Article

How Trustworthy Are Genomic Sequences of SARS-CoV-2 in GenBank?

Xuhua Xia *

Posted Date: 28 August 2024

doi: 10.20944/preprints202408.1963.v1

Keywords: SARS-CoV-2; COVID-19; GenBank; data validation; genome; genomic analysis

 doi:10.20944/preprints202408.1963.v1

Article

How Trustworthy Are Genomic Sequences of SARS-CoV-2 in GenBank?

Xuhua Xia [1,2]*

[1] Department of Biology, University of Ottawa, Marie-Curie Private, Ottawa, ON K1N 6N5, Canada
[2] Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada
* Correspondence: xxia@uottawa.ca

Abstract: Well-annotated gene and genomic sequences serve as a foundation for making inferences in molecular biology and evolution, and can directly impact public health. The first SARS-CoV-2 genome was submitted to GenBank and used to develop the two successful vaccines. Conserved protein domains are often chosen as targets for developing antiviral medicines or vaccines. Mutation and substitution patterns provide crucial information not only on functional motifs and genome/protein interactions but also for characterizing phylogenetic relationships among viral strains. These patterns, together with the collection time of viral samples, serve as the basis for addressing the question of when and where the host-switching event occurred. Unfortunately, viral genomic sequences submitted to GenBank undergo little quality control, and critical information in the annotation is frequently changed without being recorded. Researchers often have no choice but to hold blind faith in the accuracy of the sequences. There have been reports of incorrect genome annotation but no report that casts doubt on the genomic sequences themselves because it seems theoretically impossible to identify genomic sequences that may not be authentic. This paper takes an innovative approach to show that some SARS-CoV-2 genomes submitted to GenBank cannot be possibly authentic. Specifically, some SARS-CoV-2 genomic sequences deposited in GenBank with collection time in 2023 and 2024, isolated from saliva, nasopharyngeal, sewage, and stool are identical to the reference genome of SARS-CoV-2 (NC_045512). The probability for such occurrence is effectively 0. I also compile SARS-CoV-2 genomes with changed sample collection time. One may led astray in bioinformatic analysis without being aware of errors in sequences and sequence annotation.

Keywords: SARS-CoV-2; COVID-19; GenBank; data validation; genome; genomic analysis

1. Introduction

Databases housed in NCBI/EMBL/DDBJ are the most important bioinformatic resources for modern biological and biomedical research worldwide. GenBank as one of the databases has been the most frequently used resource for functional and comparative genomics. Well-annotated gene and genomic sequences pave the way for a variety of inferences about gene functions as well as interactions among genes and their products. The first SARS-CoV-2 genome was submitted to GenBank [1] and immediately used to develop two successful COVID-19 vaccines [2,3]. The genomic resources also facilitated critical evaluation of the mRNA optimization in the development of the two vaccines [4] and a detailed understanding of the domain structure and function of the viral spike protein [5]. The many submitted SARS-CoV-2 genomes enabled many studies to date the most recent common ancestor (MRCA) of the sequenced SARS-CoV-2 genomes [6-11]. To facilitate this endeavor, NCBI staff have assembled very large phylogenies based on aligned SARS-CoV-2 genomes using NCBI's C++ toolkit [12]. Such trees, with the collection time for each genome, have been used to date the common ancestor of the sampled SARS-CoV-2 genomes and to estimate their evolutionary rate with unprecedented resolving power [10,11,13]. The aligned genomes also showed that the SARS-CoV-2 exhibited extreme CpG deficiency, leading to the inference that the virus is under the selection

 doi:10.20944/preprints202408.1963.v1

of human zinc-finger antiviral proteins [14]. This inference was quickly substantiated by experimental evidence [15-17].

Almost all the inferences above require high-quality sequences and accurate annotations. While wrong annotations can often be detected, it is far more difficult to validate the authenticity of a genomic sequence. If one takes an existing sequence, makes a few random nucleotide replacements, and resubmits to GenBank as a new sequence, it is theoretically impossible to discriminate between this fake sequence and a real sequence.

In this paper, I take an innovative but admittedly low-power approach to detect unreal sequences. I also compile a partial list of SARS-CoV-2 genomes in GenBank with altered collection times, as well as some genomes that have been submitted but withdrawn. The results highlight the urgency of quality control sequence submission to GenBank.

1.1. Rationale for Identifying Unreal SARS-CoV-2 Genomes in GenBank

I illustrate the rationale with the reference genome of SARS-CoV-2 (NC_045512) which was sampled at time T (= December 26, 2019), and an evolutionary rate of 0.05526/genome/day estimated from a phylogeny of 83,688 full-length and high-quality SARS-CoV-2 genomes [10]. The evolutionary rate r has also been estimated in several studies [18-21] using other methods, with a clock changing linearly over time [11,13] or various uncorrelated relaxed clock models [22-24]. The estimated evolutionary rate in these studies is expressed as the number of changes per site per year, and varies from low values such as 0.0006 [18] and 0.000605 [19] to substantially higher values of 0.001793 [20] and 0.0024 [21]. One needs to multiply a factor of (30000/365) to obtain the number of changes per genome per day. The two slow rates would become 0.0493 and 0.0497/genome/day, and the two high rates become 0.1474 and 0.1973/genome/day.

Suppose a genome S identical to NC_045512 was sampled at time $T+\delta$ (e.g., Jan. 18, 2024, so $\delta =$ 1484 days). Given the evolutionary rate r = 0.05526/genome/day, the expected number of nucleotide differences between the two genomes over the period of δ is

$$\boldsymbol{\lambda = r\delta = 0.05526 \times 1484 \ = \ 82.0058} \quad (1)$$

Assuming that mutations are random, we can use the Poisson distribution to find the probability of no nucleotide differences between genome S and the reference genome. This probability mass is

$$\boldsymbol{f(0|\lambda) = e^{-\lambda} = 2.4284 \times 10^{-36}} \quad (2)$$

This calculation shows that the probability of getting a SARS-CoV-2 genome S on Jan. 18, 2024, that is identical to NC_045512 is effectively 0 even if trillions of SARS-CoV-2 genomes were sequenced. Such a genome S, identical to NC_045512 but sampled on Jan. 18, 2024, would be deemed unreal. Note that the probability in Eq. (**1**) could be even smaller for two reasons. First, a viral genome can change not only through point mutations but also throught insertions and deletions (indels). The formulation of this probability in Eq. (**2**) considered only point mutations. If indels also occur, then the chance of finding an exact copy of NC_045512 on Jan. 18, 2024, would be even smaller. Second, the formulation in Eq. (**2**) assumes that all SARS-CoV-2 strains are descendants of the reference genome NC_045512. If the subsequently dominant SARS-CoV-2 strains are not direct descendants of NC_045512, then the probability that we would get a genome at time $T+\delta$ that is identical to NC_045512 would be smaller.

1.2. Identifying SARS-CoV-2 Genomes in GenBank with Altered Collection Time

No record is kept when annotations of sequences submitted to GenBank are changed. This includes the collection time of viral samples. If the originally reported collection time was subsequently modified, only the modified collection time will appear in the GenBank sequence file. This creates difficulties in identifying which SARS-CoV-2 genomes have a modified collection time.

Fortunately, NCBI has routinely compiled full-length high-quality SARS-CoV-2 genomes and built phylogenetic trees as a service to the public (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/precomptree). The OTU names in the tree include

 doi:10.20944/preprints202408.1963.v1

GenBank accession and collection time. One may download two trees at times T1 and T2. If a SARS-CoV-2 genome appears in both trees but with different collection times, then a modification of the collection time has occurred during the interval between the compilation of the two trees.

2. Materials and Methods

2.1. Identify Unreal Sequences

I downloaded early SARS-CoV-2 genomes and searched for identical sequences in GenBank. A stringent criterion of identity was used, i.e., two sequences are considered identical if they are exact copies of each other (identical in both sequence length and nucleotide sequence). For those identical genomes thus identified, I calculated the probability of their occurrence according to Eq. (**1**) as the basis for judging the authenticity of these sequences based on the probability.

2.2. Identify Genomes with Altered Collection Time

NCBI released phylogenetic trees of SARS-CoV-2 genomes continuously. I downloaded seven trees on Apr. 3, Apr. 25, May 29, Jul. 12, Sept. 4, and Nov. 8, 2021, and May 7, 2022. These trees are hereafter referred to as Apr3_21, Apr25_21, May29_21, Jul12_21, Sept4_21, Nov8_21, and May7_22, respectively, and contain 86582, 142591, 183347, 304221, 459944, 633995, and 978217 SARS-CoV-2 genomes, respectively. SARS-CoV-2 genomes in an early tree do not represent a subset of those in a late tree. For example, 4850 genomes in the Apr3_21 tree are absent in the Sep4_21 tree, 2412 genomes in the Sep4_21 tree are not present in the Nov8_21_tree, and 4473 genomes in the Nov8_21 tree are not present in the May7_22 tree. This is partly because some SARS-CoV-2 genomes submitted to GenBank were subsequently withdrawn by the submitters (e.g., FR988889, FR988892, FR988974, FR989034, FR988093).

For SARS-CoV-2 genomes present in two trees, their collection times were compared, and the difference between the collection times was recorded. For those SARS-CoV-2 genomes that are present in only one tree, whether their collection dates have altered was not checked. For example, the SARS-CoV-2 genome from Utah, USA (MW795884) was present in the Nov8_21 tree but not in the May7_22 tree. Although this sequence has changed from a collection time of 2020-01-13 to 2021-01-13, the change will not be detected by the described comparisons between the two trees. The SARS-CoVo2 genome (OK244698) is similar, changing the collection date from 2020-01-14 to 2021-12-30. I have included these genome in the result.

3. Results

3.1. Unreal SARS-CoV-2 Genomes in GenBank

While most SARS-CoV-2 genomes identical to the reference genome (NC_045512) were sampled in early 2020, at least nine such SARS-CoV-2 genomes were collected from 2021 to 2024 (Table 1). They are all exact copies of NC_045512 with a sequence length of 29903. As shown in the last column of Table 1, the probability for such occurrences is effectively 0.

Table 1. At least nine SARS-CoV-2 genomes deposited in GenBank in 2021-2024 are identical to the reference genome NC_045512.

ACCN[(1)]	Country	T[(2)]	δ[(3)]	λ[(4)]	$f(\mathbf{0}\|\lambda)$[(5)]
OM094978	USA	3/24/2021	454	25.0880	1.2718E-11
OM108445	India	7/1/2021	553	30.5588	5.3517E-14
OP022337	USA	10/20/2021	664	36.6926	1.1603E-16
OP268178	Mexico	8/19/2022	967	53.4364	6.2067E-24
PP434597	India	4/10/2023	1201	66.3673	1.5034E-29
PQ008636	India	12/10/2023	1445	79.8507	2.0955E-35
PQ008633	India	1/11/2024	1477	81.6190	3.5753E-36

 doi:10.20944/preprints202408.1963.v1

PQ008634	India	1/11/2024	1477	81.6190	3.5753E-36
PQ008635	India	1/18/2024	1484	82.0058	2.4284E-36

(1) GenBank accession number; (2) sample collection time; (3) time interval in days between Dec. 26, 2019 (collection time for NC_045512) and T. (4) expected number of nucleotide replacements during the period δ. (5) the probability that the genome was sampled at the collection time.

One can appreciate such probability statements intuitively. The reference genome NC_045512 belongs to the CCCA lineage (where CCCA stands for the four nucleotides at sites 241, 3037, 14408, and 23403, respectively, following the numbering of the reference genome NC_045512) [25,26]. This lineage was rapidly replaced by the D614G lineage characterized by TTTG at the four sites mentioned above [25]. There are 1262 SARS-CoV-2 genomes of length 29903 sampled between Apr. 1, 2023, to Jan. 31, 2024. The chance of the exact original NC_045512 being sampled in 2021-2024, even just once, is extremely small, let alone multiple times as shown in Table 1. It is odd that all such genomes are from India. In fact, during the period from Apr. 1, 2023 to Jan. 31, 2024, all five genomes of length 29903 from India are exactly copies of NC_045512.

One genome from the USA (OM094978) was sampled on Mar. 24, 2021. USA contributed a total of 863 genomes of length 29903 to GenBank in March 2021. Thus, the chance of getting a genome like OM094978 is effectively 0. Similarly, USA contributed four genomes of length 29903 to GenBank in October 2021, which also implies an extremely small probability of getting a genome like OP022337 (Table 2).

There are also multiple SARS-CoV-2 genomes identical to NC_045512 that were sampled in late 2020 (Table 2). A total of 98218 SARS-CoV-2 genomes of length 29903 were sampled between Oct. 1, 2020 to Dec. 30, 2020. Thus, getting even just one sequence identical to NC_045512 in November and December of 2020 is very small, let alone 13 sequences in Table 2. Two genomes sampled in December 2020 are from Pakistan Table 2, out of eight genomes of length 29903 submitted from Pakistan in December 2020. USA contributed 13 genomes identical to NC_045512 (Table 2). From Oct. 1 to Dec. 30, 2020, the USA contributed 1887 SARS-CoV-2 genomes of length 29903 to GenBank, which is far from sufficient to explain the 13 genomes identical to NC_045512. Thus, even a technologically advanced country could contribute SARS-CoV-2 genome sequences that are unlikely authentic.

Table 2. At least nine SARS-CoV-2 genomes deposited in GenBank in 2021-2024 are identical to the reference genome NC_045512. Column headings are the same as in Table 1.

ACCN	Country	T	δ	λ	$f(0\|\lambda)$
OM095202	USA	10/8/2020	287	15.8596	1.2950E-07
MZ722043	USA	10/25/2020	304	16.7990	5.0614E-08
OM095001	USA	11/25/2020	335	18.5121	9.1264E-09
OM095004	USA	11/25/2020	335	18.5121	9.1264E-09
OM095010	USA	11/25/2020	335	18.5121	9.1264E-09
OM095127	USA	12/11/2020	351	19.3963	3.7697E-09
MW960278	Pakistan	12/11/2020	351	19.3963	3.7697E-09
MZ722192	USA	12/14/2020	354	19.5620	3.1938E-09
OP278726	Pakistan	12/17/2020	357	19.7278	2.7059E-09
OM095142	USA	12/21/2020	361	19.9489	2.1693E-09
MZ722000	USA	12/21/2020	361	19.9489	2.1693E-09
MZ722615	USA	12/21/2020	361	19.9489	2.1693E-09
MZ722630	USA	12/21/2020	361	19.9489	2.1693E-09
MZ722702	USA	12/21/2020	361	19.9489	2.1693E-09
OP022336	USA	12/30/2020	370	20.4462	1.3193E-09

The presence of those unreal genomes in Table 1 could dramatically affect the dating of the common ancestor of sequenced SARS-CoV-2 genomes and the estimation of the evolutionary rate. For example, the genome PQ008635 sampled on Jan. 18, 2024 (Table 1) implies the possibility of an

 doi:10.20944/preprints202408.1963.v1

infectious SARS-CoV-2 strain without any nucleotide substitution or indels over more than four years. No commonly used bioinformatics tools automatically filter out such unreal sequences, which could lead to highly biased estimates.

3.2. Changes In Viral Sample Collection Time

Many changes in the collection dates are minor, with the date discrepancy smaller than five days. I list those date changes for SARS-CoV-2 genomes with date discrepancy equal to or greater than five days in Table 3**.** Most of the changes in the collection dates were due to the wrong entry of the year, i.e., 2021 entered as 2020 (Table 3). Of the two genomes submitted by Iranian scientists, the discrepancy in the original and the modified dates was attributed to the usage of different calendars. I should mention that many changes in collection dates may not be revealed by the comparison of collection dates between NCBI-generated phylogenetic trees as described in the methods.

Table 3. A partial list of SARS-CoV-2 genomes deposited in GenBank with modified collection time that differs from the original by ≥ 5 days.

ACCN	Country	T1[(1)]	T2[(2)]	Tree1..Tree2[(3)]	T1 - T2
MW795884	USA	1/13/2020	1/13/2021		-366
OK244698	USA	1/14/2020	12/30/2021		-716
MW585340	USA	1/5/2020	1/5/2021		-366
MZ028629	USA	2/18/2020	2/18/2021	7/12/2021..5/7/2022	-366
MZ436887	Sierra_Leone	1/14/2020	1/14/2021	11/8/2021..5/7/2022	-366
MZ436896	Sierra_Leone	1/14/2020	1/14/2021	11/8/2021..5/7/2022	-366
MZ469886	US	1/12/2020	1/12/2021	11/8/2021..5/7/2022	-366
MZ469887	US	1/6/2020	1/6/2021	11/8/2021..5/7/2022	-366
MZ473469	US	2/17/2020	2/17/2021	11/8/2021..5/7/2022	-366
MW786995	USA	3/10/2020	3/10/2021	4/3/2021..5/7/2022	-365
MW921831	USA	3/15/2020	3/15/2021	4/25/2021..5/7/2022	-365
MZ021503	India	3/1/2020	3/1/2021	11/8/2021..5/7/2022	-365
MZ021504	India	3/6/2020	3/6/2021	11/8/2021..5/7/2022	-365
MZ021505	India	3/6/2020	3/6/2021	11/8/2021..5/7/2022	-365
MZ021506	India	3/3/2020	3/3/2021	11/8/2021..5/7/2022	-365
MZ278198	US	4/21/2020	4/21/2021	11/8/2021..5/7/2022	-365
MZ397171	Myanmar	5/28/2020	5/28/2021	11/8/2021..5/7/2022	-365
MZ397172	Myanmar	5/28/2020	5/28/2021	11/8/2021..5/7/2022	-365
MZ397173	Myanmar	5/28/2020	5/28/2021	11/8/2021..5/7/2022	-365
MZ397174	Myanmar	5/28/2020	5/28/2021	11/8/2021..5/7/2022	-365
MZ397175	Myanmar	6/2/2020	6/2/2021	11/8/2021..5/7/2022	-365
MZ397176	Myanmar	6/2/2020	6/2/2021	11/8/2021..5/7/2022	-365
MZ397177	Myanmar	5/26/2020	5/26/2021	11/8/2021..5/7/2022	-365
MW591579	USA	1/18/2020	12/17/2020	4/25/2021..5/7/2022	-334
MW750862	USA	5/22/2020	3/2/2021	4/3/2021..5/7/2022	-284
MW750906	USA	5/23/2020	1/14/2021	4/3/2021..5/7/2022	-236
MW737421	Iran	10/25/2019	2/11/2020	4/3/2021..5/7/2022	-109
MW898809	Iran	12/12/2019	2/29/2020	4/25/2021..5/7/2022	-79
MZ077094	USA	4/14/2021	4/20/2021	7/12/2021..5/7/2022	-6
MW093534	USA	6/6/2020	6/11/2020	4/3/2021..9/4/2021	-5
MW883366	USA	3/29/2021	3/22/2021	4/25/2021..5/7/2022	7
MW883371	USA	3/27/2021	3/16/2021	4/25/2021..5/7/2022	11
MW883363	USA	3/29/2021	3/11/2021	4/25/2021..5/7/2022	18
MW883370	USA	3/27/2021	3/8/2021	4/25/2021..5/7/2022	19
MW883364	USA	3/29/2021	1/21/2021	4/25/2021..5/7/2022	67

 doi:10.20944/preprints202408.1963.v1

(1) Sample collection dates recorded in an earlier tree. (2) Sample collection dates in a later tree. (3) Two trees downloaded at two dates, shown in the form of "Date1..Date2". The first three genome were from my communication with submitters of the GenBank genomes (i.e., not from the comparisons of collection time of genomes between NCBI-generated trees).

The first 26 genomes in Table 3 are all typical D614G strains, with TTTG present at sites 241, 3037, 14408, and 23403, respectively, following the numbering of the reference genome NC_045512. The original wrong dates in these genomes would lead one to infer that the D614G strains occurred quite early, almost simultaneously circulating with the CCCA strain. Had one included these genomes with the original wrong dates in tip-dating, one would tend to date the common ancestor to a date earlier than it should.

Sometimes the submitter would want to replace a submitted SARS-CoV-2 genome with another genome, e.g., MT276328.2 by MT304487. The two may have different sample collection times, e.g., MT276328.2 with a collection time of 2020-02-27 replaced with MT304487 with a collection time of 2020-03-01. GenBank does not keep a record of such changes, nor does it ask for reasons for change. This causes not only confusion but also discrepancies in results of genomic sequence analysis.

3.3. NCBI Is Slow To Correct Annotation Errors

I will use SARS-CoV-2 genomes derived from minks to illustrate the slowness in making corrections to genomic sequence annotation. There are many mink-derived SARS-CoV-2 genomes [27]. In many of these mink-derived genomes (e.g., MT457390 to MT457401), the host was annotated as *Mustela lutreola* (European mink). However, all farmed minks are American mink (*Neovison vison*), so the infection of many European minks would represent a significant transmission event. I contacted one of the submitters on Sept. 7, 2021, and the submitter replied that they would correct the error. I waited until today (Aug. 20, 2024) and the error remains uncorrected. NCBI needs to have more resources to address the data curation problem.

4. Discussion

This is the first paper that casts doubt on the genome sequences themselves. I originally suspected that those sequences in Table 1 are likely from frozen meat, i.e., an original SARS-CoV-2 in Wuhan was frozen in their evolution but was isolated more than four years later by food inspectors. Unfortunately, this was not true. For example, the last two SARS-CoV-2 genomes in Table 1 were isolated from four different sources: saliva, nasopharyngeal, sewage, and stool. It has to take multiple miracles for them to be identical to the reference genome NC_045512. One cannot help asking how many sequences in GenBank are not authentic, given that even a low-power analysis can detect so many impossible sequences. Can we still trust GenBank? NCBI has to find more human resources to implement quality control, otherwise there will be a huge number of incorrect conclusions in publications.

Statistical inference and bioinformatic analysis depend heavily on the quality of data. As I have shown, there are errors and uncertainties in the submitted SARS-CoV-2 genomes. Uncertainty in genome annotation can dramatically affect our conclusions. For example, two SARS-CoV-2 genomes from Japan (MW219695, BS001049) have the same collection time of 2/1/2020 (as of today, Aug. 20, 2024), but MW219695 belongs to the CCCA clade and BS001049 to the TTTG/D614G clade. The two differ by 28 nucleotides. If the collection dates are correct, then we can infer that the TTTG/D614G lineage must have been co-circulating with the Wuhan CCCA lineage simultaneously around the time of the Wuhan outbreak. This would suggest that most published papers on SARS-CoV-2 evolution are incorrect. However, if we cannot be certain of the collection date, it is possible that BS001049 actually has a later collection date but with an incorrectly entered collection date of 2/1/2020. The conventional wisdom in the early phase of the COVID-19 pandemic is that the TTTG/D614G lineage is a late derivative, descending from the early CCCA lineage that caused the Wuhan outbreak [20,28,29]. In this framework, the TTTG/D614G genomes such as BS001049 with an early collection time is typically assumed to have a wrong collection time (i.e., the true collection time

 doi:10.20944/preprints202408.1963.v1

is some time later). However, the TTTG/D614G lineage and the intermediate forms between the CCCA lineage and the TTTG/D614G lineage were subsequently isolated in China and Germany as early as January 2020 [26,30], and increasing evidence favors the hypothesis of the CCCA and TTTG/D614G lineages co-circulating before the Wuhan outbreak [31,32]. All these uncertainties would disappear if SARS-CoV-2 genomes in GenBank have accurate sample collection time.

5. Conclusions

This paper revealed many errors in both sequences and sequence annotation in SARS-CoV-2 genomes submitted to GenBank. Because the method could only detect a small fraction of errors in sequences and sequence annotation, the real amount of error could be much greater and revealed in this paper. There is an urgency for NCBI to implement quality control in genome submissions, especially when public health depends on the quality of such sequences and sequence annotations.

Author Contributions: Everything is done by X.X.

Funding: This research was funded by Discovery Grant from Natural Science and Engineering Research Council (NSERC) of Canada, grant number RGPIN-2024-05641.

Data Availability Statement: Data are contained within the article.

Acknowledgments: I thank Xia Lab members for discussion and comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y., *et al.*, A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265-269.
2. Polack, F.P.; Thomas, S.J.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J.L.; Pérez Marc, G.; Moreira, E.D.; Zerbini, C., *et al.*, Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med* **2020**.
3. Corbett, K.S.; Edwards, D.K.; Leist, S.R.; Abiona, O.M.; Boyoglu-Barnum, S.; Gillespie, R.A.; Himansu, S.; Schäfer, A.; Ziwawo, C.T.; DiPiazza, A.T., *et al.*, SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature* **2020**, *586*, 567-571.
4. Xia, X., Detailed Dissection and Critical Evaluation of the Pfizer/BioNTech and Moderna mRNA Vaccines. *Vaccines* **2021**, *9*, 734.
5. Xia, X., Domains and Functions of Spike Protein in SARS-COV-2 in the Context of Vaccine Design. *Viruses* **2021**, *13*, 109 doi: 110.3390/v13010109.
6. MacLean, O.A.; Lytras, S.; Weaver, S.; Singer, J.B.; Boni, M.F.; Lemey, P.; Kosakovsky Pond, S.L.; Robertson, D.L., Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLOS Biology* **2021**, *19*, e3001115.
7. Wang, H.; Pipes, L.; Nielsen, R., Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evol* **2021**, *7*, veaa098.
8. Boni, M.F.; Lemey, P.; Jiang, X.; Lam, T.T.-Y.; Perry, B.; Castoe, T.; Rambaut, A.; Robertson, D.L., Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature microbiology* **2020**, *5*, 1408–1417.
9. Lytras, S.; Xia, W.; Hughes, J.; Jiang, X.; Robertson, D.L., The animal origin of SARS-CoV-2. *Science* **2021**, *373*, 968-970.
10. Xia, X., Dating the Common Ancestor from an NCBI Tree of 83688 High-Quality and Full-Length SARS-CoV-2 Genomes. *Viruses* **2021**, *13*, 1790.
11. Xia, X., Improved method for rooting and tip-dating a viral phylogeny. In *Handbook of Statistical Bioinformatics*, Lu, H.H.-S.; Scholkopf, B.; Wells, M.T.; Zhao, H., Eds. Springer: Berlin, 2022; pp 397-410.
12. Vakatov, D., *The NCBI C++ Toolkit Book*. National Center for Biotechnology Information (US) https://ncbi.github.io/cxx-toolkit/ (accessed on Sept. 1, 2021): Bethesda (MD), 2009.
13. Xia, X., Rooting and Dating Large SARS-CoV-2 Trees by Modeling Evolutionary Rate as a Function of Time. *Viruses* **2023**, *15*, 684.
14. Xia, X., Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Molecular Biology and Evolution* **2020**, *37*, 2699–2705.
15. Nchioua, R.; Kmiec, D.; Müller, J.A.; Conzelmann, C.; Groß, R.; Swanson, C.M.; Neil, S.J.D.; Stenger, S.; Sauter, D.; Münch, J., *et al.*, SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans. *MBio* **2020**, *11*.

 doi:10.20944/preprints202408.1963.v1

16. Zimmer, M.M.; Kibe, A.; Rand, U.; Pekarek, L.; Ye, L.; Buck, S.; Smyth, R.P.; Cicin-Sain, L.; Caliskan, N., The short isoform of the host antiviral protein ZAP acts as an inhibitor of SARS-CoV-2 programmed ribosomal frameshifting. *Nature communications* **2021**, *12*, 7193.
17. Kmiec, D.; Lista, M.J.; Ficarelli, M.; Swanson, C.M.; Neil, S.J.D., S-farnesylation is essential for antiviral activity of the long ZAP isoform against RNA viruses with diverse replication strategies. *PLOS Pathogens* **2021**, *17*, e1009726.
18. van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T., *et al.*, Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* **2020**, *83*, 104351.
19. Gómez-Carballa, A.; Bello, X.; Pardo-Seco, J.; Martinón-Torres, F.; Salas, A., Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res.* **2020**, *30*, 1434-1448.
20. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G., A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology* **2020**, *5*, 1403-1407.
21. Chaw, S.-M.; Tai, J.-H.; Chen, S.-L.; Hsieh, C.-H.; Chang, S.-Y.; Yeh, S.-H.; Yang, W.-S.; Chen, P.-J.; Wang, H.-Y., The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J. Biomed. Sci.* **2020**, *27*, 73.
22. Drummond, A.J.; Ho, S.Y.; Phillips, M.J.; Rambaut, A., Relaxed phylogenetics and dating with confidence. *PLoS Biol* **2006**, *4*, e88.
23. Lepage, T.; Bryant, D.; Philippe, H.; Lartillot, N., A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **2007**, *24*, 2669-2680.
24. Rannala, B.; Yang, Z., Inferring speciation times under an episodic molecular clock. *Syst Biol* **2007**, *56*, 453-466.
25. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B., *et al.*, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812-827 e819.
26. Yurkovetskiy, L.; Wang, X.; Pascal, K.E.; Tomkins-Tinch, C.; Nyalile, T.P.; Wang, Y.; Baum, A.; Diehl, W.E.; Dauphin, A.; Carbone, C., *et al.*, Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **2020**, *183*, 739-751 e738.
27. Oude Munnink, B.B.; Sikkema, R.S.; Nieuwenhuijse, D.F.; Molenaar, R.J.; Munger, E.; Molenkamp, R.; van der Spek, A.; Tolsma, P.; Rietveld, A.; Brouwer, M., *et al.*, Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **2021**, *371*, 172.
28. Worobey, M.; Levy, J.I.; Malpica Serrano, L.; Crits-Christoph, A.; Pekar, J.E.; Goldstein, S.A.; Rasmussen, A.L.; Kraemer, M.U.G.; Newman, C.; Koopmans, M.P.G., *et al.*, The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science* **2022**, *377*, 951-959.
29. Pekar, J.E.; Magee, A.; Parker, E.; Moshiri, N.; Izhikevich, K.; Havens, J.L.; Gangavarapu, K.; Malpica Serrano, L.M.; Crits-Christoph, A.; Matteson, N.L., *et al.*, The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science* **2022**, *377*, 960-966.
30. Volz, E.; Hill, V.; McCrone, J.T.; Price, A.; Jorgensen, D.; O'Toole, Á.; Southgate, J.; Johnson, R.; Jackson, B.; Nascimento, F.F., *et al.*, Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **2020**, *184*, 64-75.
31. Xia, X., Sequence evidence that the D614G clade of SARS-CoV-2 was already circulating in northern Italy in the fall of 2019. Qeios: 2022.
32. Ruan, Y.; Wen, H.; Hou, M.; He, Z.; Lu, X.; Xue, Y.; He, X.; Zhang, Y.-P.; Wu, C.-I., The twin-beginnings of COVID-19 in Asia and Europe—one prevails quickly. *National science review* **2022**, *9*, nwab223.