A probabilistic approach to evaluate the likelihood of artificial genetic modification and its application to SARS-CoV-2 Omicron variant

Hideki Kakeya* and Yoshihisa Matsumoto+

* Graduate School of Science and Technology, University of Tsukuba, Japan

+ Institute of Innovative Research, Tokyo Institute of Technology, Japan

* Corresponding author: kake@iit.tsukuba.ac.jp

Abstract

A method to find a probability that a given bias of mutations occur naturally is proposed to test whether a newly detected virus is a product of natural evolution or artificial genetic modification. The probability is calculated based on the neutral theory of molecular evolution and binominal distribution of non-synonymous (N) and synonymous (S) mutations. Though most of the conventional analyses, including dN/dS analysis, assume that any kinds of point mutations from a nucleotide to another nucleotide occurs with the same probability, the proposed model takes into account the bias in mutations, where the equilibrium of mutations is considered to estimate the probability of each mutation. The proposed method is applied to evaluate whether the Omicron variant strain of SARS-CoV-2, whose spike protein includes 29 N mutations and only one S mutation, can emerge through natural evolution. The result of binomial test based on the proposed model shows that the bias of N/S mutations in the Omicron spike can occur with a probability of 1.6×10^{-3} or less. Even with the conventional model where the probabilities of any kinds of mutations are all equal, the strong N/S mutation bias in the Omicron spike can occur with a probability of 3.7×10^{-3}, which means that the Omicron variant is highly likely a product of artificial genetic modification.

Author summary

In the method the authors propose to find a probability that a given bias of mutations occur naturally, equilibrium of mutations is considered to estimate the probability of each mutation, whereas most of the conventional genetic analyses assume that point mutations from a nucleotide to another nucleotide occurs with the same probability. The proposed method is applied to evaluate whether the Omicron variant strain of SARS-CoV-2, whose spike protein includes 29 nonsynonymous mutations and only one synonymous mutation, can emerge through natural evolution. The result shows that the bias of mutations in the Omicron spike can occur with a probability of 1.6×10^{-3} or less, which means that the Omicron variant is highly likely a product of artificial genetic modification. Compulsory investigations into related laboratories may be justified to see whether they are the source of the pathogen in focus or not when the likelihood of natural emergence is below a certain threshold like 1%, which holds true in case of the Omicron variant.

Introduction

Since the outbreak of SARS-CoV-2 infection in Wuhan in December of 2019, many kinds of variant strains have emerged one after another. Among them, Omicron VOC (variant of concern) is notably different from the other VOC strains, for it has as many as 30 or more mutations in the spike protein alone [1], while the others have around 10 spike mutations. Phylogenetic analysis shows that the Omicron variant clearly did not emerge from the other precedent VOCs [2].

There are three major hypotheses to explain the emergence of this unique variant [3,4]. The first hypothesis presupposes that the Omicron variant slowly evolved in a region with little viral surveillance. Given the attention COVID-19 pandemic has received globally, however, it is unlikely that the Omicron variant could have evolved, escaping from detection over the period of months.

The second hypothesis postulates that it evolved in a non-human host before spilling over into a human again with a new set of massive mutations. Indeed, one recent paper suggests that the Omicron variant may have evolved in mice [5]. It is known, however, that the original strain of SARS-CoV-2 do not infect mice [6]. Therefore, it is unlikely that an early strain of SARS-CoV-2 infected from human to mice and back from mice to human. Even if the intermediate host had been an animal other than mice, a strain adapted to the animal with a long incubation period could not infect human better than the variants evolved in human-human transmission from the early stage of its emergence.

The third hypothesis is based on the idea that the Omicron variant arose in an immunocompromised patient, chronically infected with a SARS-CoV-2 early strain, which evolved into a distant variant through immune escape. In fact, a SARS-CoV-2 variant with around 10 non-synonymous mutations in the spike protein was found in an immunocompromised patient [7]. However, this quantity of mutations is much smaller than that in the Omicron variant.

Besides the above hypotheses, some argue that the Omicron variant was a product of artificial genetic modification in a laboratory, which leaked accidentally and spread globally. The main basis of this argument is that all the point mutations in the spike of the Omicron variant are non-synonymous mutations except for one, which is extremely unnatural from a statistical point of view. Others argue that this bias toward non-synonymous mutation can naturally occur by pointing out the case of immune escape described above, though it has markedly fewer mutations than the Omicron spike. No statistical evaluations have been made to see whether the case of immune escape can justify natural emergence of mutation bias in the Omicron variant.

In biology, it is often the case that probabilities of events are estimated on a subjective basis, without any firm mathematical calculations. For example, there has been a huge debate on whether the original SARS-CoV-2 virus came from natural spill-over or accidental lab-leak [8-13]. The e-mails to and from Anthony Fauci, which have

been made public through the Freedom of Information Act (FOIA), has revealed that many virologists, including Kristian Andersen, Edward Holmes, Mike Farzan, and Robert Garry, had the idea that SARS-CoV-2 was not likely a product of natural evolution, some of them referring to the insertion of furin cleavage site (See S1 Table). It has also been revealed in the book by Jeremy Farrar that Andersen was "60 to 70%" convinced the virus came from a laboratory, while Holmes was "80% sure this thing had come out of a lab" in a video conference that took place on February 1, 2020 [14]. In an e-mail to Francis Collins and Anthony Fauci, which has also been made public through FOIA, Farrar quoted an opinion by Farzan that estimated ratio of accidental release to natural event was "70:30 or 60:40."

Among these virologists, Andersen, Holmes, and Garry flip-flopped their opinions and co-authored a paper insisting that SARS-CoV-2 came from natural spill-over, which was published as early as in mid-March, 2020 [8]. Though the conclusion of this paper is contrary to their original thoughts, the authors have given little explanation on what proof made them change their minds during this short period of time.

What is consistent throughout their debate on the origin of SARS-CoV-2 is that they have not given any mathematical basis on the probability or the conclusions they reached. Lack of mathematical discussion often leads to subjective speculations without any firm evidence.

Many lab-leak accidents have happened historically and they are occuring more often due to the recent spread of genetic engineering [15,16]. After the emergence of No See'm technology [17], clear traits of artificial genetic modification cannot be found even if a newly detected virus is a product in a laboratory. In case of SARS-CoV-2 lab-leak in Taiwan, the incident was identified as a lab-leak because the virus infection had been subdued there due to the strict quarantine policy. Had a lab-leak taken place in a city populated with many infected patients, the incident would not have been detected.

In the age of elaborate genetic engineering with almost no trait in each mutation, statistical analysis of genetic sequences can be an important tool to evaluate the origin of new viruses. Few trials, however, have been tried to find a probability based on rigid mathematical analyses. One of the exceptions is the work by Steven Quay, where Bayesian approach was applied to find the probability that SARS-CoV-2 was originated from a laboratory [18]. This study, however, allocates a-priori probabilities to many circumstantial evidences, which means that the final probability value can vary depending on what evidences are selected for evaluation.

A more certain approach to evaluate the probability is to depend only on the genetic information. Depending only on a single genetic feature, such as the emergence of 12 nucleotide insertion to form a furin cleavage site or the rare CGG-CGG codon in case of SARS-CoV-2, however, is not robust enough to draw a conclusion. Though cumulative approaches are taken to evaluate statistical bias in the types of mutation, such as dN/dS (Ka/Ks) ratio [19,20], a reliable method has not been developed to find a probability that a certain mutation bias emerges.

In this study, the authors propose a method to find a probability that a given bias of mutations occurs naturally to test whether a newly detected virus is a product of natural evolution or artificial genetic modification. If the obtained upper-bound probability of natural evolution is small enough, it suggests that the possibility of lab-leak should be taken seriously. We apply the proposed method to evaluate whether the Omicron variant of SARS-CoV-2 can emerge through natural evolution.

Methods

Here we focus on the bias of synonymous (S) and non-synonymous (N) point mutations to evaluate whether a variant strain is a result of accumulative random mutations or not. To enhance the function of viruses, such as transmissibility or virulence, N mutation should be introduced to change an amino acid in the protein. The neutral theory of molecular evolution states that changes are given by random genetic drift, most of which do not alter the fitness of an organism [21]. Therefore, a virus evolved naturally should accompany neutral mutations, such as silent mutations, with the mutations that enhance fitness to the environment, while an artificially enhanced virus does not need to accompany neutral mutations.

Since the goal of this paper is to find the upper-bound probability that a newly emerged virus is a product of natural evolution, the basic premises set here are neutral or in favor of natural evolution. First, the probability of each point mutation is set so that it may be equal or in favor of N mutations, which increases the likelihood of natural origin even when the count of N mutation is larger than normal. Second, we neglect the effect of natural selection. It is known that dN/dS exceeds unity if natural selection promotes changes in the protein sequence, which means that more N mutations are observed than the neutral hypothesis predicts. However, it is also known that dN/dS ratio is relatively insensitive to the strength of selective pressure when applied to sequences from a single population [22], which is the case in this analysis. Actually, dN/dS is smaller than unity in the early variants of SARS-CoV and SARS-CoV-2 [23], which means neutral hypothesis is in favor of N mutation and natural origin hypothesis.

The probability that a certain bias emerges in the count of S and N mutations is calculated as follows. In the following discussion we assume that the genome sequence is long enough compared with the count of mutations. Here we assign a number i to each triplet codon from 1 to 64. Each codon has nine kinds of point mutations in total, for each nucleotide in the triplet can be altered into one of the three other nucleotides.

Let s_i be the kinds of S point mutation from codon i, t_i be the kinds of point mutation to a stop codon from codon i, c_i be the counts of codon i in the sequence in focus, and p_{ij} be the probability of point mutation from codon i to codon j. The parameters s_i and t_i are obtained from the codon table, as shown in Table 1. For example, among nine point mutations from UUA, two kinds (mutations to UUG and CUA) are synonymous, while other two kinds (mutations to UAA and UGA) are mutations to stop codons. Among nine point mutations from CGA, four kinds (mutations to CGU, CGC, CGG, and AGA) are synonymous, while the only point mutation that creates

a stop codon is a replacement of the first nucleotide from C to U.

Table 1. Kinds of S point mutations and point mutations to a stop codon for each codon

Codon	AA	s_i	t_i	Codon	AA	s_i	t_i	Codon	AA	s_i	t_i	Codon	AA	s_i	t_i
UUU	F	1		UCU	S	3		UAU	Y	1	2	UGU	C	1	1
UUC				UCC				UAC				UGC			
UUA	L	2	2	UCA			2	UAA				UGA			
UUG			1	UCG			1	UAG				UGG	W	0	2
CUU		3		CCU	P	3		CAU	H	1		CGU	R	3	
CUC				CCC				CAC				CGC			
CUA		4		CCA				CAA	Q	1	1	CGA		4	1
CUG				CCG				CAG				CGG			
AUU	I	2		ACU	T	3		AAU	N	1		AGU	S	1	
AUC				ACC				AAC				AGC			
AUA				ACA				AAA	K	1	1	AGA	R	2	1
AUG	M	0		ACG				AAG				AGG			
GUU	V	3		GCU	A	3		GAU	D	1		GGU	G	3	
GUC				GCC				GAC				GGC			
GUA				GCA				GAA	E	1	1	GGA			1
GUG				GCG				GAG				GGG			

The parameter c_i is obtained just by counting the codons you want to analyze. When a point mutation is randomly selected with the same probability (p_{ij} is constant for all *i* and *j*), which is also the case in dN/dS analysis, the probability P_s^C that a random point mutation is synonymous is given by

$$P_s^C = \frac{C_s}{C_a - C_t}, \quad (1)$$

$$C_a = \sum_{i=1}^{64} 9c_i, \quad (2)$$

$$C_s = \sum_{i=1}^{64} c_i s_i, \quad (3)$$

$$C_t = \sum_{i=1}^{64} c_i t_i, \quad (4)$$

where C_a is the number of all possible point mutations, C_s is the number of all possible S point mutations, and C_t is the number of all possible mutations to a terminal codon in the sequence. Note that the probability P_n^C that a point mutation is non-synonymous is given by $1 - P_s^C$.

When the probability of each point mutation is not uniform, the probability of synonymous mutation P_s^D is given by

$$P_s^D = \frac{D_s}{D_a - D_t}, \quad (5)$$

$$D_a = \sum_{i=1}^{64} \sum_{j=1}^{64} c_i p_{ij}, \quad (6)$$

$$D_s = \sum_{i=1}^{64} \sum_{j=1}^{64} c_i \sigma_{ij} p_{ij}, \quad (7)$$

$$D_t = \sum_{i=1}^{64} \sum_{j=1}^{64} c_i \tau_j p_{ij}, \quad (8)$$

where $\sigma_{ij} = 1$ if the mutation from codon *i* to codon *j* is synonymous, else $\sigma_{ij} = 0$. In the same way $\tau_j = 1$ if codon *j* is the stop codon, else $\tau_j = 0$.

Here the parameters p_{ij} are not easy to obtain with a high certainty. One possible way is to estimate p_{ij} based on the number of point mutations in the strains observed in the virus in focus so far. We can estimate the probability q_{XY} of point mutation from nucleotide X to nucleotide Y based on the observed data, where the kind of point mutation is 12, from four nucleotides to the remaining three nucleotides. The parameter p_{ij} can be represented by q_{XY} when codon *i* turns into codon *j* by replacing one of the three nucleotides from X to Y.

The problem here is that the estimation of q_{XY} becomes unreliable when the size of observed data is small. To cope with this problem, we can use the distribution of four nucleotides in the sequence. If the ratio of nucleotide X is significantly higher/smaller than 1/4, it means that a point mutation to X is more/less often than a point mutation from X. This distribution is reflected also in the count of each codon c_i, which becomes larger when the ratio of the nucleotide included in the codon is higher.

When the mutation from nucleotide X is more often and the ratio of X becomes small enough, the state of equilibrium is attained. First, we consider the case where only two kinds of nucleotides are available for simplicity. Then the equilibrium is attained when

$$q_{\mathrm{XY}} c_{\mathrm{X}} = q_{\mathrm{YX}} c_{\mathrm{Y}} \quad (9)$$

holds, where c_{X} and c_{Y} are the counts of nucleotides X and Y respectively. When all the nucleotides are considered,

$$\sum_{\mathrm{Y}} q_{\mathrm{XY}} c_{\mathrm{X}} - \sum_{\mathrm{Y}} q_{\mathrm{YX}} c_{\mathrm{Y}} = 0 \quad (10)$$

holds for all X in the equilibrium. Since the probability of mutation and the remaining count of nucleotides has an inversely proportional relationship, their product becomes constant. In case we take into account this factor in

equations (5-8), the product of c_i and p_{ij} cancels out, which leads to the approximated equations:

$$P_s^E = \frac{E_s}{E_a - E_t}, \quad (11)$$

$$E_a = (64 - 3) \times 9, \quad (12)$$

$$E_s = \sum_{i=1}^{64} s_i, \quad (13)$$

$$E_t = \sum_{i=1}^{64} t_i. \quad (14)$$

In equation (12), 64 stands for all possible codons and 3 stands for the different kinds of stop codons. We use these equations to evaluate the bias between S and N mutations. Since these equations are independent of c_i, P_s^E is obtained only with the codon table, which calculates as 134/(549-23) $\cong$ 0.255.

Results

We apply the method proposed above to evaluate the bias between S and N point mutations in the Omicron variants of SARS-CoV-2, which has notably more mutations than other variants.

Table 2 shows the counts of S and N point mutations from the Wuhan-derived reference genome (GenBank accession no. NC_045512.2) to the Omicron strain of SARS-CoV-2 submitted by a team of researchers from the Botswana-Harvard HIV Reference Laboratory on November 22, 2021 (GISAID accession. EPI_ISL_6752027). Here mutations around insertion and deletion are omitted to focus only on point mutations (See S2 Table for full list of mutations). As Table 2 shows, the spike protein has 30 point mutations, of which 29 are non-synonymous and only one is synonymous, while it has 15 N mutations and 8 S mutations in the remaining genome, which is much longer than the spike genome.

Table 2. Counts of S and N point mutations in the Omicron variant

	N	S	Total
Spike	29	1	30
Other	15	8	23
Total	44	9	53

The probability that 30 point mutations include one or fewer S mutation is given by

$$p = \binom{30}{1} \times (P_n^E)^{29} \times P_s^E + \binom{30}{0} \times (P_n^E)^{30} \cong 1.6 \times 10^{-3} \quad (15)$$

based on the binomial distribution, which is much smaller than 1% level of statistical significance. Also, the bias

of N and S mutations is significantly different between the spike and the remaining sequence when Chi-squared test is applied ($p \cong 2.5 \times 10^{-3}$).

Table 3 shows the number of each codon in the spike of the Wuhan-derived reference strain of SARS-CoV-2. Here the start codon and the stop codon are excluded. As this table shows, uracil (U) and adenine (A) are more prevalent than cytosine (C) and guanine (G). It is well known that C is easily converted to U by deamination, which leads to the unbalance between the number of nucleotides. Table 4 shows the counts of each nucleotide in the spike and the whole sequence of SARS-CoV-2, both of which show similar ratio of nucleotides.

Table 3. Counts of codons in the spike protein of SARS-CoV-2 (start and stop codons excluded)

Codon	AA	c_i	Codon	AA	c_i	Codon	AA	c_i	Codon	AA	c_i
UUU	F	59	UCU	S	37	UAU	Y	40	UGU	C	28
UUC		18	UCC		12	UAC		14	UGC		12
UUA	L	28	UCA		26	UAA		0	UGA		0
UUG		20	UCG		2	UAG		0	UGG	W	12
CUU		36	CCU	P	29	CAU	H	13	CGU	R	9
CUC		12	CCC		4	CAC		4	CGC		1
CUA		9	CCA		25	CAA	Q	46	CGA		0
CUG		3	CCG		0	CAG		16	CGG		2
AUU	I	44	ACU	T	44	AAU	N	54	AGU	S	17
AUC		14	ACC		10	AAC		34	AGC		5
AUA		18	ACA		40	AAA	K	38	AGA	R	20
AUG	M	13	ACG		3	AAG		23	AGG		10
GUU	V	48	GCU	A	42	GAU	D	43	GGU	G	47
GUC		21	GCC		8	GAC		19	GGC		15
GUA		15	GCA		27	GAA	E	34	GGA		17
GUG		13	GCG		2	GAG		14	GGG		3

Table 4. Counts of nucleotides in the spike (start and stop codons excluded) and the whole sequence of SARS-CoV-2

	A	U	G	C	Total
Spike	1122 (29.4%)	1269 (33.3%)	702 (18.4%)	723 (18.9%)	3816
All	8954 (29.9%)	9584 (32.1%)	5863 (19.6%)	5492 (18.4%)	29903

When we calculate the probability of S and N mutations in the Omicron spike by applying eqns. (1-4) based on the counts of codons in Table 3, P_S^C becomes as small as 0.232 and $p \cong 3.7 \times 10^{-3}$ is obtained with the binomial test, which is larger but still has a statistically significant difference.

The probability given by eqns. (11-14) postulates that the state of equilibrium is attained among different kinds of point mutations. To check whether an equilibrium is attained, point mutations from C to U and the other mutations are counted, as shown in Table 5. As this table shows, point mutation from C to U is outstandingly more often than the other eleven kinds of point mutations, which means that the equilibrium is not yet attained in this virus. When only mutations from C to U take place, the ratio of S mutation to N mutation is 1:2 (the probability of S mutation is 1/3), for the mutations in the third nucleotide in the codon are all synonymous, which means that emergence of extreme bias toward N mutations becomes all the less likely.

Table 5. Counts of C to U mutation and the other mutations in the spike and the remaining sequence of the Omicron variant.

	C to U	Other (X to C)	Total
Spike	7	23 (4)	30
Other	10	15 (3)	25
Total	17	38 (7)	55

Another conspicuous point in Table 5 is the notable difference in the spike and the remaining sequences. In the sequence other than the spike, point mutations from C to U take place more often, while this tendency becomes weak in the spike, though the bias is not statistically significant because of the small data size ($p \cong 0.18$ when Chi-squared test is applied).

Discussion

The counts of 29 N mutations and one S mutation in the Omicron spike is extremely unlikely to emerge through a random process. Even when every mutation is assumed to occur with the same probability, which is not plausible, the probability that 30 mutations include 29 N mutations or more is 3.7×10^{-3}. Under the assumption of equilibrium, the probability goes down to 1.6×10^{-3}, while the count of point mutations, which is strongly leaning toward C to U mutations even though C in the sequence is already low, shows that mutations in SARS-CoV-2 have not reached the equilibrium. When C to U mutations occur more often than the other mutations, the above probability goes down further. Thus, it is extremely unlikely that only one S mutation emerges while 29 N mutations take place.

Since intentional point mutation to change an amino acid to another amino acid to gain function is non-synonymous, the strong bias toward N mutation suggests that the spike protein of the Omicron variant is highly likely to have been manipulated in a laboratory, where it was leaked by an accident and has spread worldwide.

Some people deny the laboratory origin of Omicron variant on the ground that 10 consecutive N mutations (including two deletions, meaning eight N point mutations) was observed in an immunocompromised patient. However, the probability of eight consecutive N point mutations given by eqns. (11-14) is as high as $(0.745)^8 \cong 0.095$, which is not comparable to the probability in the case of the Omicron spike.

To design site-directed mutagenesis, catalogs of functional amino acid changes in spike protein are required, which can be obtained with current technologies. For example, Starr et al. surveyed the effect of all the possible single mutations in the receptor binding domain [24], which is quite informative to realize a mutation with higher transmissibility. Gain-of-function experiments in cultured cells and animals can clarify the effect of amino acid changes on protein function [25].

As stated in the introduction, however, any footprint of genetic modification cannot be found after the emergence of No See'm technology. Therefore, statistical analysis as shown here has become crucial to detect artificial modification of viruses. It is true that statistical bias alone, even if it is extreme, cannot be a definitive proof of lab-leak. Direct proofs of genetic modification in a laboratory are needed for final judgement. The problem is lack of transparency in the current culture of life science. The Wuhan Institute of Virology (WIV) took its virus database offline in September 2019 and has never shared the data since. It also has never accepted full inspection of its facilities by a third party.

It is true that there are no definitive proofs to conclude that the original or the Omicron variant of SARS-CoV-2 are the product of genetic modification in one of the laboratories in the world. Therefore, related laboratories should be granted the benefit of the doubt. In case of forensic investigation, however, prosecutors are allowed to raid the suspect when an investigation warrant is issued by the judiciary. A compulsory investigation is admitted when circumstantial evidences strongly support the prosecutor's claim. In the world of life science, such investigation is not practiced, which makes it easier for scientists to conceal accidents. A typical example is the Sverdlovsk anthrax leak in 1979 [26], which took 15 years to be accepted officially as a lab-leak event.

One possible solution to this problem is to establish an international organization with an authority to inspect laboratories when a detected pathogen is unlikely to have emerged naturally. Since the method applied in this paper gives a likelihood of natural emergence, it can be a useful and objective tool to decide whether a compulsory investigation is justified or not. Compulsory investigations into related laboratories may be justified to see whether they are the source of the pathogen or not when the likelihood of natural emergence is below a certain threshold like 1%, which holds true in the case of the Omicron variant.

One possible concern is that a researcher may artificially add S mutations to pass the tests applied in this paper. In that case, the number of mutations becomes extremely large, whose bias is detected by testing whether the number of mutations can emerge naturally. Even in the sequence data used here, an unusually larger ratio of point mutations is observed in the spike protein. Should a researcher add more mutations in the remaining sequence to balance the ratio of mutation, the count of mutations would contradict with the evolutionary clock. Thus, various statistical approaches can be applied to detect the possibility of artificial genetic modification in genome sequences.

Jureidini and McHenry warn that evidence-based medicine has been corrupted by corporate interests, failed regulation, and commercialization of academia [27], which holds true in life science in general. Regarding the SARS-CoV-2 issue, the letter published in February 2020, condemning the lab-leak theory as a conspiracy theory without showing any scientific evidence [28], was found to be orchestrated by a researcher who had long collaborated with the WIV, without declaring the existing conflict of interest [29]. To prevent the next pandemic, the origins of the virus and its variants need to be unraveled [30], which is attainable only through transparent, objective, and data-driven investigations [31]. Probabilistic evaluation as shown in this paper is a prerequisite and indispensable tool to support such investigations.

Acknowledgement

The authors thank Prof. Hiroshi Tauchi (Ibaraki University, Japan) and Dr. Hiroshi Arakawa (IFOM, Italy) for comments and suggestions on the manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

[1] Callaway E. Heavily mutated Omicron variant puts scientists on alert. Nature 2021;600(7887), 21. doi: 10.1038/d41586-021-03552-w

[2] Jung C, Kmiec, D, Koepke L, et al. Omicron: what makes the latest SARS-CoV-2 variant of concern so concerning? J Virology 2022; in press doi: 10.1128/jvi.02077-21

[3] Kupferschmidt K. Where did 'weird' Omicron come from? Science 2021;374(6572), 1179. doi: 10.1126/science.acx9738

[4] Mallapaty C. The hunt for the origin of Omicron, Nature 2022;602(7898), 26-28. doi: 10.1038/d41586-022-00215-2

[5] Wei C, Shan KJ, Wang W, et al. Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genomics 2021;48(12), 1111-1121.
doi: 10.1016/j.jgg.2021.12.003

[6] Piplani S, Singh PK, Winkler DA, et al. In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. Science Report 2021;11, 13063.
doi: 10.1038/s41598-021-92388-5

[7] Choi B, Choudhary MC, Regan J, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. The New England Journal of Medicine 2020;383(23):2291-2293.
doi: 10.1056/NEJMc2031364

[8] Andersen KG, Rambaut A, Lipkin, WI, et al. The proximal origin of SARS-CoV-2. Nature Medicine 2020;26(4), 450-452.
doi: 10.1038/s41591-020-0820-9

[9] Maxmen A, Mallapaty S. The COVID lab-leak hypothesis: what scientists do and don't know. Nature 2021;594(7863), 313-315.
doi: 10.1038/d41586-021-01529-3

[10] Holmes E, Goldstein SA, Rasmussen A, et al. The origins of SARS-CoV-2: A critical review. Cell 2021;184(19), 4848-4856.
doi: 10.1016/j.cell.2021.08.017

[11] Sallard E, Halloy J, Casane D, et al. Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. Environ Chem Lett 2021;19(4), 769–785.
doi: 10.1007/s10311-020-01151-1

[12] Segreto R, Deigin Y, McCairn K, et al. Should we discount the laboratory origin of COVID-19? Environ Chem Lett 2021;Mar 15, 1-15 (2021).
doi: 10.1007/s10311-021-01211-0

[13] Chan A, Ridley M. Viral: The search for the origin of COVID-19. Harper; 2021.

[14] Farrar F, Ahuja A. Spike: The Virus vs. The People - the Inside Story. Profile Books Ltd; 2021.

[15] Butler D. Fears grow over lab-bred flu. Nature 2011;480(7378), 421–422.
doi:10.1038/480421a

[16] Biosafety in the balance. Nature 2014;510(7506), 443.
doi:10.1038/510443a

[17] Yount B, Denison MR, Weiss SR, et al. Systematic assembly of a full-length Infectious cDNA of mouse hepatitis virus strain A59. J. Virology 2002;76(21), 11065–11078.
doi: 10.1128/JVI.76.21.11065-11078.2002

[18] Quay SC, A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory, Zenodo 2021.
doi: 10.5281/zenodo.4477081

[19] Miyata T, Yasunaga T. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J

Molecular Evolution 1980;16(1), 23–36.
doi: 10.1007/BF01732067

[20] Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 1985;2(2), 150−174.
doi: 10.1093/oxfordjournals.molbev.a040343

[21] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217(5129), 624-626.
doi: 10.1038/217624a0

[22] Kryazhimskiy S, Plotkin JB, The population genetics of dN/dS, PLOS Genetics 2008;4(12), e1000304.
doi: 10.1371/journal.pgen.1000304

[23] Zhan SH, Deverman BE, Chan YA. SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? bioRxiv 2020.
doi: 10.1101/2020.05.01.073262

[24] Starr TN, Greaney AJ, Hilton SK, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 2020;182(5), 1295-1310.
doi: 10.1016/j.cell.2020.08.012

[25] Menachery VD, Yount BL, Jr., Debbink K, et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nature Medicine 2015;21(12), 1508-1513.
doi: 10.1038/nm.3985

[26] Meselson M, Guillemin J, Hugh-Jones M. The Sverdlovsk Anthrax Outbreak of 1979. Science 1994;266(5188), 1202-1208.
doi: 10.1126/science.7973702

[27] Jureidini J, McHenry LB. The illusion of evidence based medicine. BMJ 2022;376:o702
doi: 10.1136/bmj.o702

[28] Calisher C, Carroll D, Colwell R, et al. Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. Lancet 2020;395(10226), E42-E43.
doi: 10.1016/S0140-6736(20)30418-9

[29] Addendum: competing interests and the origins of SARS-CoV-2. Lancet 2021;397(10293), 2449-2450.
doi: 10.1016/S0140-6736(21)01377-5

[30] Relman DA. To stop the next pandemic, we need to unravel the origins of COVID-19. PNAS 2020;117 (47), 29246-29248.
doi: 10.1073/pnas.2021133117

[31] Bloom JD, Chan YA, Baric RS, et al. Investigate the origins of COVID-19, Science 2021; 372(6543), 694.
doi: 10.1126/science.abj0016

Supplemental Materials

S1 Table. Initial comments by virologists on the origin of SARS-CoV-2 [S1, S2]

Andersen	The unusual features of the virus make up a really small part of the genome (<0.1%) so one has to look really closely at all the sequences to see that some of the features (potentially) look engineered.
	Eddie, Bob, Mike, and myself all find the genome inconsistent with expectations from evolutionary theory.
Farzan	a likely explanation could be something as simple as passage SARS-live CoVs in tissue culture on human cell lines (under BSL-2) for an extended period of time, accidently creating a virus that would be primed for rapid transmission between humans via gain of furin site (from tissue culture) and adaption to human ACE2 receptor via repeated passage.
	So, I think it becomes a question of how do you put all this together, whether you believe in this series of coincidences, what you know of the lab in Wuhan, how much could be in nature – accidental release or natural event? I am 70:30 or 60:40.
Garry	I really can't think of a plausible natural scenario where you get from the bat virus or one very similar to it to nCoV where you insert exactly 4 amino acids 12 nucleotide that all have to be added at the exact same time to gain this function – that and you don't change any other amino acid in S2? I just can't figure out how this gets accomplished in nature.

S2 Table. List of mutations in the Omicron variant (T is used in place of U to comply with the database)

ORF1ab		**S**		**ORF3a**	
NSP3		**C21762T**	A67V	**C25584T**	synonymous
A2832G	K38R	**21765-71**	del	**E**	
C3037T	synonymous	**C21846T**	T95I	**C26270T**	T9I
T5386G	synonymous	**21987-95**	del	**M**	
C5730T	T1004I	**22193-209**	subst, del, ins	**A26530G**	D3G
6513-15	del	**G22578A**	G339D	**C26577G**	Q19E
G8393A	A1892T	**T22673C**	S371L	**G26709A**	A63T
NSP4		**C22674T**		**ORF6**	
C10029T	T492I	**T22679C**	S373P	**A27259C**	synonymous
NSP5		**C22686T**	S375F	**ORF7b**	
C10449A	P132H	**G22813T**	K417N	**C27807T**	synonymous
NSP6		**T22882G**	N440K	**N**	
11283-91	del	**G22898A**	G446S	**C28311T**	P13L
A11537G	I189V	**G22992A**	S477N	**28362-70**	del
NSP10		**C22995A**	T478K	**G28881A**	R203K
T13195C	synonymous	**A23013C**	E484A	**G28882A**	synonymous
NSP12		**A23040G**	Q493R	**G28883C**	G204R
C14408T	P323L	**G23048A**	G496S	**UTR & IR**	
C15240T	synonymous	**A23055G**	Q498R	**C241T**	
NSP14		**A23063T**	N501Y	**A28271T**	
A18163G	I42V	**T23075C**	Y505H		
		C23302A	T547K		
		A23403G	D614G		
		C23525T	H655Y		
		T23599G	N679K		
		C23604A	P681H		
		C23854A	N764K		
		G23948T	D796Y		
		C24130A	N856K		
		A24424T	Q954H		
		T24469A	N969K		
		C24503T	L981F		
		C25000T	synonymous		

Supplemental References

[S1] https://www.documentcloud.org/documents/20793561-leopold-nih-foia-anthony-fauci-emails [cited 2022 March 23]

[S2] https://republicans-oversight.house.gov/wp-content/uploads/2022/01/Letter-Re.-Feb-1-Emails-011122.pdf [cited 2022 March 23]