

Day 2 slides shown at debate

The difficulty of arguing against the lab leak theory is that there are so many different lab leak objections for every topic.

I need to put in extensive work to debunk every objection.

As the goalposts keep shifting, new objections crop up.

For instance, here's what we went through in the first debate:

Stages of Market Denial:

- ~~• The first case wasn't at the market~~
- ~~• The first two cases were visitors to the market~~
- ~~• There are actually lots of early cases, besides the market~~
- ~~• The data is biased~~
- ~~• Cases won't be centered on the market anyways~~
- ~~• It's actually just the mahjong room~~
- ~~• The market is actually a super likely place for a cluster to start~~
- It's the ventilation
- Sure it's centered on the market, but that's because it's the Wuhan CDC
- Even if the raccoon dogs were sick, it's still a lab leak

← (I think maybe we made it to here?)

Inside the market, I think it went something like this

Stages of Denial, inside market:

- ~~• There were no animals at the market (after WHO report)~~
- ~~• There were raccoon dogs in 2014, but what about 2019? (after eddie holmes photo)~~
- ~~• There were raccoon dogs in 2019, but what about December 2019? (after xiao xiao paper)~~
- ~~• There was raccoon dog DNA, but the human DNA was removed (Alina Chan)~~
- ~~• There was raccoon dog DNA, but those samples were negative (Steven Quay)~~
- ~~• There was raccoon dog DNA in positive samples, but it's not correlated to covid RNA~~
- There was raccoon dog DNA in positive samples, but not enough covid RNA reads ← we are here, maybe?
- Even if the raccoon dogs were sick, it's still a lab leak

We're going to have a similar conversation today

Stages of Denial for two spillovers:

- Lineage B may have started at the market but Lineage A came first
- Actually, proCov2 came even earlier than Lineage A
- The Lineage A sample at the market is mutated/fake/unimportant/something
- There are actually intermediate genomes
- Even if there were 2 lineages, the 2 lineages came from the lab.

Let's start with an easy one. Could the lineage A sample at the market have mutated from B to A?

A is ancestral to B, and the market is dominated by B

Huanan had only 1 lineage A sample but:

- it was an environmental sample, found on a glove
- had additional mutations (G26262T, C6145T, and possibly T24979C)
- the lineage A genome was recovered only after passaging the A20 sample in culture, while direct sequencing of the A20 sample yielded only 22 SARS2 reads and no reads covering positions 8782 or 28144

Thus it is possible the lineage A genome in A20 was not present originally but was introduced during viral passaging of the sample in culture

Given the number of samples, and A being a popular strain early in the pandemic, it's very reasonable that there will be a few people not infected in the market coming in and leaving a trace

It could even have been a health professional (hence the glove) contaminating it from outside the market

← Yuri said a few slides before this that he thinks that each reversion mutation has a frequency of 3%. Here he's saying that 2 such reversions happened in one market sample, during culturing. The odds of 2 reversions would be 1 in 1,100.

But the odds of this happening are actually much lower – 1 in 1,100 would be the odds of 2 random reversions.

Here they're saying it's 2 specific reversions that happen to match the exact 2 reversions that are important.

There are 30,000 possible nucleotides for each mutation.

The odds of seeing the right two are more like $(1 \text{ in } 30,000)^2$, or **1 in 900 million**.

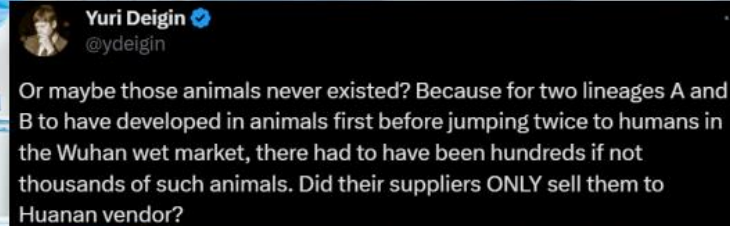
This is one of the few claims in this debate that we can dismiss outright as statistically impossible.

Next up, could there have been two spillovers at the lab?

Two Jumps in this Pattern are Likelier in a Lab Leak

Even if there were two separate jumps, since they occurred in the same location within a short time frame, they don't strengthen the zoonosis case:

Two spillovers can well happen in a lab; One of the three SARS1 lab leaks had two jumps from the same lab.

A screenshot of a tweet from Yuri Deigin (@ydeigin). The tweet text reads: "Or maybe those animals never existed? Because for two lineages A and B to have developed in animals first before jumping twice to humans in the Wuhan wet market, there had to have been hundreds if not thousands of such animals. Did their suppliers ONLY sell them to Huanan vendor?"

Importantly, two spillovers from wildlife imply many infected animals in contact with humans, which would make it much more unlikely that Wuhan will be the only outbreak.

I claim the odds of one infected person at the Wuhan lab making it across town to the market to cause the first cluster there are about 1 in 10,000. We can, of course, argue about the exact number.

If Rootclaim wants to say that lineage A and lineage B both leaked from the lab, and then were found centered on the market, I get to square the odds of that, to **1 in 100 million against that happening**.

I don't know why they are presenting this option, it's an improbable thing that hurts their case.

Now, the next points of discussion are much more interesting, and these will be good conversations to have:

- Which came first, lineage A or lineage B?
- What's the deal with proCov2? Could that have come before lineage A?
- Were there intermediate genomes, or not?

Let's start with proCoV2, that's the easiest

Summary of the claim:

Lineage A is 2 mutations closer to known bat viruses than Lineage B, so maybe it's the ancestor

Some Lineage A genomes also have either mutation C18060T or C29095T, both of which are even closer to known bat viruses.

First off, Jesse Bloom didn't actually find 18060T genomes

proCov2 is assumed to have 18060T, but these genomes were only sequenced from 21,570 to 29,550, so we don't actually know what they have at position 18060.

When they say "no substitutions from proCov2", that's meaningless. Likely, most or all of these have 18060C.

A is ancestral to B, and the market is dominated by B

Table 1.

Samples for which the SARS-CoV-2 sequence could be called at $\geq 90\%$ of sites between 21,570 and 29,550, and the substitutions in this region relative to the putative SARS-CoV-2 progenitor proCov2 inferred by Kumar et al. (2021).

Sample	Fraction sites called (21,570-29,550)	Patient group	Substitutions relative to proCov2
A4	0.9266	Early outpatient	<u>None</u>
C1	0.9396	Early outpatient	G22081A (A=924, C=4, G=9), C28144T (C=6, T=1185), T29483G (C=1, G=45, T=1)
C2	0.9397	Early outpatient	C29095T (C=1, G=1, T=751)
C9	0.9005	Early outpatient	C28144T (C=3, T=823), G28514T (G=1, T=36)
D9	0.9051	Early outpatient	C28144T (C=4, T=1653)
D12	0.9400	Early outpatient	C28144T (C=8, T=2400)
E1	0.9223	Early outpatient	C28144T (T=125)
E5	0.9227	Early outpatient	<u>C24034T (A=5, C=3, T=74), T26729C (C=12), G28077C (C=142, G=4)</u>
E11	0.9321	Early outpatient	C25460T (C=2, T=246), C28144T (C=1, T=412)
F11	0.9054	Early outpatient	T25304A (A=9, T=1), C28144T (C=6, G=1, T=1328)
G1	0.9396	Early outpatient	<u>None</u>
G11	0.9112	Early outpatient	<u>None</u>
H9	0.9381	Early outpatient	C28144T (C=2, T=1254)
R11	0.9422	Hospital patient (Feb)	C21707T (T=401), C28144T (A=1, C=18, T=4265)

Numbers in parentheses after each substitution give the deep sequencing reads with each nucleotide identity.

Sample C2 is missing C28144T, meaning it is lineage A. There are a total of 4 mutations in the 5 lineage A samples, making a reversion possible but unlikely (3% x 4 times)

Note: Only mutations above 21,570 are shown

But there were some genomes with 18060T, even if Bloom didn't find any.

These genomes have either 18060T or 29095T.

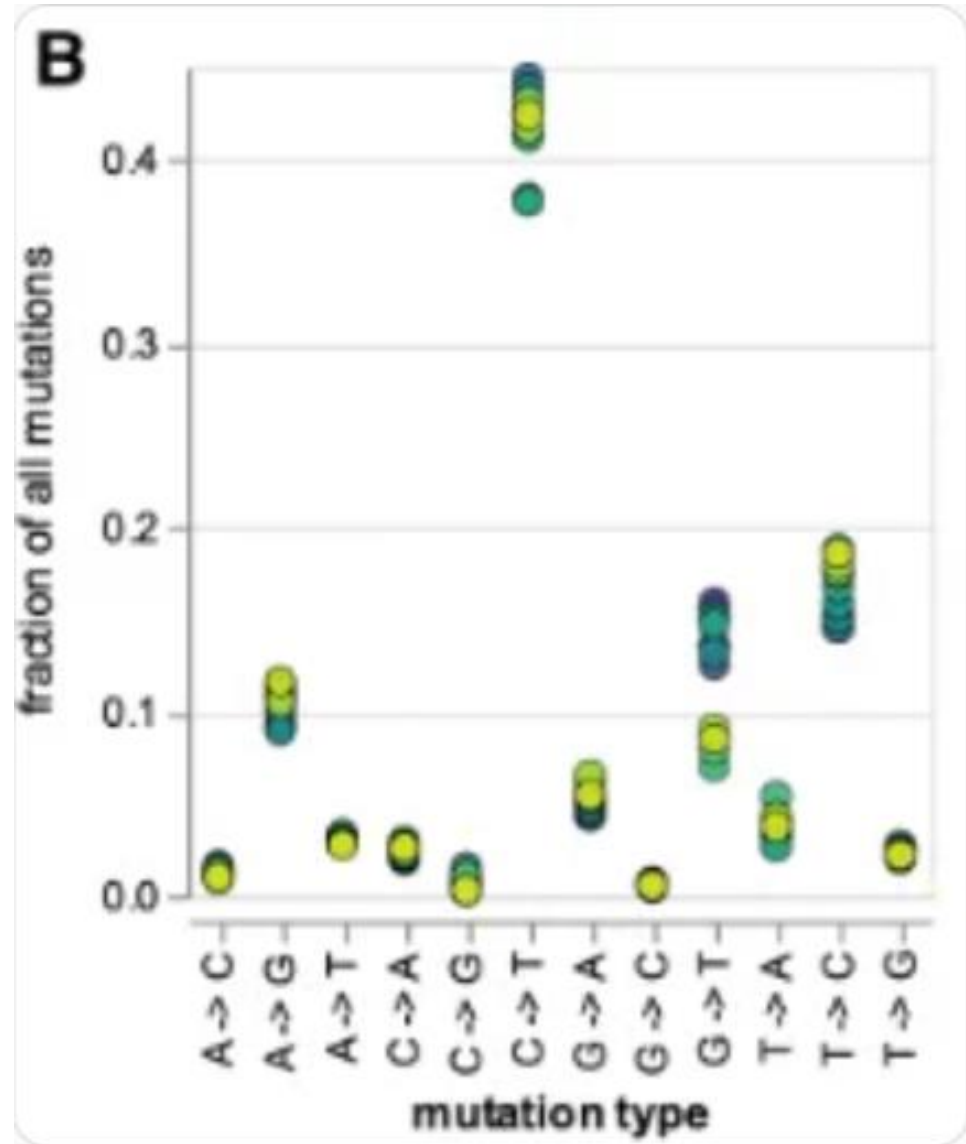
So, only one of them could be the progenitor virus and the other one must be a reversion.

That alone should prove that reversions are possible.

But, what are the odds of one reversion? What are the odds of two reversions?

Here's one clue – the odds of each possible mutation in SARS2.

C -> T is the most common mutation, and all these reversions are C -> T.



We can also calculate the odds of 2 reversions

There are 1,200 nucleotide differences between SARS-CoV-2 and RATG-13.

The odds one mutation would affect one of these: $1,200/30,000 = 4\%$

There are at least 41 observed mutations off the lineage A root. (depending on how you count the double mutations, if you counted those as 1, it would only be 28 mutations).

The odds of zero reversions = $0.96^{41} = 18\%$

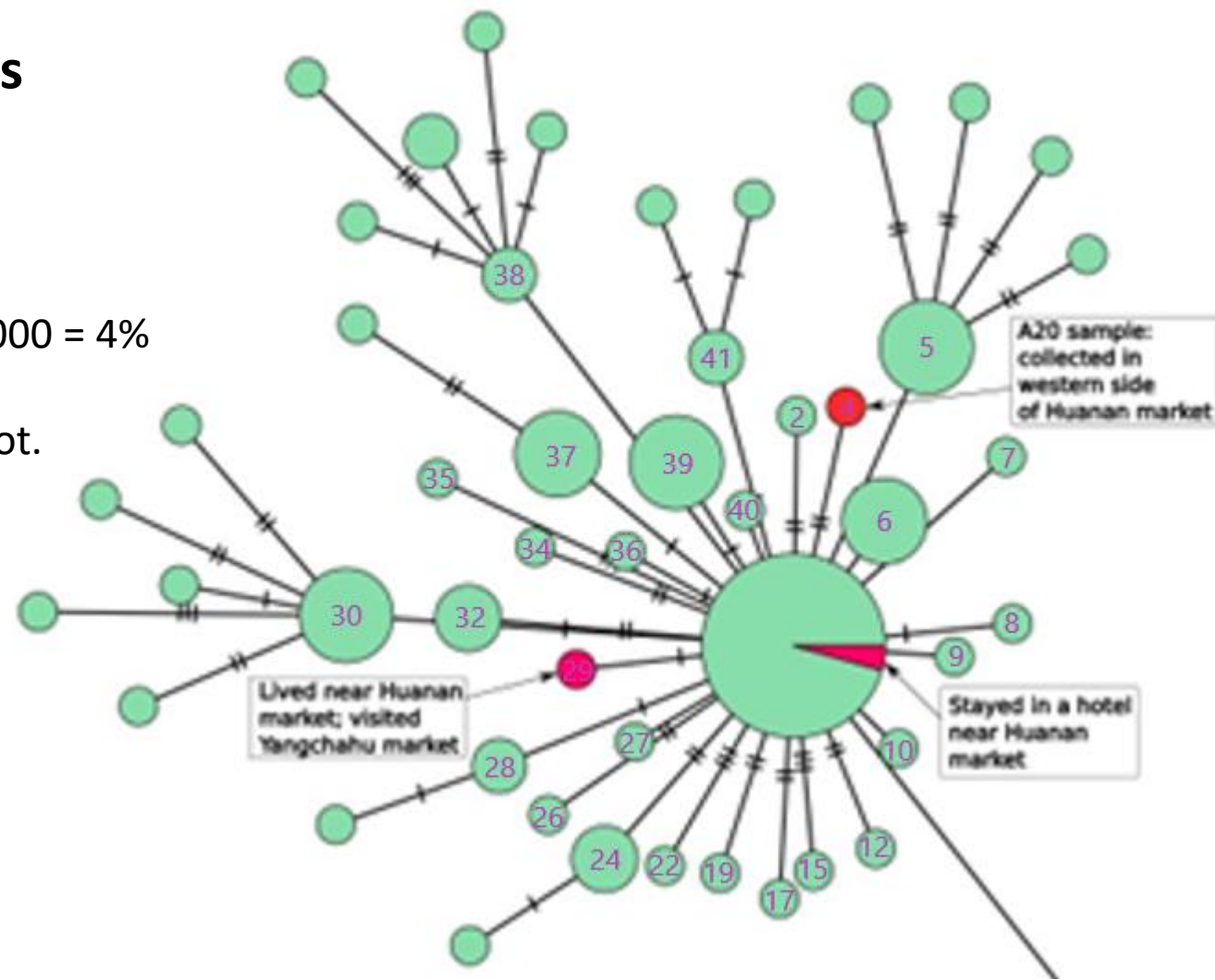
The odds of one reversion = 82%

The odds of two reversions = 67%

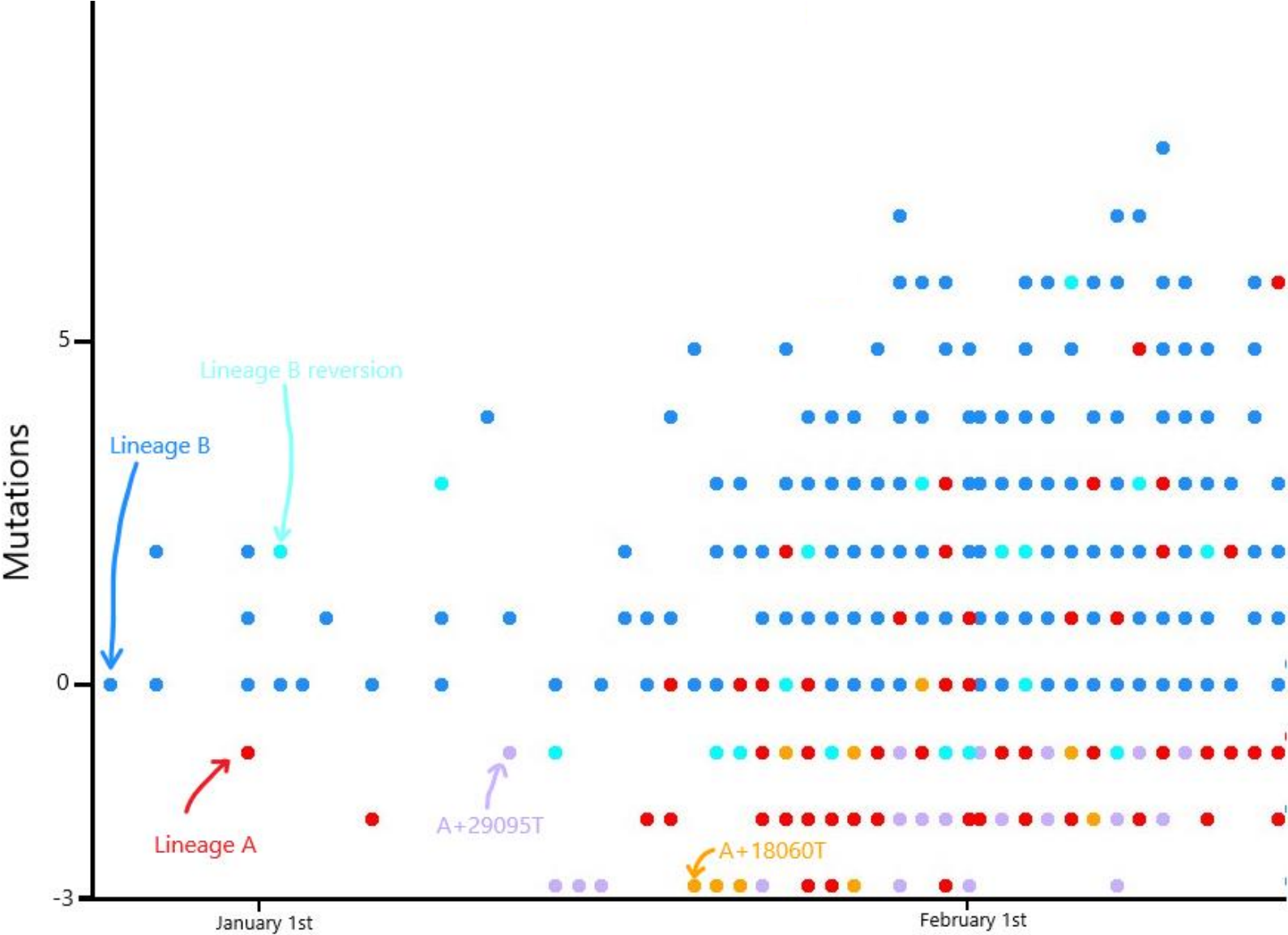
Strictly speaking, that's just the odds of a mutation at one of those sites, not a reversion. But C->T is the most common mutation, so the odds of a C->T reversion won't be much lower than that. Yuri suggested it was 3% odds of a reversion.

You can change the math to use 3%, then it's one reversion = 71%, two reversions = 50%

Any way you look at it, 2 reversions is not something unlikely.

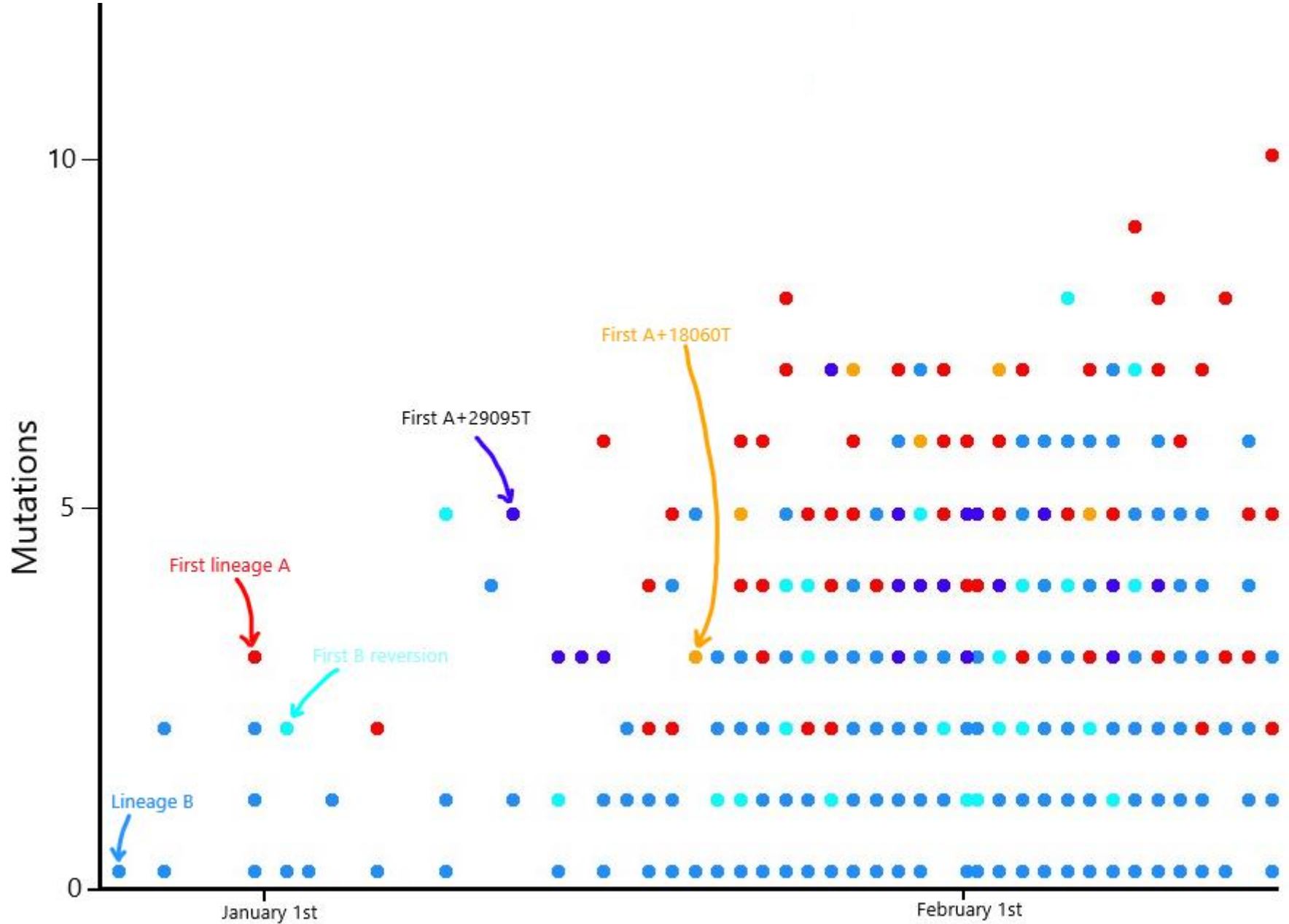


I recreated the graph and drew reversions into it. Lineage B has 2 or 3 reversions before the Lineage A reversions show up. Here's a plot of mutations over time. The Y axis is mutations relative to the outgroup (I used RATG13). If lineage B is set at 0 mutations, then lineage A is -2. Lineage B reversions show up as -1, Lineage A reversions as -3.



How we measure mutations is arbitrary, though.

We can also graph mutations relative to Lineage B, instead of the outgroup, and now all the reversions are positive numbers.



And then, I read Pekar 2022 to see how he thought about this, and he also catalogued lots and lots of reversions off of lineage A and lineage B. This is what he wrote:

“Most reversions were C-to-T mutations (19 of 23, 82.6%), matching the mutational bias of SARS-CoV-2 ([15–17](#)). Genomes with C-to-T reversions can be found within lineage A, including C18060T (lineage A.1; for example, WA1) and C29095T (for example, 20SF012), as well as C24023T, C25000T, C4276T, and C22747T in mid-late January and February 2020. Hence, triple revertant genomes, such as WA1 and 20SF012, are neither unique nor rare.”

Pekar also found one genome from Malaysia which had another reversion even closer to the bat virus outgroup. But that doesn't mean it's the origin of Covid, because it's one random genome:

“We also identified a lineage A genome (Malaysia/MKAK-CL-2020-6430/2020), sampled on 4 February 2020 from a Malaysian citizen traveling from Wuhan whose only four mutations from Hu-1 are all reversions (lineage A.1+T6025C) ([Fig. 1](#)). Therefore, no highly revertant haplotype can automatically be assumed to represent the MRCA of SARS-CoV-2, especially when these reversions are most often the result of C-to-T mutations. We continue to observe these reversion patterns throughout the pandemic, including in the emergence of World Health Organization (WHO)–named variants (figs. S15 and S16).”

Instead of just assuming that whichever genome is closest to the bat virus outgroup is the origin of covid, Pekar built a Bayesian model to estimate the odds of each of these lineages being the origin of covid.

His model selects against 18060T and 29095T with extremely high odds.

Haplotype	Mutations from Hu-1 reference	Representative genome	Phylodynamic analysis	
			Unconstrained (%)	No market (%)
B (C/T)	N/A	Hu-1	80.85 [†]	62.96 [†]
A (T/C)	C8782T+T28144C	WH04	1.68 ^{**}	5.73 ^{**}
C/C	T28144C	N/A	10.32 [*]	23.02
T/T	C8782T	N/A	0.92 ^{**}	1.68 ^{**}
A+C29095T (T/C)	C8782T+T28144C+C29095T	20SF012	<0.01 ^{***}	<0.01 ^{***}
A.1 (T/C)	C8782T+T28144C+C18060T	WA1	<0.01 ^{***}	<0.01 ^{***}

For an intuitive explanation – if 18060T or 29095T were the base lineages, there should be a lot more genomes like that, and those should show up a lot earlier.

Also, C → T is more common than T → C, so you'd need to have the less common mutation happen right away without leaving much trace. That's unlikely.

[Pekar's model](#) shows that the most common type of evolutionary tree is a polytomy, which is the graph on the left:

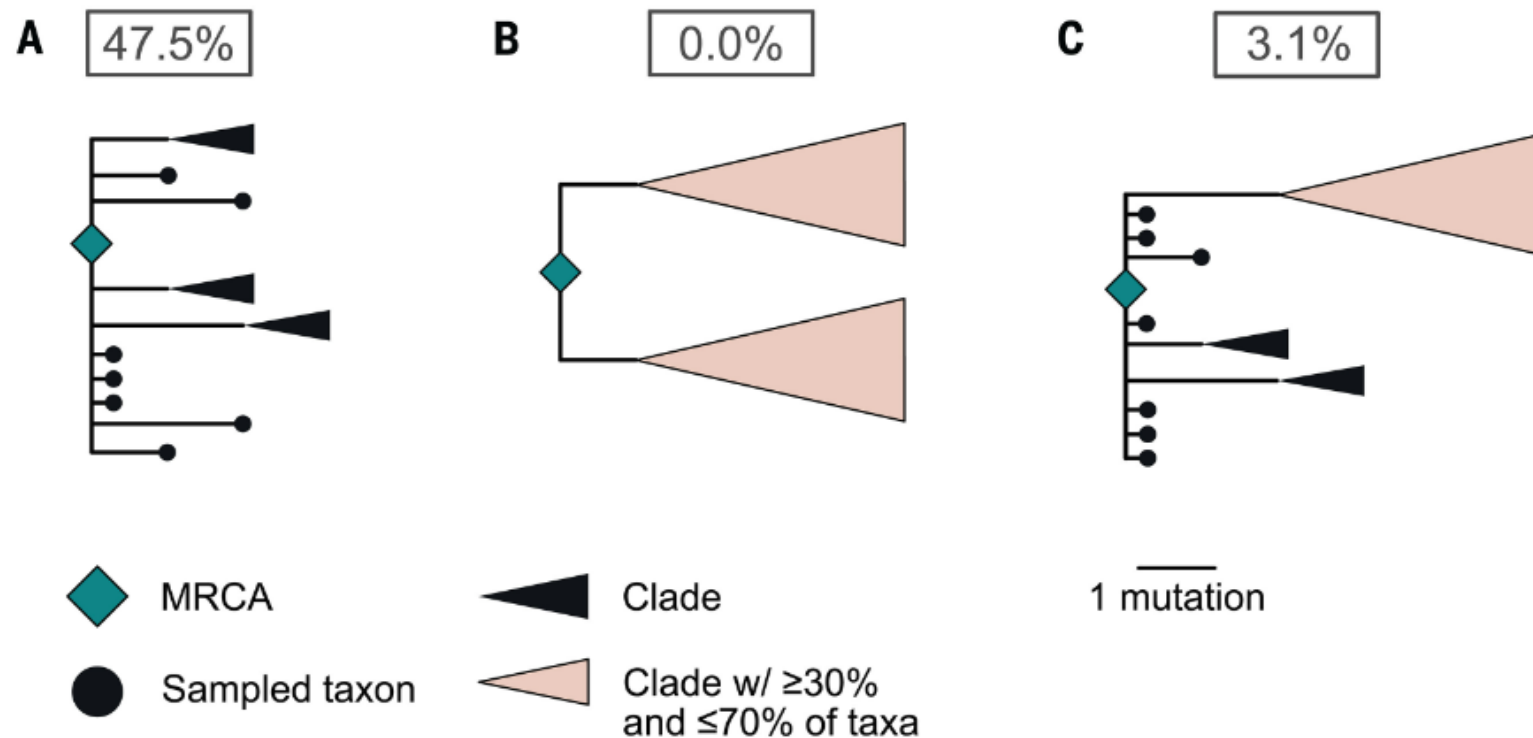
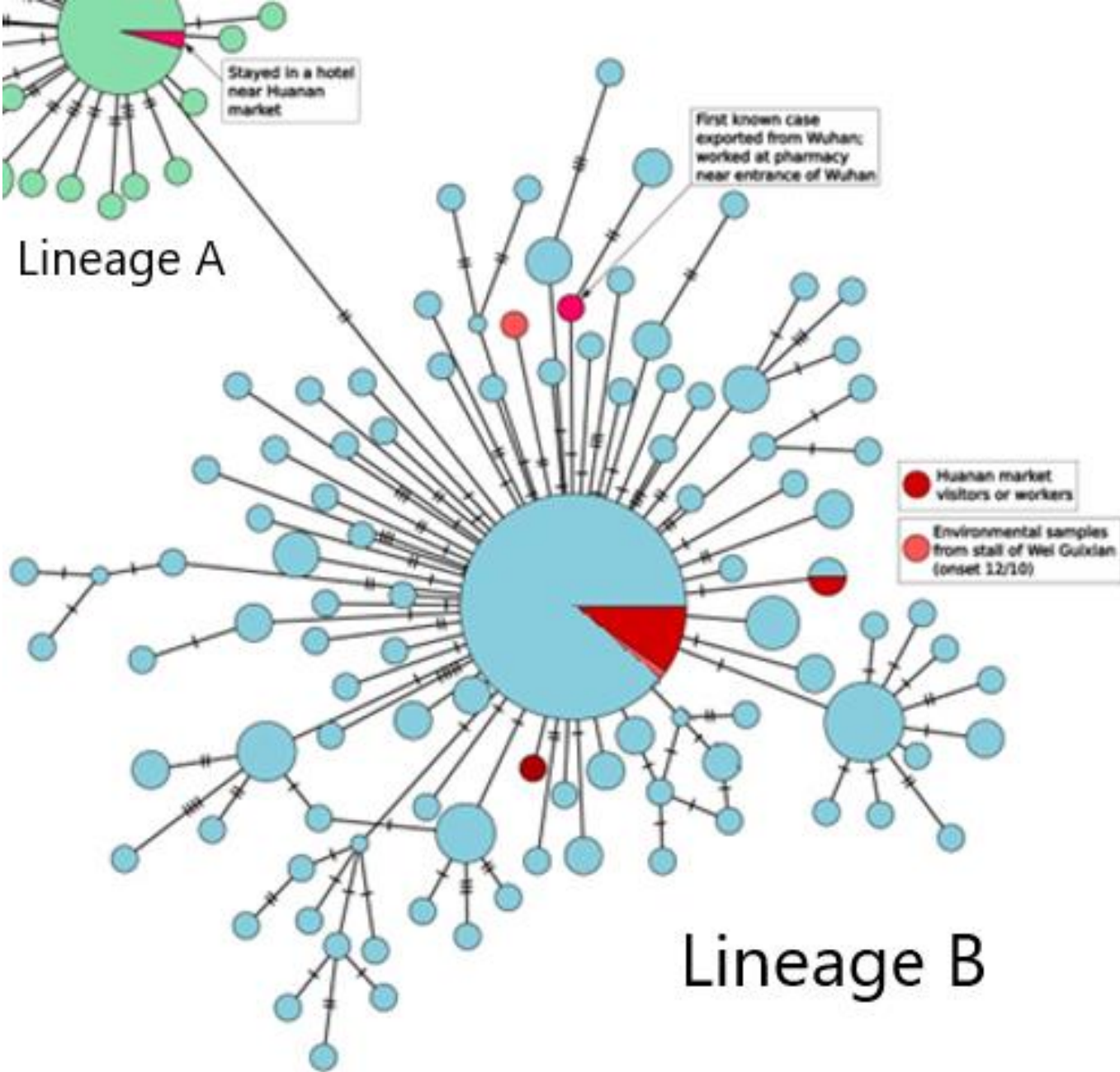


Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations.

The middle graph is what rules out a C/C ancestor. And the final graph is what makes 2 lineages unlikely.

You can also draw a polytomy like this, with the base lineage in the center, and all the descendant lineages sticking out from it.

(Lineage A and B are polytomies)

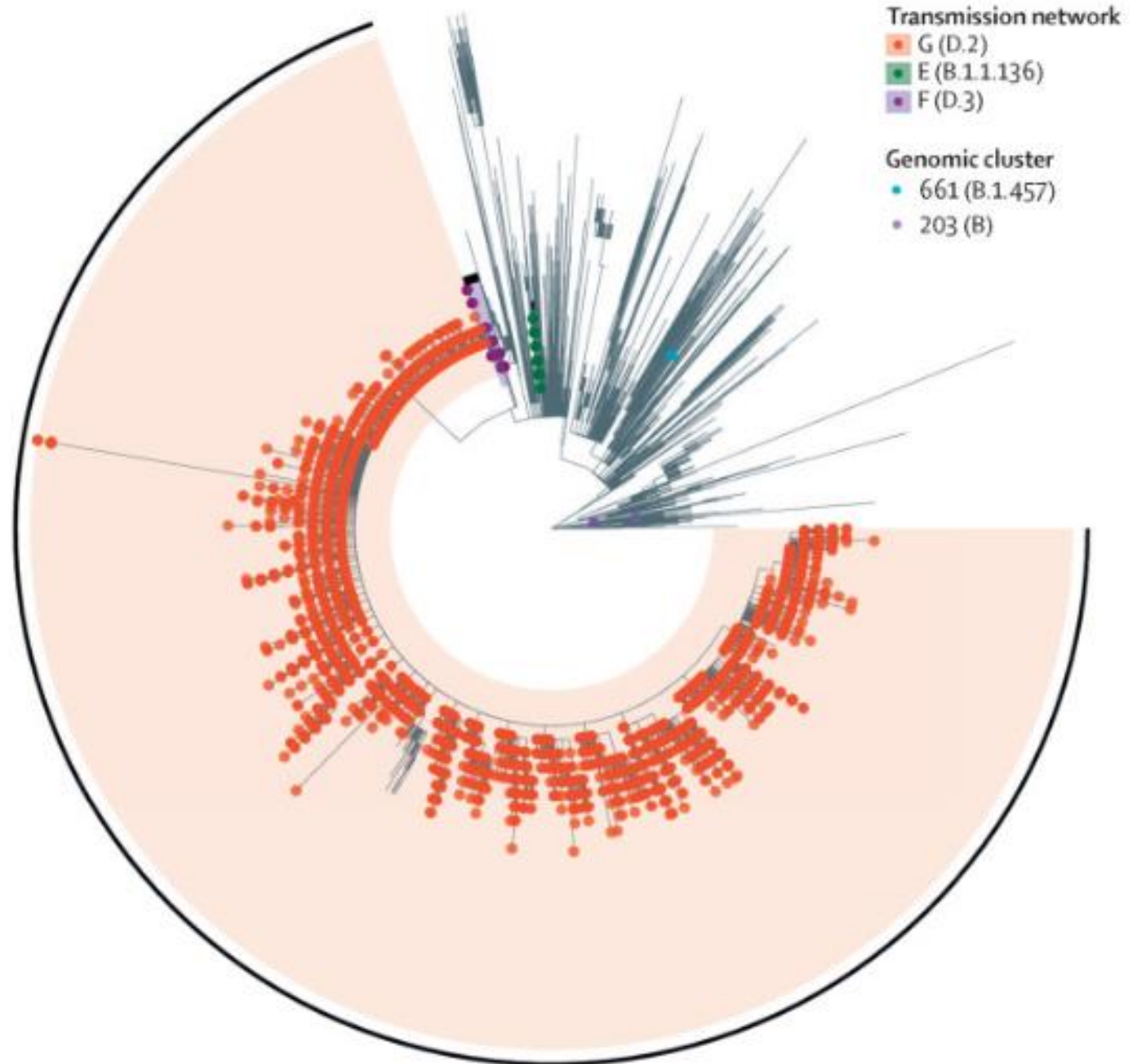


Another example is Victoria, Australia, mid-2020.

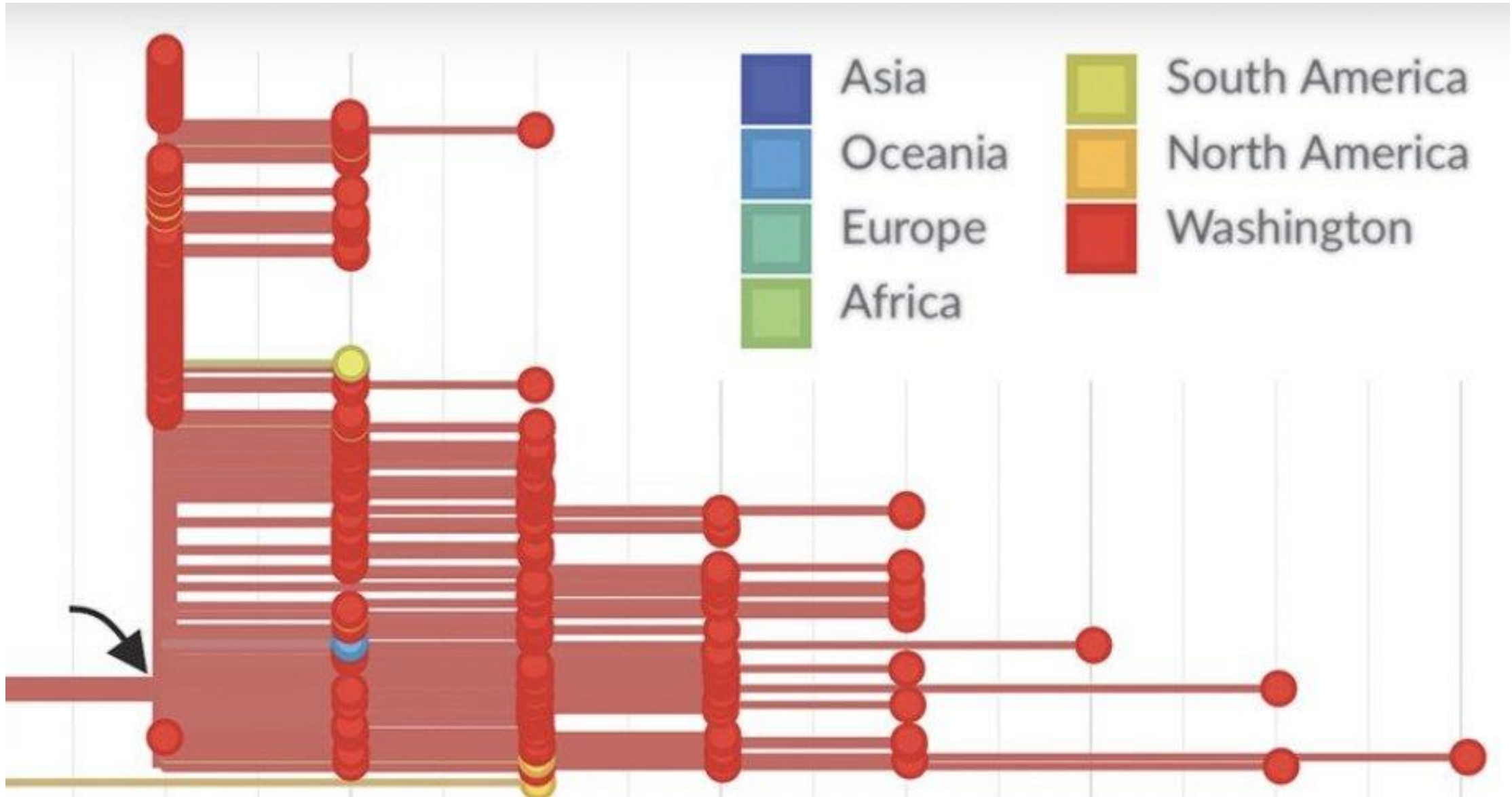
A single escape from a quarantine hotel [caused a single polytomy](#).

(marked as “G” in this diagram).

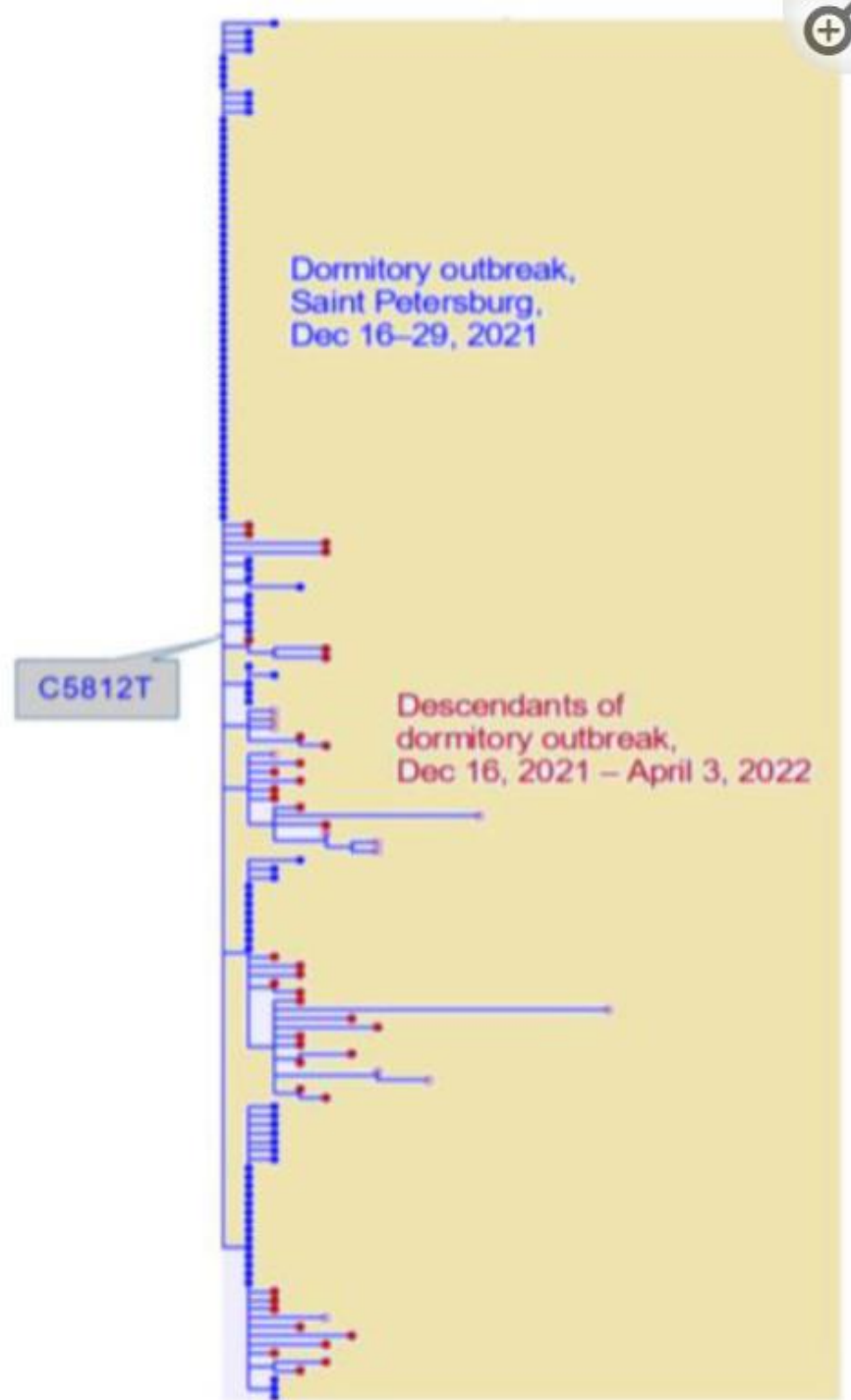
A



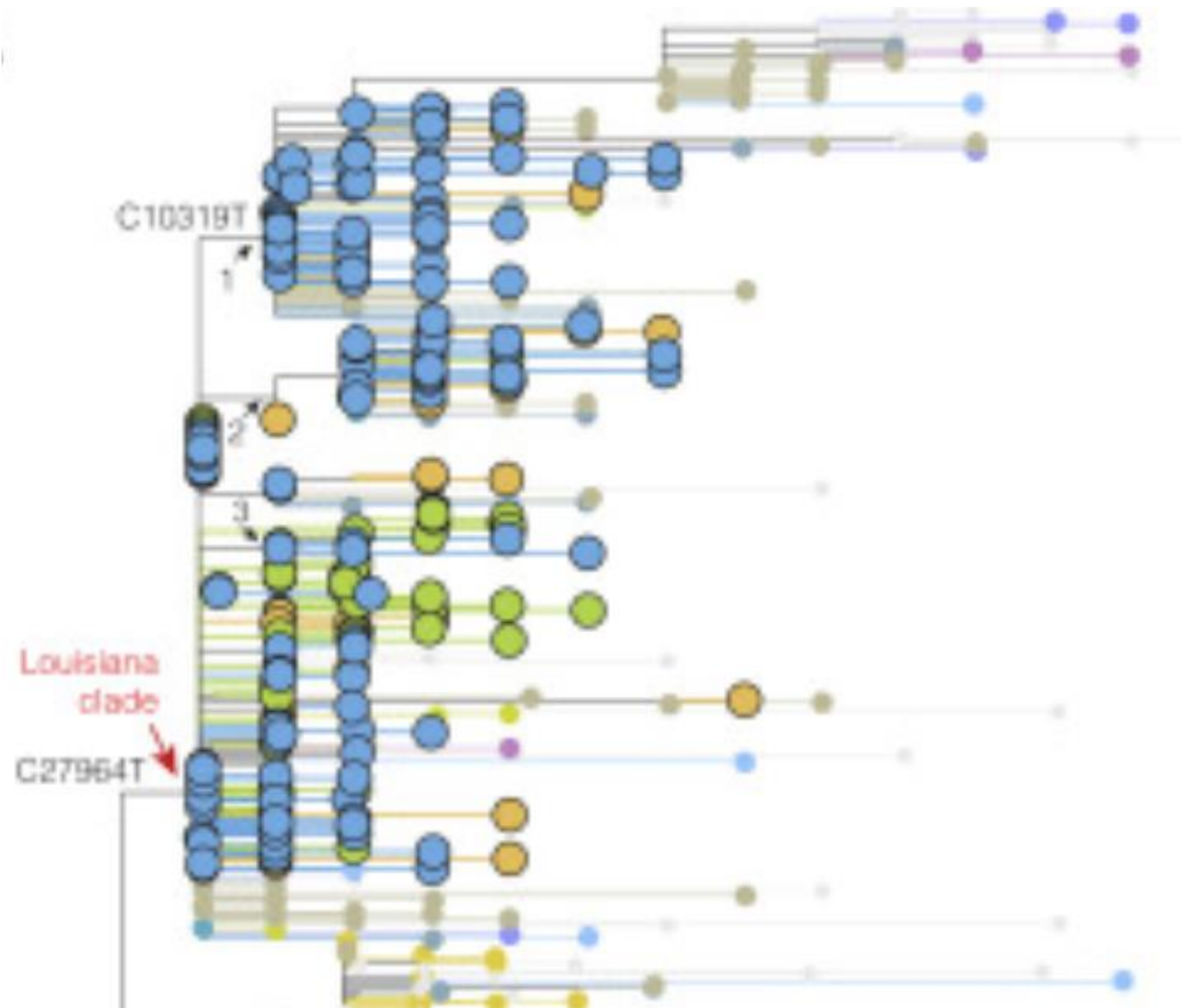
Another is the [Seattle](#) covid outbreak, in March 2020



Another is a [single introduction of covid in Russia](#):



Another example is the [2020 Louisiana mardi gras](#)



Another is the diamond princess cruise ship.

That's a single introduction, and a single polytomy.

And it's the same thing in many other places:

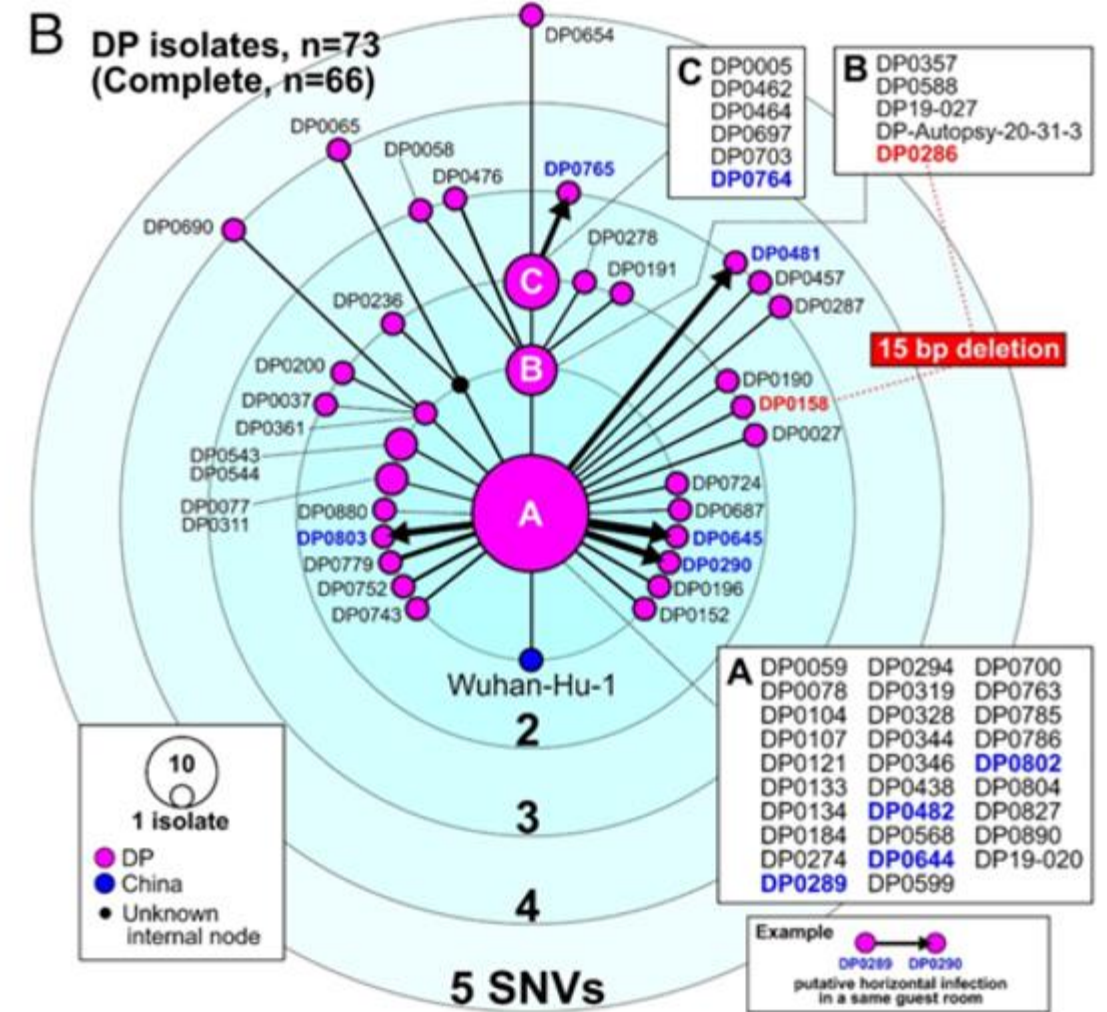
[Lombardy, Italy 2020](#)

Delta wave in [New Zealand](#)

Single introductions to [California nursing homes](#)

Several outbreaks [in Minnesota](#)

Diamond Princess viral genomes (single introduction)



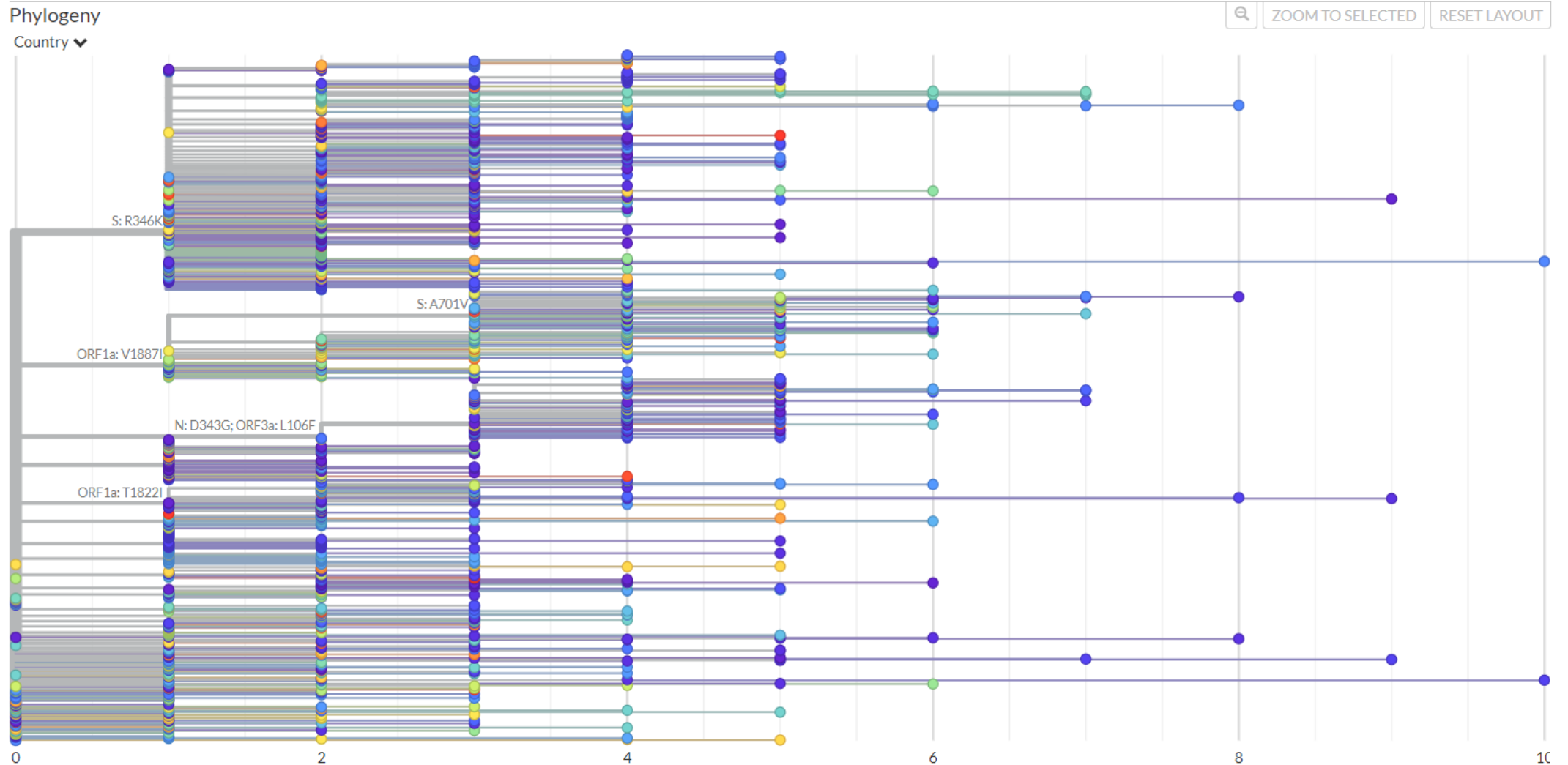
The emergence of Omicron also appears to be a single polytomy. [Data here](#), [discussion here](#).
(of course, it's less clear how omicron emerged, whether that was a single immunocompromised patient or what)

Subsampled phylogenetic analysis of within-clade 21K (Omicron) diversity



Built with [corneliusroemer/ncov-simplest](#). Maintained by [Cornelius Roemer](#) and [Richard Neher](#). Enabled by data from [GISAID](#).

Showing 1458 of 1458 genomes.



Are there any real world counter-examples?

Alex Washburne cites [Austria in spring 2020](#) as an example of multiple polytomies. But this actually confirms the point, as most of these were separate introductions.

Tyrol-1 was introduced from one person from North America, and it's a single polytomy.

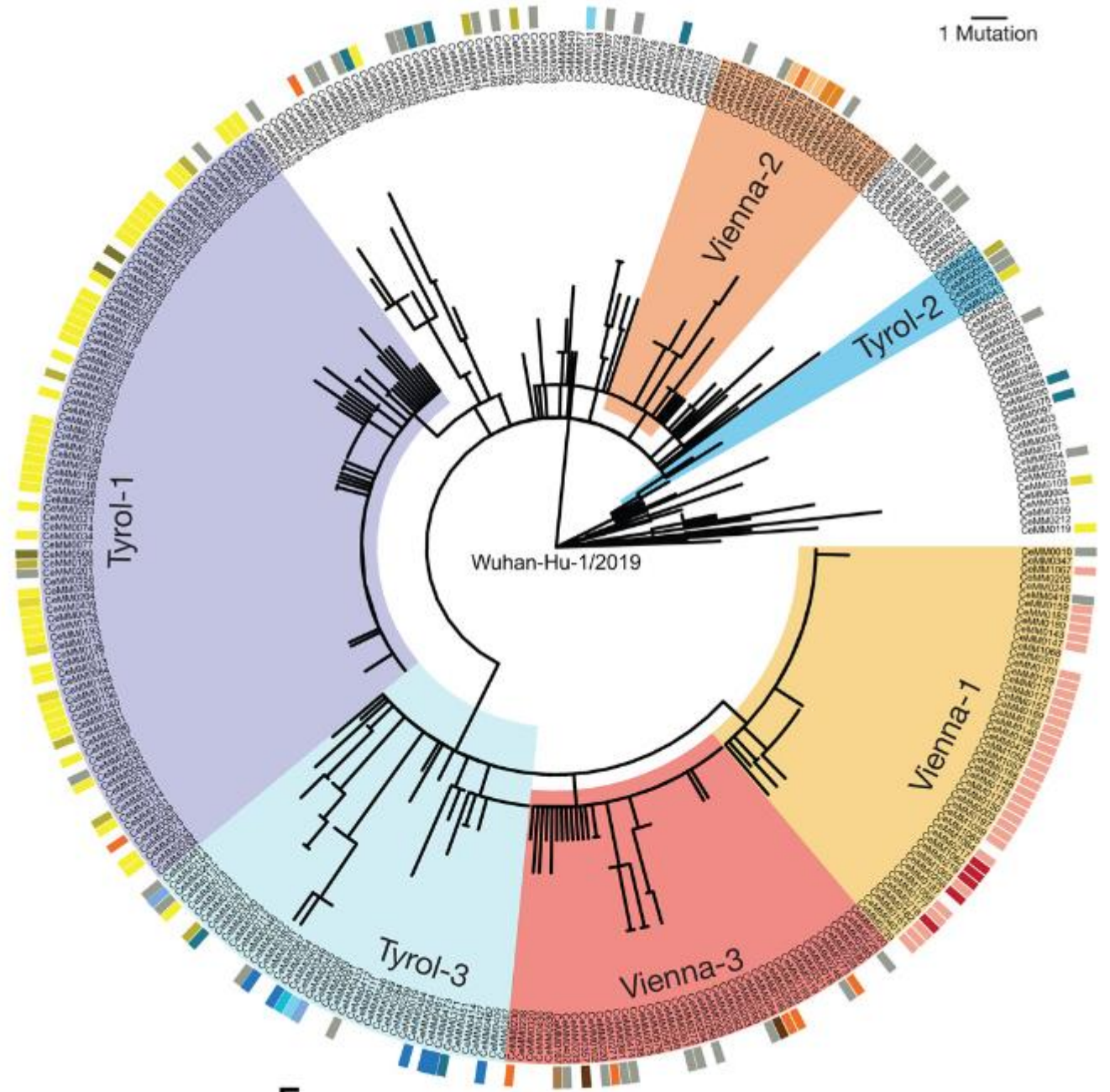
Vienna-2 and Tyrol-2 are separate introductions.

The bottom 3 clusters here look possibly connected (by one mutation, not two). But those are likely separate introductions, as well:

Vienna-1 is from an index patient from Italy.

Vienna-3 is connected to cluster OG, an independent travel-associated cluster.

Tyrol-3 is connected to cluster D, another independent travel-associated cluster.



So, single polytomies have continued showing up through out the pandemic.

Reversions have also continued happening, throughout the pandemic

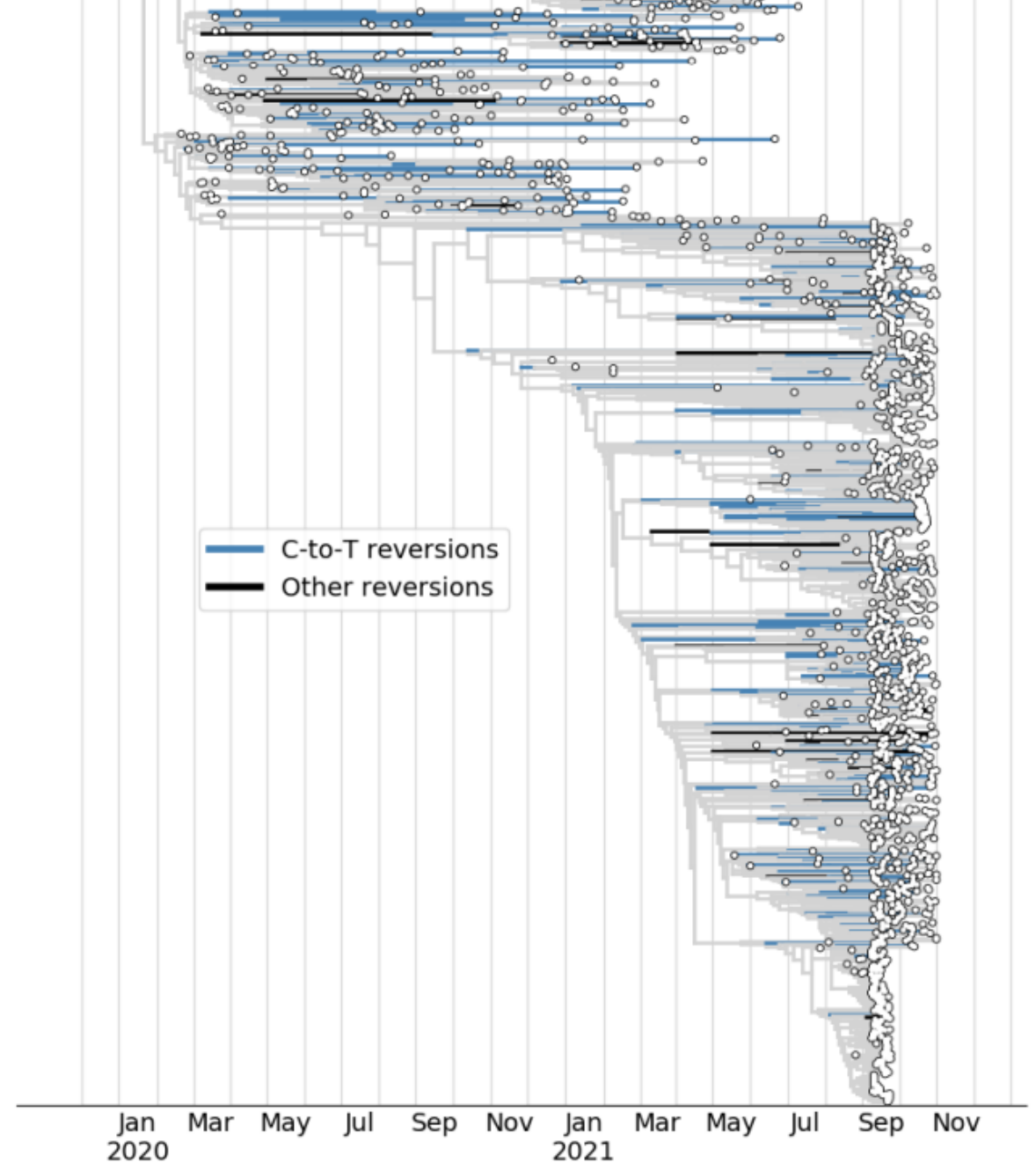


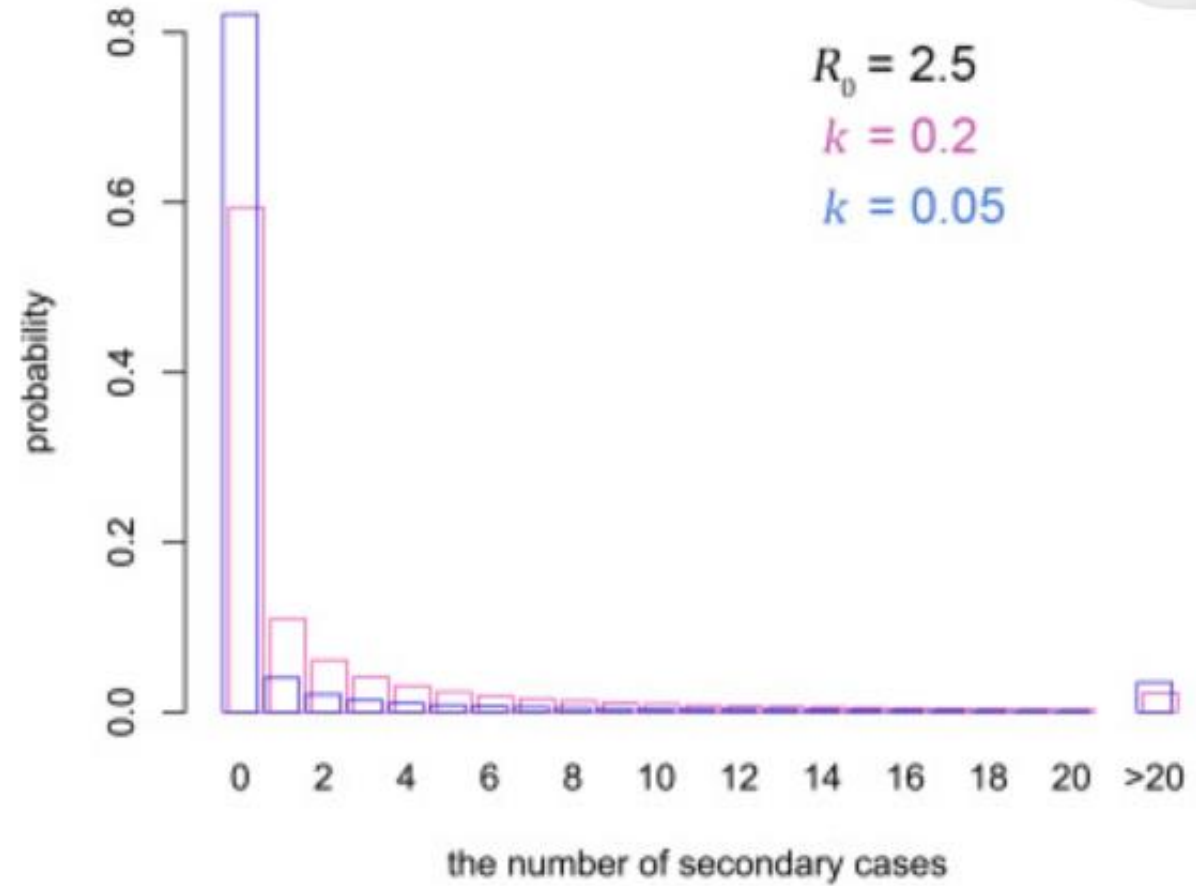
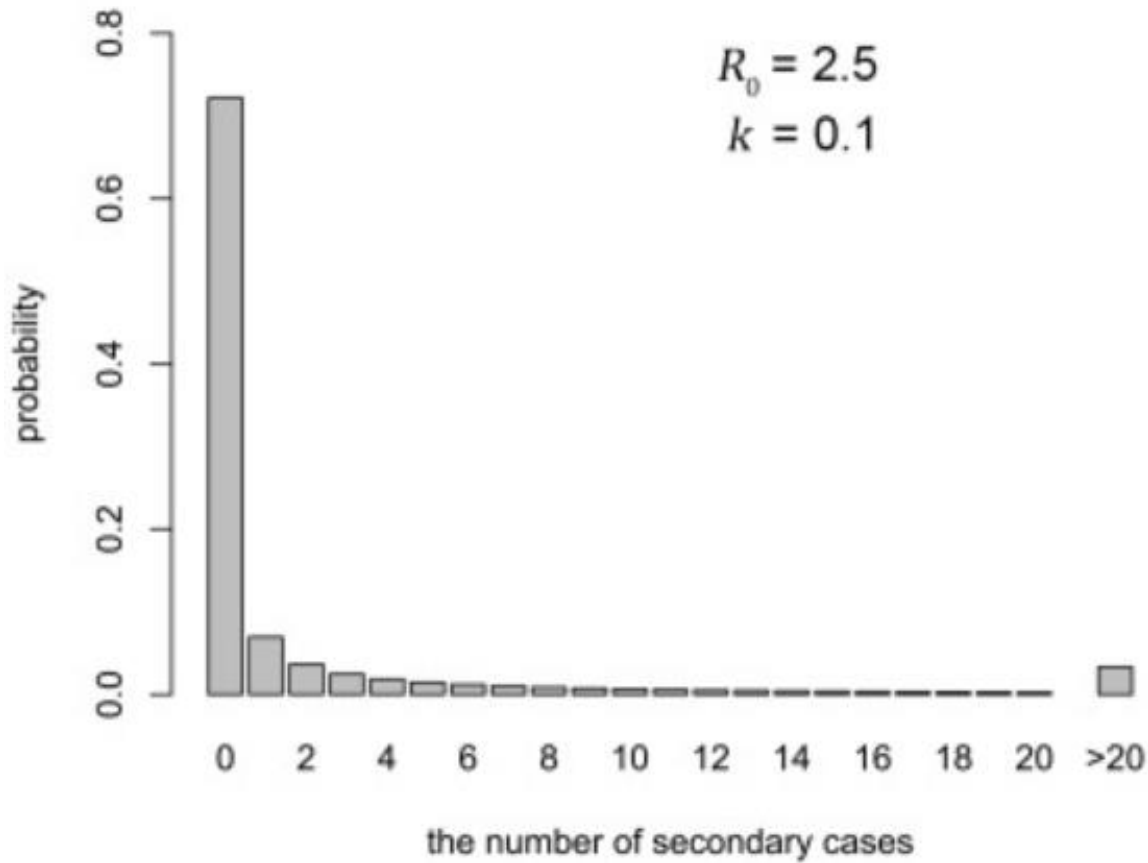
Figure from Pekar et al, 2022

Figure S16. **Subsampled global phylogeny showing reversions.** Subsampled SARS-CoV-2 time-resolved phylogeny from Nextstrain, with reversions colored blue if a C-to-T reversion and black otherwise.

Covid transmission is overdispersed

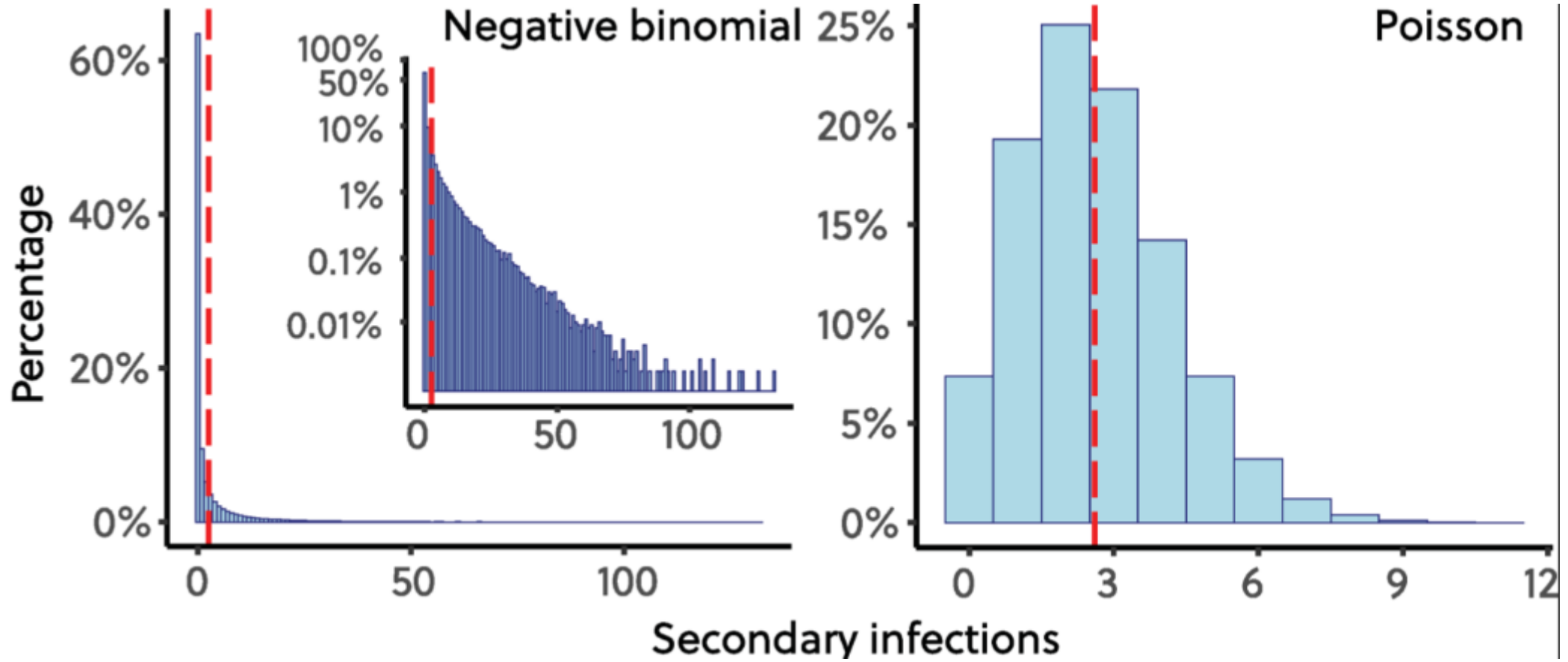
Early in the pandemic, we all heard about R_0 : the average number of people each covid patient infects

But there's another important parameter called k , which is a measure of how uneven the transmission is.



[Image source](#)

When you think of $R_0 = 2-3$, you think that means that most people transmit covid to 2-3 people. But most people transmit covid to 0 people, while a few transmit it to 100 people. Here's a negative binomial vs a poisson distribution, both these have $R_0 = 2.6$



[Image source](#)

Negative binomial distributions can sometimes grow faster, but growth rate evens out after some time

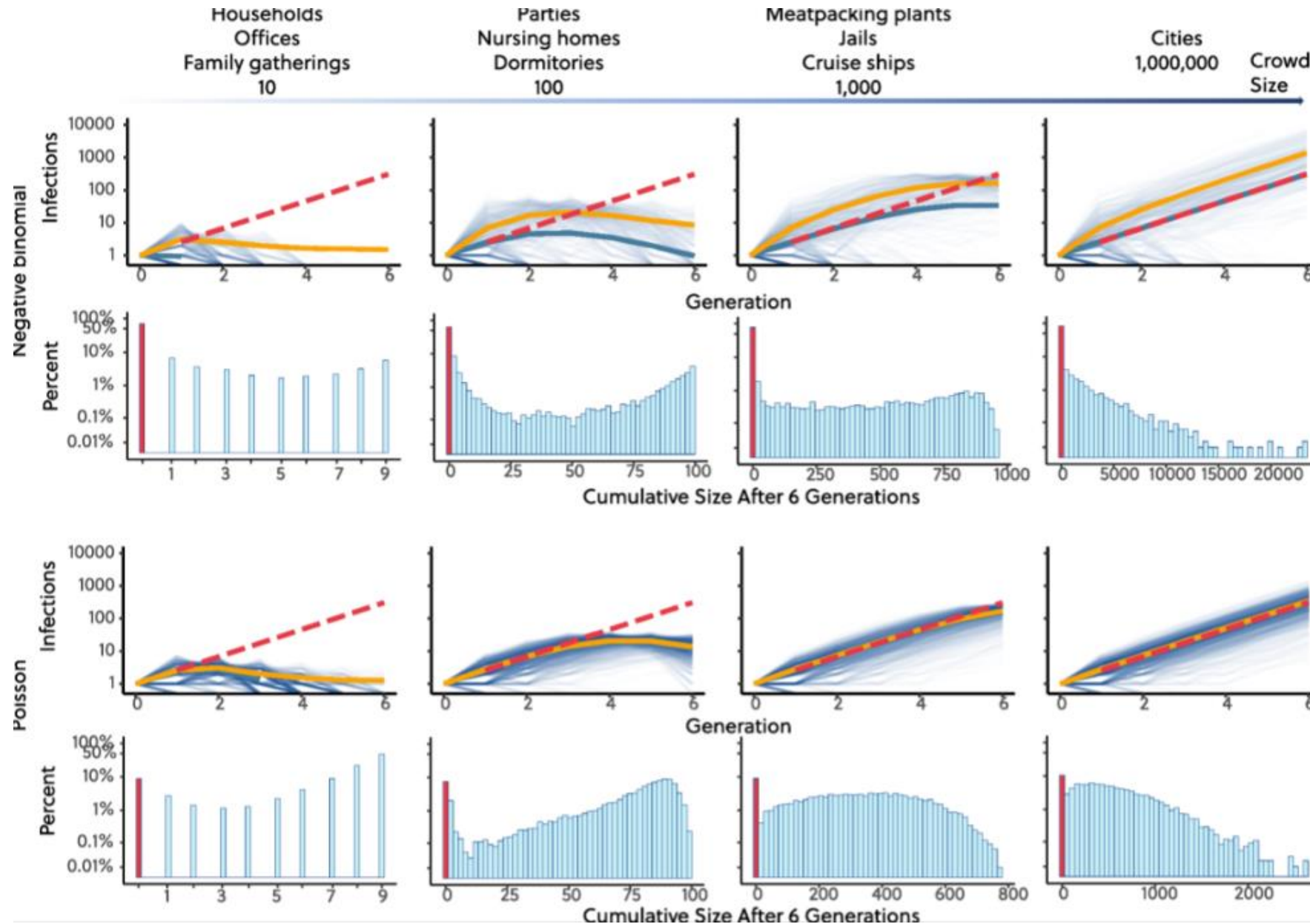
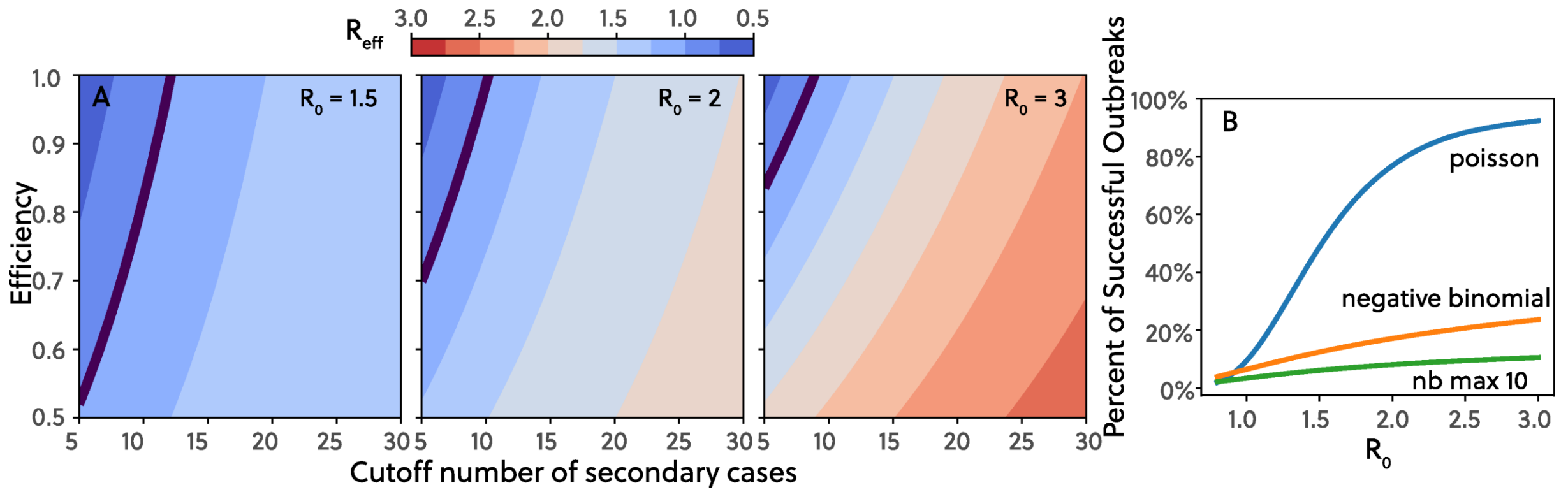


Fig 2. Example trajectories of NB and Poisson branching processes.

Figure shows example trajectories (in number of active infections versus generation) of NB and Poisson branching processes and cumulative infection sizes after 6 generations of spread. Both simulations start with 1 infection and have the same $R_0 = 2.6$. For NB branching process, we assume dispersion parameter $k = 0.16$, same as SARS-CoV-1. We run all simulations 10,000 times. Dashed red lines represent theoretical values in the large-population limit, where I is number of active infections, and n is number of generations. Solid blue lines are the mean values of all simulations including those that have not taken off, which overlap with the theoretical values when the susceptibles are not depleted. Solid orange lines are the mean value for simulations that took off, and the outbreaks appear more explosive in the first few generations in the NB simulations. Both number of active cases and cumulative infections are in log10 scale.

Negative binomial distributions cause a higher extinction rate than poisson distributions:



[Image source](#)

Overdispersion helps explain a few things:

- Why it's more likely for a pandemic to start in a big city.
- Why certain countries did better during the pandemic (i.e. Japan took measures to prevent clusters)
- Why Rootclaim's idea that a lab leak would show up at the market is very unlikely:
Covid tends to either go extinct or spread widely. It's hard for it to transmit at low levels for a long period of time. It doesn't get to spread person to person slowly and try lots of different places in the city to figure out which one is the best superspreader location (until it finds the magical mahjong closet). It will either go extinct or it will start a cluster at the first suitable location it hits (either the lab or perhaps some random location nearby).

proCov2 history

The history of the proCov2 theory is kind of fun.

Kumar originally wrote a paper saying that there's a different proCov2, with 4 mutations relative to Lineage A, but that ultimately got downgraded to 1 mutation

Extended Data Table 1. SARS-CoV-2 variants and their molecular types and first timing and location.

Mutant (major)	Mutant (minor)	Gene	Genomic Position	Nucleotide change	Amino acid change	Time (days)	Variant Frequency	Genomes mapped	First location
μ_1		ORF1ab	2416	U>C		0	98.1%	18	China, Asia
μ_2		ORF1ab	19524	U>C		0	98.6%	0	China, Asia
μ_3		S	23929	U>C		0	98.4%	0	China, Asia
α_1		ORF1ab	18060	U>C		0	95.1%	849	China, Asia
	α_{1a}	N	28657	C>U		63	1.3%	2	France, Europe
	α_{1b}	ORF1ab	9477	U>A	F>Y	63	1.2%	8	France, Europe
	α_{1c}	N	28863	C>U	S>L	63	1.2%	0	France, Europe
	α_{1d}	ORF3a	25979	G>U	G>V	63	1.2%	344	France, Europe
α_2		ORF1ab	8782	U>C		0	91.0%	47	China, Asia
α_3		ORF8	28144	C>U	S>L	0	90.8%	1116	China, Asia
	α_{3a}	ORF1ab	1606	U>C		43	1.7%	501	United Kingdom, Europe
	α_{3b}	ORF1ab	11083	G>U	L>F	24	9.2%	377	China, Asia
	α_{3c}	N	28311	C>U	P>L	64	1.9%	3	South Korea, Asia
	α_{3d}	ORF1ab	13730	C>U	A>V	71	1.8%	3	Taiwan/Malaysia, Asia
	α_{3e}	ORF1ab	6312	C>A	T>K	71	1.7%	483	Taiwan/Malaysia, Asia
	α_{3f}	ORF3a	26144	G>U	G>V	28	5.1%	452	China, Asia
	α_{3g}	ORF1ab	14805	C>U		54	6.0%	3	United Kingdom, Europe
	α_{3h}	ORF1ab	17247	U>C		64	2.0%	580	Switzerland, Europe
	α_{3i}	ORF1ab	2558	C>U	P>S	54	1.7%	26	United Kingdom, Europe
	α_{3j}	ORF1ab	2480	A>G	I>V	54	1.6%	462	United Kingdom, Europe
β_1		ORF1ab	3037	C>U		31	77.0%	40	China, Asia
β_2		S	23403	A>G	D>G	31	77.1%	9	China, Asia
β_3		ORF1ab	14408	C>U	P>L	41	76.9%	3032	Saudi Arabia, Middle East
	β_{3a}	ORF1ab	20268	A>G		64	5.7%	1213	Italy, Europe
	β_{3b}	N	28854	C>U	S>L	29	3.1%	34	China, Asia
	β_{3c}	ORF1ab	15324	C>U		29	2.3%	669	China, Asia
	β_{3d}	ORF3a	25429	G>U	V>L	77	1.7%	484	United Kingdom, Europe
	β_{3e}	N	28836	C>U	S>L	74	1.6%	3	Switzerland, Europe
	β_{3f}	ORF1ab	13862	C>U	T>I	74	1.6%	50	Switzerland, Europe
	β_{3g}	ORF1ab	10798	C>A	D>E	86	1.4%	414	United Kingdom, Europe

Kumar, [2020 version](#),

he says that proCov2 has the mutations:
C2416T
C19524T
C23929T
C18060T

Table 1.

SARS-CoV-2 variants in 29KG dataset.

Kumar, [2021 version](#),

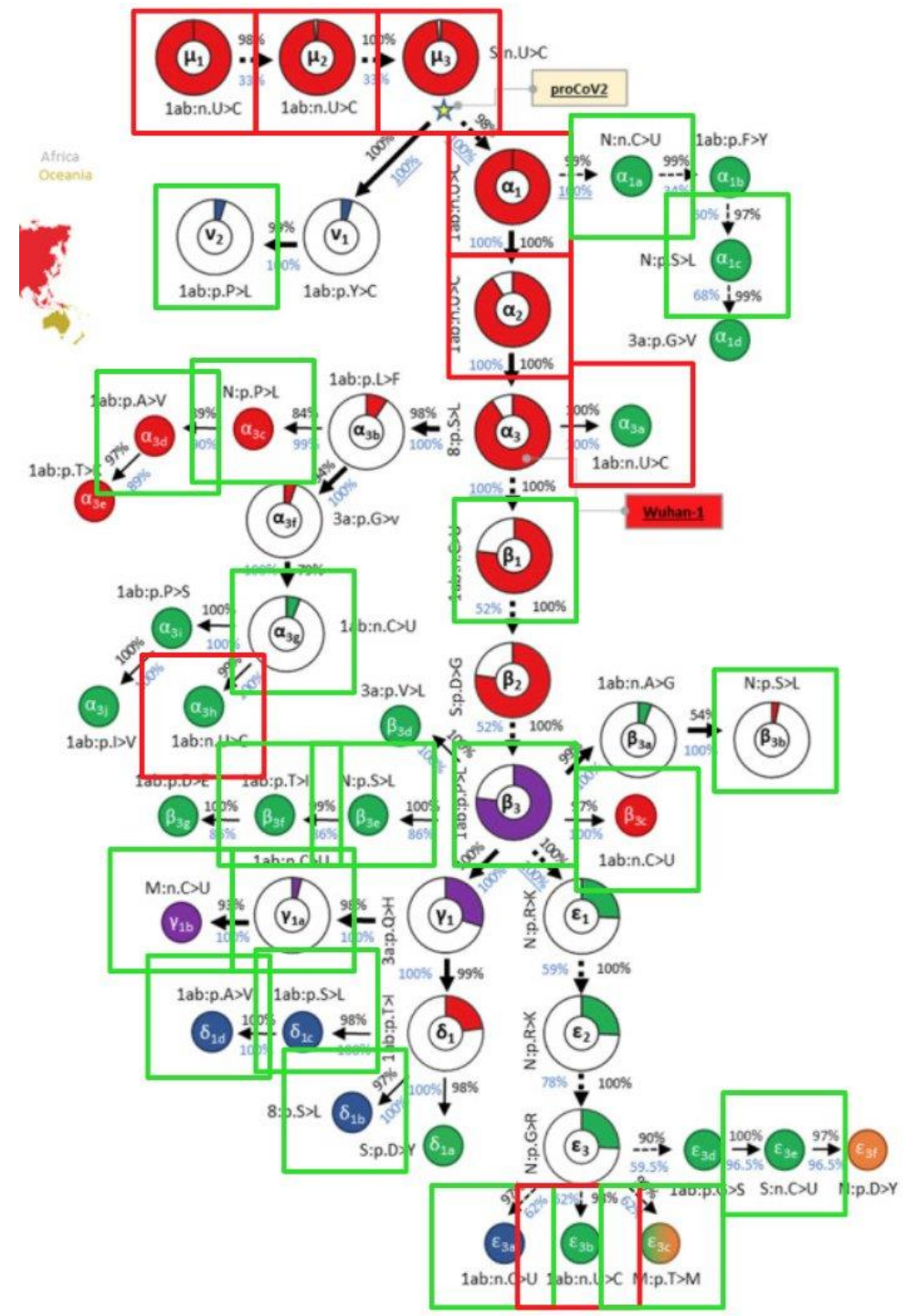
It looks like maybe those first 18 genomes were misplaced somehow, now it's just: C18060T

Mutant (major)	Mutant (minor)	Gene	GenomicPosition	Nucleotide change	Amino acid change	Time(days)	VariantFrequency	Genomesmapped	Firstlocation
μ_1		ORF1ab	2416	U>C		0	98.1%	0	China, Asia
μ_2		ORF1ab	19524	U>C		0	98.6%	0	China, Asia
μ_3		S	23929	U>C		0	98.4%	18	China, Asia
α_1		ORF1ab	18060	U>C		0	95.1%	849	China, Asia
	α_{1a}	N	28657	C>U		63	1.3%	2	France, Europe
	α_{1b}	ORF1ab	9477	U>A	F>Y	63	1.2%	3	France, Europe
	α_{1c}	N	28863	C>U	S>L	63	1.2%	5	France, Europe
	α_{1d}	ORF3a	25979	G>U	G>V	63	1.2%	344	France, Europe
α_2		ORF1ab	8782	U>C		0	91.0%	47	China, Asia
α_3		ORF8	28144	C>U	S>L	0	90.8%	1115	China, Asia
	α_{3a}	ORF1ab	1606	U>C		43	1.7%	501	United Kingdom, Europe
	α_{3b}	ORF1ab	11083	G>U	L>F	24	9.2%	376	China, Asia
	α_{3c}	N	28311	C>U	P>L	64	1.9%	3	South Korea, Asia
	α_{3d}	ORF1ab	13730	C>U	A>V	71	1.8%	3	Taiwan/Malaysia, Asia
	α_{3e}	ORF1ab	6312	C>A	T>K	71	1.7%	483	Taiwan/Malaysia, Asia
	α_{3f}	ORF3a	26144	G>U	G>V	28	5.1%	121	China, Asia
	α_{3g}	ORF1ab	14805	C>U		54	6.0%	334	United Kingdom, Europe

*Amino acid change is shown only for non-synonymous change.

You can see the unlikeliness of Kumar's theory by just graphing C -> T (green, more likely) and T -> C (red, less likely) in his diagram.

His scenario has a lot of red (unlikely) mutations, right at the beginning, and then all the subsequent covid evolution was green.



Zach Hensel
@alchemytoday

...

Plotting U>C (red) and C>U (green) on Fig. 1 is a good illustration of why the proCoV2 scenario is implausible.

6:15 PM · Mar 2, 2022

The 2020 version (but not 2021 version) can be disproven, via this [Spyros Lytras thread](#)

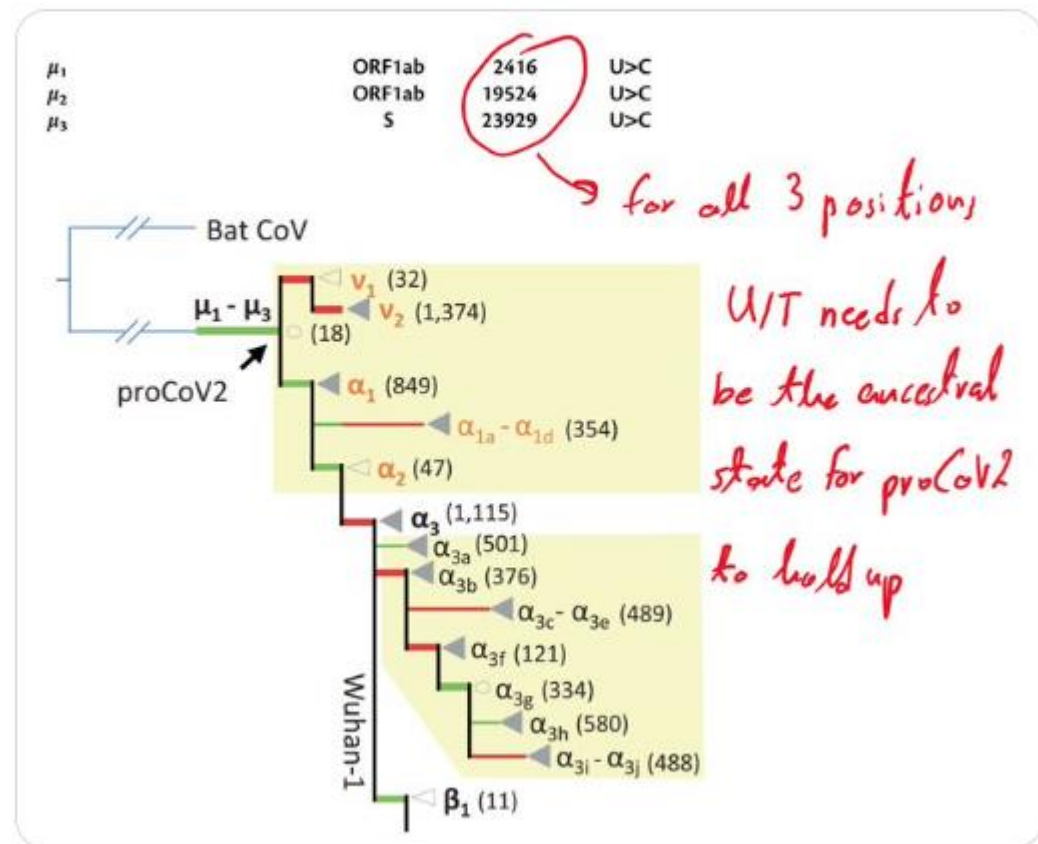
Spyros Lytras @SpyrosLytras · Mar 2, 2022
 proCoV2 is a hypothetical SARS-CoV-2 progenitor sequence reconstructed using a 'Mutation Order Analysis' approach published here: doi.org/10.1093/molbev... and also used here: doi.org/10.1093/molbev...

2 1 4

Spyros Lytras @SpyrosLytras · Mar 2, 2022
 The proCoV2 analysis was done before publication of the Laos BANAL viruses (now the closest known relatives to SARS-CoV-2, and assumes a number of mutational steps to go from proCoV2 to sampled SC2 sequences (lineages A and B)...

1 1 3

Spyros Lytras @SpyrosLytras · Mar 2, 2022
 The first mutations assumed are termed μ_1 - μ_3 in positions 2416, 19524 and 23929 and U/T should be the ancestral state of all 3 positions if proCoV2 is a correct reconstruction of the SC2 progenitor



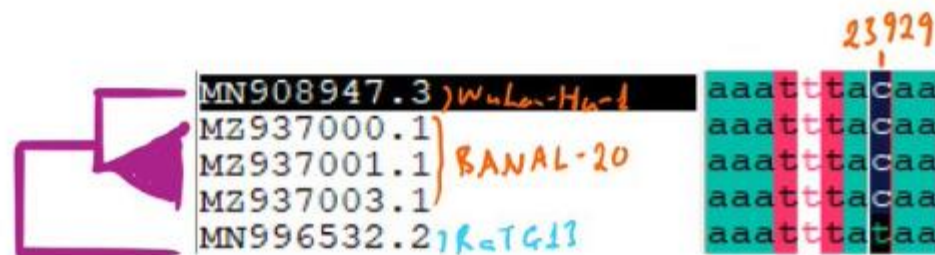
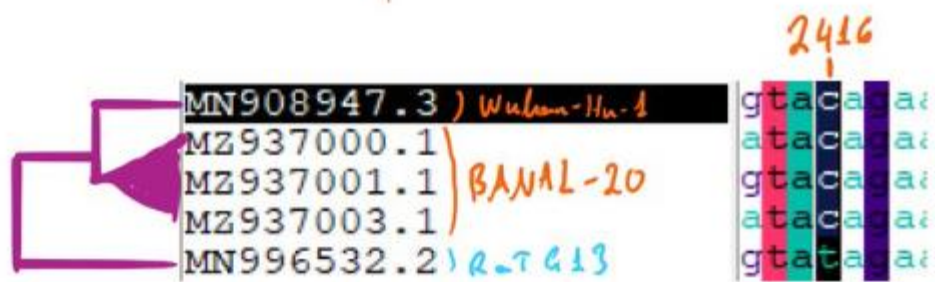
2 1 3



Spyros Lytras @SpyrosLytras · Mar 2, 2022

...

However, in light of the BANAL CoVs that's certainly not the case! for sites 2416 & 23929 3 BANALs are the closest known relatives to SC2 (not RaTG13 that was used for proCoV2) and guess what base they've got in the corresponding positions!



proCoV2 scenario doesn't hold up!
ancestral states 2416 (μ_2) and 23929 (μ_3)
certainly a C!

1

4

3

|||

↑

The closest bat viruses to SARS-CoV-2 are still ~1,000 mutations away, so we don't actually know what nucleotides the ancestor virus had at those positions.

We don't know whether it had any of these proCov2 mutations.

We don't know for sure whether it was closer to lineage A or lineage B.

Lineage A / Lineage B

Two spillovers at the market

Any theory of lineage A and B has to account for several facts:

Lineage A is 2 mutations closer to known bat viruses.

Lineage B was found before A.

Lineage B has more diversity than Lineage A, over time, showing that it did start/spread earlier.

Lineage B was found at the market in early December.

The earliest Lineage A cases were found very near the market, later in December

Both lineages were found in environmental samples at the market.

Two spillovers at the market neatly explains all of these facts. Other theories are more complicated:

- If lineage A didn't start at the market, why are the December lineage A cases found so close?
- The earliest lineage A case was found before the search could have been biased.
- If you think lineage A was widespread before the market, why is the viral diversity lower?
- If you think lineage A was low prevalence before, then why was the market the one and only superspreader event?

If lineage A is the root, it looks like it hasn't evolved enough, compared to lineage B.

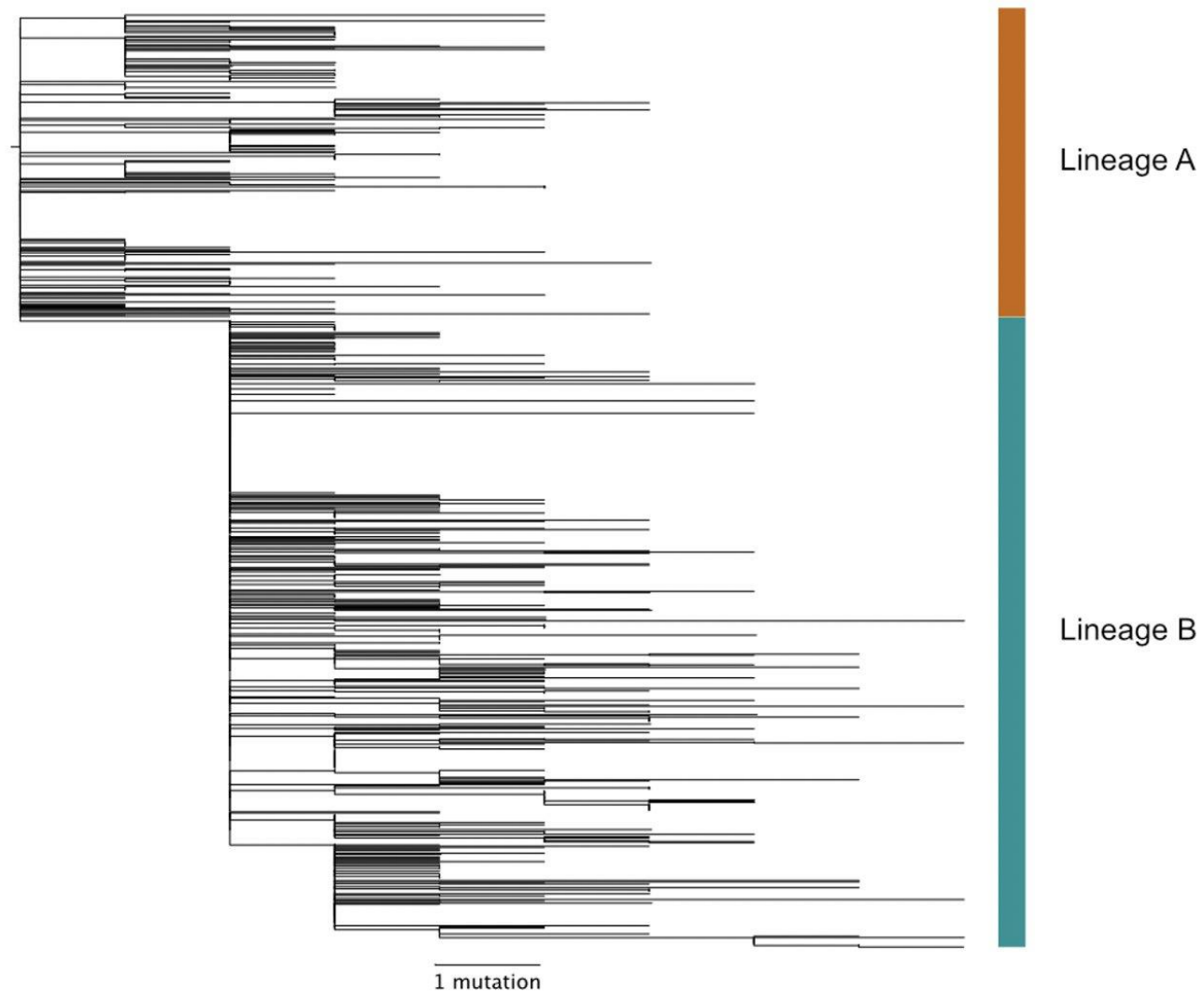


Figure S19. SARS-CoV-2 maximum likelihood tree rooted on lineage A (n=787 taxa, through 14 February 2020).

Pekar plotted the lineages with a 2 nucleotide gap, which is correct if you think A started first.

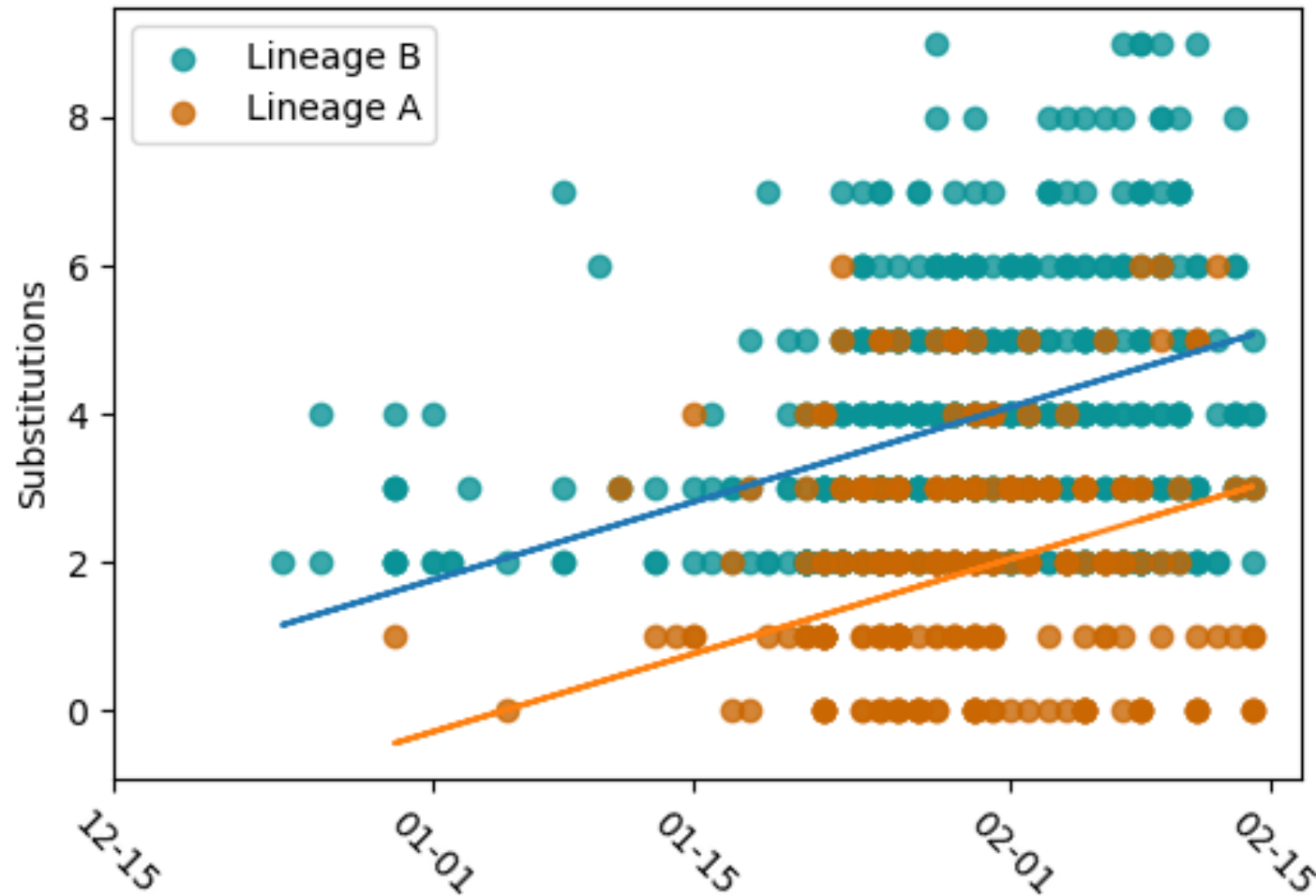


Figure S20. Substitution counts of SARS-CoV-2 genomes through 14 February 2020 from the root of the maximum likelihood tree when rooted on lineage A (Fig. S19). The plotted lines have a slope of 27.51 substitutions/year, are fit to their respective lineages, and are separated by 2.04 substitutions, showcasing the greater divergence of lineage B than lineage A when the tree is rooted on lineage A.

Lineage B has more diversity than Lineage A, over time

531 lineage B genomes sampled up to mid February

256 lineage A genomes sampled

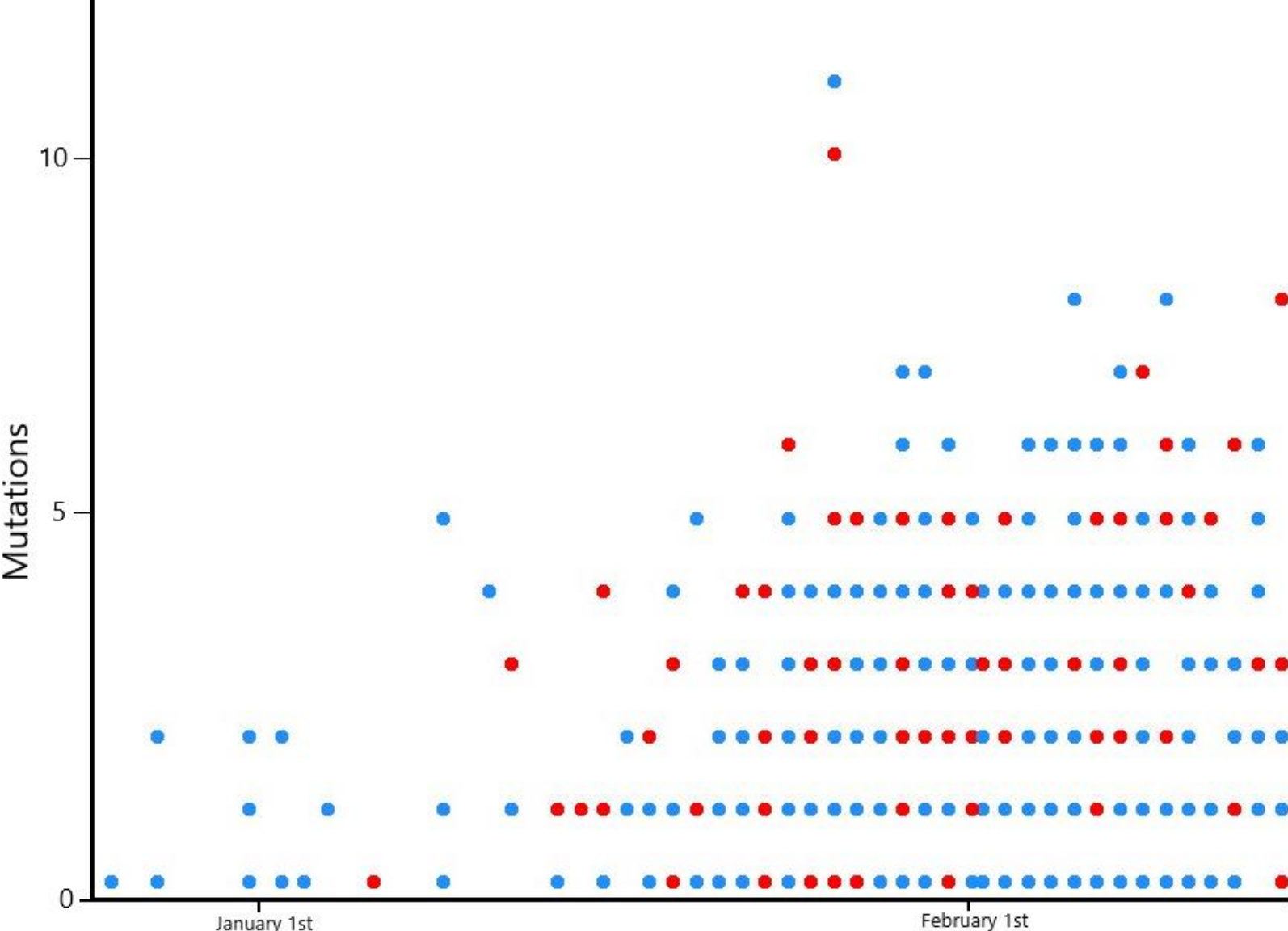
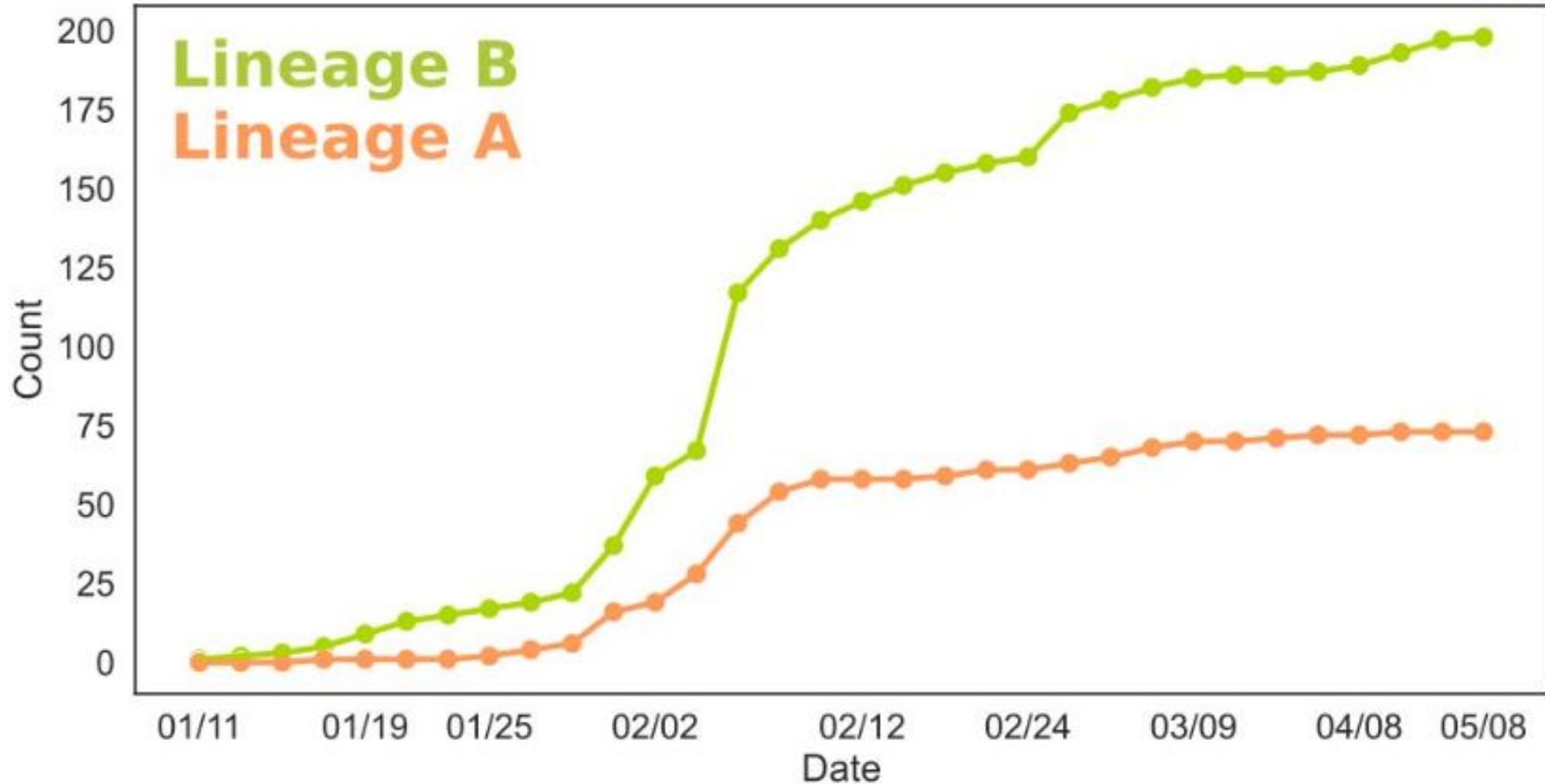


Image plotted without the 2 nucleotide gap between A and B, to show relative diversity

(a few overlapping points disappear)

This is true across multiple data sources.

Jesse Bloom wrote his paper on “deleted early sequences”. His paper confirmed what was true in all other datasets – there are [more lineage B than lineage A cases](#).



Here's some early data from Zhongnan hospital:

A few things to notice:

Zhongnan hospital is the one right next to the Wuhan institute of virology.

There are twice as many Lineage B genomes as there are Lineage A.

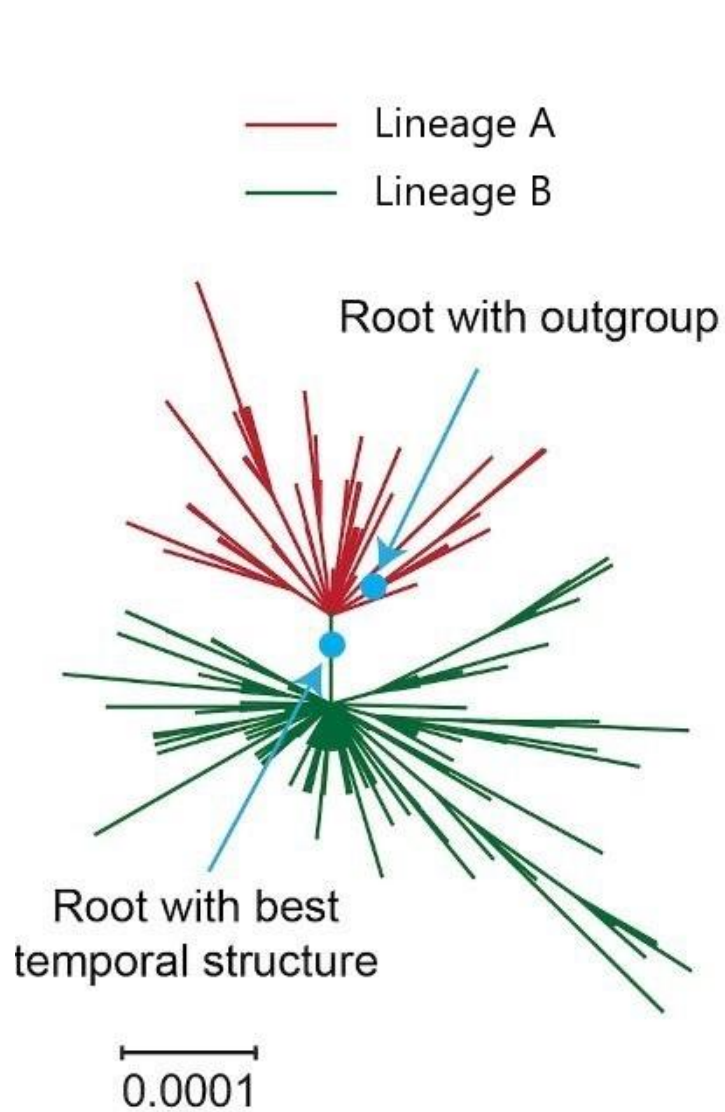
There are no intermediates.

A proCov2 genome was found in here and used as the outgroup, but it looks like only one sample.

A few of these samples were already evolving D614G.

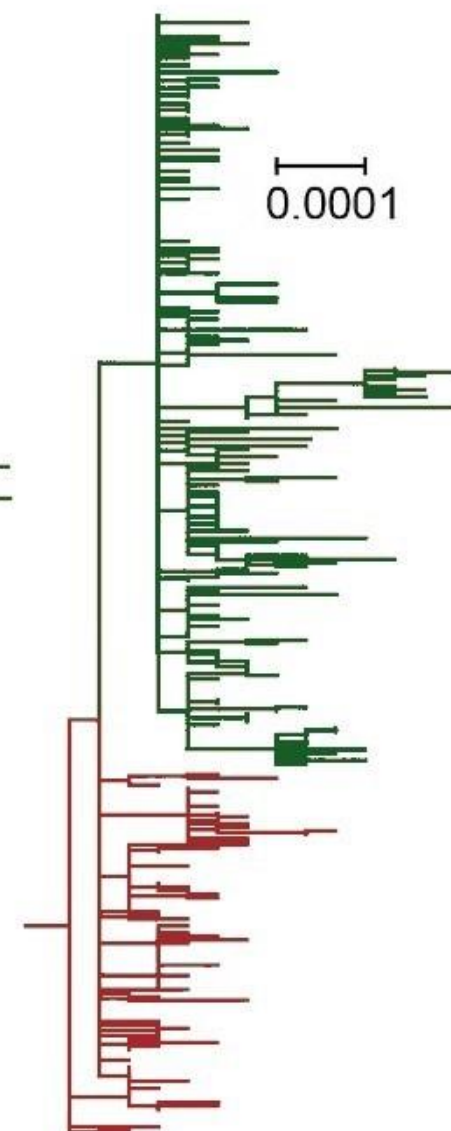
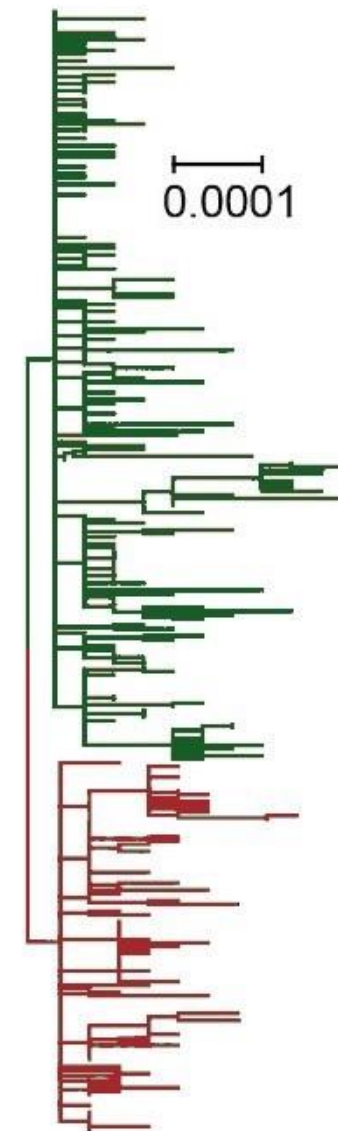
[Data from Eddie Holmes.](#)

Note that I converted S/L lineages into A/B. Also, the original diagram had colors swapped on left and right for confusing reasons, so I fixed it. You can tell which colors are correct based on where the outgroup rooting is placed in each diagram.



Root with best temporal structure

Root with outgroup



Probability

With 1 lineage, the odds are 1 in 10,000 that the market would be the first cluster of cases, if this was a lab leak.

With 2 lineages, the odds are 1 in 100 million that the virus would come from the lab to the market twice.

Since I want to steelman the lab leak theory, I will instead consider the possibility that it looks like 2 spillovers by chance.

Pekar's paper says there's a 3% chance it would look like 2 lineages by chance. (He says it's bayes factor 4.2)

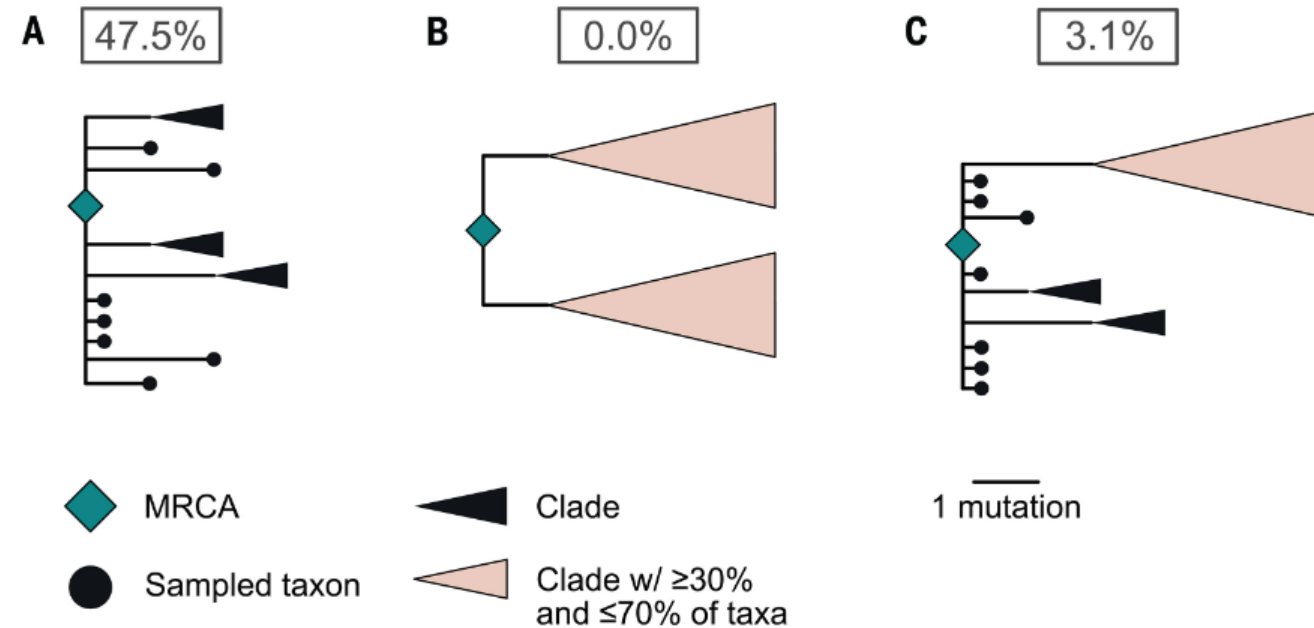


Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations.

(A) A large polytomy of at least 100 descendent lineages, which is consistent with the base of both lineages A and B. (B) Topology matching a C/C ancestral haplotype: two clades, each one mutation from the ancestor, both with polytomies of at least 100 descendent lineages. (C) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 100 descendent lineages, including a large clade separated by two mutations, also possessing a polytomy of at least 100 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the respective boxes.

Figure from [Pekar et al, 2022](#)

Why C/C ancestor is unlikely

If C/C is the ancestral haplotype, then SARS-CoV-2 is characterized by two clades: lineages A and B, each one mutation from the root with no transitional genomes (Fig. 2B). This topology, where there are only two clades of any size, each one mutation from the root, was present in 10.5% of phylogenies from our simulated epidemics. However, both lineages A and B are large clades, comprising 35.2% and 64.8% of the early SARS-CoV-2 genomes, respectively, and the smaller clade in these simulations was rarely this large. If we require our simulated clades to more realistically comprise at least 1% of the taxa, only 6.7% of the simulations match the C/C topology. If we require both clades to comprise $\geq 30\%$ of the taxa—better reflecting empirical genomic diversity—only 1.5% of the simulations match the C/C topology. Finally, both lineages A and B comprise large polytomies. When we require each of these clades to have a basal polytomy of at least 100 descendant lineages—a conservative reflection of the 108- and 231-lineage polytomies characterizing lineages A and B, respectively—none of the simulations still match the C/C topology. These results indicate that a single introduction of C/C virus would not be expected to give rise to lineages A and B with no surviving ancestral C/C lineages.

Why 2 lineages are unlikely from a single introduction

If lineage A or B is the ancestral haplotype, then SARS-CoV-2 is characterized by a large basal polytomy with the largest clade in the tree separated by two mutations from the root (lineage B is the descendant clade if lineage A is the root, and vice-versa) (Fig. 2C). Importantly, our simulations permit these two mutations to occur either within a single individual or during successive infected hosts (107), reflective of multiple mutations of SARS-CoV-2 occurring within the serial interval between transmission partners (108). We see a large clade comprising a substantial fraction of the sampled taxa (*i.e.*, between 30% and 70%, reflecting either lineage A or B prevalence) in 10.8% of the epidemic simulations. When we require the large clade separated by at least two mutations from the basal polytomy of at least 100 descendant lineages, we observe this topology in 4.1% of epidemic simulations. However, if we also require the large clade to have at least a 100-lineage polytomy at its base, only 0.5% of the simulations match the topology if there were a single introduction of lineage A or B without any surviving transitional C/C lineages.

Probabilities:

The odds of a lab leak are even lower than 3%.

If you think lineage A came from the lab, you need to explain why B looks older than A, and has more diversity.

That's bayes factor 48. Even if you ignore all the market genomes, it's bayes factor of 11.

Haplotype	Mutations from Hu-1 reference	Representative genome	Phylodynamic analysis	
			Unconstrained (%)	No market (%)
B (C/T)	N/A	Hu-1	80.85 [†]	62.96 [†]
A (T/C)	C8782T+T28144C	WH04	1.68 ^{**}	5.73 ^{**}
C/C	T28144C	N/A	10.32 [*]	23.02

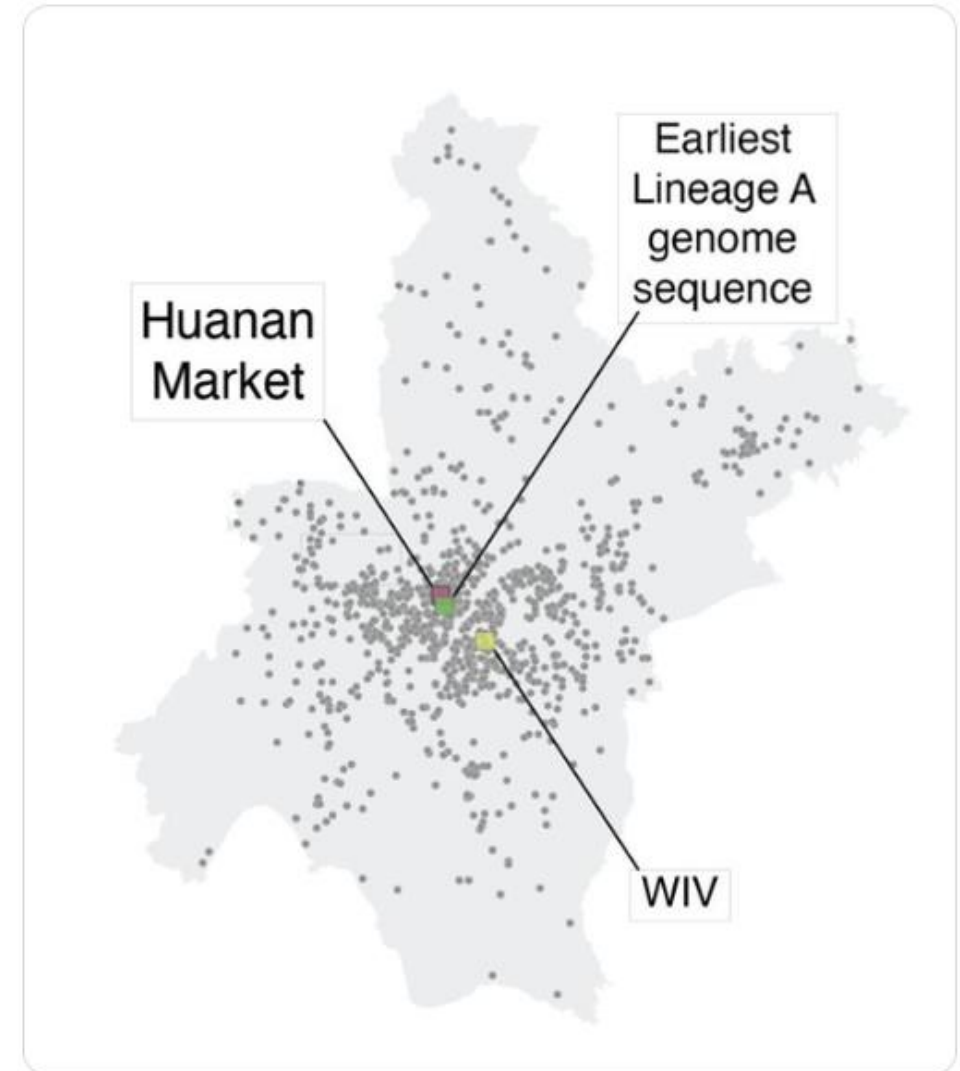
Probabilities:

The first 2 lineage A cases were found closer to the market than expected by chance, if you think covid was actually all over town. ($p = .001$)

Maybe that's a bayes factor of 50? Or 100?

(depends on your exact model of how close cases should be distributed for a market origin)

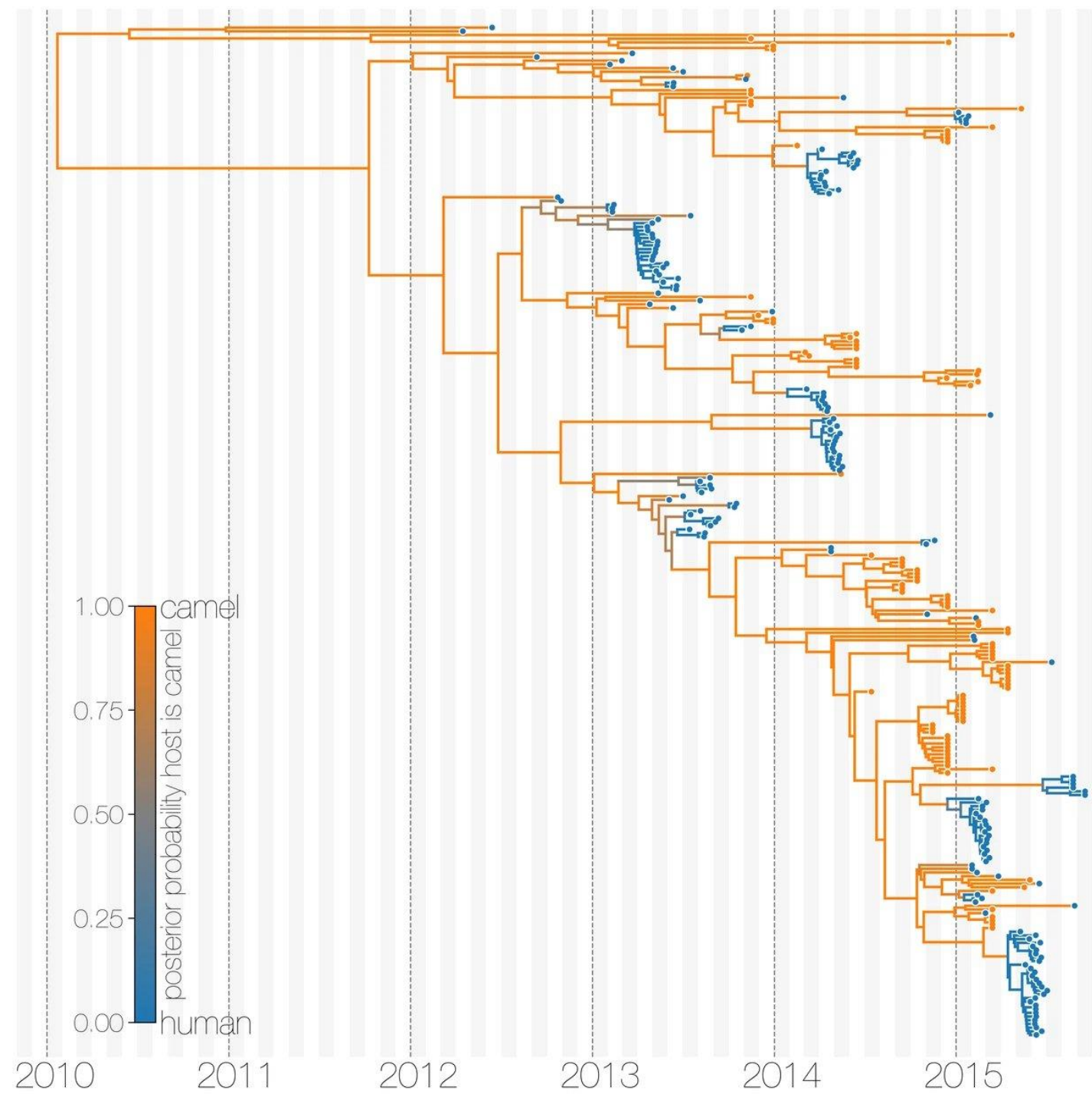
Remember that one of these two cases was diagnosed before the connection was known between the market and covid, so that can't be ascertainment bias.



Multiple spillovers are normal.

SARS had > 10 spillovers.

MERS had [multiple spillovers](#).



When covid infected mink farms, it spilled over back into humans, multiple times.

Covid infected hamsters also reinfected humans, multiple times.

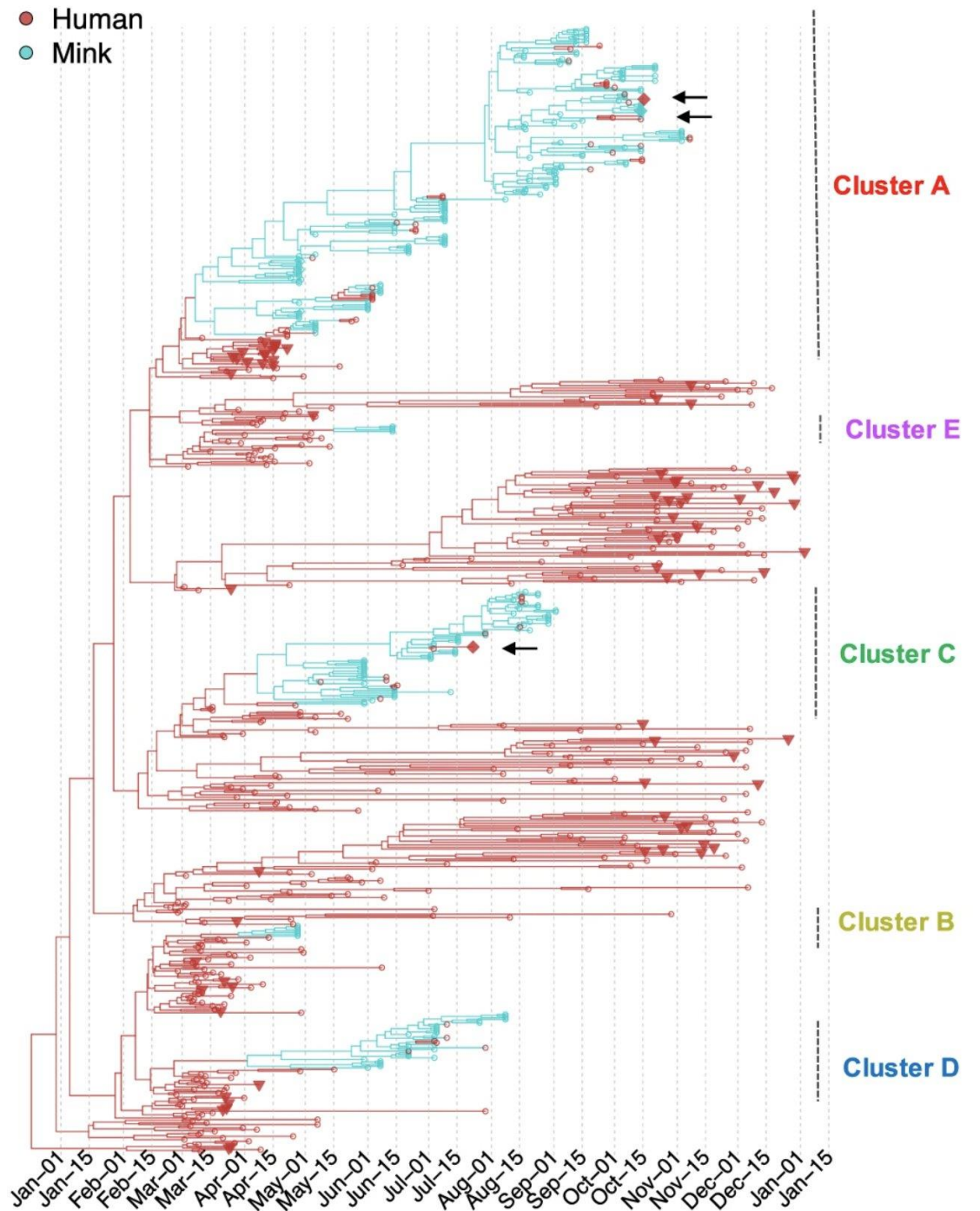


Figure from [Lu et al, 2021](#)

First Lineage A sample had one mutation:

A is ancestral to B, and the market is dominated by B

The first known lineage A case also had an extra mutation, meaning that it likely was already circulating for some time before being sampled.

hCoV-19/env/Wuhan/IVDC-HBA20/2020|EPI_ISL_10497477|2020-01-01
Sequence ID: Query_56425 Length: 29854 Number of Matches: 4

Range 1: 477 to 16373 [Graphics](#) [Next Match](#) [Prev](#)

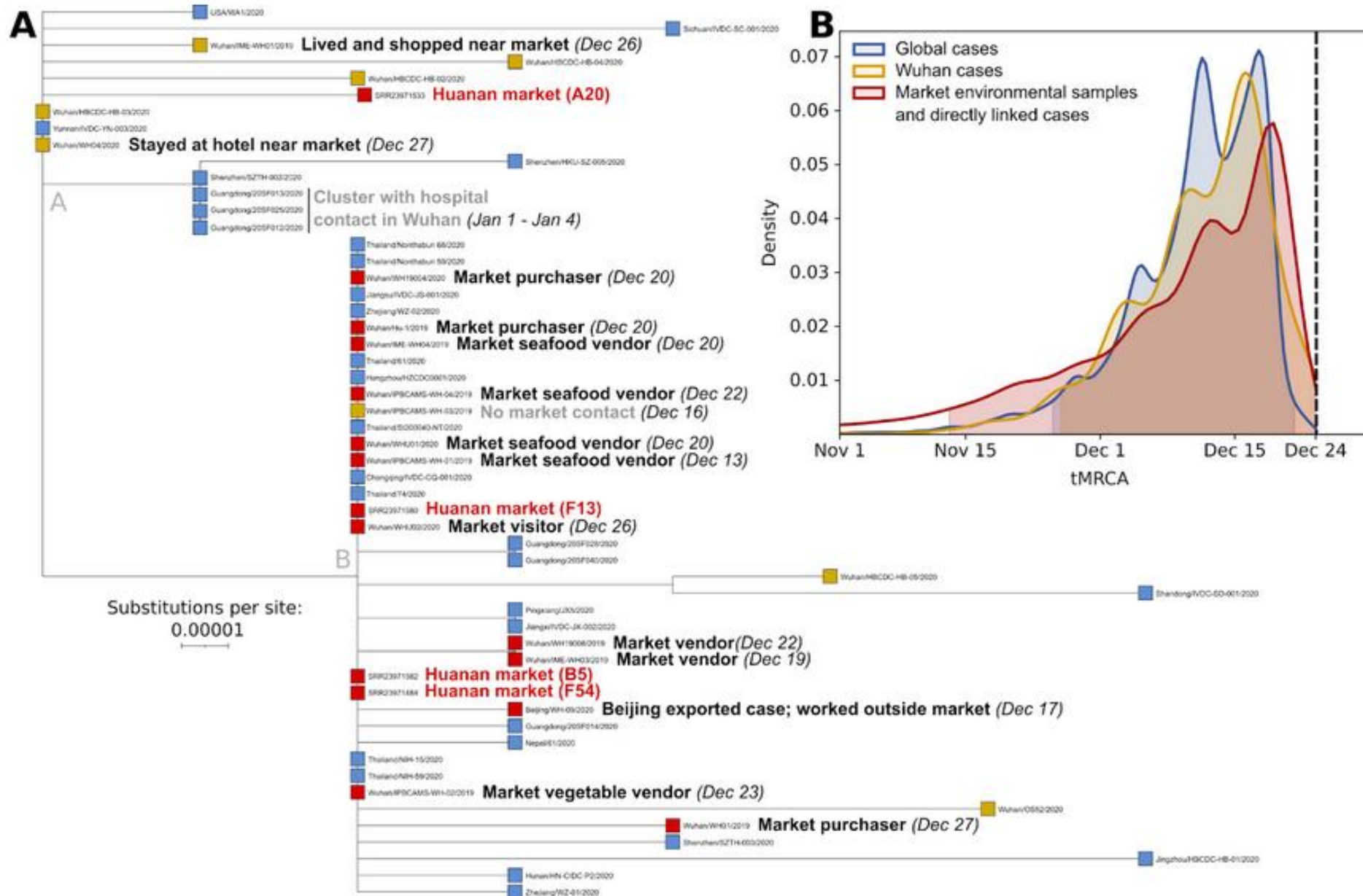
Score	Expect	Identities	Gaps	Strand
28755 bits(15571)	0.0	15739/15899(99%)	4/15899(0%)	Plus/Plus
Query 493	TCGAACTGCACCTCATGGTCATGTTATGTTGAGCTGGTAGCAG--AACTCGAAGGCATT	550		
Sbjct 477NNNNNNN..--....	534		
Query 6131	AAAGTTACATTTTTCCCTGACTTAAATGGTGATGGTGGCTATTGATTATAAACTACTAC	6190		
Sbjct 6115T.....	6174		
Query 8771	ACATGGTTTAGCCAGCGTGGTGGTAGTTATACTAATGACAAAGCTTGCCCATTTGATTGCT	8830		
Sbjct 8755T.....	8814		
Query 20215	AATTTACAAGAATTTAAACCCAGGAGTCAAATGGAAATTGATTCTTAGAATTAGCTATG	20274		
Sbjct 20199NA.....	20258		
Query 26228	CTGATGAGTACGAACCTATGTACTCATTGTTTCGGAAGAGACAGGTACGTTAATAGTTA	26287		
Sbjct 26212T.....	26271		
Query 28088	TGGTTCTAAATCACCATTGACATCGATATCGGTAATTATACAGTTTCCTGTTTACC	28147		
Sbjct 28072	..NN.....C...	28131		
Query 28148	TTTTACAATTAATTGCCAGGAACCTAAATGGGTAGTCTTGTAGTGCCTTGTTCGTTCTA	28207		
Sbjct 28132	28191		

Note the inconsistency here – Yuri says 2 mutations is likely, when he wants it to not be 2 spillovers,

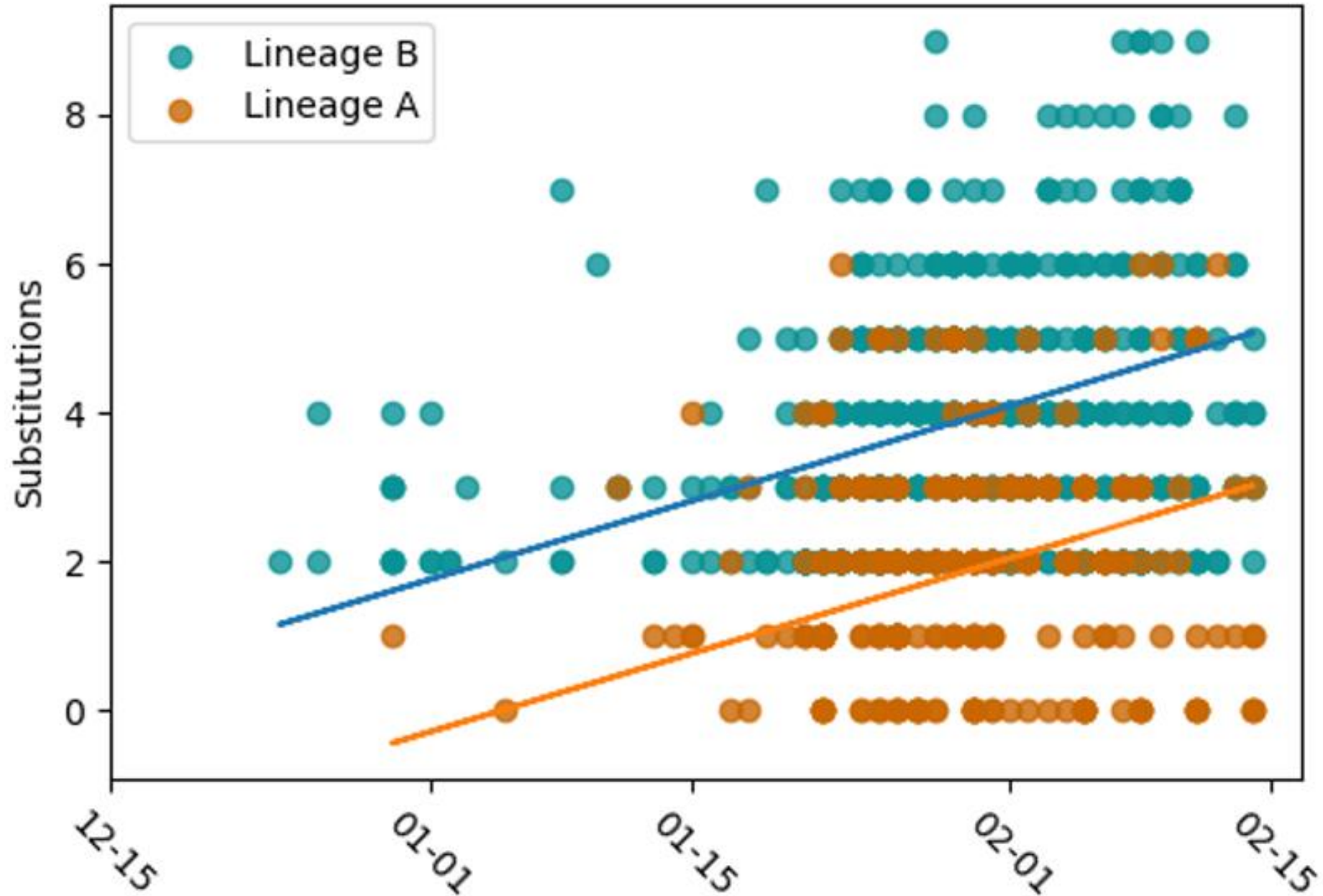
But he says 1 mutation is unlikely, when he wants to prove that lineage A has been around for a while.

Statements like this need to be quantified and integrated into a model of the outbreak. That’s what Pekar 2022 did.

Also, some Lineage B patients at the market had 1-2 mutations:



We can also just look it over time: by the time lineage A has had one mutation, lineage B has had two:



Pekar erratum

This is [from the preprint](#):

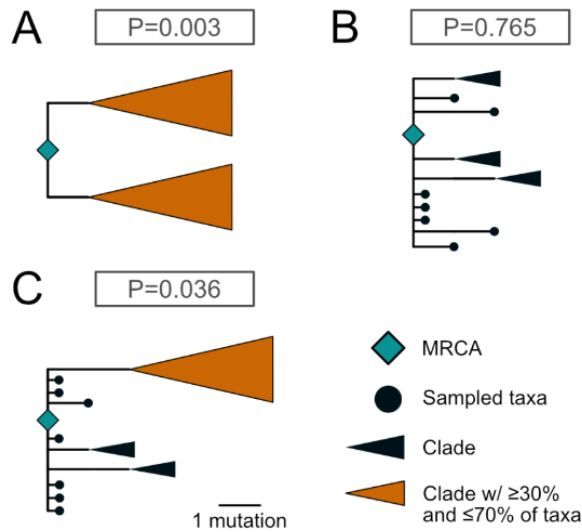


Figure 5. Probability of potential phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. (A) Topology matching a C/C ancestral haplotype: large polytomy, consistent with the base of both lineages A and B. (C) Topology matching lineage A or lineage B ancestral haplotype. Basal taxa have short branch lengths for clarity. Probability of each phylogenetic structure after a single introduction is reported in the box.

From [the published paper](#):

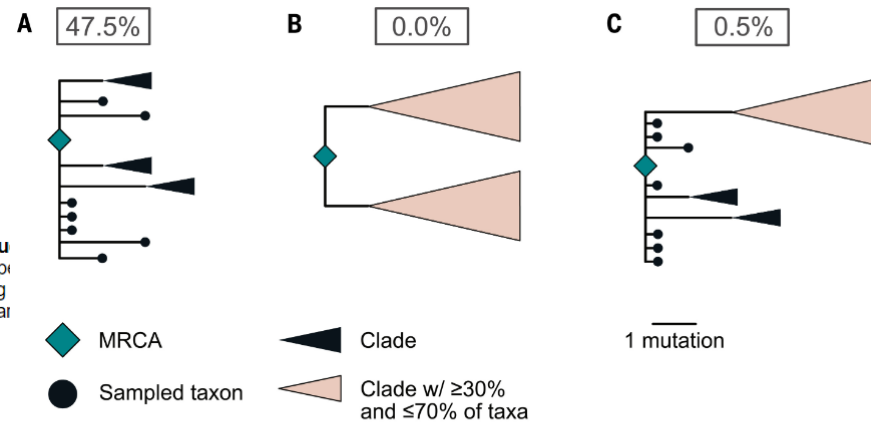


Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations.

(A) A large polytomy of at least 100 descendent lineages, which is consistent with the base of both lineages A and B. (B) Topology matching a C/C ancestral haplotype: two clades, each one mutation from the ancestor, both with polytomies of at least 100 descendent lineages. (C) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 100 descendent lineages, including a large clade separated by two mutations, also possessing a polytomy of at least 100 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the respective boxes.

This is the latest version:

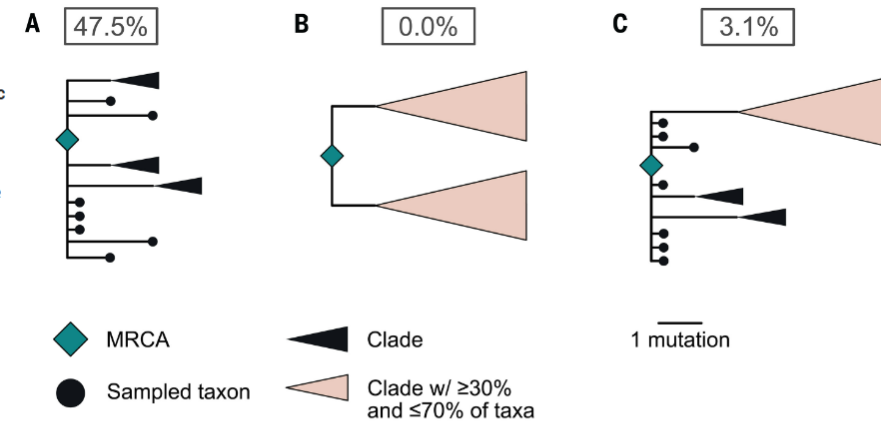


Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations.

(A) A large polytomy of at least 100 descendent lineages, which is consistent with the base of both lineages A and B. (B) Topology matching a C/C ancestral haplotype: two clades, each one mutation from the ancestor, both with polytomies of at least 100 descendent lineages. (C) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 100 descendent lineages, including a large clade separated by two mutations, also possessing a polytomy of at least 100 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the respective boxes.

The erratum raises some interesting questions about peer review.

Science is falsifiable. If a paper is wrong, it can get corrected by simply pointing out the flaws.

The lab leak theory is decentralized, so there's no one place to try to correct it. Much of it is not published or peer reviewed, so there's no easy way to correct it when it is wrong.

I've asked lab leak theorists a lot of questions, when I disagreed with something or did not understand their arguments. Usually, one of three things happens:

- I get no response.
- I'm told to reread the argument that I think is wrong.
- My account gets blocked.

On the rare occasion that I see one lab leak supporter change their mind about something disproven, I also see many others still citing the same disproven arguments.

A wise or sensible person might eventually conclude that lab leak is not a theory that you can argue with.

I'm neither wise, nor sensible, so I instead decided to bet \$100,000 against it.

Intermediate genomes

There are 787 near-full-length A or B genomes sampled by February 14, 2020.

There are also 20 genomes of intermediates: C/C or T/T

The intermediates can be excluded for a few different reasons.

Where do false intermediate genomes come from?

1. It's sometimes an issue of low read depth.

Pekar excluded one C/C genome from South Korea with low sequencing depth (< 10X) at position 28144 (it also shared three mutations with non-intermediates)

A T/T genome sampled in Singapore had low coverage at both 8782 and 28144 ($\leq 10\times$)

Three T/T genomes from Wuhan had low depth and indeterminate assignment at position 8782:

Table S1. Nucleotide variant calls at positions 8782 and 28144 for three SARS-CoV-2 genomes with intermediate T/T haplotypes¹.

GISAID accession	8782									28144								
	Depth	Count				Proportion				Depth	Count				Proportion			
		A	C	G	T	A	C	G	T		A	C	G	T	A	C	G	T
EPI_ISL_493179	64	0	39	1	24	0.000	0.609	0.016	0.375	61361	121	3784	195	57261	0.002	0.062	0.003	0.933
EPI_ISL_493180	40	0	24	1	15	0.000	0.600	0.025	0.375	95374	226	5709	293	89146	0.002	0.060	0.003	0.935
EPI_ISL_493182	29	0	10	0	19	0.000	0.345	0.000	0.655	69369	153	4051	232	64933	0.002	0.058	0.003	0.936

¹Variant calls and depths provided by Di Liu and Yi Yan.

Where do false intermediate genomes come from?

2. Early in the pandemic a few people around the world used a bioinformatics pipeline that called the reference base at positions with no read coverage.

This means that many lineage A genomes would have coverage at one position, but not the other, which would then get the B reference genotype, creating an intermediate.

One study in Sichuan, China found 12 C/C intermediates, but these were actually not full sequences, the extra positions were just filled in by that software.

If only one of the two positions gets sequenced,

Where do false intermediate genomes come from?

3. Pekar 2022 can exclude some intermediates based on shared mutations.

Those "intermediate" genomes share multiple other exact mutations of known lineage A viruses. So either there was perfect convergent evolution of those mutations in a hypothetical intermediate lineage and lineage A, or the "intermediate" genome was just lineage A with a miscall at one of the two sites. The first possibility is extremely improbable.

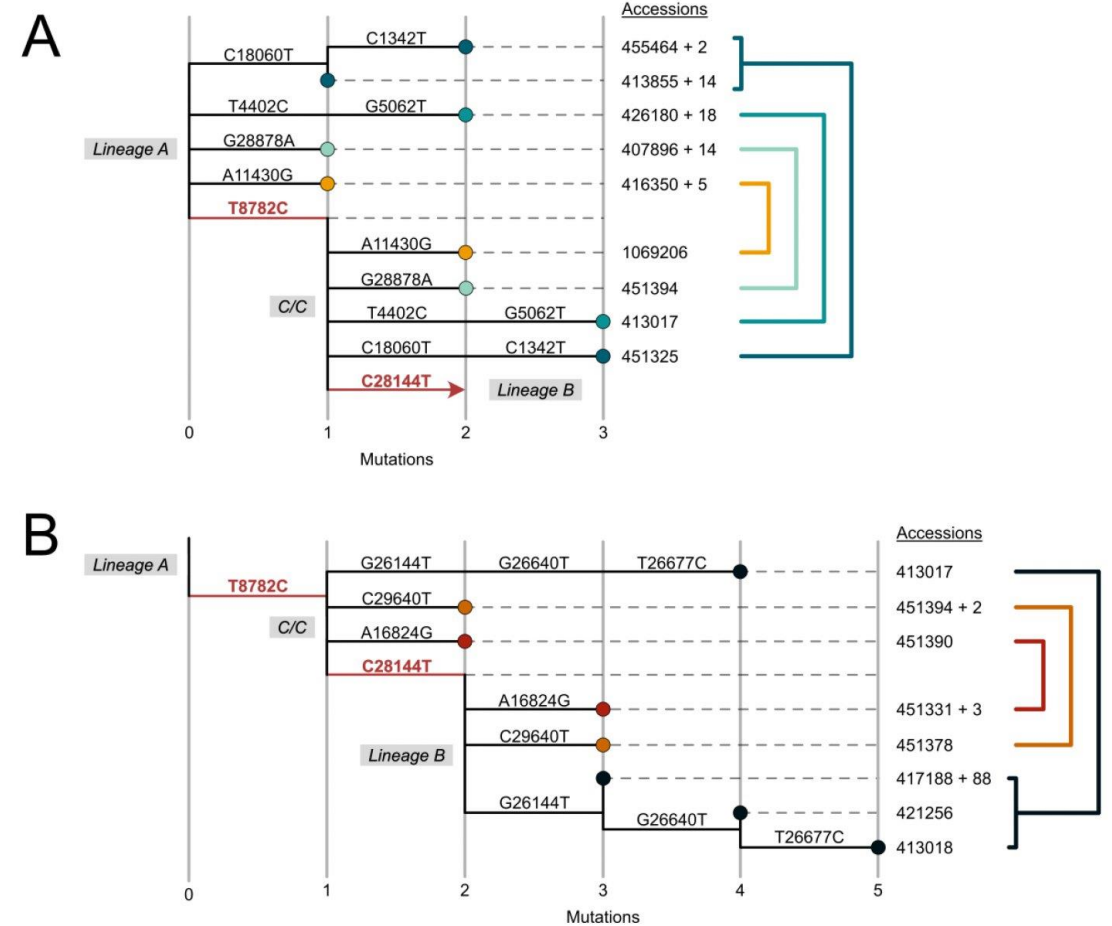


Figure 1. Phylogeny of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineages A and B. (A) Shared mutations across lineage A and C/C. (B) Shared mutations across lineage B and C/C. Mutations relative to the Hu-1 reference genome are shown above each branch. Lineage-defining mutations (8782 and 28144) are colored in red. Derived mutations not shared by both lineages are excluded. The taxon names are GISAID accession numbers, and the total number of additional matching homoplastic sequences are indicated. Sequences that share derived mutations are connected by the lines on the right, and brackets indicate that a group of sequences share the derived mutations that cannot be individually resolved.

Washburne, Massey, and Yuri made a list of more genomes, they're mostly from the same Sichuan study:

Accession (GISAID and NCBI)	Intermediate genotype	Location	Sampling date
EPI_ISL_451351	C/C	Sichuan	27 Jan 2020
EPI_ISL_451332	C/C	Sichuan	30 Jan 2020
EPI_ISL_453783*	C/C	Wuhan	31 Jan 2020
EPI_ISL_451333*	C/C	Sichuan	1 Feb 2020
EPI_ISL_451317*	C/C	Sichuan	3 Feb 2020
EPI_ISL_451342*	C/C	Sichuan	11 Feb 2020
EPI_ISL_451355*	C/C	Sichuan	12 Feb 2020
EPI_ISL_451318	C/C	Sichuan	19 Feb 2020
EPI_ISL_454952	C/C	Wuhan	19 Feb 2020
EPI_ISL_454973	C/C	Wuhan	22 Feb 2020
EPI_ISL_455365	C/C	Wuhan	23 Feb 2020
EPI_ISL_455366	C/C	Wuhan	23 Feb 2020
EPI_ISL_455370	C/C	Wuhan	24 Feb 2020
OM065349	T/T	Lu'an, Anhui	30 Jan 2020

Table 1: A-B intermediate genomes. 13 C/C and 1 new T/T intermediate genome were identified. The asterisks denote those genomes that adhere to Pekar et al.'s inclusion criteria.

I looked these all up on [GISAID](https://gisaid.org).

They're all from the same Sichuan lab.

The same lab that Pekar already talked to and confirmed that their software filled in partial reads with Lineage B data.

Also, none of these are early genomes.

And most aren't even from Wuhan.

If C/C was the original virus, then you would see early C/C genomes in Wuhan. And you would see them across various labs and papers, not just at only one lab that used misleading software.

Acknowledgement of Data Contributors

EPI_ISL_451351

Virus name: hCoV-19/Sichuan/SC-PHCC1-030/2020

Collection date: 2020-01-27

Originating Lab: West China Hospital of Sichuan University

Submitting Lab: State Key Laboratory of Biotherapy of Sichuan University

Authors: Baowen Du, Minjin Wang, Chao Tang, Chuan Chen, Yongzhao Zhou, Mingxia Yu, Hancheng Wei, Weimin Li, Jing-wen Lin, Jia Geng, Binwu Ying, Lu Chen

Acknowledgement of Data Contributors

EPI_ISL_454952

Virus name: hCoV-19/Wuhan/HB-WH4-200/2020

Collection date: 2020-02-19

Originating Lab: Wuhan Chain Medical Labs (CMLabs)

Submitting Lab: State Key Laboratory of Biotherapy of Sichuan University

Authors: Baowen Du, Minjin Wang, Chao Tang, Chuan Chen, Yongzhao Zhou, Mingxia Yu, Hancheng Wei, Weimin Li, Jing-wen Lin, Jia Geng, Binwu Ying, Lu Chen

Acknowledgement of Data Contributors

EPI_ISL_454973

Virus name: hCoV-19/Wuhan/HB-WH5-229/2020

Collection date: 2020-02-22

Originating Lab: Wuhan Chain Medical Labs (CMLabs)

Submitting Lab: State Key Laboratory of Biotherapy of Sichuan University

Authors: Baowen Du, Minjin Wang, Chao Tang, Chuan Chen, Yongzhao Zhou, Mingxia Yu, Hancheng Wei, Weimin Li, Jing-wen Lin, Jia Geng, Binwu Ying, Lu Chen

Other scientists have already pointed out this exact problem:



Zach Hensel 

a year ago edited

Most of the C/C sequences discussed in this manuscript come from a single study (Lin et al 2021 DOI: 10.1016/j.chom.2021.01.015) that reports methods inconsistent with Washburne et al concluding that associated GISAID records represent complete, full-length sequences. For example, the very first sequence shown in Table 1 in Washburne et al, EPI_ISL_451351, corresponds to sample SC-PHCC1-030. Table S2 shows that this sample has only 89.4% coverage with at least 1 read and only 63.2% coverage with at least 10 reads. Yet, the associated GISAID record is full length with zero Ns. Clearly these are consensus Wuhan-Hu-1 genomes modified by detected variations, and this is confirmed in the manuscript by Lin et al that is cited by Washburne et al:

For Nanopore sequencing data, the ARTIC bioinformatics pipeline for COVID (<https://artic.network/ncov-...>) was used to call single nucleotide changes, deletions and insertions relative to the reference sequence. The final consensus genomes were generated for each sample based on the variants called in each position.

This is not limited to Sichuan sequences, but also to Wuhan samples from the same study.

Furthermore, Table 1 in Washburne et al includes a sample that was, in fact, considered in Pekar et al. EPI_ISL_453783 is a second record for EPI_ISL_452363 (identical sample ID, patient age, sampling date, and sequence).

Multiple authors of this manuscript have promoted their claimed discovery of new intermediate genomes on social media for the past several weeks and have been repeatedly been informed of these and other errors in their claims and have yet to make any corrections.

Massey might have tipped his hand to some of the deception here, he [changed the font for one of these samples](#):



Steve Massey
@stevenemassey

Sure, here is a table

I don't know why Pekar et al missed / didn't consider these intermediates - they've sitting right there in the database ! Certainly the five that conformed to their inclusion criteria should have been picked up in their search

ID	Intermediate genotype	Location
EPI_ISL_451355	C/C	Sichuan
EPI_ISL_451333	C/C	Sichuan
EPI_ISL_451317	C/C	Sichuan
EPI_ISL_451342	C/C	Sichuan
EPI_ISL_451318	C/C	Sichuan
EPI_ISL_451351	C/C	Sichuan
EPI_ISL_451332	C/C	Sichuan
EPI_ISL_453783	C/C	Wuhan
EPI_ISL_454973	C/C	Wuhan
EPI_ISL_454952	C/C	Wuhan
EPI_ISL_455370	C/C	Wuhan
EPI_ISL_455366	C/C	Wuhan
EPI_ISL_455365	C/C	Wuhan
OM065349	T/T	Anhui

6:29 AM · Sep 21, 2022



Zach Hensel
@alchemytoday

Is the different font for EPI_ISL_453783 because you briefly thought it was a good idea to not include the same sample twice and then changed your mind and decided to keep it anyway?

	Passage de	Accession ID
DT-WH04/2020	Original	EPI_ISL_453783
DT-WH04/2020	Original	EPI_ISL_452363

10:52 AM · Sep 21, 2022

← (EPI_ISL_453783 = EPI_ISL_452363, which was already considered in Pekar's paper, but it had two sample numbers and Massey decided to cite the other number)

Where do false intermediate genomes come from?

4. Only C/C or T/T could be the intermediate.

DRASTIC often claims examples of both.

T/T is the most unlikely because of C->T mutation bias.

C/C is more likely, but Pekar's model finds that a C/C ancestor is highly unlikely to cause 2 equal polytomies.

A20 sample

Stages of Denial for A20 sample:

- It's not important
- It's only 1 sample
- ~~It turned from lineage B into lineage A during culturing~~
- It's fake
- The glove came from the lab

Known Lab leak explanations for the lineage A sample: It's meaningless, it's fake, or "the glove came from the lab"



@mbalter — investigations and commentary
@mbalter



Is anyone seriously suggesting that the finding of lineage A on a glove in the market says something about Covid origins? I guess so, but hope not.

9:07 AM · Apr 6, 2023 · 48 Views



Dr Steven Quay ✓
@quay_dr



The latest market data is missing the A20 specimen, a glove with the only Lineage A virus in the market.

Why is it missing?

Probably because I showed over a year ago the specimen was bogus.



youtube.com
The Huanan Market Origin of SARS-CoV-2 is Unlikely
The Huanan Market Origin of SARS-CoV-2 is Unlikely: The ancestral lineage-containing specimen appears to have ...

6:05 AM · Mar 21, 2023 · 1,212 Views



Florin
@Florin_Uncovers



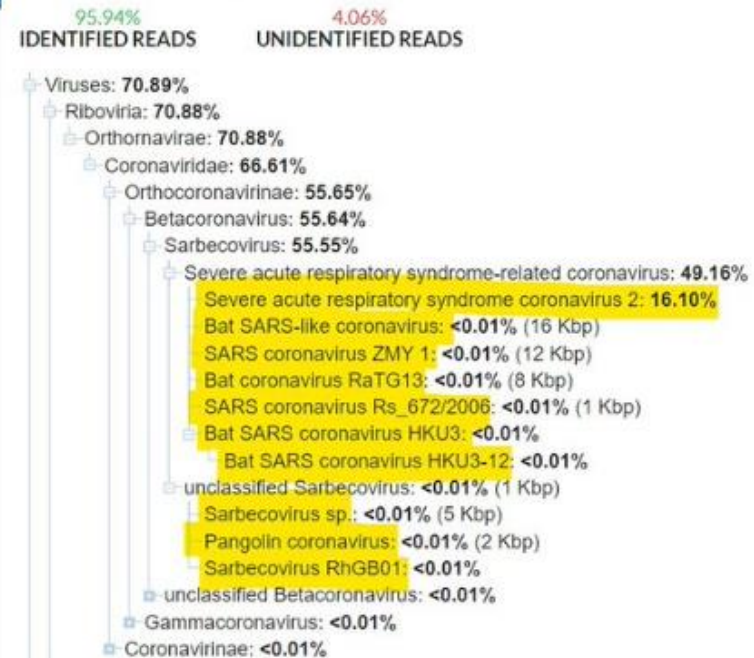
Sample A20 is from a vendor's glove at Huanan on Jan 1 2020. Beside SARS-CoV-2 it has bits of: RaTG13, pangolin SARS-L, HKU3, ZMY1, Rs672, RhGB01 & 2 unidentified SARS-like viruses!
Either this is Shi Zhengli's glove or it indicates A20 is a lab contamination! Sorry Worobey! 😊



RNA sequencing of total nucleic acids from environmental swabs for viral whole-genome assembly (Env_0020_seq01) (SRR23971533)

Metadata Analysis Reads Data access FASTA/FASTQ download

Taxonomy Analysis



8:20 AM · Mar 30, 2023 · 37.2K Views

Yuri says that it's only 1 of 69 positive samples.
That's misleading because only 5 were sequenced fully.



Yuri Deigin ✓
@ydeigin



1. Only one sample out of 69 positive SARS2 samples from the market is lineage A.

Some people argued that vendor never used gloves.
Babar [found a picture](#) of gloves in his shop:



babar
@babartelephant



I guess some gloves there.



2:56 AM · May 12, 2023 · 190 Views

Rootclaim was confused on how many samples were sequenced:

What does 4 refer to? There were 923 samples collected

Many samples were PCR+, but most don't have full coverage of the SARS2 genome:

Lab code	Sequencing run	SARS2 paired read count	SARS2 covered bases	Sample type
A20	SRR23971533	10197098	29516	sars2_amplicon
B5	SRR23971484	2121910	29641	sars2_amplicon
F54	SRR23971416	1921583	29640	sars2_amplicon
F13	SRR23971473	510287	29789	sars2_amplicon
F13	SRR23971591	263085	29784	sars2_amplicon
F13	SRR23971580	197351	29683	market_metagenome
F54	SRR23971582	46392	29680	market_metagenome
F100	SRR23971579	3235	2302	market_metagenome
B5	SRR23971573	689	25994	market_metagenome
B17	SRR23971572	183	11672	market_metagenome
A61	SRR23971510	148	12311	market_metagenome
F98	SRR23971583	137	12386	market_metagenome
F46	SRR23971581	103	6602	market_metagenome
A18	SRR23971505	102	6594	market_metagenome
A15	SRR23971504	83	3811	market_metagenome
A55	SRR23971509	79	7951	market_metagenome
D32	SRR23971574	69	5550	market_metagenome
A87	SRR23971567	55	805	market_metagenome
E7	SRR23971578	35	3770	market_metagenome
A101	SRR23971502	28	1123	market_metagenome
E61	SRR23971576	28	3785	market_metagenome
A2	SRR23971506	26	3067	market_metagenome
A63	SRR23971512	24	2629	market_metagenome
A20	SRR23971507	22	2101	market_metagenome
A33	SRR23971508	14	1698	market_metagenome
RLC-3	SRR23971517	13	1328	other_metagenome
A90	SRR23971570	10	1538	market_metagenome
A88	SRR23971569	8	1156	market_metagenome
Q68	SRR23971451	7	301	market_metagenome
E48	SRR23971575	5	830	market_metagenome

Market samples: I said at the last debate that only 4 were sequenced. After checking, it's actually five samples: F13, F54, B5, A61, and A20.

But there are some which are partially sequenced and you can guess the genome if they have a few reads which are at either position 8,782 or 28,144.

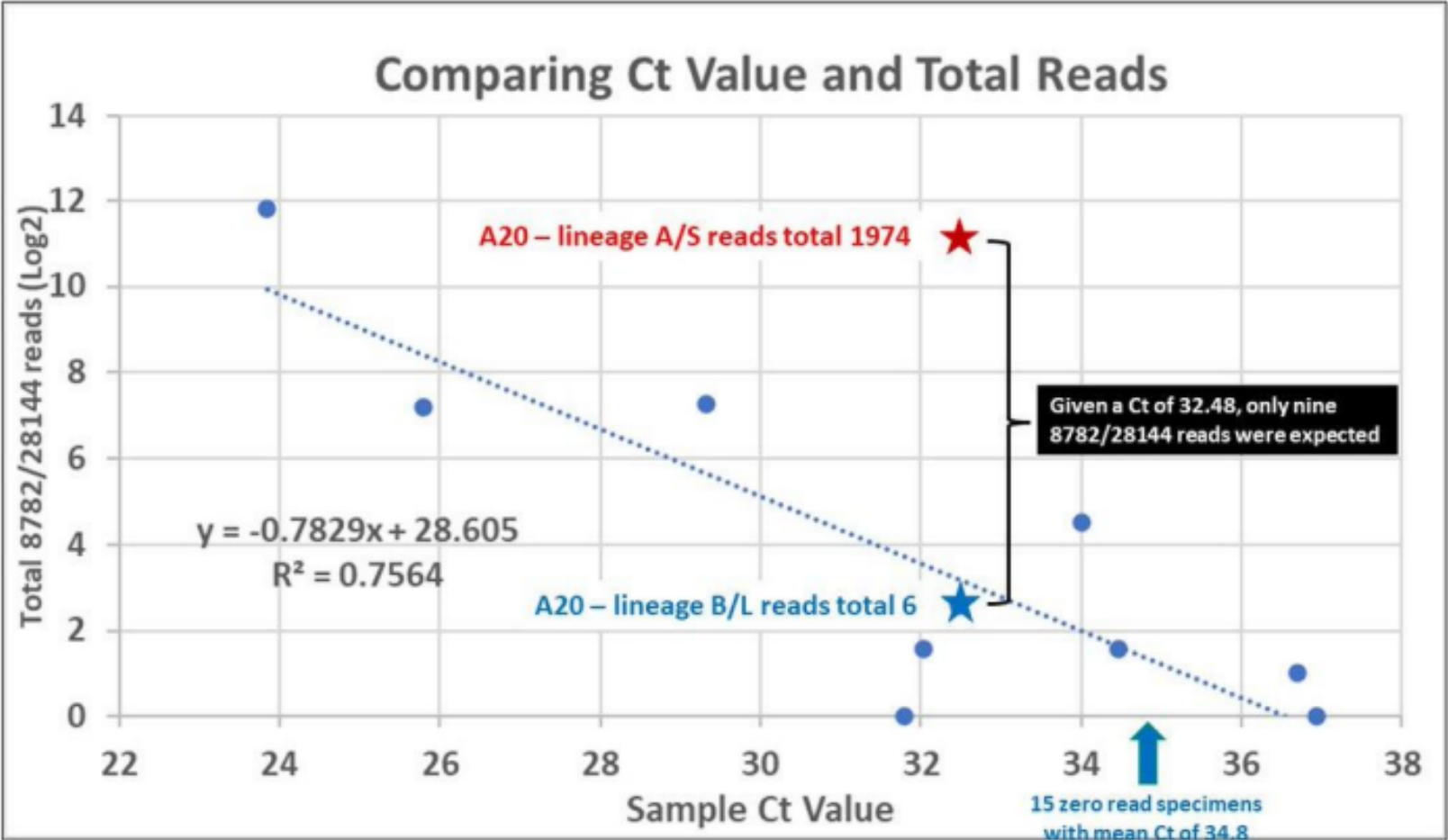
Most of those have only a few reads, so it's hard to say for sure.

But one of them looks like it's lineage A.

Overall, 2 out of 11 samples are lineage A, 9 are lineage B.

C	D	E	F	G	H	I	J	K	L	M	N
bases_covered_covern	8782 C	8782 T	28144 C	28144 T	Env_name	Sample ID	Lab code	Sampling date	Sampling location	Street No.a	Vendor No.a
29845	814	0	0	1508	Env_0313	Env_0313	F13	01/01/20	West Wine of HSM	11	15
29827	180	0	2	333	Env_0354	Env_0354	F54	01/01/20	West Wine of HSM	2	14
26666	3	0	0	7	Env_0126	Env_0126	B5	01/01/20	West Wine of HSM	5	6-8
15127	1	0	0	0	Env_0213	Env_0213	D32	01/01/20	West Wine of HSM	15	15
14930	0	0	0	1	Env_0398	Env_0398	F98	01/01/20	West Wine of HSM	4	X6-X4
14199	2	0	0	3	Env_0061	Env_0061	A61	01/01/20	West Wine of HSM	7	20-22-24
13078	0	0	0	1	Env_0138	Env_0138	B17	01/01/20	West Wine of HSM	15	X44
9534	0	0	0	3	Env_0055	Env_0055	A55	01/01/20	West Wine of HSM	7	25
6869	0	2	0	0	Env_0346	Env_0346	F46	01/01/20	West Wine of HSM	2	24
4277	2	0	0	0	Env_0063	Env_0063	A63	01/01/20	West Wine of HSM	7	16-18

Steven Quay made this diagram showing that he thinks the A20 read counts are too high for the PCR cycle threshold

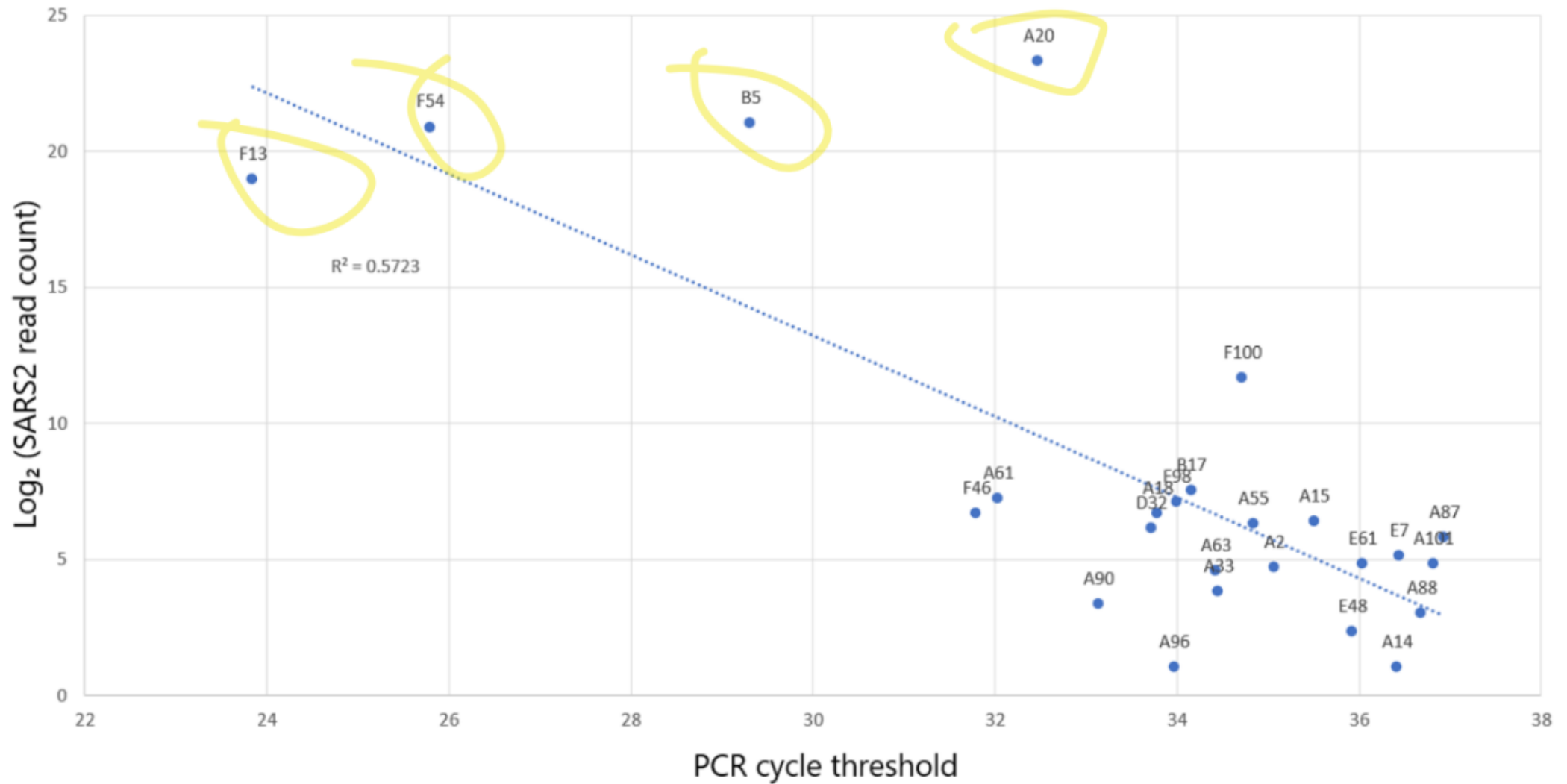


Supp. Fig. 38. Log₂ of the total sequencing depth at 8782/28144 in all sequenced Huanan market environmental samples plotted against the qRT-PCR Ct value of each sample. After Quay (2022).

I tried to reproduce Steven Quay's diagram, but I used total SARS2 reads, not just reads at those 2 positions.

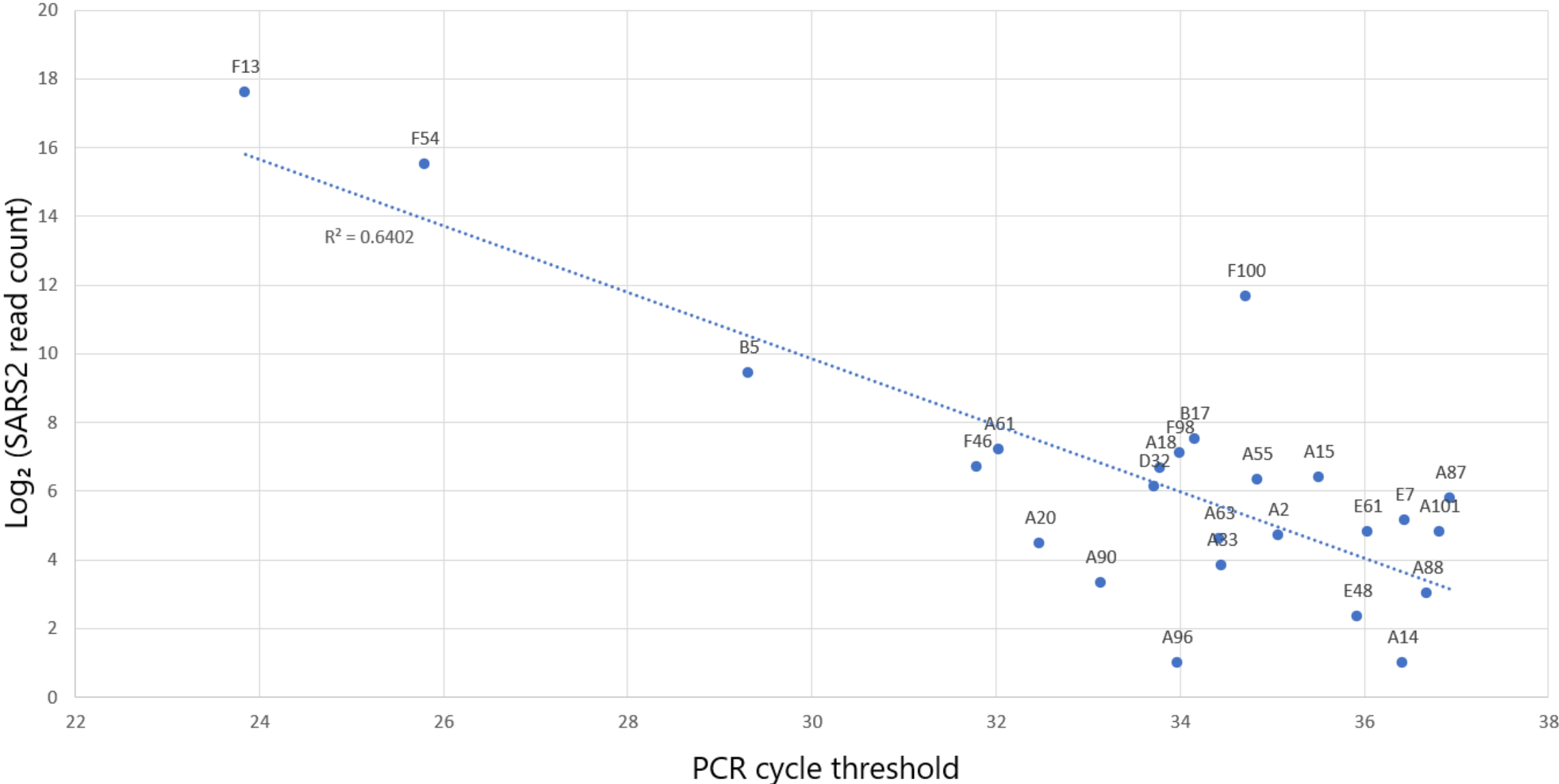
4 points stand out with lots more reads than the others.

Those points stand out because [they were sequenced with a different method](#) (SARS2 amplicon sequencing)

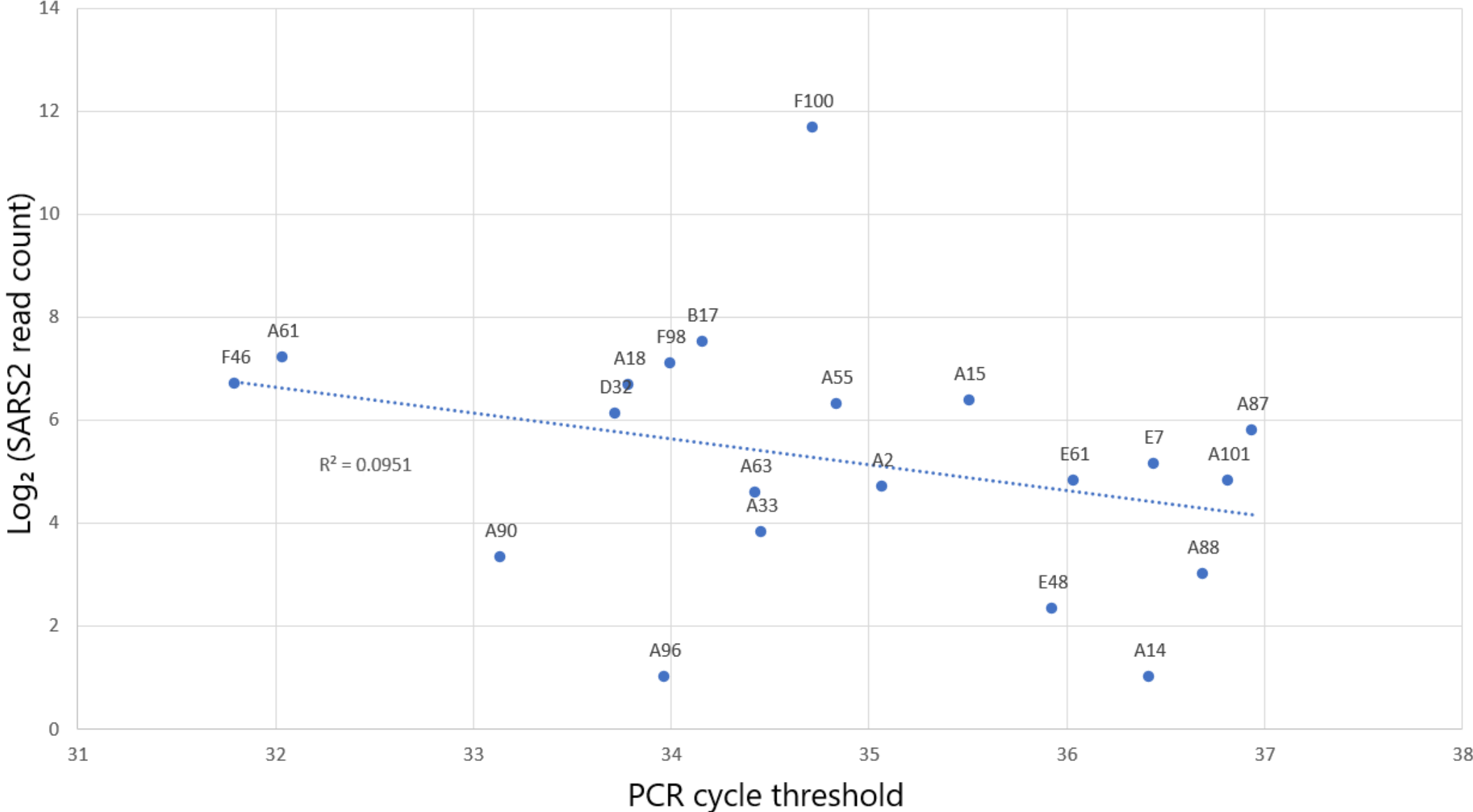


I searched around a little more and found that those 4 samples had also been sequenced the same way.

If you use an apples to apples comparison, A20 now fits right in the trendline.



Incidentally, the R^2 value gets very low if you take out the 3 samples with a lower cycle threshold, so I'm not entirely convinced that this method of analysis is robust, to begin with. But Quay's mistake is obviously that he compared samples sequenced with different methods.



Sample A20 has 2 mutations: C6145T and G26262T

Quay suggested these mutations might also be suspicious in some way, perhaps contamination.

Sample A20 carried 2 additional mutations: C6145T and G26262T (Supp. Fig. 39). Both mutations have been found in isolates of SARS-CoV-2 in humans, whereas while C6145T is of uncertain ancestry due to it being a hypervariable site in Sarbecoviruses (both C and T

I scanned through the first 787 other covid genomes and zero of them had these 2 mutations.

Quay also suggested that these mutations could have arisen during culturing or sequencing of the sample:

It is worth mentioning that the sample F54 accumulated two mutations compared to the original at the third passage in VERO E6 cells. We cannot rule out the possibility, however unlikely, that contamination by cultured SARS-CoV-2 sequences within the same laboratory during the sequencing of sample A20 in 2021 could have led to the appearance of mutations C6145T and G26262T within the final assembled genome. Access to raw data is important to confirm all samples.

That's possible, but seems harmful to the lab leak theory, since it places the lineage A root at the market. In other circumstances, [Quay has argued](#) that the earliest lineage A root case was very important, because that was detected at a hotel near the market, not at the market itself.

For the record, it looks like the A20 sample was not cultured:

204 We further performed high-throughput sequencing (Supplementary Table 3) and
205 successfully obtained seven complete or near complete SARS-CoV-2 genome
206 sequences, including three sequences from three environmental samples (Env_0313,
207 Env_0354 and Env_0020), and four sequences from cell supernatants of Env_0313,
208 Env_0354 and Env_0126 (Fig. 3, Supplementary Table 4). A few samples were re-
209 sequenced using a multiplex PCR approach, including Env_0020_seq01,

Env_0313 = F13

Env_0354 = F54

Env_0126 = B5

Env_0020 = A20

On more of a meta level, one hint that Quay may be a bad faith actor is the way he presents himself



**The Five Undisputed
Facts Favoring a
Lab Origin of COVID**

Steven C. Quay, MD, PhD, FCAP

May 24, 2021

Steven@DrQuay.com

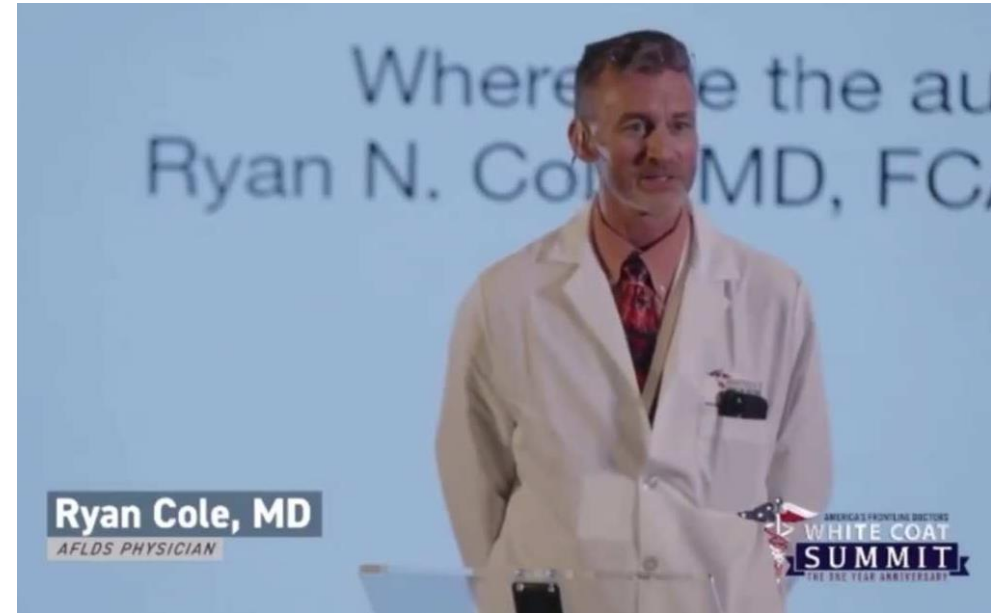
www.DrQuay.com



Anti-vaxxers also use white coats and microscopes as props to look like reputable doctors



“Ivermectin is a wonder drug with miraculous effectiveness against COVID-19”



“Covid-19 vaccine Causes Turbo Cancer”



The hydroxychloroquine people also used white coats as props:



And this practice has been copied in other countries. Here's an [Eastern European anti-vax protest](#). Their protest name translates to “frosty silence of white coats”.



Yuri lacks any skepticism for the sick WIV researchers claims



Yuri Deigin 
@ydeigin



🌟🌟🌟 Holy shit. If true, this is THE SMOKING GUN.

Loss of smell and ground glass lung opacities in WIV researchers in November 2019?! This will raise my estimate of a lab leak to 99.999%.



hotair.com

Josh Rogin: The sick researchers from the Wuhan Institute of Virology lost their sense of smell

8:17 AM · Aug 24, 2021

Yuri was slightly more cautious when Ben Hu was named as patient zero:



Yuri Deigin @ydeigin

...

If the latest leak is true and Ben Hu was indeed one of the 3 WIV staffers with Covid-like symptoms in Nov 2019, it would be VERY suspicious, as Ben Hu was named on the infamous DEFUSE grant that proposed creating novel furin cleavage sites in SARS-like CoVs:

Yuri Deigin @ydeigin · Mar 24, 2022

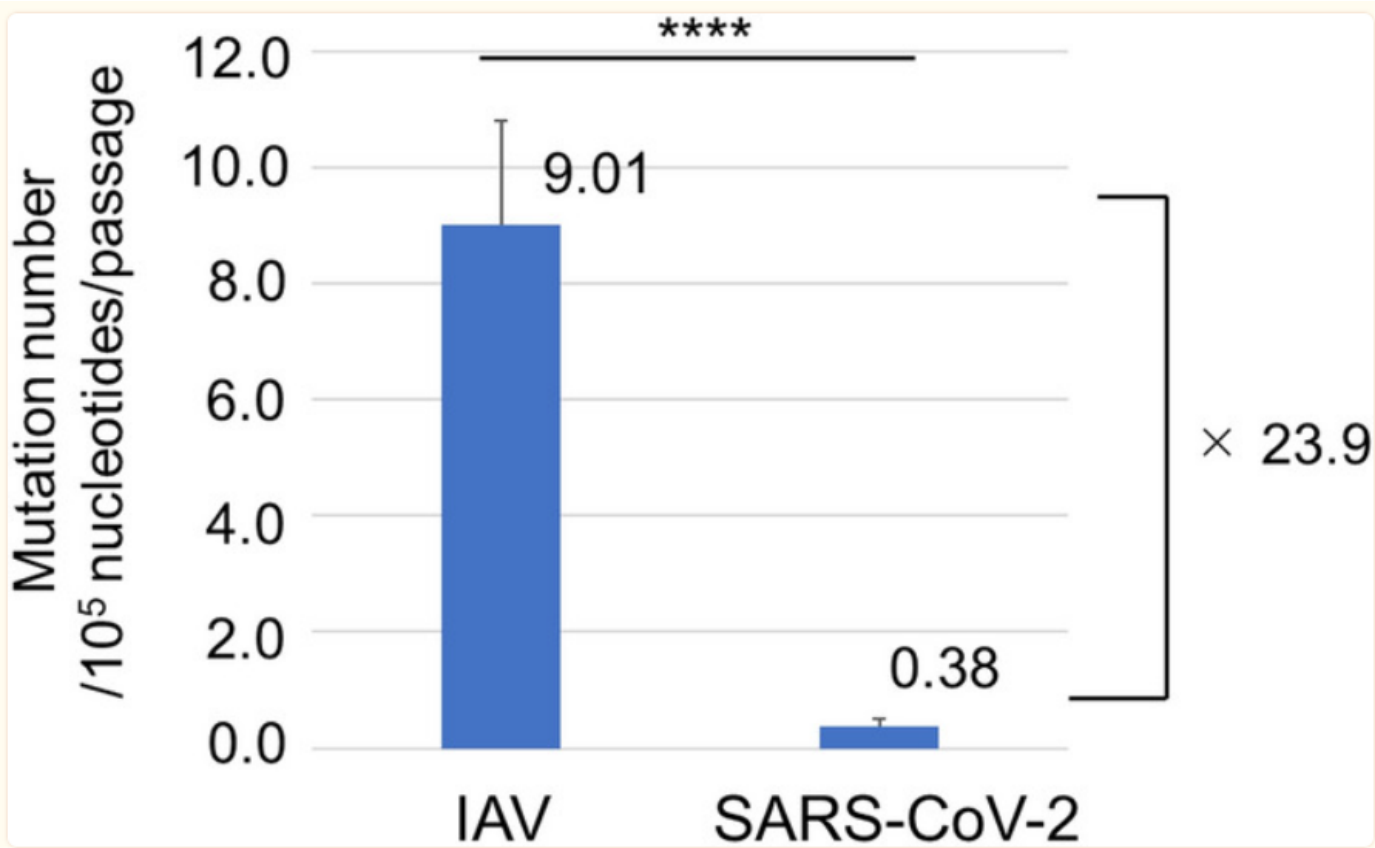
An interesting detail: the management plan for the infamous DEFUSE proposal named Ben Hu from the WIV side. In 2019 he received a Chinese grant to investigate two new SARS-like CoVs in humanized mice. Still implausible to think Ben might've thought to insert an FCS?

The screenshot shows a document titled "DEFUSE proposal" with a project management chart. The chart is organized into several sections: "Administration", "TAL: Host-Pathogen Prediction", and "Wuhan Institute of Virology, Pathogen Discovery". Under "Administration", Dr. William Karesh is listed as the lead. Under "TAL: Host-Pathogen Prediction", Dr. Peter Daszak is the lead, with sub-teams for "Host-pathogen dynamics" (led by Dr. Kevin Olival and Dr. Noam Ross) and "Predictive models" (led by Dr. Carlos Zambrana-Torrealo, Dr. Alice Lattin, and Toph Allen). Under "Wuhan Institute of Virology, Pathogen Discovery", Dr. Zhengli Shi is the lead, with sub-teams for "Host immunity" (led by Dr. Peng Zhou and Dr. Ben Hu) and "Virology" (led by Dr. Jon Epstein, Dr. Leticia Gutiérrez-Jiménez, Dr. Guangjian Zhu, and Dr. Yunzhi Zhang). The text to the right of the chart discusses funding and Ben Hu's role, with several phrases highlighted in yellow: "addition to sub-grants from Health, research at the WIV was supported by Chinese funding. Ben Hu researcher at the WIV, was awarded a one-year grant from the Youth Science and Technology Award for a project to investigate pathogenicity of Two New Bat SARS-like Coronaviruses to Transgenic mice Expressing Human ACE2 receptor."⁷³ Hu has been a member of [unclear]'s group at the WIV since 2015.⁷⁴

2:58 PM · Jun 13, 2023 · 16.1K Views

Clean Insertion of the Furin Cleavage Site

[Kawasaki et al 2023](#) cultured Influenza A and SARS-CoV-2 in Calu-3 cells, and found something interesting: Influenza mutates 23 times faster than SARS-CoV-2, per passage:



The authors write:

“These variants are produced through replication errors of the viral genome by viral RNA-dependent RNA polymerase (RdRp)...

The mutation rate of SARS-CoV-2 was 23.9-fold lower than that of IAV because of the proofreading activity of the SARS-CoV-2 RdRp complex.”

But they also found that that SARS-CoV-2's proofreading mechanism does not work for insertions and deletions. Those are equally common for influenza and SARS-CoV-2.

“There was no significant difference in the frequency of indels between IAV and SARS-CoV-2...

Our results revealed that the fidelity of SARS-CoV-2 genome replication was 23.9-fold higher than that of IAV. This higher fidelity of the SARS-CoV-2 RdRp complex is thought to be mainly due to the proofreading activity of the 3'-to-5' exoribonuclease activity of the viral protein, nsp14.

In contrast, there was no significant difference in the frequencies of indels between IAV and SARS-CoV-2, suggesting that SARS-CoV-2 does not have a special mechanism to prevent insertion and deletion in its genome replication and the process works as well as that in IAV.”

That might be a simple explanation for why the furin cleavage site looks inserted.

Mink evolution

Since Yuri and I disagreed on this, I did a quick review of all the mink evolution literature:

Table. Early evolutionary rates of SARS-CoV-2 in mink vs. humans				
Study	Host	Country	subst/site/year	mutations/year
Lu et al. (2021) , <i>Nature Communications</i>	Mink	Netherlands (Cluster A)	1.41×10^{-3} (95% HPD of 1.2×10^{-3} to 1.75×10^{-3})	42.2 (35.8 to 52.3)
	Mink	Netherlands (Clusters A-E)	7.9×10^{-4} (95% HPD of 7.2×10^{-4} to 8.4×10^{-4})	23.6 (21.5 to 25.1)
Porter et al. (2023) , <i>Virus Evolution</i>	Mink	Netherlands	1.83×10^{-3} (95% HPD of 1.3×10^{-3} to 2.41×10^{-3})	54.7 (38.9 to 72.1)
	Mink	Denmark	2.43×10^{-4} [95% HDP of 1.76×10^{-4} to 3.17×10^{-4}]	7.3 (5.3 to 9.5)
Tan et al. (2022) , <i>Nature Communications</i>	Mink, deer, and humans	Denmark, Latvia, Netherlands, and Poland	$\sim 6.45 \pm 0.4 \times 10^{-4}$	$\sim 19.3 \pm 1.2$
McBride et al. (2023) , <i>Nature Communications</i>	Human	China	1.3×10^{-3} (95% HPD of 1.1 to 1.6×10^{-3})	38.9 (32.9 to 47.8)
Li et al. (2020) , <i>Journal of Medical Virology</i>	Human	China	1.19 to 1.31×10^{-3}	35.5 to 39.2
Chaw et al. (2020) , <i>Journal of Biomedical Science</i>	Human	Worldwide	2.4×10^{-3} (95% HDP of 1.5×10^{-3} to 3.3×10^{-3})	71.7 (44.9 to 98.7)

Yuri's values?

6.59×10^{-3}

198

Yuri cited the rate from [Porter et al 2023](#), which cites several very different numbers.

I haven't read the paper well enough to understand the range they're giving.

But it's pretty clear that he picked the highest possible value you can find in the literature, which is a clear outlier from the rest of the published research.

Table 3.

Estimates generated from local clock (FLC) models with a gamma prior on the clock rate. Estimates include the evolutionary rates (substitution/site/year) estimated for the whole phylogeny, and the Netherlands and Denmark foreground branches. The 95 per cent HPD interval is shown in brackets.

Model	Estimated evolutionary rate (mean)	Netherlands evolutionary rate	Denmark evolutionary rate
FLC (stem*)	4.54×10^{-4} [4.13×10^{-4} , 4.93×10^{-4}]	1.83×10^{-3} [1.3×10^{-3} , 2.41×10^{-3}]	2.43×10^{-4} [1.76×10^{-4} , 3.17×10^{-4}]
FLC (shared, stem*)	4.78×10^{-4} [4.36×10^{-4} , 5.2×10^{-4}]		6.59×10^{-3} [3×10^{-3} , 1.05×10^{-2}]

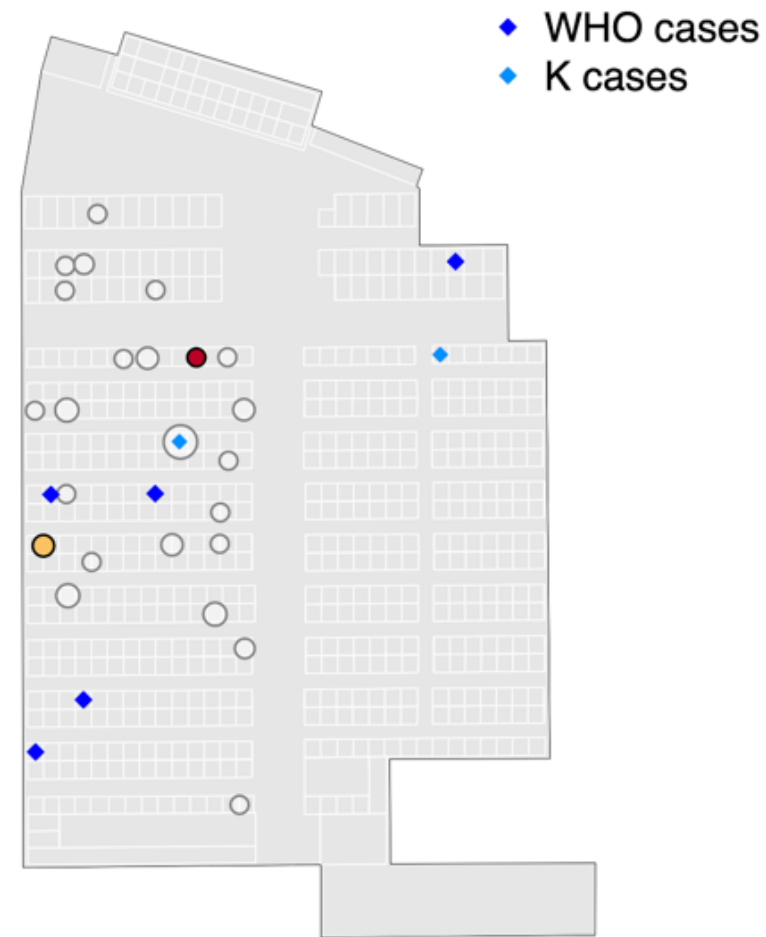
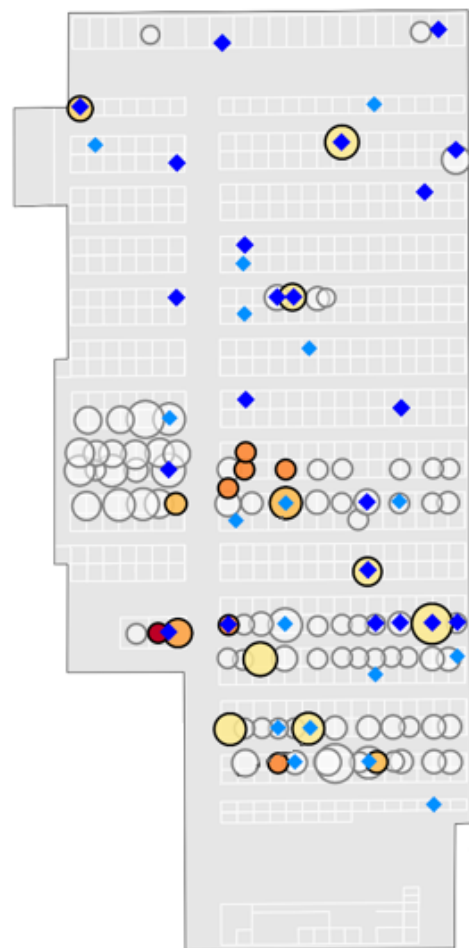
[Open in a separate window](#)

Sampling maps

01 January samples

Liu et al 2023 data includes positive and negative sample numbers

Jan 1st sampling focused on stalls with known cases and blocks near these cases



- ◆ WHO cases
- ◆ K cases

Proportion of positive samples



Number of samples



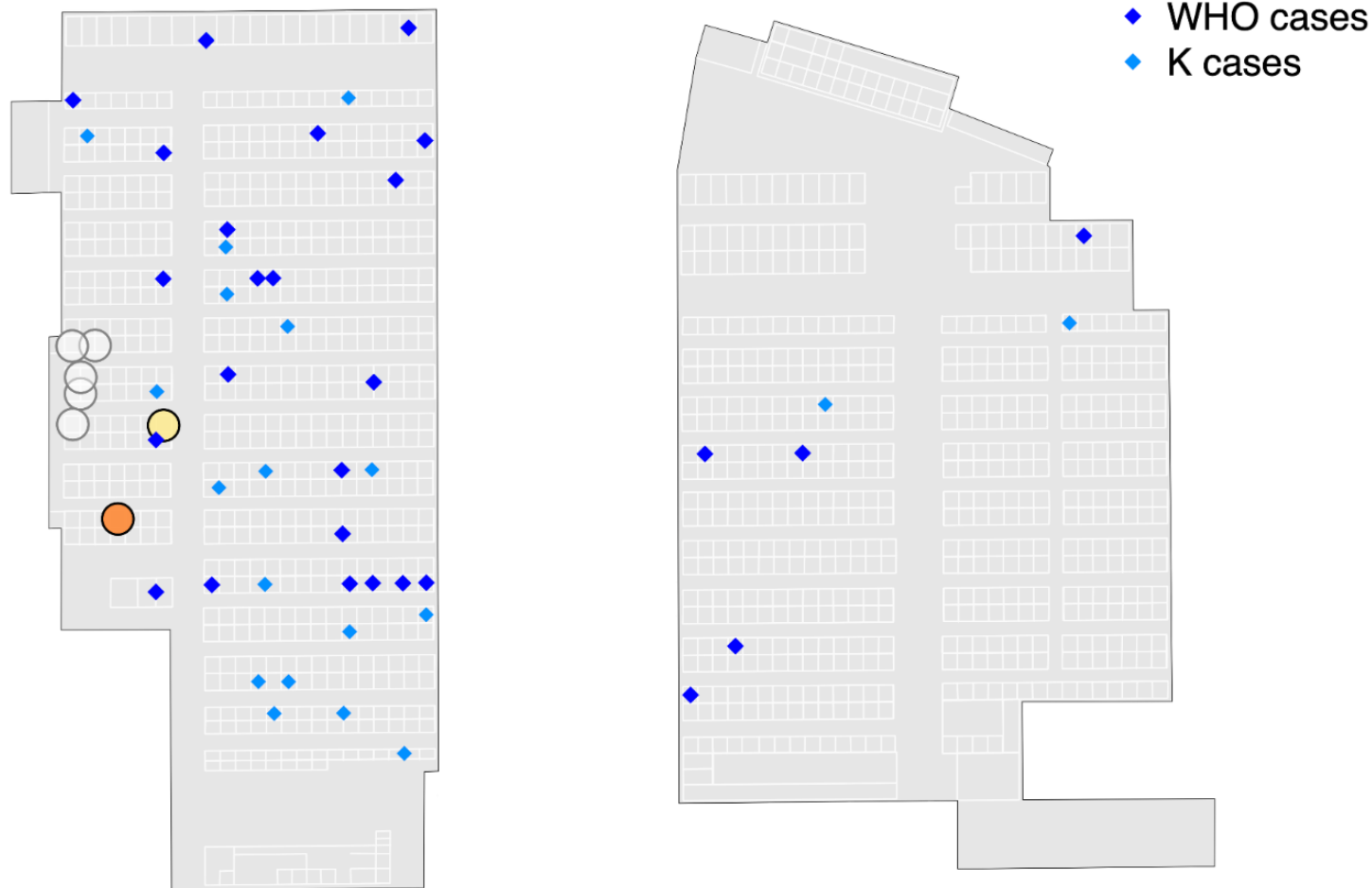
Jan 12th testing focused on the 7 wildlife shops

2 shops tested positive

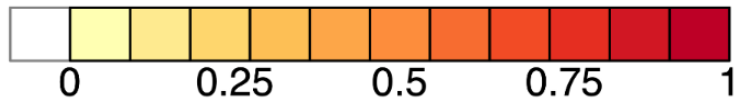
6-29: raccoon dog shop, 5 positive samples. 2 positives taken later from water drains.

8-25: hedgehog shop, 1 positive sample this day. More positive samples taken in February. Positive warehouse samples associated with this shop.

12 January samples



Proportion of positive samples



Number of samples



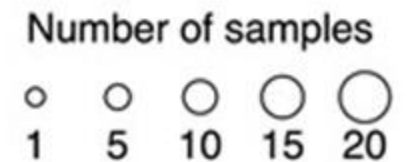
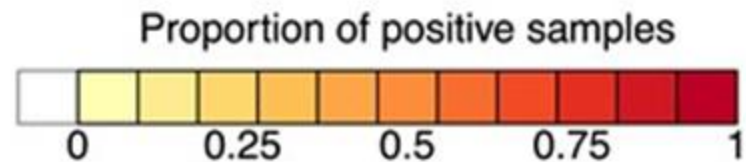
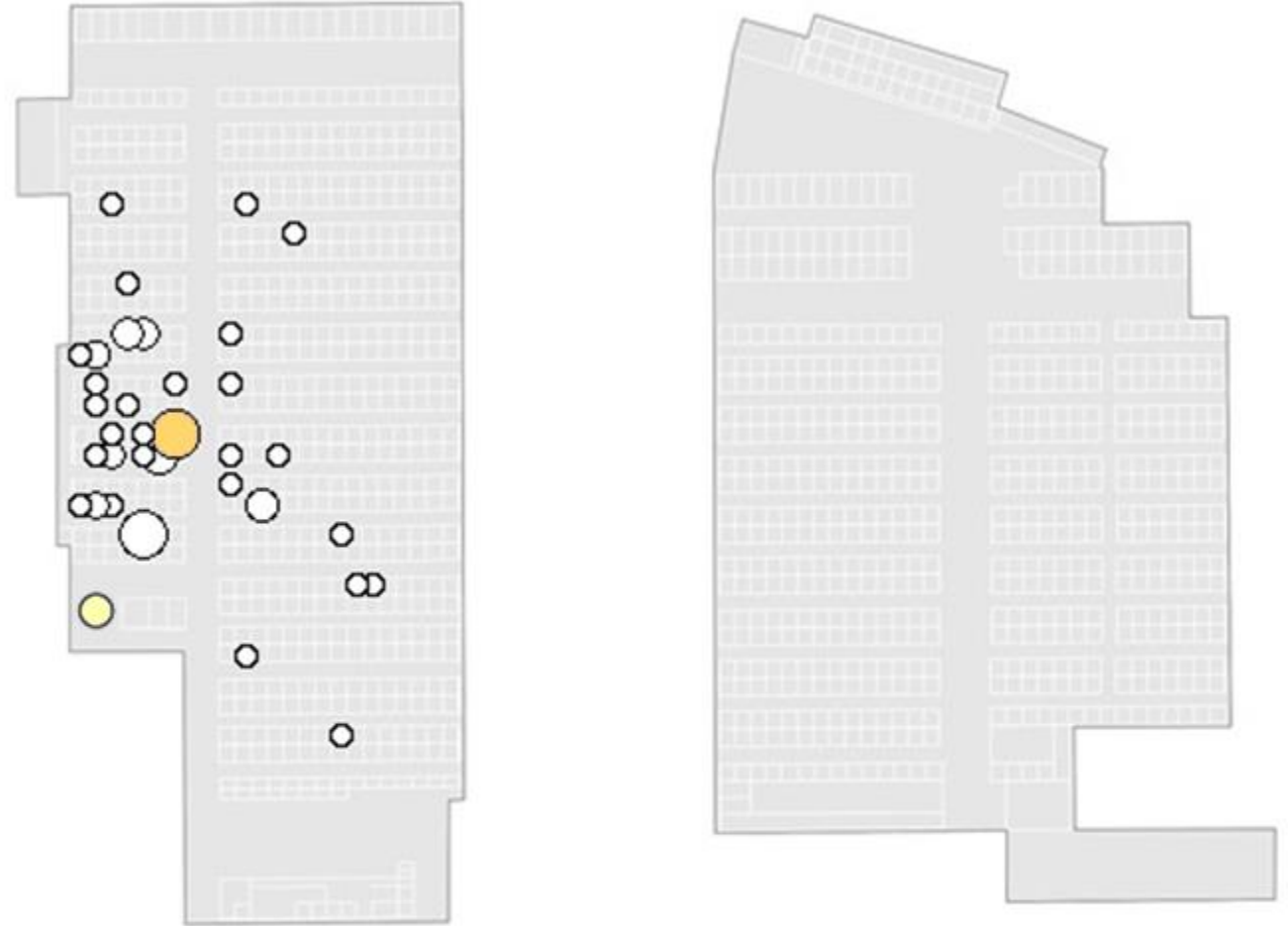
Jan 23rd to Mar 2nd

A number of shops were retested, with an emphasis on 6/29 and 8/25.

Shop 6/29 retests were negative after Jan 12th (but the 6/29 drains were still positive until February 15th)

6 positive samples in Shop 8/25. Tests were positive until Feb 15th.

One other positive test in the market: 5th street stairs between floor 1 and 2. That could be stairs up to other shops or it could be the stairs up to the Mahjong room. Samples within the Mahjong room itself were negative.



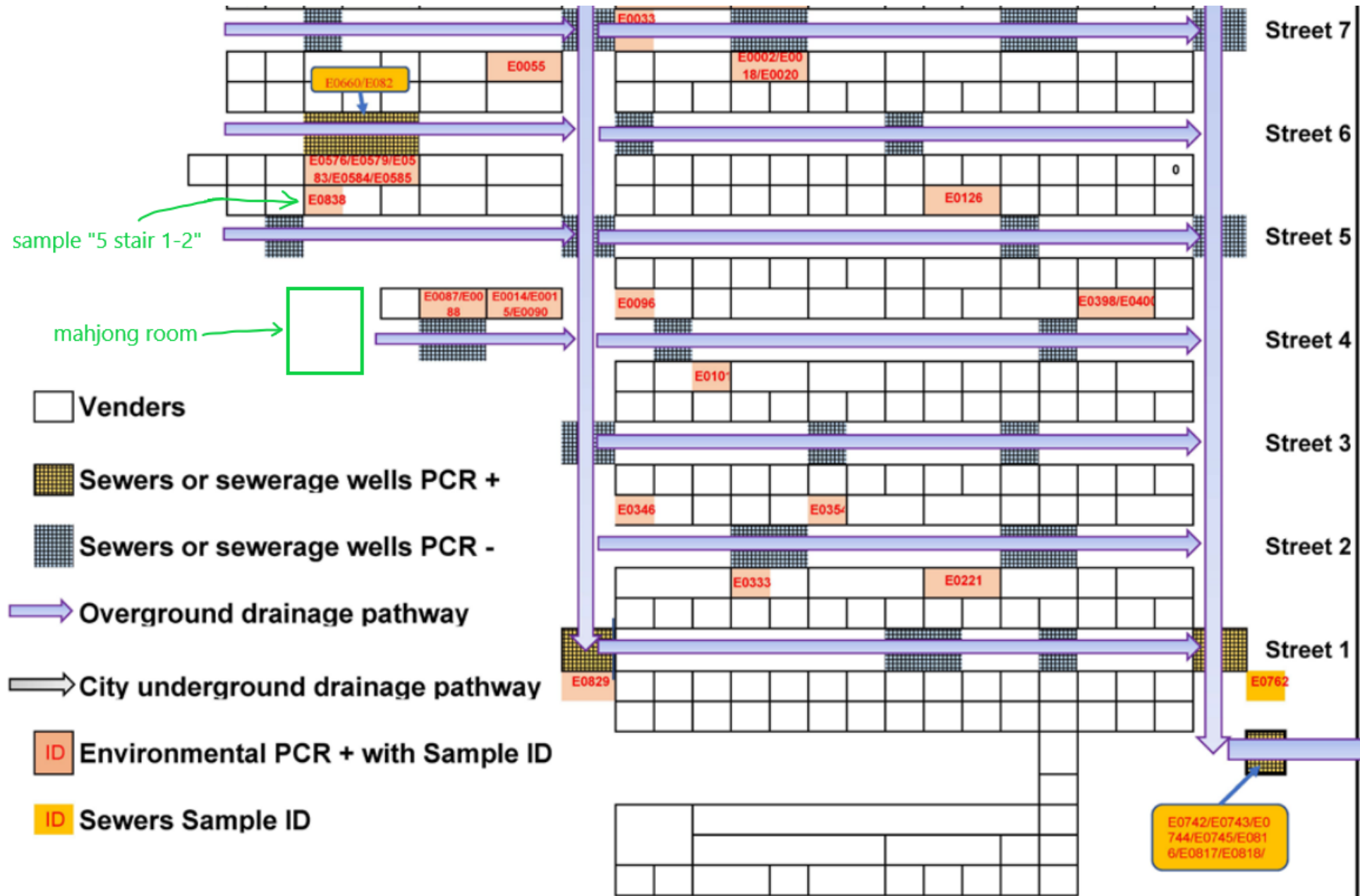
Sample "5 stair1-2":

Some people think that's the stairs to the mahjong room, others think it's [stairs up to the second floor](#).

A store called Eyeglass city, on the second floor, was still open after the market downstairs was closed.



[Liu et al 2023](#) says that the staircase sample is not the mahjong room stairs, but online opinions vary.

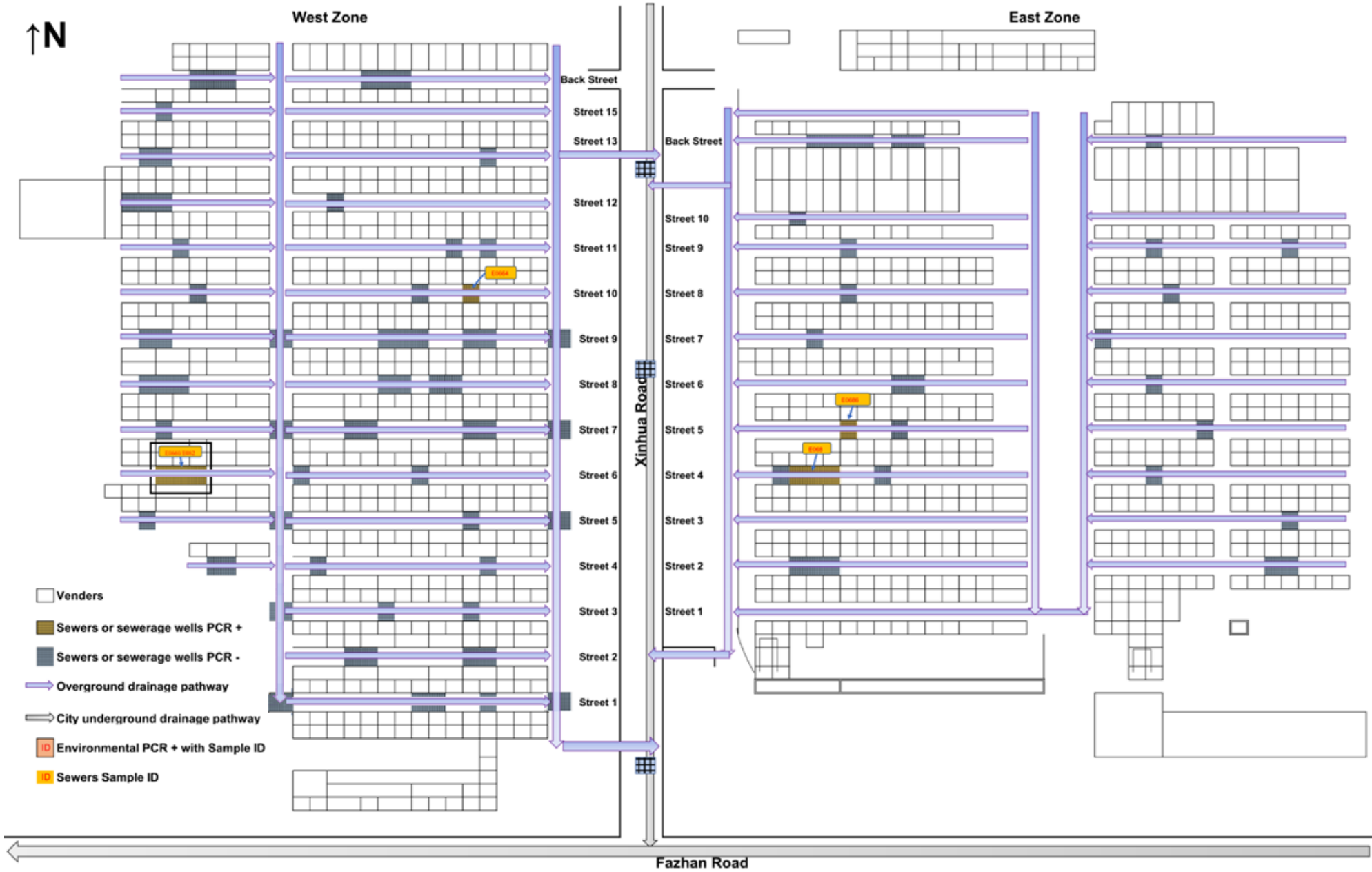


Drain sampling in the market points to shop 6/29

Jan 27th - 29th

4 out of 60 drains test positive.

One is shop 6-29.



Drain sampling in the market points to shop 6/29

Feb 9th - 15th

3 of 17 drains test positive: only shop 6-29 drain and two places downstream.

(these two downstream drains were not sampled in January)



Inconsistency with saying whether or not spatial distribution is important



Yuri Deigin  @ydeigin · Nov 10, 2022



Replying to @ydeigin

And no, if WIV or some other lab was the source of the virus, it is not logical to claim that the spatial distribution of cases would be centered around it — despite what **Worobey** and Rasmussen have claimed:



Yuri Deigin  @ydeigin · Oct 30, 2022

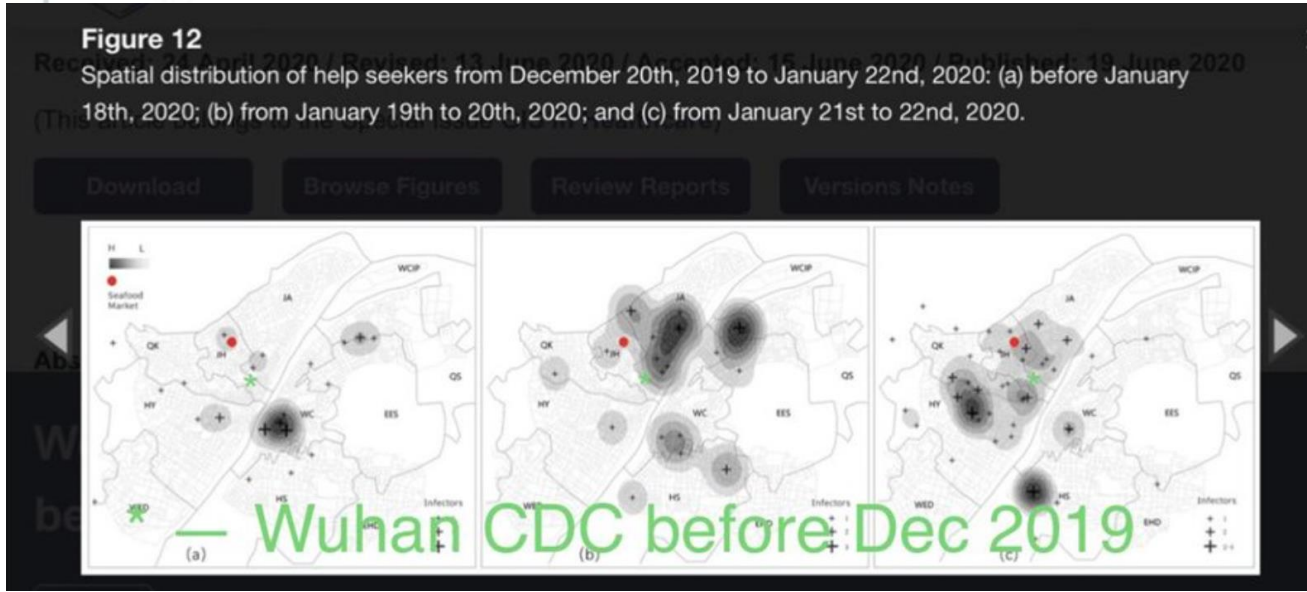
Replying to @ydeigin

Another logical fallacy that Worobey and Rasmussen commit in their follow-up popular article is claiming that if WIV was the source of the virus, one should also expect early SARS2 cases to spatially center around it:

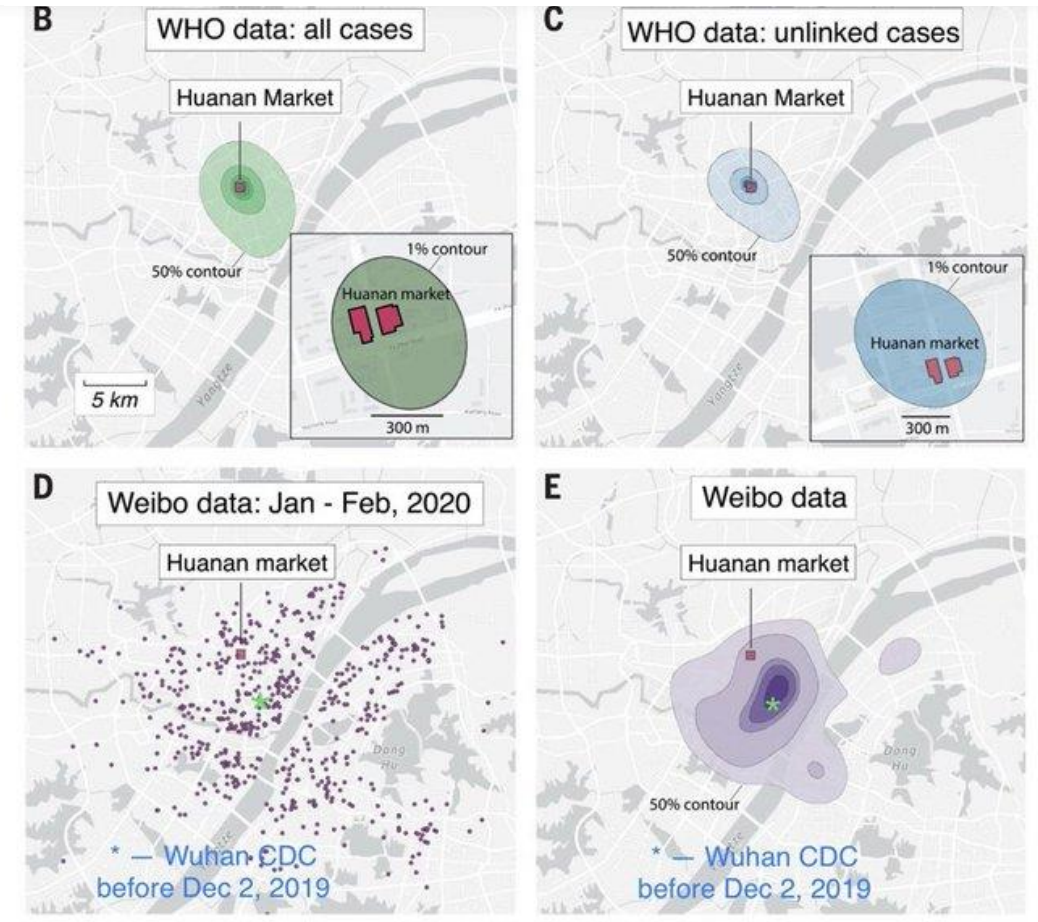
7/n

But people use the Weibo data to point to the WIV and also to point to the “old location of the Wuhan CDC”

Yuri Deigin @ydeigin · Mar 5
12/n
Looking at more granular temporal slices of the Jan-Feb Weibo data, the old Wuhan CDC location remains a better fit for the Worobey et al. “centroid hypothesis” than the Huanan market:
mdpi.com/2220-9964/9/6/...



Yuri Deigin @ydeigin
10/n
Moreover, based on the Worobey et al. “centroid hypothesis” that early cases must cluster around the outbreak source, the original Wuhan CDC location is actually a great candidate, as it sits right at the epicenter of the Jan-Feb 2020 Covid cases self-reported on Weibo:

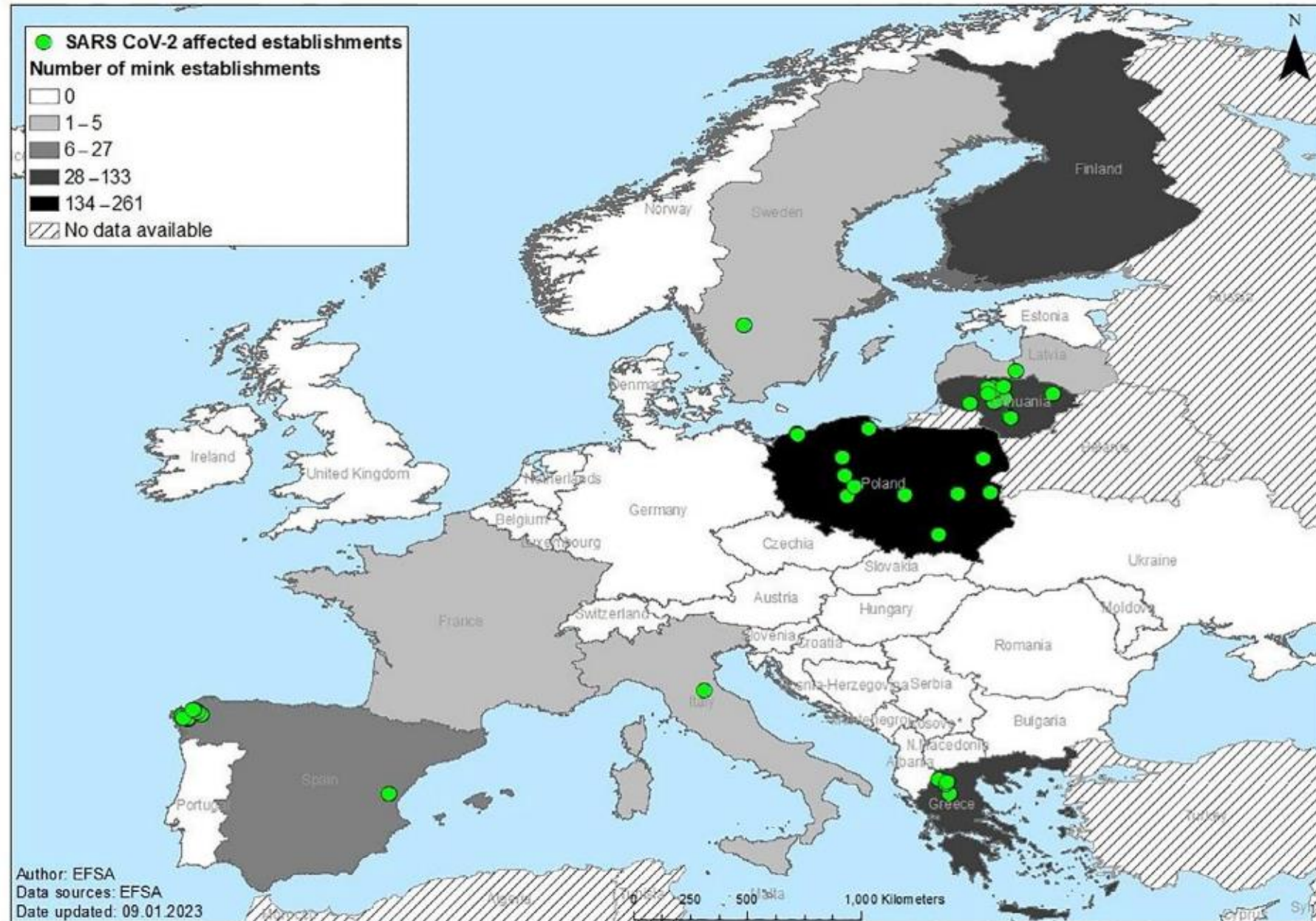


And they also made a big deal about that December 8th case near the lab, before he ended up on December 16th.

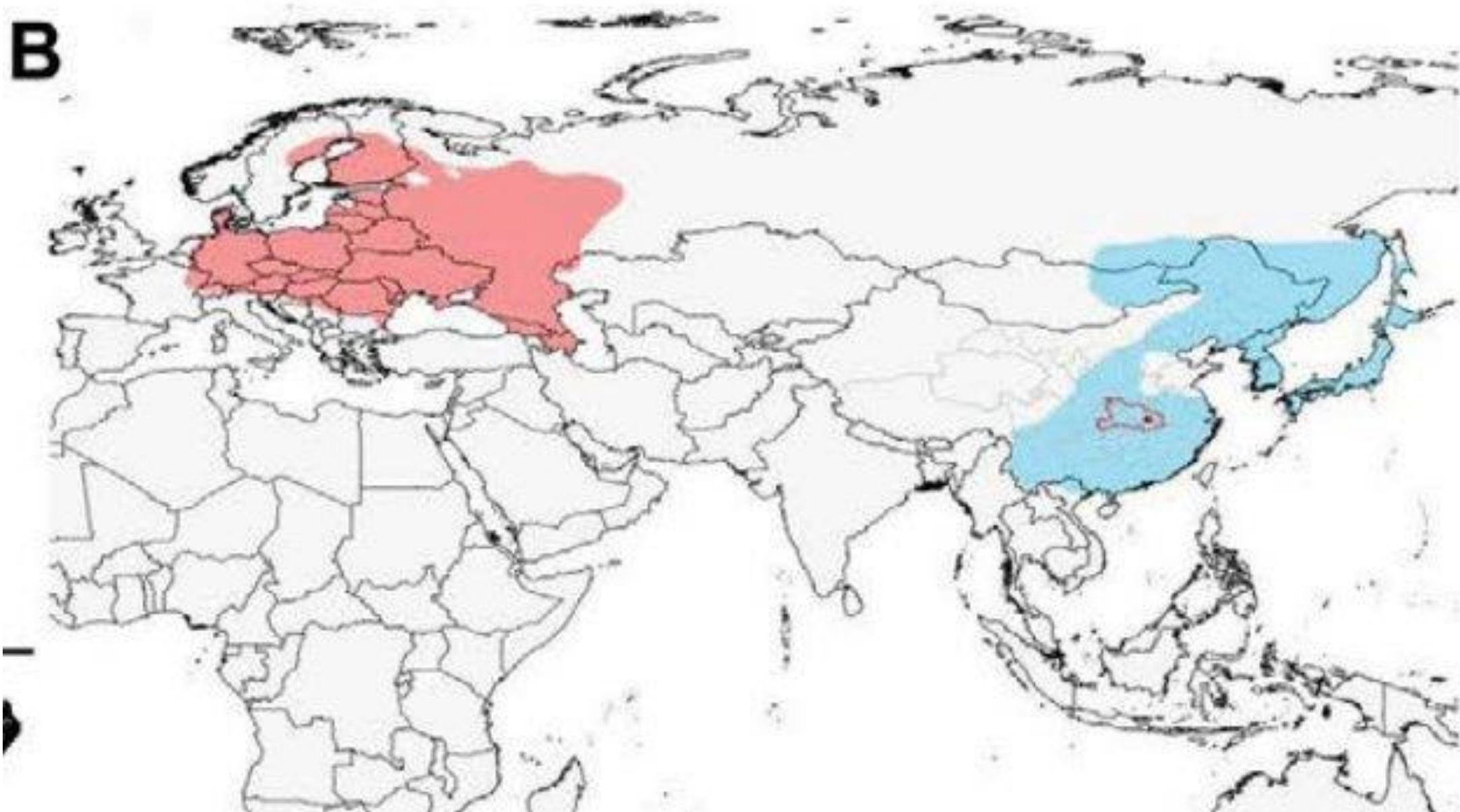
Fur Farms

Outbreaks on [fur farms in the EU](#), 2021 and 2022.

There was one reported outbreak on a Polish raccoon dog farm, in 2021.



Raccoon Dog population map:

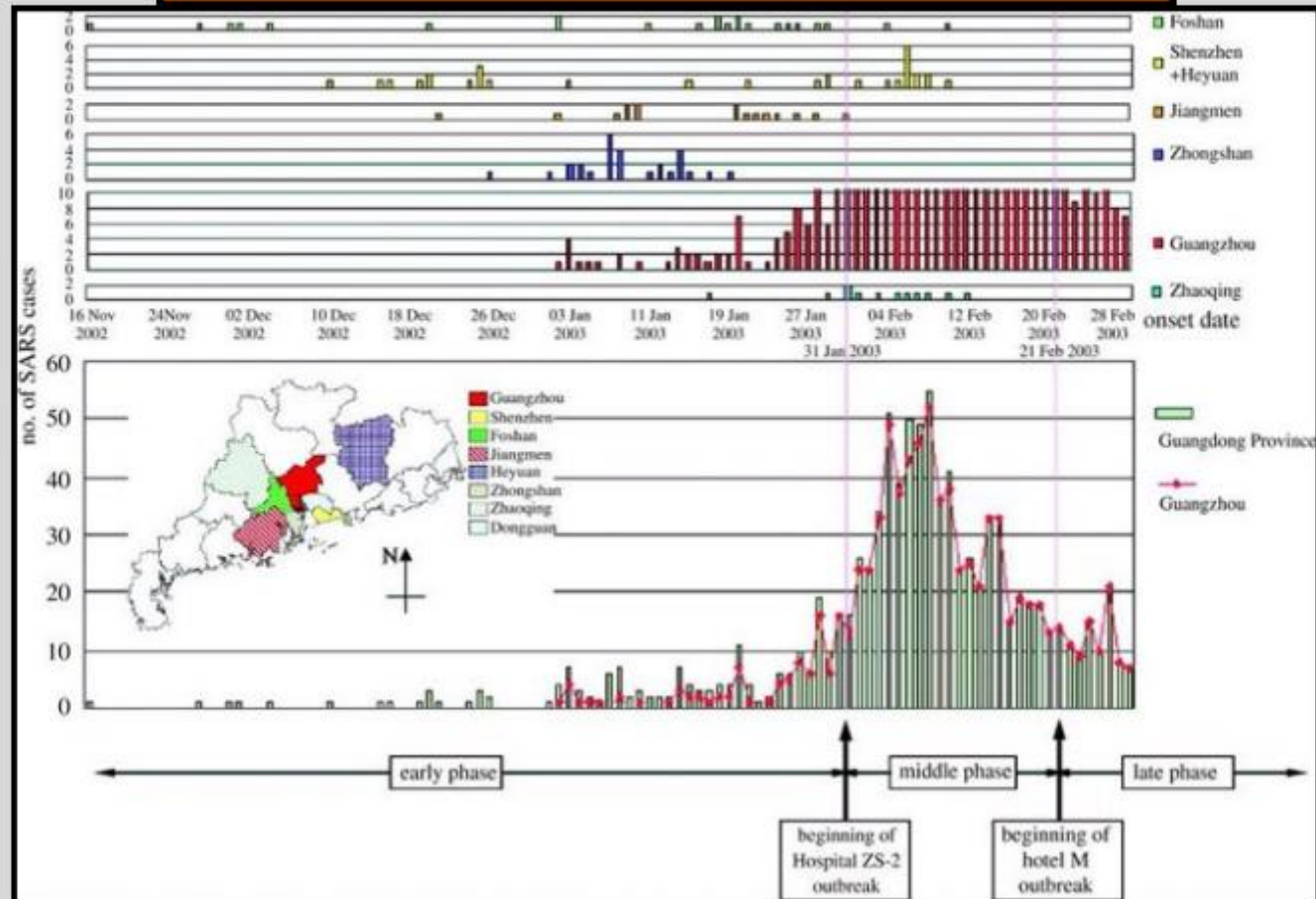


Comparing to SARS

Rootclaim asks why SARS1 made it to multiple cities and Covid only spilled over in Hubei

Comparing to SARS1

SARS Infections Over Time



This is what SARS would look like, with the speed of China's 2020 response:

It wouldn't even be in another town by the time the market closed, it would barely be in 4 by the Wuhan lockdowns.

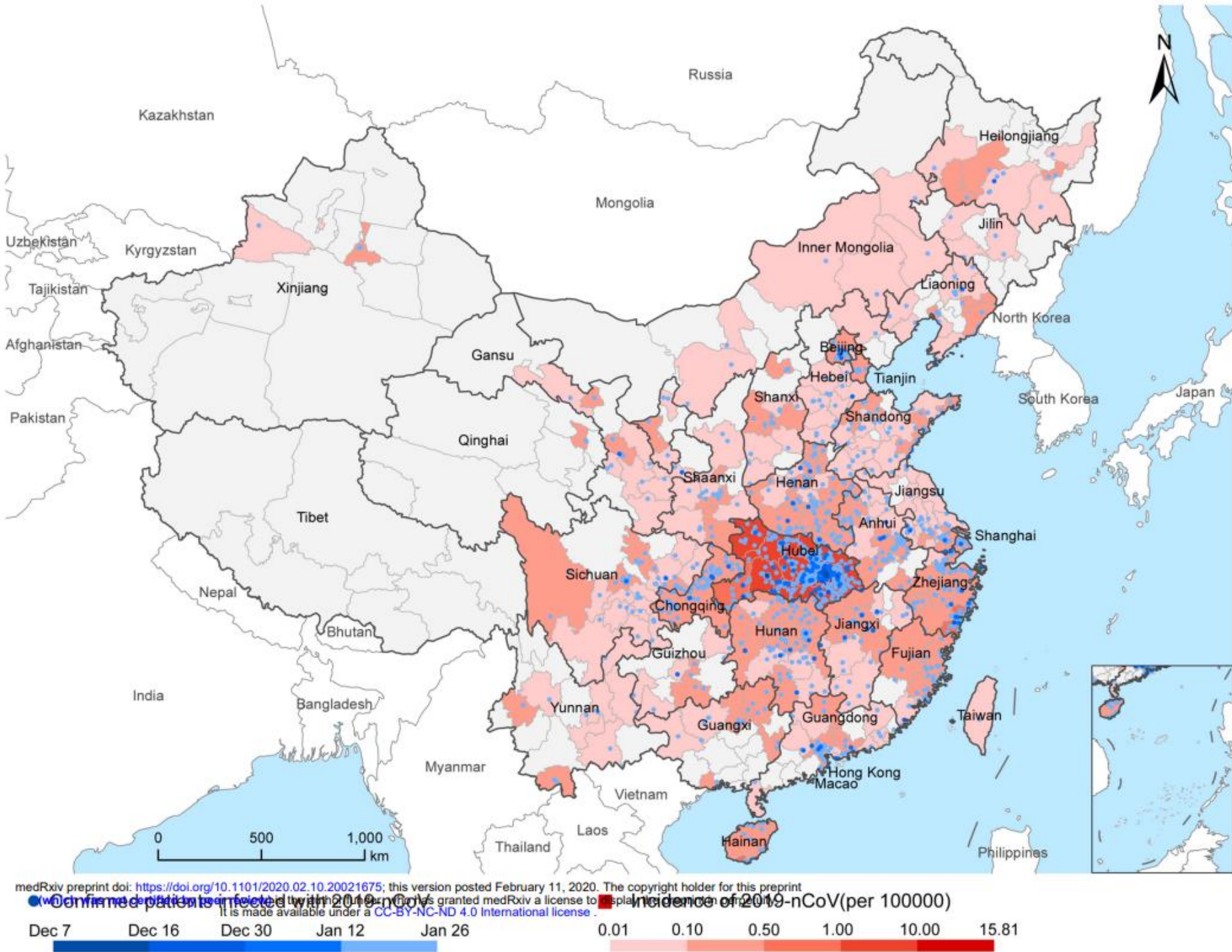


Here's what actually happened with Covid during that time period.

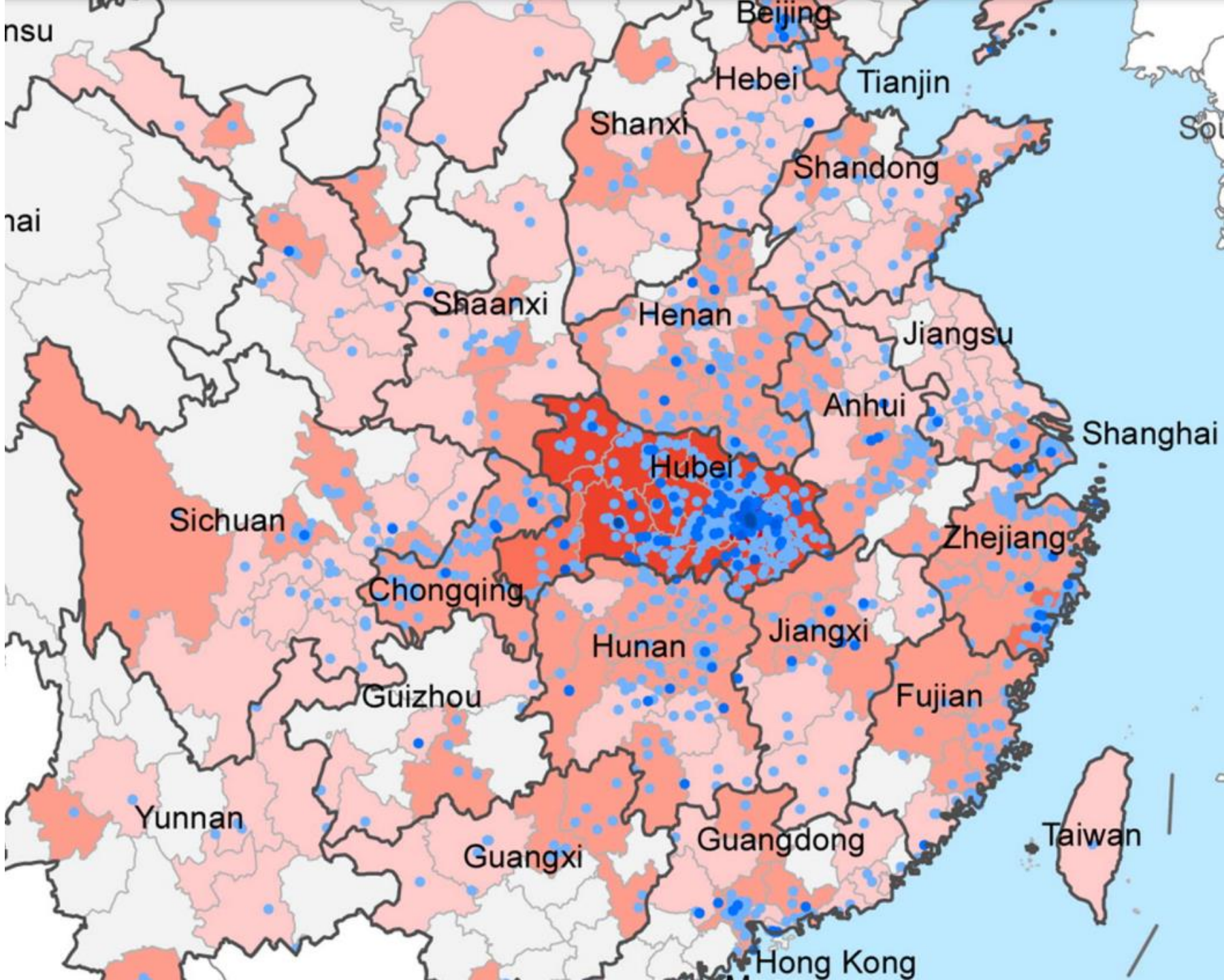
It's all over China because it's so contagious.

If there was a second, smaller spillover 4-6 weeks later, would you even notice that?

Figures from [Yang et al, 2020](#)



Zoomed in a bit for detail



Mapped along with the railway network.



Days from disease onset of the first patient in each city to Dec.7,2019

- 0
- 1-10
- 11-20
- 21-30
- 31-40
- 41-50

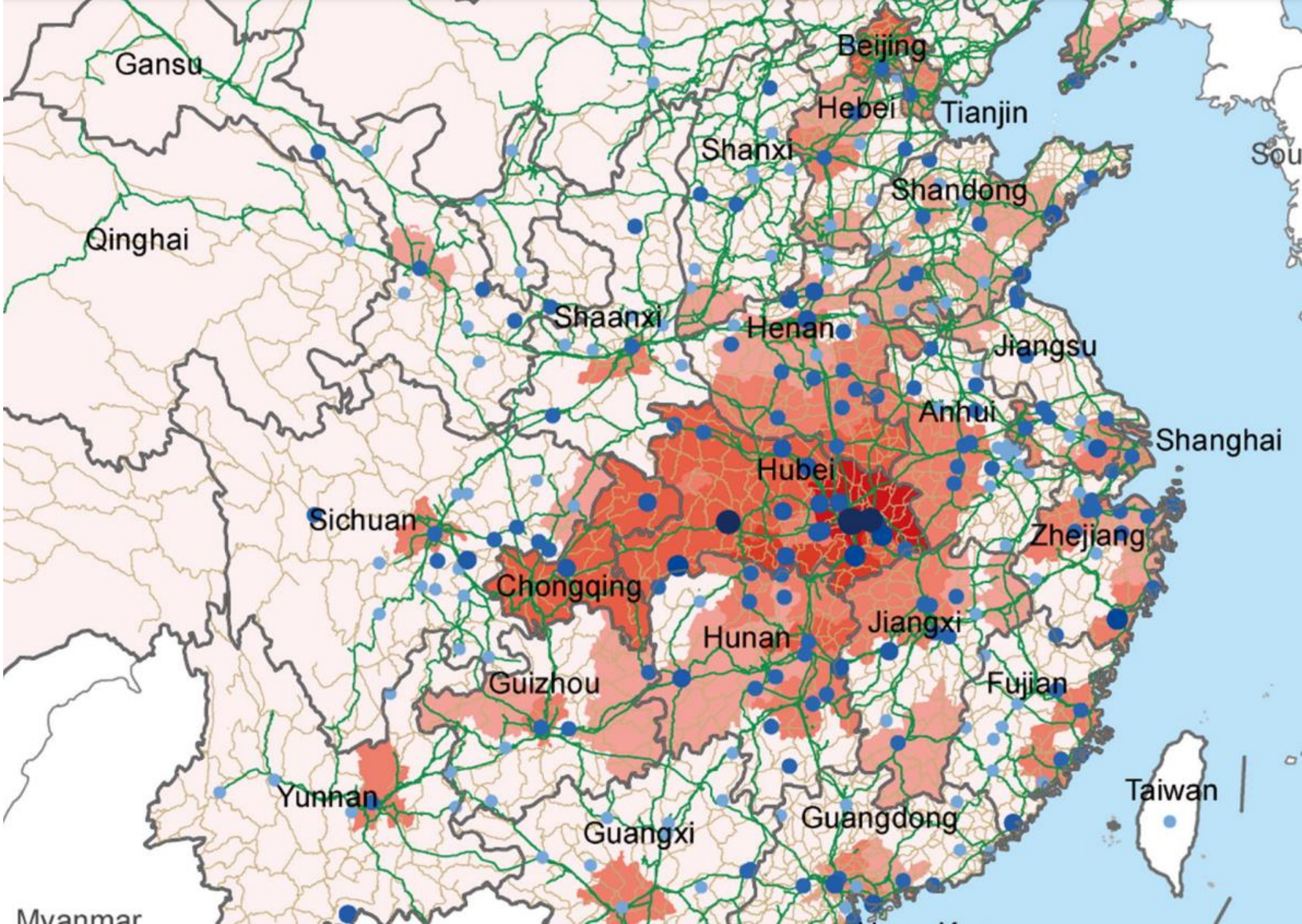
The proportion of receiving travellers from Wuhan (%)

- 0.00-0.01
- 0.01-0.10
- 0.10-1.00
- 1.00-5.00
- 5.00-10.00
- 10.00-16.00

Transportation

- Railway
- National and provincial highway and freeway

Zoomed In:



Antarctica soil samples:

Soil samples sent to China for sequencing [picked up sequences of Covid](#) sequenced on the same machine.

Jesse Bloom [wrote a thread](#) suggesting this provided evidence of earlier covid strains in Wuhan.

Kristian Andersen [debunked it](#) – the sequences show unique mutations seen only in later covid strains.

Alex Crits Christoph also left a [comment on the paper](#).

Also, these were sequenced at a different lab than the WIV uses.

The authors also have a history of [misinterpreting sequencing data](#).

Tabloid news summary:



Incidentally, there was also a [previous effort](#) to look at all SRA samples to look for pre-pandemic covid contamination. It didn't turn up anything, anywhere in the world:



Prof Francois Balloux

@BallouxFrancois



In April 2020, we performed an exhaustive screening of all 'pre-pandemic' metagenomic submissions on the SRA to search for SARS-CoV-2 reads. We found nothing at the time, but our search predates the submission of those samples.

7:18 AM · Feb 9, 2022

Probabilities

Probabilities:

Odds DEFUSE grant happened secretly at the WIV (40% – this is Rootclaim’s number, I think it’s lower, but I’m steelmanning)

they had a suitable secret virus * (1 in 1,000 – based on Latinne FOIA, 2018 paper, sampling rates. This could be lower)

they recognized the spike was interesting * (1 in 10? It’s not much like SARS, but maybe they could measure ACE2 binding)

they made a reverse genetics system for it, instead of using an existing backbone * (1 in 100 – no good reason)

they inserted a furin cleavage site * (1 in 1 – probably lower, but I’m steelmanning here, I’ll just give lab leak this one)

they put the site at S1/S2, not S2’ * (1 in 2 – maybe not a huge deal)

they chose RRAR * (1 in 10 – A is weird, but not highly detrimental. K works much better)

they chose PRRAR * (1 in 20 – This one is really weird and hard to explain)

they inserted it out of frame * (1 in 6 – let’s assume that’s in the secret virus, 6 different codons for serine)

they did the experiments with live virus, not pseudovirus (1 in 2? Unclear what DEFUSE intended, probably lower)

they found some effective way to culture it * (1 in 10? – most cultures/animals fail to make SARS2, assume they’re lucky)

they never published any of the work leading up to this * (1 in 10? could be lower/higher, hard to guess here)

what they created leaked * (1 in 50 – normally 1 in 500, but adjust generously upwards to steelman – BSL-2, live virus, etc)

the leak started an outbreak * (1 in 3)

it only showed up at the market * (1 in 10,000 – use ratio of Wuhan vendors to Wuhan population, or use traffic analysis)

it showed up at the market twice * (1 in 2,000 – it could look like 2 lineages by chance, but that’s very unlikely)

this all happened in the same month the SARS outbreak started * (1 in 6? or 1 in 4, or ignore seasonality, not a big deal)

the most positive samples happened to be in a shop selling susceptible animals * (1 in 68)

that shop was one of the only three (in town) previously fined for selling illegal wildlife * (3 in 10)

the cover-up was so good that neither DRASTIC nor the US government has solved this (1 in 10 – could be lower or higher)

Uses real data.

Can be estimated from other experiments

Hard number to guess

Odds of a secret starting virus?

DEFUSE grant says they'll collect 3,000 samples (page 31). That's going to yield ~30 sarbecoviruses.

The odds of finding one like BANAL-52 are ~1 in 100, if you look at exactly the right location in Laos.

But they're much lower elsewhere. The WIV plans to sample in their known bat cave. They've already sampled there, 100? 200? times, without finding as SARS2 family virus. Looking for 30 more viruses there isn't likely to get one.

We've never found a virus that's 99+% similar to SARS-CoV-2, among hundreds of sarbecoviruses. So maybe the odds get even lower if you need it to be a very specific one.

Full inventory of bat SARSr-CoV QS at our test cave sites, Yunnan, China. To provide data to train and validate our modeling, and as baseline for our immune modulation trial (TA2), DEFUSE fieldwork will target the high-risk cave site in Yunnan Province, SW China (Fig. 4, red triangle) where we will conduct our field trial, and where we have previously identified and isolated high-risk SARSr-CoVs^{2,11,33,34}. At three cave sites (one designated for our trial, two as controls), we will determine the baseline QS₀ risk of SARSr-CoV spillover. We will conduct longitudinal surveillance of bat populations to detect and isolate SARSr-CoVs, determine changes in viral prevalence over time, and measure bat population demographics and movement, definitively characterizing their SARSr-CoV host-viral dynamics. Field data will allow us to test the accuracy

On the high end, assume the WIV found 1 virus (RATG-13) 96% similar to SARS2 out of ~200 previous viruses.

Assume they find 30 new viruses. $30 * (1/200) = 15\%$ odds of finding another like RATG-13.

But, RATG-13 isn't close enough to make SARS2, it needs the correct RBD.

Maybe 5 sarbecoviruses have that (3 BANAL viruses, 1 from Yunnan, 1 from vietnam) out of 1,500 total.
 $30 * (5/1,500) = 10\%$.

But 4 of those aren't close to SARS2, outside the RBD, you really need something like BANAL-52:
 $30 * (1/1,500) = 2\%$

BANAL is still only 97%, it really needs to be 99+%
Maybe you need to adjust downward for that. By how much?

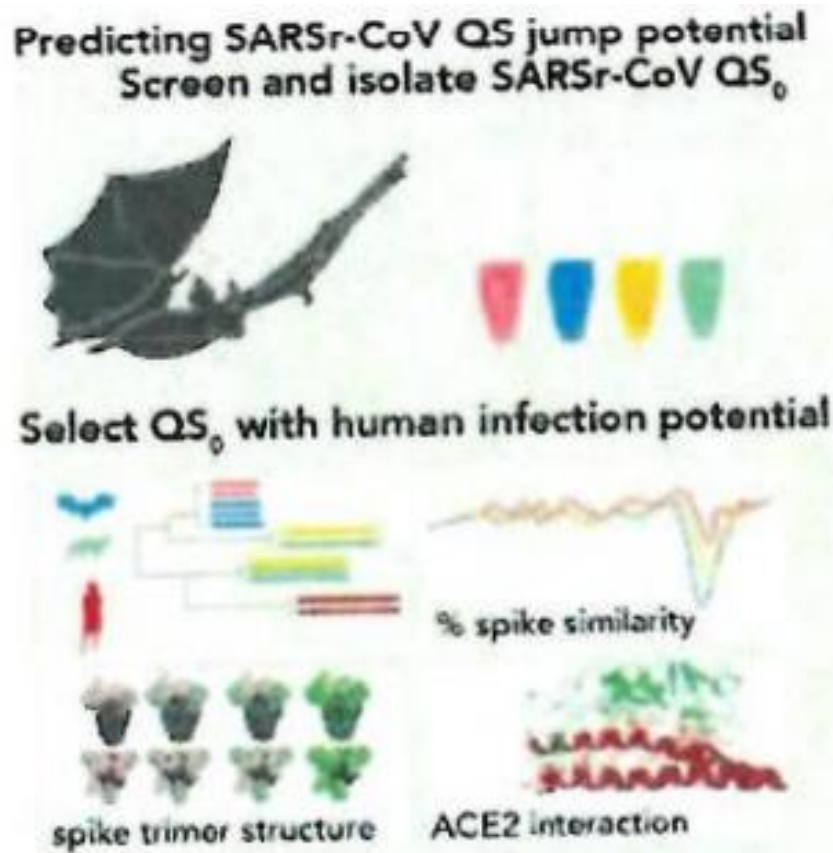
And then adjust downwards again for the fact that all these FOIA attempts and uncovered papers don't have any relevant viruses, and no evidence has shown up of secret sampling trips. Adjust downwards for that.

I went with 1 in 1,000.

But it's definitely hard to know, when you're claiming secret research programs and secret viruses.

Odds they would find a SARS2 precursor interesting?

DEFUSE was interested in ACE2 interaction,
But they were also interested in spike similarity (to SARS)



DEFUSE says they were interested in “closely related strains” with “< 5% nucleotide variation”, presumably measured from SARS.

SARSr-CoV QS detection, sequencing, and recovery. We will screen samples for SARSr-CoV nucleic acid using our pan-CoV consensus one-step hemi-nested RT-PCR assay targeting a 440-nt fragment in the RNA-dependent RNA polymerase gene (RdRp) of all known α - and β -CoVs^{1,53}, and specific assays for known SARSr-CoVs^{2,21,33,34}. PCR products will be gel purified, sequenced and qPCR performed on SARSr-CoV-positive samples to determine viral load. Full-length genomes or S genes of all SARSr-CoVs will be high-throughput sequenced followed by genome walking^{2,3,34}. We will analyze the S gene for its ability to bind human ACE2 by Biocore or virus entry assay. **Synthesis of Chimeric Novel SARSr-CoV QS:** We will commercially synthesize SARSr-CoV S glycoprotein genes, designed for insertion into SHC014 or WIV16 molecular clone backbones (88% and 97% S-protein identity to epidemic SARS-Urbani). **These are BSL-3, not select agents or subject to P3CO** (they use bat SARSr-CoV backbones which are exempt) and are pathogenic to hACE2 transgenic mice. Different backbone strains increase recovery of viable viruses identification of barriers for RNA recombination-mediated gene transfer between strains³⁴. Recombinant viruses will be recovered in Vero cells, or in mouse cells over-expressing human, bat or civet ACE2 receptors to support cultivation of viruses with a weaker RBD-human ACE2 interface. **Recovery of Full length SARSr-CoV:** We will compile sequence/RNAseq data from a panel of closely related strains (<5% nucleotide variation) and compare full length genomes, scanning for unique SNPs representing sequencing errors⁵⁴⁻⁵⁶. Consensus candidate genomes will be synthesized commercially (e.g. BioBasic), using established techniques and genome-length RNA and electroporation to recover recombinant viruses^{28,57}.

Odds of a WIV1 backbone vs full-length backbone?

Let's assume that what they're doing is likely to leak, because you think the lab is highly unsafe.

So, first they're going to create 180 chimeras in a WIV1/WIV16/SHC014 backbone, to categorize the spikes of all the viruses they already have.

Then they're going to make up to 30 more, if they find 30 new viruses.

Then they're possibly going to make 3-5 full length viruses per year, but it's hard to understand which ones and what they would prioritize. It sounds like those are the ones 95% similar to SARS. You need some odds for which one they picked.

That's ~210 chances for a virus to leak, before you even get to making the full length viruses.

So the first virus that leaks is going to only be 3-5/200 odds that it's one of the full length recreated ones (1-2% range).

And the full length ones probably aren't going to be SARS-CoV-2 anyways.

Suppose you think that SARS-CoV-2 is the only one likely to leak, because it has special features (i.e. the right RBD).

You'd still expect the SARS2 spike chimera to leak before the full length version. (wouldn't WIV1 still infect people, if it had a SARS2 like RBD and a furin cleavage site?)

Suppose you also think, like Yuri did, that they put an optimal FCS in first (RRKR), then tried different ones.

Then the RRKR one is likely to leak first, before they get to something weird like PRRAR.