of 0.0008 subs/site/year (eg. 2 mutations per month) which is the rate observed during the year of 2020 [Note 20].

## MRCA in late November

We made a tree of these 19 consensus genomes with Iqtree2+Treetime, using the onset dates for the tip dates. We included the putative recombinant bat virus from [13a] as an outgroup which makes the lineage A genetically ancestral (see Supplementary Data - on ancestrality of A for a discussion).
The MRCA was then estimated to be in late November.

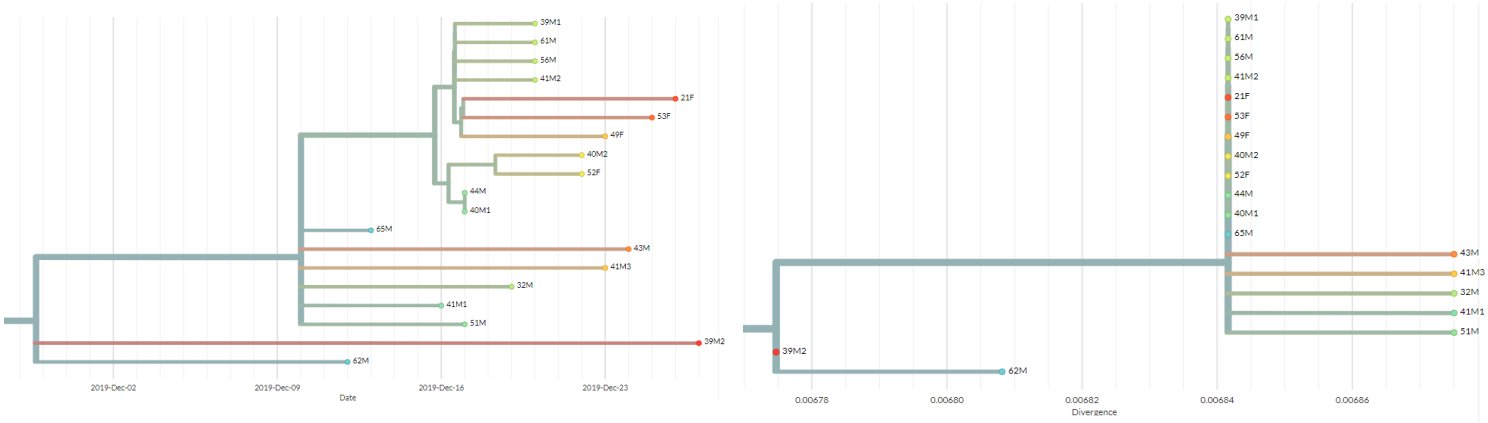



*Figure 2 - Tree of the 19 consensus sequences obtained for each patient, dated with onset dates and rooted by adding the putative recombinant bat virus from [13a] link*

With the same dataset, BEAST bayesian skygrid HKY 0.0008 subs/year/site gave an MRCA in late November (confidence interval 10-29, 12-06) and early December for the lineage B (11-26, 12-11).

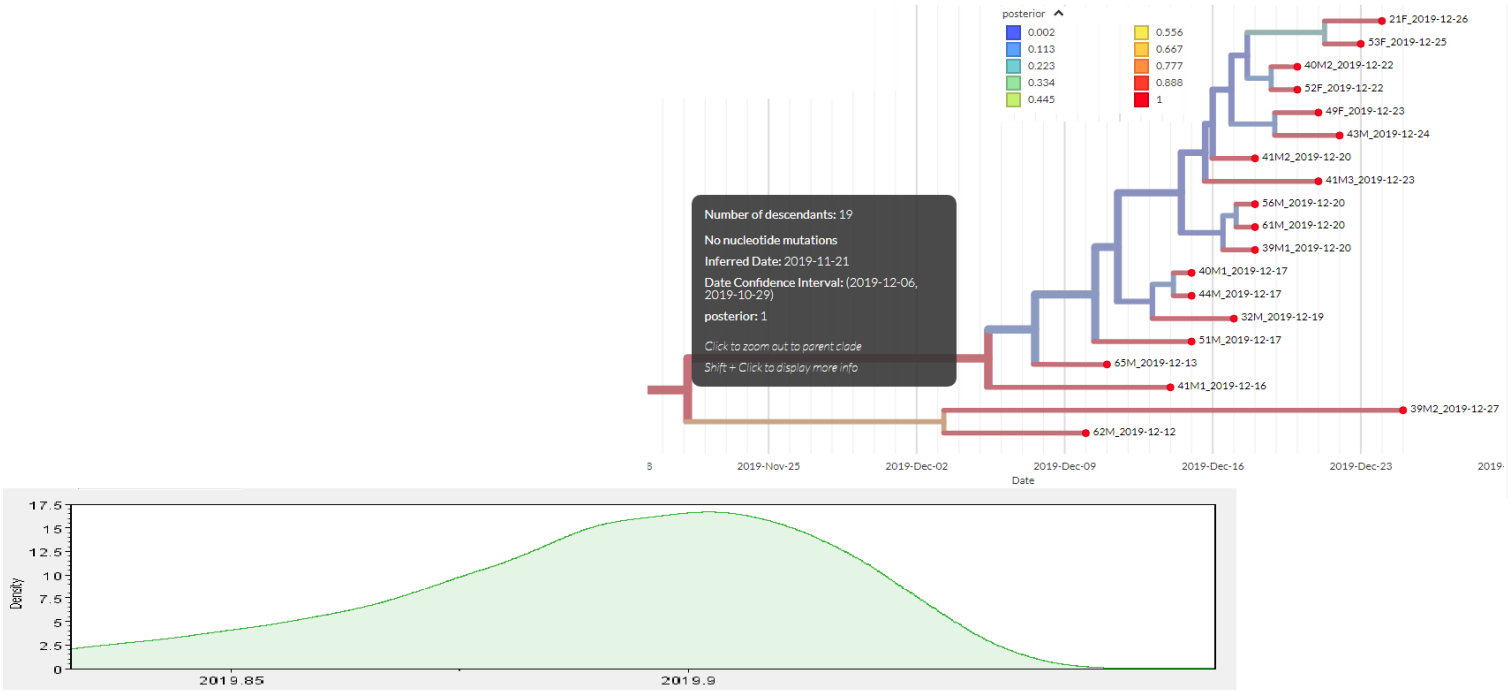



*Figure 3 - BEAST skygrid HKY 0.0008 subs/site/year (ESS > 500) link*

The MRCA dating stayed roughly the same when 62M's onset date was changed to December 24. Note 11

As expected, the early sequences alone do not imply a narrow interval for the tMRCA.

The onset curves gathered from several retrospective studies (see [10]) are clearly not in favor of an MRCA before November 2019.

The MRCA stayed roughly the same on November 26 (10-30, 12-09) when we added 162 January sequences – randomly selected among those with less than 7 mutations, less than 3000 N and either CT or TC at the lineage A/B defining sites 8782/28144. Note 17, 18

Of note, in table S2 [14] some mean tMRCA are shown, going from December 4 to December 13 depending on the phylogenetic model.

Based on the data shown here and the onset curves in [10] (mostly severe cases) we believe that it is more reasonable to propose a MRCA in late November.

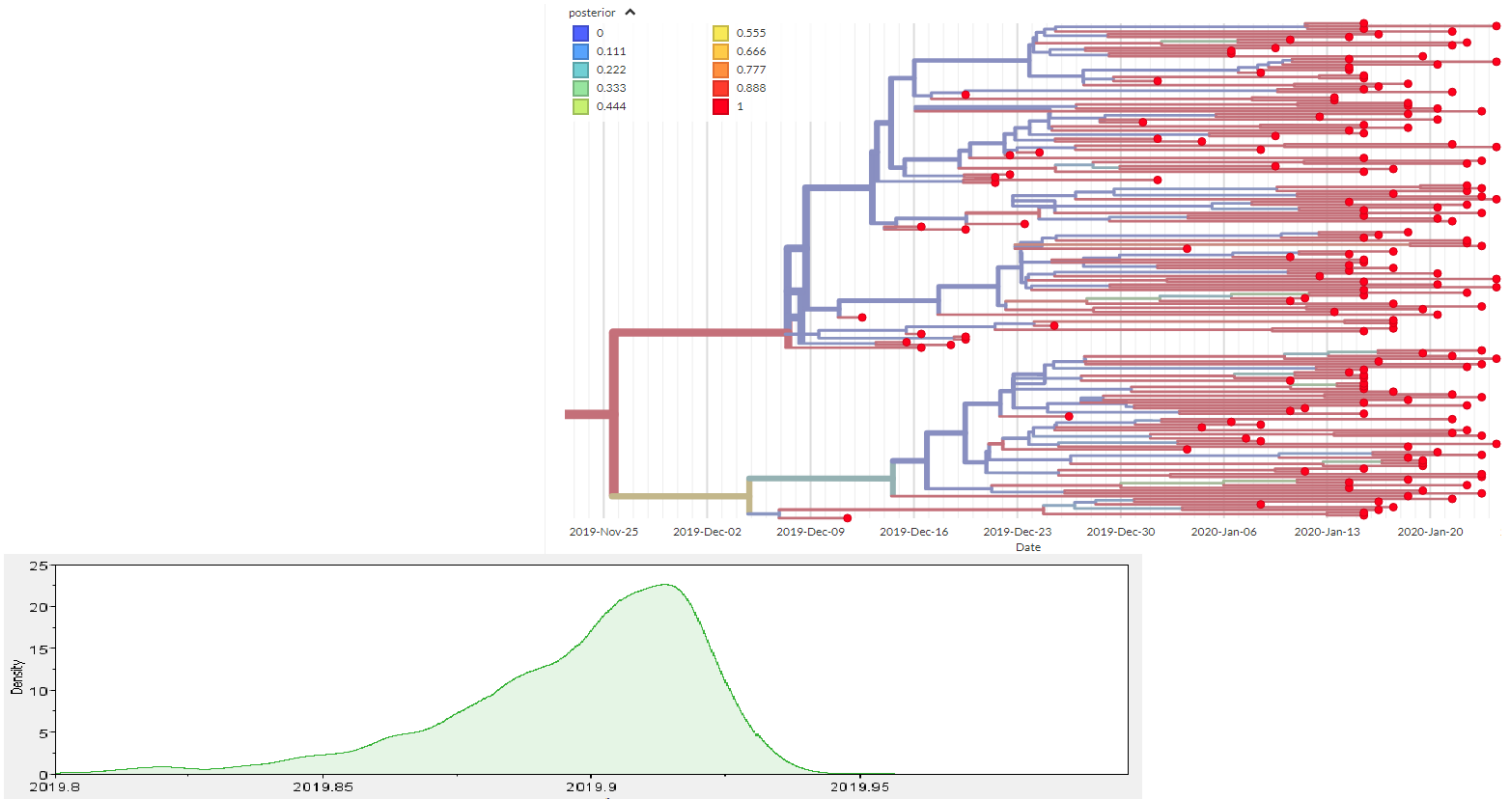



*Figure 4 - BEAST skygrid HKY 0.0008 subs/site/year (ESS > 200) link*

For all of these 162 January sequences we decreased the collection date by 6 days to avoid a large drop in tip dates between our 19 early patients and the January sequences, to fit with the idea of a phylogeny by onsets, and to take in account that we used a dataset curated for many sequencing errors.

In the dataset, 62 sequences are lineage A and 119 are lineage B.

We got similar estimates when using the coalescent exponential population prior, instead of the coalescent skygrid: November 25 (Nov 13, Dec 5). In this experiment we had removed the putative recombinant bat virus, as the ancestral bat lineages do not follow the exponential growth model.

## Taking in account the onset curve

We attempted to write a BEAST2 package (which reduced to 10 lines of java code – link) to take into account the onset curve from [23] as some kind of prior on the population sizes, which indirectly impacts the calculation of the likelihood of the tree and its tMRCA.

We chose to modify the likelihood calculation of the coalescent exponential population model in BEAST2, adding the log-likelihood of the onset curve given the effective population sizes generated by the exponential growth model.

Some fixed parameters were needed:

- the generation time, to convert the effective population sizes into number of infectees, was set to 5 days as in [24],
- the ascertainment rate, linking the onset curve with the total number of infections, was set to 0.15 as in [13].

The putative recombinant bat virus in these experiments was not included, as the ancestral bat lineages do not fit the exponential growth model. We used the onset curve from [23] from December 1 to January 11, removing the December 2 case (see p.46 of [3]) and moving the December 8 case to December 16.

Running BEAST2 with this modified model, the tMRCA was found to be on November 18 (11-11, 11-25). As we expected, the confidence interval was narrower when the observed onset curve was used to inform the parameters of the exponential model.

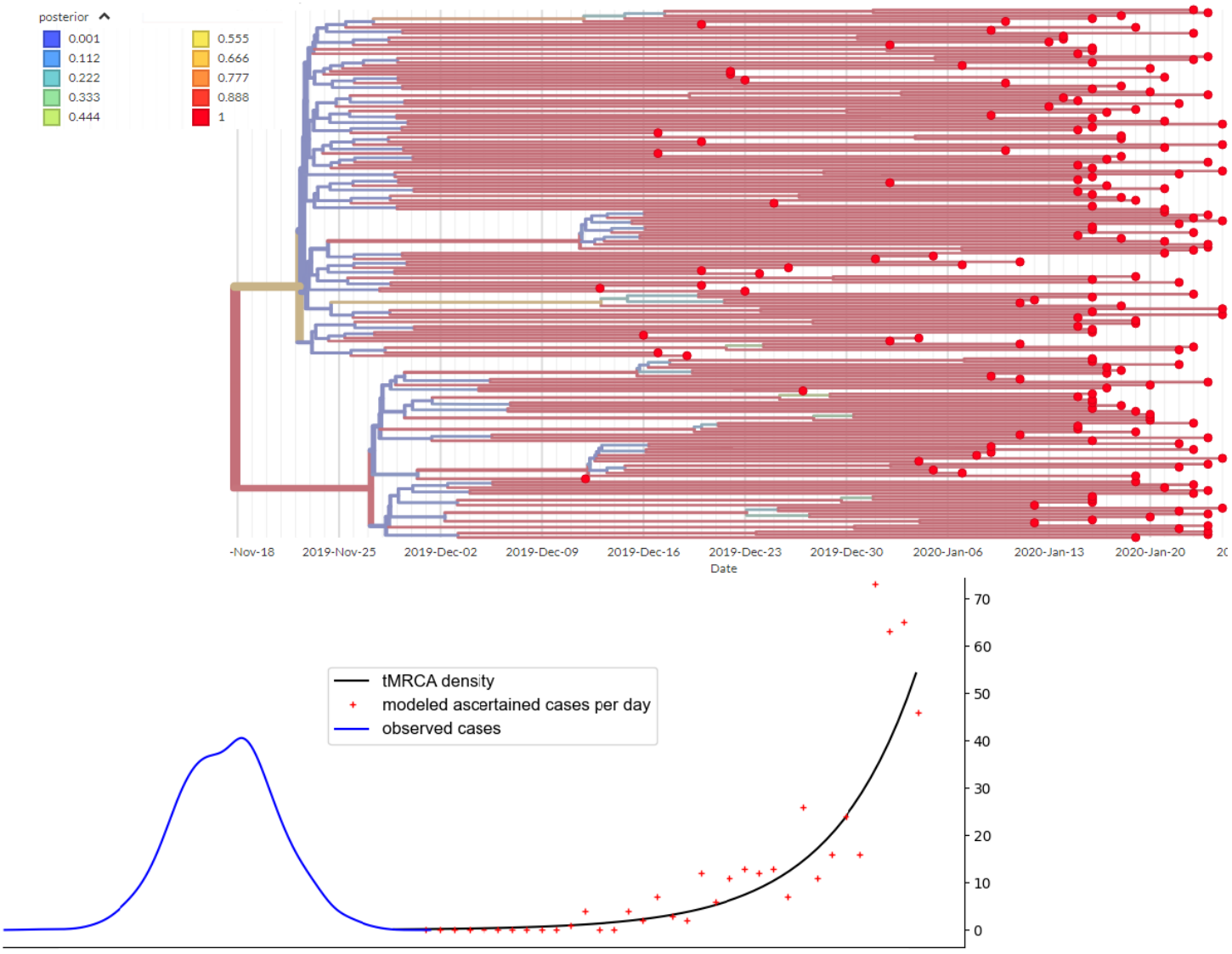



*Figure 5 - BEAST2 exponential growth HKY 0.0008/subs/site/year and likelihood of the onset curve (ESS > 200) link*

Finally we ran the same analysis after moving the onset date of 62M to Dec 15 and masking his T4946C mutation, we obtained the same tMRCA on November 18 (Nov 13, Nov 24) - link.

## The 37 early sequences don't give rise to well-supported lineages

Next, we wondered whether some of the mutations found in the 37 early sequences were supported by lineages sampled later.

- T13270C (found in 51M, Beijing-01) is found in 4 sequences collected in late January and early February, from Shanghai and Zhejiang. But all of them have 3 additional mutations (C3885A, C12778T, G29449T) and two late January Shanghai sequences have C3885A, G29449T but lack T13270C, so the picture is unclear.

- C28253T (found in IVDC-HB-04 but not in the consensus of 61M) is found in two late January Wuhan sequences, three C8782 C28144 late January sequences from Beijing and Sichuan, as well as in two February sequences from Henan and Singapore.

- T6996C is found in one late January Guangdong sequence having 4 additional mutations, including C17373T which defines a well-supported lineage.

Overall the early patient sequences lack some well-supported descendance, which is indeed surprising.

# Discussion and Conclusion

We note that among the sequenced early patients:

- 39M2 had the latest onset whereas his genome is the most ancestral.

- the genetic diversity of lineage B patients (all but one linked to the market) is very low, which suggests a recent introduction in the market followed by exponential growth.

Spread of lineage A in December is confirmed by several exports from Wuhan in early and mid-January (to Guangdong, Sichuan, Yunnan, Chongqing, USA, Fujian, Jiangxi, Zhejiang, Shandong, Thailand, South Korea, Vietnam, Taiwan) accounting for approximately one third of the sequenced cases.

Therefore, it appears that lineage A – more generally the non-market patients – were under-sequenced during the early days of the epidemic, which makes any strong conclusion on the origins hard to reach.

Although the analysis in [8] and the early sequences are consistent with the hypothesis that lineage B originated at the Huanan seafood market, lineage A remains mysterious.  Overall, the spread of SARS-CoV-2 outside of the market stays poorly understood.

- If lineage A was confirmed not to be associated with the Huanan market, then it would weaken the hypothesis that the virus was the result of a spillover at the Huanan market as proposed in [8].

- We note that A20, a lineage A environmental sequence from the Huanan market, was recently announced in [7], but it remains to be seen if the other two environmental samples from the same stall [17] confirm this finding or suggest otherwise. In [7] among the 9 Huanan environmental samples that can be typed (all collected on January 1) 8 are lineage B and 1 is lineage A. Even if A20 is confirmed, under the assumption that lineage A originated at the market, the number of lineage A samples found at the market is smaller than expected when compared to our assumption that lineage A represented about 1/3 of the early cases.[Note 19]

- If lineage A was finally found to be associated with the market then the MRCA of the whole SARS-CoV-2 tree would logically be located in the market.

Therefore, we believe that sequencing more early patients should be the top priority as some crucial questions remain unanswered.

Understanding the origins of SARS-CoV-2 is a question of understanding the context – location, host, date – of the common ancestor of lineage A and B, which is unfortunately not possible with the currently available data.

# Supplementary Data

## Methods

Each BAM was trimmed with iVar default params, except WH01 and WH02 from [14] (merging several DNBSEQ runs) which were trimmed with fastp (after increasing the maximum number of N to 12 in the options).

The consensus genome for each BAM from [1b] was then obtained with

```
samtools consensus -m simple
     --use-qual --min-MQ 10 --call-fract 0.75 -a
     --min-depth 10 --show-ins no --show-del yes
```

For WH19016 the minimum depth was decreased to 5.

Then the reads were manually inspected (see manual curation) setting more sites to N for one of the runs (WH19003 40M1) to obtain a final consensus - link

The WH02 sequence we reassembled from [14] contained 12045 N and no mutation. The WH01 sequence had 787 N and didn't contain the two mutations C6968A T11764A of the Gisaid WH01 sequence – we decided to replace the Gisaid sequence by this new assembly identical to Hu-1.

Other curations were made in the Gisaid sequences, for example IPBCAMS-WH-02 is identical to Hu-1 with the exception of 6 mutations at neighboring sites 104,111,112,119,120, 124 that we chose to mask.

The WHO report reanalyzed several of the early sequences and found that the three mutations of IPBCAMS-WH-01 were not supported by the raw reads (which are not publicly available). We therefore chose to mask these three mutations in this genome. The same was achieved for:

- two mutations of WIV05 (G7016A and A21137G, in the raw reads available on ncbi, 3 reads cover the site 7016 (A;2,G:1) and two cover the site 21137 (A:3))

- one of the two mutations of WIV07 (A8001C, this site is not covered by the raw reads available on ncbi, in contrast to the other mutation C9534T which is well-supported (T:16,C:1))

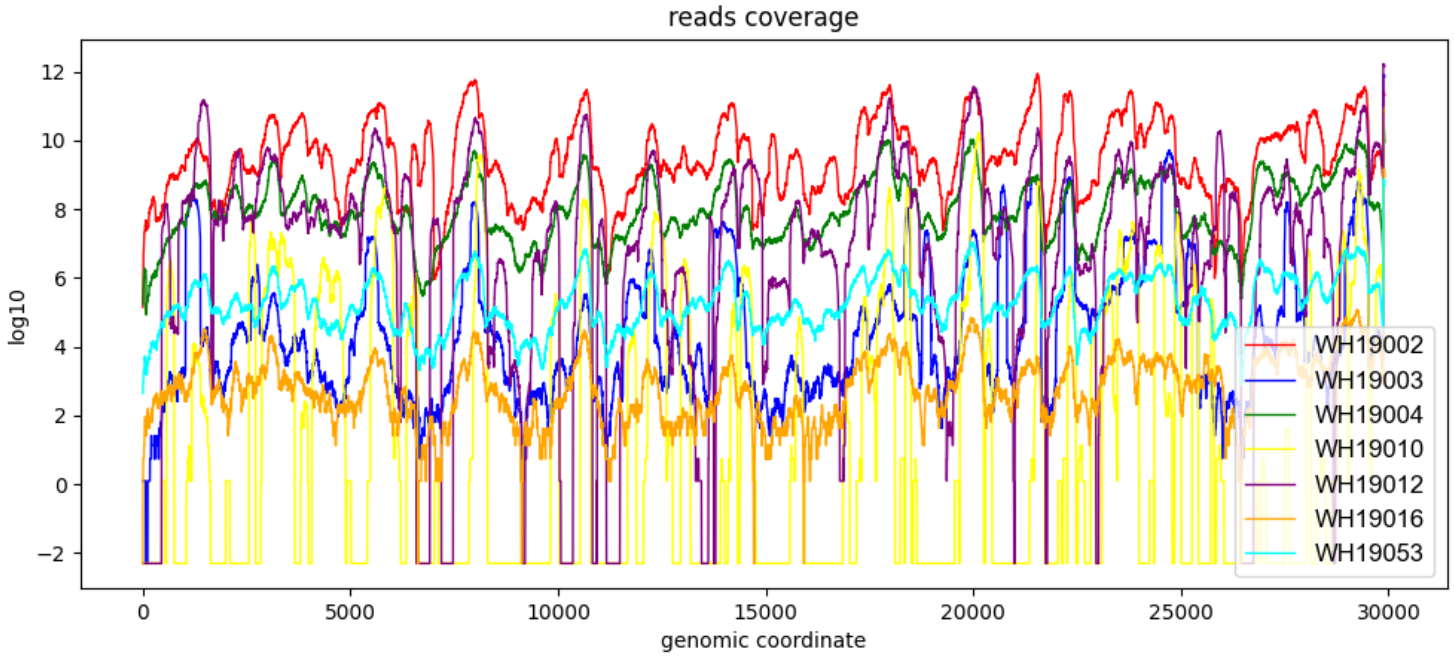
We also masked the first and last 100 sites in each sequence.

Not included in the trees:

- The low quality sequences: WH19010, WH19012, HBCDC-HB-04/2019.
- W19005, cultured from WH19001.

- The low-quality environmental sequence IVDC-HBF54 from the stall of 57F (onset Dec 11) which has yet to be sequenced otherwise. That environmental sequence may have been improved and would now be identical to Hu-1 according to [7].
- A20, A2, A18, three environmental sequences from a single stall, A20 is lineage A, but the genomes have not yet been published [7, 16].

The coverage in the raw reads from [1b] is variable

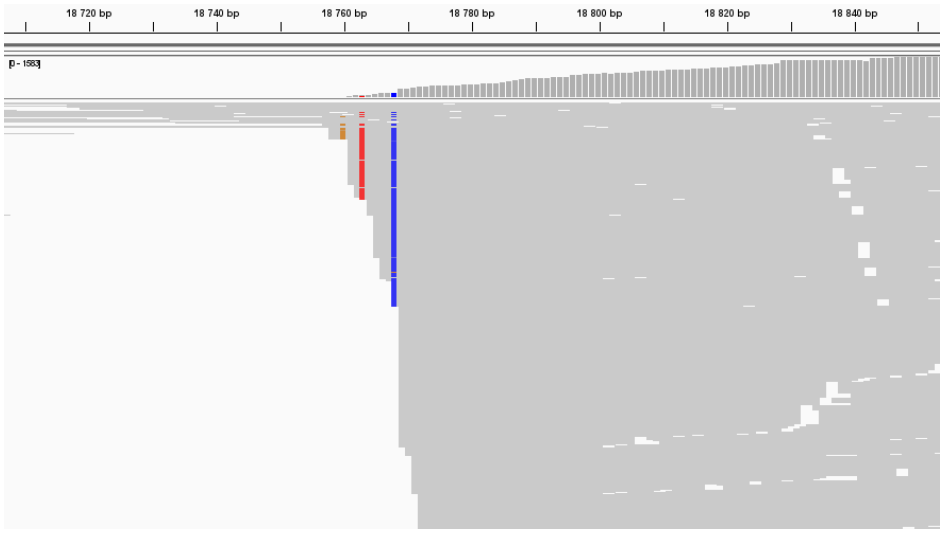



## Manual curation of WH19003 (patient 40M1)

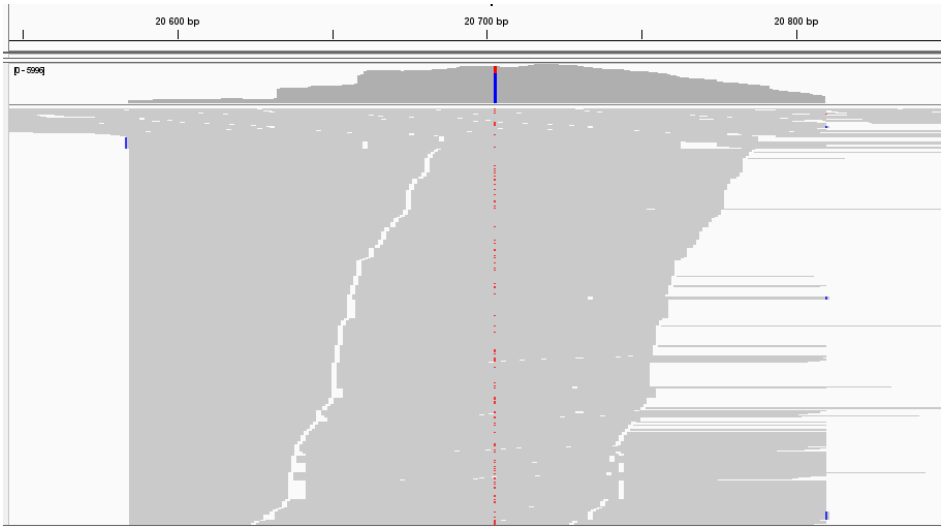
Some likely artifactual variants/mutations were found:

A18763T (A:15,T:70)
T18768C (T:17,C:167,G:1)

These mutations are clustered and appear only at the beginning of reads – which are preceded by a coverage drop. This fits better with a problem of primer/amplification than with intra-host variants.




On the other hand C20703T looks like a real intra-host variant, masked with a N.




Likely artifactual, clustered and appearing only at the end of the reads:

T24870C (T:155,C:854)
G24872C (G:149,C:580,A:1)
T24873A (T:147,A:571,G:1,C:1)

The same for T18974A (A:107,T:23,C:1).

## Confidence in the ancestrality of A

It has been claimed in [13] that the topology of the early SARS-CoV-2 tree, with two large polytomies separated by two mutations, was rather unlikely assuming only one introduction (or spillover).

If we assume that this oddity occurred, then the question remains if A is really ancestral or if it descends from B.

A naive computation is able to give part of the answer:

About 1.65% of the mutations observed in the global subsampled SARS-CoV-2 tree available at https://nextstrain.org/ncov/gisaid/global/all-time are indeed reversions to bat viruses (we mean the proportion of

mutations from the base of Hu-1 to a base conserved in RpYN06, RaTG13, BANAL-52, among all the unique mutations from the base of Hu-1 appearing in the tree).

The proportion increases to 3.5% when discarding the non-internal branches or when doing the above calculation with https://nextstrain.org/ncov/gisaid/reference that we annotated with reversions there. It is 3.9% in the early SARS-CoV-2 tree.

We can thus state with a high degree of confidence that lineage A does not descend from Lineage B.

## Linking BEAST population size parameters with the onset curve

This is the reasoning used for the code of our BEAST package

- Constant population coalescent

$$Pr(a, b \text{ coalesce at k th generation}) = (1 - 1/N)^{k-1}/N$$

where $N$ is the number of infected people, assumed to be constant.

$$Pr(a, b \text{ coalesce before k th generation}) = \sum_{l \leq k} (1 - 1/N)^{l-1}/N = \frac{1}{N} \frac{1 - (1 - 1/N)^k}{1 - (1 - 1/N)} = 1 - (1 - 1/N)^k$$

$k$ is in number of generation, so with a time variable $t$ (in year) it becomes

$$Pr(a, b \text{ coalesce before time t}) = 1 - (1 - 1/N)^{t/g}$$

with $g$ the generation time in year (ie. $g = 5/365$)

- BEAST' code says that

$$Pr(a, b \text{ coalesce before time t}) = 1 - e^{-t/N_e}$$

where $N_e$ is some parameter called the effective population size.

so

$$(1 - 1/N)^{1/g} = e^{-1/N_e} \implies \frac{1}{N_e} = -\log\ (1 - 1/N)^{1/g} \approx \frac{1}{gN}$$

($\approx$ is only valid for $N$ large)

> ie. $N_e$ is the product of the real population size and the generation time.

In this model there are $N$ cases per generation of 5 days, ie. $N/5$ cases per day.

So the expected number of cases per day is

$$N = \frac{1}{5(1 - e^{-g/N_e})} \approx \frac{N_e}{5g}$$

# Acknowledgments

We of course acknowledge the work of all the scientists - mainly in China - who collected samples, epidemiological data, and carried out the sequencing work during the early days of the epidemic. A table with the accessions and affiliations of the sequences used in this manuscript is available on github.

We'd also like to thank @FloDebarre and @BillyBostickson for their kind help during the review of the manuscript, as well as the @Nextstrain team whose great tools are the basis of all our visualizations.

# Notes

Note 0 - Sometimes the data is very scarce and we are unable to tell if the patient was confirmed. There is also a clear bias in our data, the patients linked to the market being more easily admitted in hospitals, better diagnosed, and better documented in newspapers.

Note 1 - The wife is said to be 53 year-old in [1, 6] and she is said to be 57 year-old in [4, 5].

Note 2 - The only ambiguity would be with the patient 52F who had been sequenced many times, but 52F and 53F appear as different patients in [1a].

Note 3 - The onset date of the non-market lineage B patient 41M1 has been discussed in https://www.washingtonpost.com/world/asia_pacific/covid-wuhan-outbreak-who/2021/07/15/51e7e8a6-e2c6-11eb-88c5-4fd6382c47cb_story.html and https://www.science.org/doi/10.1126/science.abm4454 (Dissecting the early COVID-19 cases in Wuhan, Worobey et al)

Note 4 - We match 40M1 with WH19003 from the age and sex, the mention of 15 days after onset, the similarity between table [1a] and table [12a], that one mutation is found both in 32M and WH19003 (given that 40M1 and 32M are said to work together in [4a]). 40M2 is a different patient in [1a, 12a].

Note 5 - See https://new.qq.com/omn/20200103/20200103A00QPZ00.html?pc and https://baijiahao.baidu.com/s?id=1654845412197587426

Note 7 - A 43M is mentioned in https://www.biorxiv.org/content/10.1101/2020.01.24.919183v2 but it is neither clear if it is the same patient nor if the given hospitalization date is relevant to his SARS-CoV-2 infection (he seems to have visited the Central hospital a second time as an outpatient).

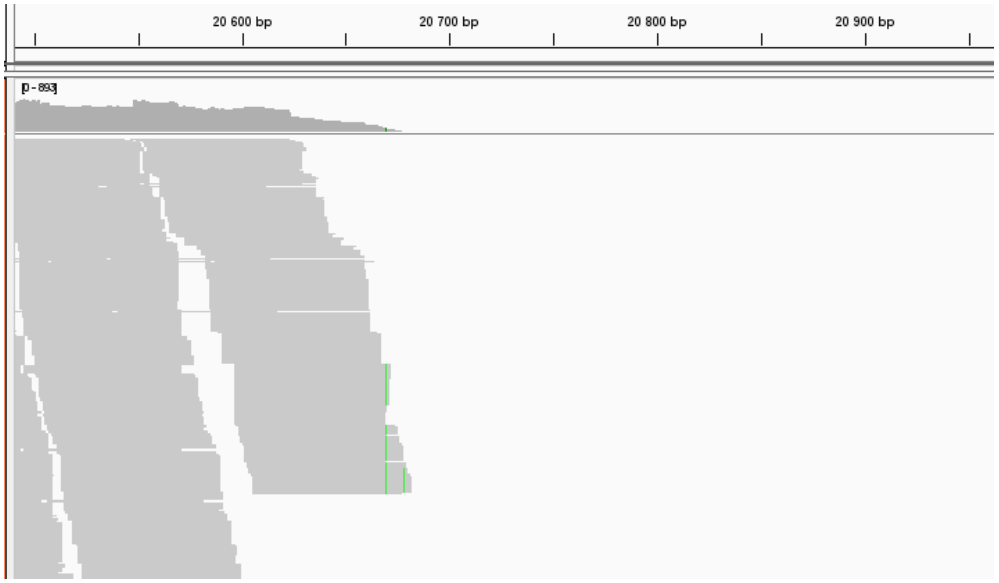
Note 8 - The same patients as in the IME-WH0x series of genomes – without any associated paper – are also found in [12] and [1]. We found that most of these patients were first admitted to Integrated (Hubei Hospital of Integrated Traditional Chinese and Western Medicine), see [22].

Note 9 - The IME-WH0x genomes do not have any metadata, but they are included in the WHO report [3b], matching each IME-WH0x genome with some sequences having enough metadata to identify the patients. Based on the mutations (one identical to Wuhan-Hu-1 and one with the A23425G mutation) we find that IME-WH03 and IME-WH02 have been inverted in [3b].

Note 10 - IVDC-HB-05 and WH19005 are identical with two mutations G20670A G20679A, and IVDC-HB-05 is said to be 32M while WH19005 is said to be an isolate cultured from the same sample as WH19001 who is 49F. Neither are present in the duplication table of the WHO report [3a]. In the available raw data [14] for WH19005 we found that these two mutations are only present at the end of the reads and are followed by a coverage drop. Similar coverage drop and mutations are present in the raw reads of WH19008 [14].

It is thus assumed that it is a primer/amplification problem and a decision was taken to mask these sites, eliminating the discrepancy of finding the same pair of mutations in some but not all of the genomes of different patients.

Once these two mutations are masked, IVDC-HB-05 becomes identical to Wuhan-Hu-1. In its metadata the admission date is said to be on December 27, but the same date is given for IVDC-HB-01 and IVDC-HB-04 so we assume that it is mistake and that this sequence does belong to the same 32M patient as the other ones carrying A23425G.




*two mutations in SAMC703641 (WH19001-5) at 20670 20679*
*present only at the end of the reads and followed by a coverage drop*

Note 11 - there is some uncertainty on the onset of 62M, see [12a].

Note 12 - we decided to trust the GISAID metadata instead of the WHO report for IPBCAMS-WH03 IPBCAMS-WH04 IPBCAMS-WH05. This affects only the consensus genome of 41M1 which would be otherwise identical to WH19053, implying that this genome was in fact from 41M1.

Note 13 - The sequence IPBCAMS-WH-05 (61M) is identical to WH19053 (41M). We assume that the metadata for these sequences is correct, so that the WHO table is wrong to associate IPBCAMS-WH-05 with 41M1. The alternative is that the WHO table is right and that both IPBCAMS-WH-05 and WH19053 are in fact 41M1, but in that case there is a clear mismatch in [1a] both for the onset date and the severity. Note that with the current data 41M1 is an early non-market patient and lineage B, which is also surprising.

Note 14 - Sometimes relatively close to the market, see [8]

Note 15

- WH19002's raw reads from [1b] are different to those analyzed in [14]. The former is a high quality genome identical to Wuhan-Hu-1, while the latter is a very low quality run matching with a very low quality sequence named WH19002 on https://nmdc.cn/resource/ncov/genome/detail/NMDC60013002-05 , all from the same patient 52F.
- For WH19004 there is one sequence on https://nmdc.cn/resource/ncov/genome/detail/NMDC60013002-09 which is identical to IVDC-HB-04 (two mutations C27493T C28253T) and 3 sets of raw reads, from [1b] (one single-end and one paired-end) and [14] (one paired-end).
  The raw reads from [1] have intra-host variants instead of C28253T (A:1,C:3191,G:3,T:421) and C27493T (C:1359,T:1327,G:1) while the raw reads from [14] lack both. Overall the source of these raw reads and the

differences with the published sequences are rather unclear. In every case the assembled sequence becomes identical to Hu-1 up to a few N.
- For WH19008 the nmdc sequence https://nmdc.cn/resource/ncov/genome/detail/NMDC60013002-06 agrees with the raw reads from [14].

Note 16 - See for example https://www.sohu.com/a/368363975_359980 and https://tech.sina.cn/2020-01-20/detail-iihnzahk5292274.d.html

Note 17 - the likely artifactual TT and CC sequences are analyzed in [13]

Note 18 - the first export abroad was a 61F who traveled from Wuhan to Thailand on January 8, so the two Thai sequences with a recorded collection date on Jan 5 are wrong. The onset was on Jan 5, and she was detected by temperature screening at the airport in Thailand. The Thai scientists had to wait for a few days to compare their preliminary partial sequences with the first published SARS-CoV-2 genomes. The sequence is lineage B and the patient is said to have no link to the market.
https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON234

The first lineage B.1 sequence (Zhejiang/HZ103/2020, EPI_ISL_422425, collected on Jan 24) is a 36M from Hangzhou, Zhejiang. It is likely that this refers to the same patient:

> Patient XII, male, 36 years old, now lives in Yuhang District (of Hangzhou city) is a close contact of a suspected case of fever from Wuhan to Hangzhou, onset of fever on January 18, temperature 38℃, with cough and sputum, and now under isolation treatment in a designated medical institution in Hangzhou.
> http://www.hangzhou.gov.cn/art/2020/1/26/art_1228998463_41942427.html

Note 19 - The probability to observe one or less lineage A is 0.00053 = (2/3)^26+26*(1/3)*(2/3)^25, if the 25 market patients and environmental samples had been selected independently and if the proportion of lineage A/B inside the market in December was the same as in the whole of Wuhan. The probability increases to 0.019 when it is assumed that the samples represent 15 independent draws.

Note 20 - https://nextstrain.org/ncov/global/2020-12-20?d=tree&l=clock&p=full

# References


[1] Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease - Shen et al - https://academic.oup.com/cid/article/71/15/713/5780800

[1a] supplementary table 1 https://academic.oup.com/cid/article/71/15/713/5780800#supplementary-data

| | ID | Gender | Age | Clinical lab | Days after onset | ICU | Outcome |
|---|---|---|---|---|---|---|---|
| WH19001 | nCov1 | female | 49 | Negative | 8 | N | Recovered |
| WH19002 | nCov2 | female | 52 | Negative | 8 | Y | Recovered |
| WH19003 | nCov3 | male | 40 | Negative | 15 | Y | Recovered |
| WH19004 | nCov4 | male | 61 | Negative | 12 | Y | Deceased |
| WH19010 | nCov5 | male | 40 | Negative | 8 | N | Recovered |
| WH19012 | nCov6 | male | 56 | Negative | 10 | Y | Recovered |
| WH19016 | nCov7 | female | 53 | Negative | 7 | N | Recovered |
| WH19053 | nCov8 | male | 41 | Negative | 4 | N | Recovered |

[1b] sequencing data
BAM file (mapped reads) https://www.ncbi.nlm.nih.gov/sra/SRR11059941
Raw reads (one paired-end run and one single-end run for each sample)
https://ngdc.cncb.ac.cn/gsa/browse/CRA002475

[1c] biosamples
WH19016 https://www.ncbi.nlm.nih.gov/biosample/SAMN14081563
WH19003 https://www.ncbi.nlm.nih.gov/biosample/SAMN14081559
WH19053 https://www.ncbi.nlm.nih.gov/biosample/SAMN14081564

[2] Pneumonia epidemic situation of new coronavirus infection on January 23
http://www.nhc.gov.cn/yjb/s3578/202001/5d19a4f6d3154b9fae328918ed2e3c8a.shtml

[3] WHO-convened Global Study of Origins of SARS-CoV-2: China Part - p.74-77
https://www.who.int/docs/default-source/coronaviruse/final-joint-report_origins-studies-6-april-201.pdf
[3a] onset curves with 174 patients (laboratory-confirmed and clinically diagnosed) - p.43 and p.45
[3b] sequences, duplication table and metadata for 13 early patients - p.73-77

[4] WHO-convened Global Study of Origins of SARS-CoV-2: China Part, Annexes - p.157-158
https://www.who.int/docs/default-source/coronaviruse/who-convened-global-study-of-origins-of-sars-cov-2-china-part-annexes.pdf
[4a] metadata for 7 clusters of early patients - p.156-160

[5] Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia - Li et al -
https://www.nejm.org/doi/full/10.1056/nejmoa2001316

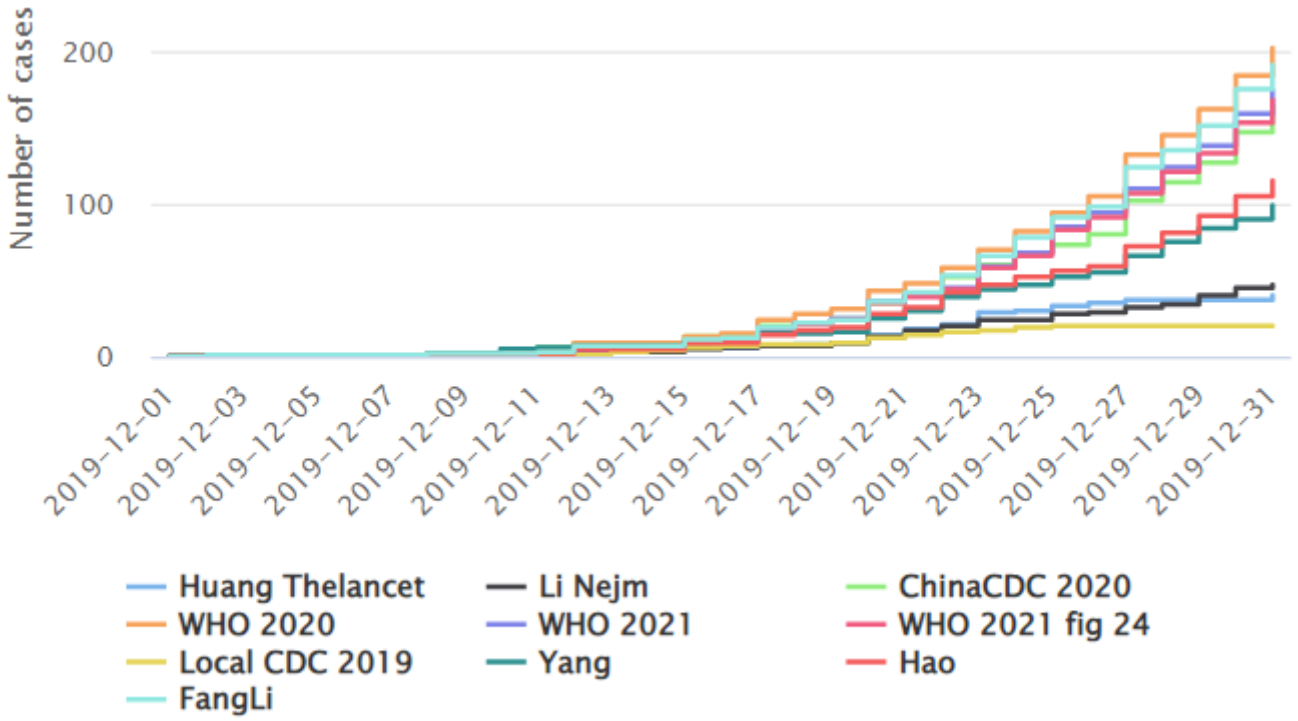
[6] Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China - Huang et al -
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext
[6a] figure 3

[7] Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market - Gao et al
https://www.researchsquare.com/article/rs-1370392/

[8] The Huanan market was the epicenter of SARS-CoV-2 emergence - Worobey et al
https://zenodo.org/record/6299116

[9] https://github.com/jbloom/SARS-CoV-2_Shen_et_al

[10] https://flodebarre.github.io/covid_firstCases/




*Supplementary Figure 1 - Onset curves of the confirmed cases with onset in December gathered from multiple sources*

[11] SARS-CoV-2 Samples from Same Early COVID-19 Patients Were Sequenced Repeatedly with Errors Distorting Phylogenetic Trees - Fuyutao
https://virological.org/t/sars-cov-2-samples-from-same-early-covid-19-patients-were-sequenced-repeatedly-with-errors-distorting-phylogenetic-trees/434

[12] A pneumonia outbreak associated with a new coronavirus of probable bat origin - Zhou et al
https://www.nature.com/articles/s41586-020-2012-7
[12a] https://www.nature.com/articles/s41586-020-2012-7/tables/1

[13] SARS-CoV-2 emergence very likely resulted from at least two zoonotic events - Pekar et al
https://zenodo.org/record/6291628
[13a] in-silico bat coronavirus recombinant
https://github.com/sars-cov-2-origins/multi-introduction/blob/main/sarbecovirus_recombination/recCA.masked.fasta

[14] analysis of the raw reads recently published for WH01 WH02 WH03 WH04 WH19001-WH19005 WH19004 WH19002 WH19008 YS8011 https://github.com/niemasd/PRJCA008874-Analysis

[15] During our writing of this manuscript we were informed of recent a preprint – SARS-CoV-2 minor variant genomes at the start of the pandemic contained markers of VoCs - Dong et al – which is also reanalyzing the raw reads of [1] https://www.biorxiv.org/content/10.1101/2022.06.10.495670v1
[15a] https://www.biorxiv.org/content/10.1101/2022.06.10.495670v1.supplementary-material

[16] Leaked document on the enviromnental samples from the Huanan market
https://www.epochtimes.com/gb/20/5/31/n12150755.htm implicitely authentified by [7]

[17] 65M is also the patient near full-length sequenced by Vision medicals on December 27, see
https://www.documentcloud.org/documents/21698235-vision_medicals_dec19_active and
https://www.washingtonpost.com/opinions/interactive/2022/china-researcher-covid-19-coverup/

We don’t know if this patient was sequenced a second time at IPBCAMS as it is possible that IPBCAMS-WH-01 is a renaming of the Vision medicals sequence.

|18] Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses - Zhou et al
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8188299/

[19] Bat coronaviruses related to SARS-CoV-2 and infectious for human cells - Temmam et al
https://www.nature.com/articles/s41586-022-04532-4

[20] Epidemiological, clinical and viral gene evolution characteristics of important emerging infectious diseases (SFTS and COVID-19) - Liu Jiluo, Chinese thesis
https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CDFD&dbname=CDFDLAST2022&filename=1020053330.nh&uniplatform=NZKPT&v=xaJppHqg1VJPX2N_WqQKNzkEvy5-sMfCznJXn2b9EMKQAtF-90vAXEmk9Atza8A2
translated at
https://docs.google.com/document/d/1FRvFae8oAgvSMmWT6g57SVsD06jGz9B61UApShrhoM4/edit?usp=sharing

[21] A tour into the Wuhan South China seafood market
http://babarlelephant.free-hoster.net/visiting-the-wuhan-seafood-market/

[22] Doctor stories, Zhang Dingyu head of Jinyintan
https://www.weibo.com/ttarticle/p/show?id=2309404483446197059980 and Zhang Jixian head of respiratory department of Integrated https://www.youtube.com/watch?v=jUHiQRJiOSQ

[23] Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: a retrospective observational study - Li et al
https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30981-6/fulltext#seccestitle150

[24] Estimates of serial interval for COVID-19: A systematic review and meta-analysis - Rai et al
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7448781/