

# Multiple probabilistic analyses suggest non-natural origin of SARS-CoV-2 Omicron variant

Hideki Kakeya<sup>\*</sup>, Hiroshi Arakawa<sup>#</sup>, and Yoshihisa Matsumoto<sup>+</sup>

<sup>\*</sup> Graduate School of Science and Technology, University of Tsukuba, Japan

<sup>#</sup> IFOM – FIRC Institute of Molecular Oncology Foundation, Italy

<sup>+</sup> Institute of Innovative Research, Tokyo Institute of Technology, Japan

<sup>\*</sup> Corresponding author: kake@iit.tsukuba.ac.jp

## Abstract

There are various approaches to evaluate probabilities of mutations, the result of which heavily depends on what kind of mathematical model is used for analysis. To make the obtained result more robust and reliable, multiple statistical analyses from different points of view are used in this paper to find the probability that the mutations of Omicron variant have emerged naturally. To be concrete, the following four approaches are taken: a Poisson test applied to the count of nonsynonymous spike mutations in the Omicron variant compared with those in the conventional variants; a binomial test applied to the counts of nonsynonymous mutations in the spike protein and synonymous mutations in the whole sequence; a binomial test applied to the count of nonsynonymous and synonymous spike mutations based on the spectrum of point mutations observed in the variants of concern; spectrum comparison of 12 kinds of point mutations in the Omicron variant and the other variants. The results of these analyses all show that the mutation pattern of the Omicron variant is extremely unlikely to emerge naturally in humans, suggesting artificial mutagenesis could have been introduced into an early strain of SARS-CoV-2, possibly cultured in transgenic mice or transgenic mouse cell lines under a genetically heterogeneous environment.

## Introduction

Among various variants of SARS-CoV-2, the Omicron VOC (variant of concern), which includes 30 or more mutations in the spike protein alone [1], is notably different from the other VOC strains, which have around 10 spike mutations. Phylogenetic analysis shows that the Omicron variant did not emerge from the other precedent VOCs [2].

There are four hypotheses to explain the emergence of the Omicron variant, including three major hypotheses mentioned on the onset of the outbreak [3,4]. The first hypothesis presupposes that it slowly evolved in an unknown human population. Though selective pressure under vaccinated population can promote mutations, early cases of the Omicron variant were reported in South Africa, where the vaccination rate was low, which makes this scenario unlikely.

The second hypothesis postulates that it evolved in a non-human host before spilling over back to human

population with a new set of massive mutations. Indeed, Wei et al. insist that the Omicron variant has evolved in mice [5], which is followed by Zhang et al. [6]. It is known, however, that the original strain of SARS-CoV-2 do not infect mice [7]. Therefore, it is unlikely that an early strain of SARS-CoV-2 infected from human to mice and back from mice to human under a natural environment. It is also known that white-tailed deer were infected with Alpha, Delta, and Omicron variants, which can be a natural reservoir of SARS-CoV-2 [8]. However, deer do not inhabit in Southern Africa. It should also be noted that a strain adapted to an 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 10 nonsynonymous mutations in the spike protein was found in an immunocompromised patient [9]. However, this quantity of mutations is much smaller than that in the Omicron variant.

Besides the above three major hypotheses, some, including the authors, argue that the Omicron variant was a product of non-natural process, which may include artificial genetic modification in a laboratory [10]. The main basis of this argument is that all the point mutations in the spike of the Omicron variant are nonsynonymous mutations except for one, which is extremely unnatural from a statistical point of view. Strong bias toward nonsynonymous mutation is observed in the spikes of other VOCs in general. Arakawa suggests the possibility of site-directed mutagenesis in the spike protein of other VOCs as well as the Omicron variant [11].

Among these four hypotheses, the third hypothesis is the most popular [1,3], though the count of mutations observed in the immunocompromised patients so far is far fewer than that observed in the Omicron variant [9,12,13]. A common metric called  $dN/dS$  ( $Ka/Ks$ ) that compares  $N$  (nonsynonymous) mutations to  $S$  (synonymous) mutations is often used to evaluate selective pressure [14,15]. 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 [16]. Wei et al. argue that  $dN/dS$  as high as 6.64, which is observed in the spike protein of the Omicron variant, is extremely unlikely to emerge in an immunocompromised patient [5]. Mutation of virus can have a  $dN/dS$  value much higher than unity only when the virus spreads among multiple species [17]. Indeed, even the HIV-1 regulatory gene *tat*, which is known for its high selective pressure, has around 1.5  $dN/dS$  ratio in the human population [18]. In SARS-CoV and SARS-CoV-2,  $dN/dS$  is usually smaller than 1 [19].

In this paper we test whether natural emergence of the Omicron variant is plausible from four perspectives. First, we apply a Poisson test to the count of  $N$  mutations in the Omicron spike compared with those in the spikes of other VOCs. Second, we apply a binomial test to the count of  $N$  mutations in the spike protein and  $S$  mutations in the whole sequence in the Omicron variant. Third we apply a binomial test to the count of  $N$  and  $S$  spike mutations in the Omicron variant based the spectrum of point mutations observed in the VOCs. Lastly, we compare spectra

of 12 kinds of point mutations in the Omicron variant and the other VOCs.

## Methods

Mutations from the prototype of Wuhan strain to those of the prototypes of Alpha, Beta, Gamma, Delta, Lambda, MuGH, and Omicron VOCs were counted following Arakawa [11]. Independent mutations observed in all the prototypes of VOCs were also counted.

In the first analysis, a Poisson test was applied to the count of mutations in the prototype of the Omicron variant based on the distribution of mutations in the prototype of the conventional VOCs and the known immunocompromised patients.

In the second analysis, a binomial test was applied to the count of N mutations in the spike protein and S mutations in the whole sequence included in the proto-Omicron based on the distribution of mutations observed in the prototypes of other VOCs.

In the third analysis, a binomial test was applied to the number of N and S spike mutations in the Omicron variant [10]. Let  $c_i$  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 expected ratio of S mutations  $P_s$  and that of N mutations  $P_n$  are obtained by

$$P_s = D_s / (D_a - D_t), \quad (1)$$

$$P_n = 1 - P_s, \quad (2)$$

$$D_a = \sum_{i=1}^{64} \sum_{j \in M_i} c_i p_{ij}, \quad (3)$$

$$D_s = \sum_{i=1}^{64} \sum_{j \in M_i} c_i \sigma_{ij} p_{ij}, \quad (4)$$

$$D_t = \sum_{i=1}^{64} \sum_{j \in M_i} c_i \tau_j p_{ij}, \quad (5)$$

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 a stop codon, else  $\tau_j = 0$ .  $M_i$  is a set of codons that can be reached with a single point mutation from codon  $i$ . The parameter  $c_i$  was obtained from the Omicron spike sequence and  $p_{ij}$  was obtained based on the spectrum of independent mutations from one of the four nucleotide to one of the other three nucleotides observed in the prototype VOCs.

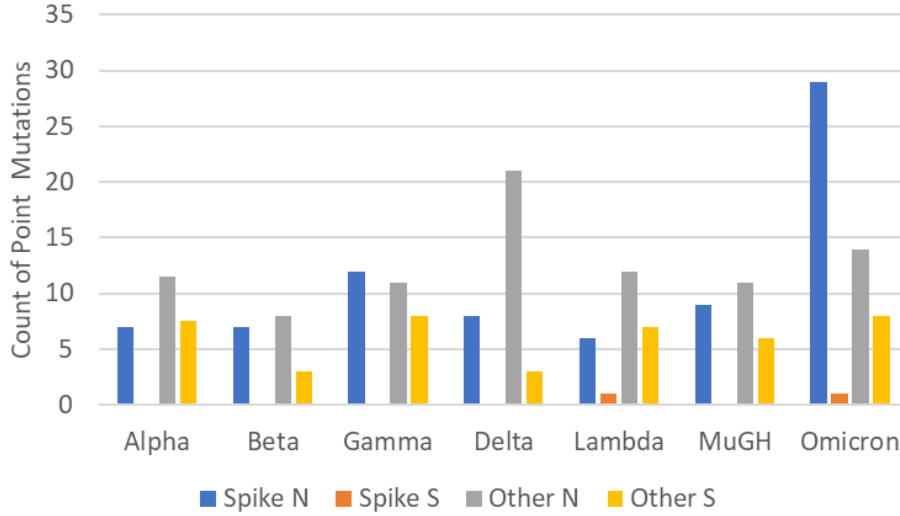
In the fourth analysis, spectra of 12 kinds of point mutations in the Omicron variant and the other VOCs were obtained. The obtained mutation spectra were compared with that of SARS-CoV-2 mutations in humans [20].

## Results

The counts of S and N mutation in the spike protein sequence and the remaining sequence of the proto variants are

shown in Figure 1. As Figure 1 shows, the Omicron variant has by far the most spike mutations, including 29 N mutations and only one S mutation. Though mutations in the spike of other VOCs are also biased toward N mutations, the count is around 10 or fewer. This tendency also holds in the reported cases of spike mutations in the immunocompromised patients [9,12,13].

It is known that the counts of relatively rare phenomena have a Poisson distribution. When the expected number of incidents is 10 in the Poisson process, the probability of observing 29 or more incidents is  $7.6 \times 10^{-7}$ , which is extremely small.



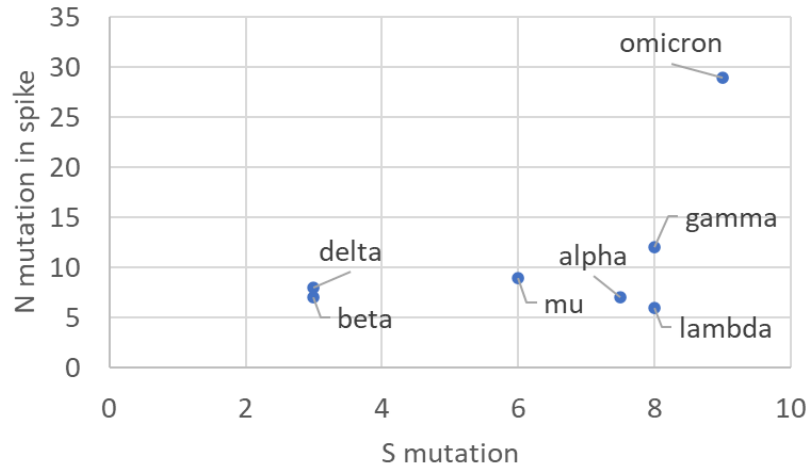
**Figure 1.** Counts of S and N mutations in the spike protein and the other regions of each VOC.

The counts of N mutations in the spike protein and S mutations in the whole sequence in the prototypes of VOCs are shown in Figure 2. The average ratio of the former to the latter is 0.58:0.42 in the VOCs except for the Omicron. When the sum of N mutations in spike and S mutations in the whole sequence is 36, the probability that 29 or more N mutations appear in spike under a binomial distribution given by 0.58:0.42 ratio is  $1.4 \times 10^{-2}$ , which is usually considered to be statistically significant.

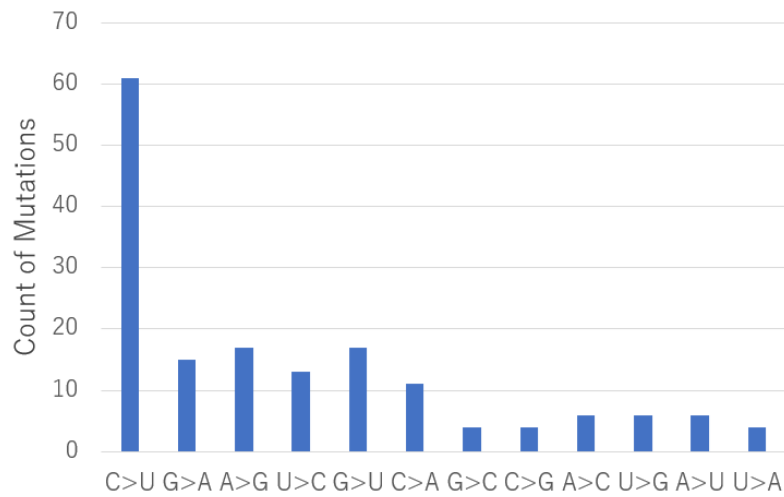
The spectrum of independent mutations from one of the four nucleotide to one of the other three nucleotides observed in the prototype VOCs is shown in Figure 3. The counts of nucleotides A, U, G, and C in the Wuhan strain of SARS-CoV-2 are 8954, 9584, 5863, and 5492 respectively [10]. The count of each mutation divided by the count of nucleotide before mutation is shown in Figure 4. Based on this spectrum,  $P_s$  and  $P_n$  are calculated to be 0.241 and 0.759, as shown in supplemental Table s1.

First, we assume the survival rates of N and S mutations are the same. Then, the probability  $p$  that the count of N mutations is  $M$  or larger when the count of all mutations is  $L$  is given by

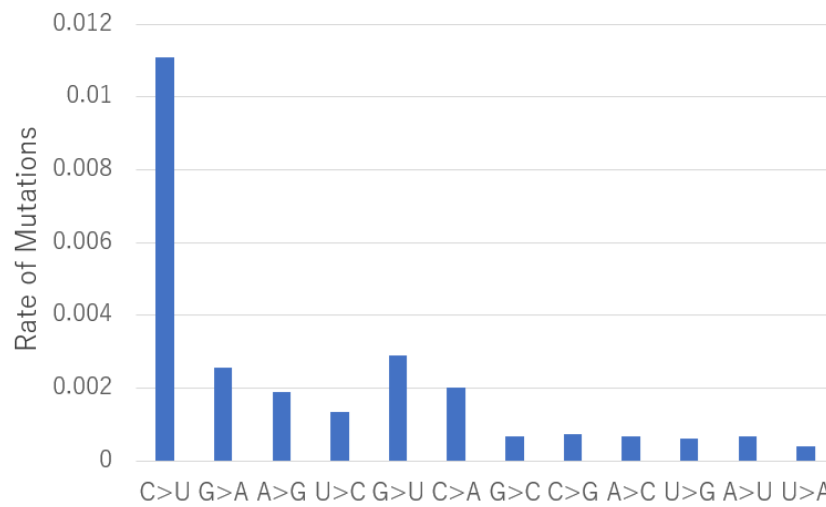
$$p = \sum_{k=M}^L \binom{L}{k} \times P_n^k \times P_s^{L-k}, \quad (6)$$



**Figure 2.** Counts of N mutations in the spike protein and S mutations in the whole sequence included in the prototypes of VOCs.



**Figure 3.** Counts of 12 kinds of nucleotide substitution in the prototypes of VOCs.



**Figure 4.** Rate of 12 kinds of nucleotide substitution in the prototypes of VOCs.

When N mutation is  $r$  times more likely to survive than S mutation, the probability  $q$  that the count of N mutations is  $M$  or larger under the selective pressure is given by

$$q = \sum_{k=M}^L \binom{L}{k} \times Q_n^k \times Q_s^{L-k}, \quad (7)$$

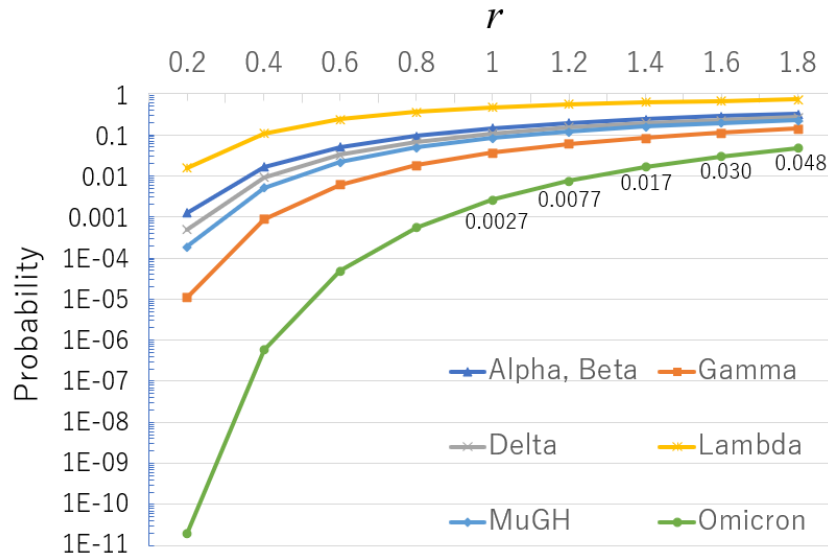
$$Q_n = rP_n / (rP_n + P_s), \quad (8)$$

$$Q_s = P_s / (rP_n + P_s). \quad (9)$$

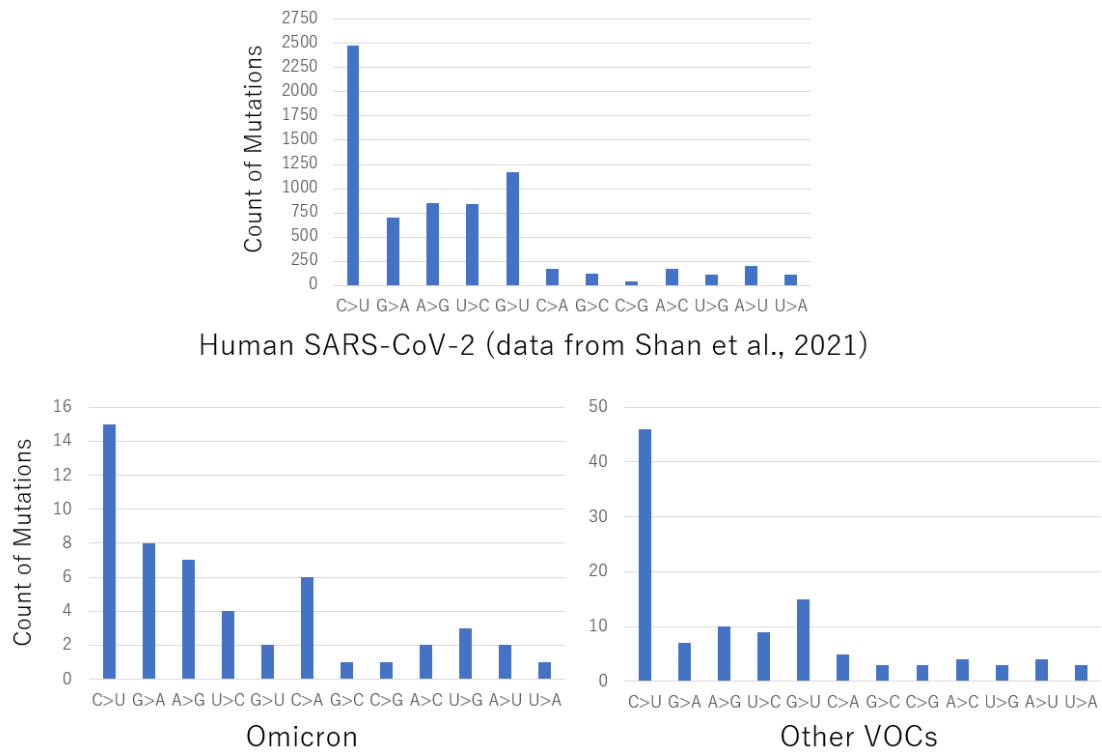
Figure 5 shows the probability that the bias toward N mutation in each variant can emerge under the selective pressures  $r$ . As the result shows, the probability that the bias in the Omicron variant emerges is smaller than the conventional significance level of 0.05 even when  $r$  is as large as 1.8.

In Figure 3, the counts of mutations in the Omicron variant and the other VOCs are merged. The counts of mutations in the Omicron variant and the other VOCs are separately shown in Figure 6. Similar results given by Wei et al. [5] are reproduced. The mutation spectrum of the Omicron variant is different from that of mutations in humans [20] with statistical significance ( $p = 0.014$  given by G-test), while that of the other VOCs is not significantly different ( $p = 0.23$  given by G-test).

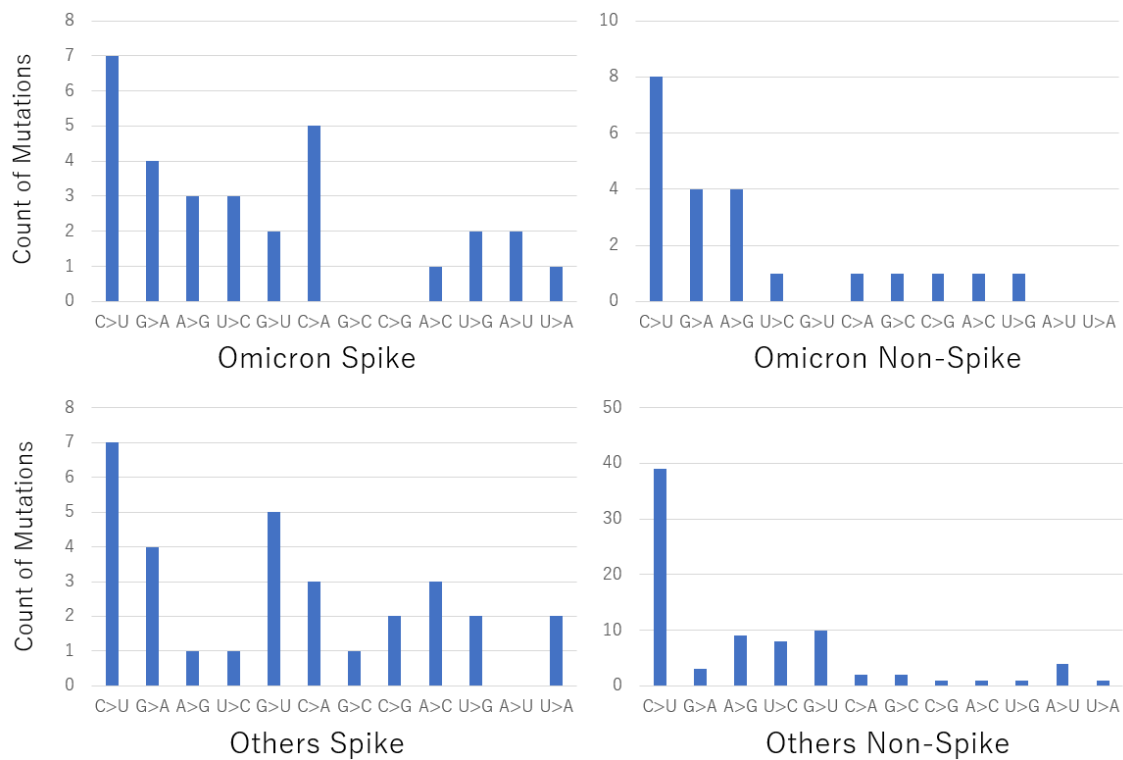
Figure 7 shows the mutation spectra of the spike and the non-spike regions of the Omicron variant and the other VOCs. (Note that the ratio of nucleotides in the spike region is almost the same as that in the whole sequence [10], which does not affect the mutation spectrum.) The mutation spectra of the spike region in the Omicron variant and the other VOCs are both different from that of humans [20] with statistical significance ( $p = 0.042$  and  $p = 0.0034$  given by G-test respectively), while those of the non-spike region in the Omicron variant and the other VOCs are not significantly different from that of humans ( $p = 0.12$  and  $p = 0.43$  given by G-test respectively).



**Figure 5.** Probability that the bias toward N mutation in each variant can emerge under various selective pressures.



**Figure 6.** Separate counts of point mutations in the Omicron variant and the other VOCs. Human SARS-CoV-2 mutation data from [20].



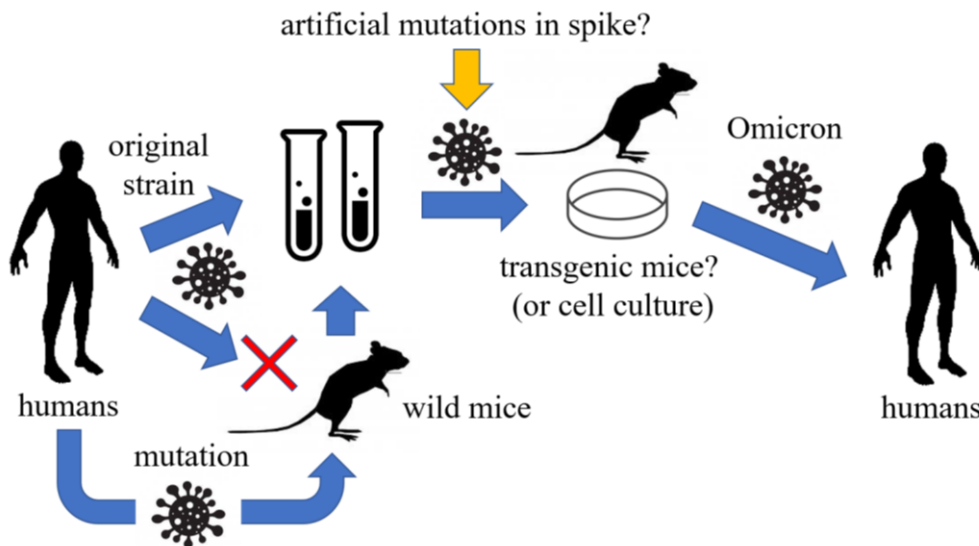
**Figure 7.** Separate counts of point mutations in the spike and the non-spike regions of the Omicron variant and the other VOCs.

## Discussion

All the results given by the four analyses in this paper suggest that it is extremely unlikely that the Omicron variant evolved in a human population including immunocompromised patients. One may say that one of the billions of mutated viruses can bear extreme number of mutations even if the probability is low. The observed mutated viruses, however, are the viruses that have won the competition among them, the variation of which is quite limited.

Mouse origin of the Omicron variant is a plausible scenario considering the mutation spectrum, which is different from that in humans and similar to that in mice [5]. In terms of genetic algorithms, a larger portion of mutations in the survived sequences influence and improve the evaluation score (analogous to nonsynonymous mutation) when the sequence is far from the optimal solution, for there remain many mutations that increase the value of evaluation function (analogous to reproducibility and infectivity). Since cross-species infection greatly changes the evaluation function, there is more room for improvement, while immune response within the same species do not change the evaluation function as much. This is the reason why dN/dS is relatively small in a single population.

Figure 8 shows possible scenarios on the emergence of the Omicron variant. One possible scenario is that transgenic mice that have human ACE2 receptors were infected with the Wuhan strain, where the virus with accumulated mutations was infected back to a human. As explained in the introduction, the original Wuhan strain does not infect wild mice.



**Figure 8.** Possible scenarios of mutation and infection.

Another possible scenario is that the Wuhan strain was mutated naturally or artificially to enable infection to mice, followed by accumulation of mutations in mice. It is known the spike mutation N501Y enables infection to mice, which is widely used for experimental purposes in laboratories [21]. Though some of the VOCs including the Omicron share this mutation and can infect mice [22], the precedent VOCs include many mutations that the Omicron variant does not have. Therefore, it is not highly likely that the other VOCs infected mice and evolved



into the Omicron variant. It is also noteworthy that the spike protein of the Omicron variant binds strongly to both human and mouse ACE2 receptors [5,6]. To realize that kind of evolution, the mice could have a heterogenetic feature with human and mouse ACE2 receptors both expressed, or natural mice and transgenic mice were kept together, where viruses that can efficiently infect both of them evolved.

A more likely scenario is that artificial mutations were introduced in the spike of the Omicron variant, which were cultured in transgenic mice or transgenic mouse cell lines. This hypothesis explains not only the peculiar mutation spectrum but also the outstanding number of N mutations in the spike protein. Indeed, the mutation spectrum of the Omicron spike is different from that of humans. It is reported that most of the mutations in the spike protein of the Omicron are known to affect infectivity through previous variants or past experiments [23], which means that virologists can have motivations to introduce the mutations included in the Omicron variant artificially for research purposes, such as development of pan-variant vaccine. It should also be noted that the main part of spike protein can be clipped without dissecting the remaining sequence by the restriction enzyme BsaI [24]. This makes it convenient to exchange the spike sequence, where mutations are concentrated in the Omicron variant.

As Figure 1 shows, bias toward N mutation is commonly observed in the spike of all VOCs. Though each bias is not statistically significant under large  $r$  (Figure 5), co-occurrence of them can have a low probability with a statistical significance, as suggested by Arakawa [11]. Indeed, the mutation spectrum of the spike in the VOCs other than the Omicron variant is also different from that of humans. Further study is needed to analyze other VOCs in detail.

Definitive footprint of genetic modification cannot be found after the emergence of No See'm technology [25]. Therefore, statistical analysis has become crucial to detect the origin of viruses. It is true that statistical bias alone cannot be a definitive proof of laboratory leak. Direct proofs of laboratory origin are needed to reach a dispositive conclusion. 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.

One of the possible reasons why immunocompromised patient origin hypothesis is popular is that it can obscure the origin. It is almost impossible to identify the patient that became the source of the virus. Therefore, it is unscientific to support this theory from the outset. Laboratory origin and animal origin theories can be tested through thorough investigations to detect the source of infection. Only after the pursuit of direct proof fails should a theory with little chance of finding direct proof be taken seriously.

In the world of life science, a global system for inspection and oversight of related facilities, like the IAEA (International Atomic Energy Agency) for atomic engineering, is missing, which makes it easier for life 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. Also, it took about 30 years to reach a consensus among virologists that the

1977 Russian influenza H1N1 originated from a frozen virus in a laboratory [27], which is still little known outside the virology field. Compulsory investigations into related laboratories should 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 5% or 1%, which holds true in the case of the Omicron variant.

To prevent the next pandemic, the origins of the virus and its variants need to be unraveled [28], which is attainable only through transparent, objective, and data-driven investigations [29]. Indeed, a laboratory origin of the Omicron variant means that the worldwide surge of infection and the loss and damage thereof could have been avoided had thorough investigations of related laboratories taken place soon after the spread of the original strain of SARS-CoV-2, which could have been possible had the virologists been honest enough. As the documents revealed by the Freedom of Information Act shows [30], laboratory origin of SARS-CoV-2 was suspected by the authors of the paper that insists natural origin of SARS-CoV-2 definitively [31].

Risky experiments are still going on at this moment. A recent study reports that a chimera virus with the Omicron spike spliced into the backbone of the original Wuhan strain regains lethality lost by the Omicron variant [32]. Should this synthetic virus be leaked from a laboratory, it could again claim millions of lives. Inspection and oversight of all the laboratories conducting gain of function research are needed for the safety of mankind.

## Summary

This paper analyzes whether natural emergence of the Omicron variant is plausible from four perspectives. First, a Poisson test is applied to the count of mutations in the Omicron variant compared with those in the other VOCs. The probability that the count of  $N$  mutations in the Omicron spike emerges naturally is  $7.6 \times 10^{-7}$ , which is extremely unlikely. Second, a binomial test is applied to the count of  $N$  mutations in the spike protein and  $S$  mutations in the whole sequence of the Omicron variant. The probability that the mutation pattern in the Omicron emerges naturally is  $1.4 \times 10^{-2}$ , which is also unlikely. Third, a binomial test is applied to the count of  $N$  and  $S$  spike mutations in the Omicron variant based the spectrum of point mutations observed in the VOCs. Probability of the emergence of bias in the Omicron variant is 0.0012/0.014/0.047 when the survival ratio of  $N$  mutations is 1.0/1.5/2.0 times higher than that of  $S$  mutations. Lastly, spectra of 12 kinds of point mutations are compared between the Omicron variant and the SARS-CoV-2 mutations in humans, which are different with statistical significance ( $p = 0.014$  by G-test).

These results all suggest that the Omicron variant is highly likely to have originated from a non-natural environment. Artificial mutation is likely to have been spliced into the spike of SARS-CoV-2, possibly cultured in mice or mouse cell lines under a heterogenetic environment where human ACE2 receptors and mouse ACE2 receptors coexist.

## Acknowledgement

The authors thank Prof. Hiroshi Tauchi (Ibaraki University), Prof. Takeshi Nitta (University of Tokyo), and Prof.

Koji Okabayashi (University of Tsukuba) for comments and suggestions on the manuscript. The authors also thank State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences for providing data.

### 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;96(6): e02077-21.  
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] Zhang W, Shi K, Geng Q, et al. Structural basis for mouse receptor recognition by SARS-CoV-2 omicron variant, *PNAS* 2022; 119 (44): e2206509119.  
doi: 10.1073/pnas.2206509119
- [7] 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
- [8] Kuchipudi SV, Surendran-Nair M, Ruden RM, et al. Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer, *Proc Natl Acad Sci U S A*. 2022 Feb 8; 119(6):e2121644119. doi: 10.1073/pnas.2121644119
- [9] 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
- [10] Kakeya H, Matsumoto Y. A probabilistic approach to evaluate the likelihood of artificial genetic modification and its application to SARS-CoV-2 Omicron variant, *ISPJ Trans. Bioinformatics* 2022;15:22-29.  
doi: 10.2197/ipsjtbio.15.22

- [11] Arakawa H, Mutation signature of SARS-CoV-2 variants raises questions to their natural origins, Zenodo 2022.  
doi: 10.5281/zenodo.6601991
- [12] Kemp SA, Collier DA, Datier RP, et al. SARS-CoV-2 evolution during treatment of chronic infection, Nature 2021;592:277-282.  
doi: 10.1038/s41586-021-03291-y
- [13] Truong TT, Ryutov A, Pandey U, et al. Increased viral variants in children and young adults with impaired humoral immunity and persistent SARS-CoV-2 infection: A consecutive case series, EBioMedicine 2021 May;67:103355.  
doi: 10.1016/j.ebiom.2021.103355
- [14] 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
- [15] 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
- [16] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217(5129):624-626.  
doi: 10.1038/217624a0
- [17] Kryazhimskiy S, Plotkin JB, The population genetics of dN/dS, PLOS Genetics 2008;4(12):e1000304.  
doi: 10.1371/journal.pgen.1000304
- [18] Hasan Z, Hasan M, Ashik AI, et al. Prediction of immune pressure on HIV-1 regulatory gene tat by human host through bioinformatics tools, J Adv Biotechnol Exp Ther. 2020 Sep;3(3):233-240.  
doi: 10.5455/jabet.2020.d129
- [19] 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
- [20] Shan KJ, Wei C, Wang Y, et al. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process, Innovation 2021;2(4), 100159.  
doi: 10.1016/j.xinn.2021.100159
- [21] Gu H, Chen Q, Yang G, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy, Science 2020;369(6511), 1603-1607.  
doi: 10.1126/science.abc4730
- [22] Chen Q, Huang XY, Liu Y, et al. Comparative characterization of SARS-CoV-2 variants of concern and mouse-adapted strains in mice, J Medical Virology 2022;94(7), 3223-3232.  
doi: 10.1002/jmv.27735

- [23] Bruttel V, Molecular evidence indicates a synthetic origin of SARS-CoV2 Omicron,  
<https://www.stopgof.com/english/omicron-origin/> [cited 2022 Nov 13]
- [24] Bruttel V, Washburne A, VanDongen A, Endonuclease fingerprint indicates a synthetic origin of SARS-CoV-2, bioRxiv 2022.  
doi: 10.1101/2022.10.18.512756
- [25] 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
- [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] Kransnitz M, Levine AJ, Rabadan R. Anomalies in the Influenza Virus Genome Database: New Biology or Laboratory Errors? *J Virol.* 2008 Sep;82(17):8947-50. doi: 10.1128/JVI.00101-08.
- [28] 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
- [29] Bloom JD, Chan YA, Baric RS, et al. Investigate the origins of COVID-19, *Science* 2021; 372(6543):694.  
doi: 10.1126/science.abj0016
- [30] <https://www.documentcloud.org/documents/20793561-leopold-nih-foia-anthony-fauci-emails> [cited 2022 Nov 13]
- [31] 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
- [32] Chen DY, Kenney D, Chin CV, et al. Role of spike in the pathogenic and antigenic behavior of SARS-CoV-2 BA.1 Omicron, bioRxiv 2020.  
doi: 10.1101/2022.10.13.512134

**Supplemental Table s1.** Calculation process of N and S mutation ratio. 1X means the rate ( $\times 10^{-4}$ ) that the first nucleotide of the codon is mutated to nucleotide X in Figure 4. Yellow means synonymous mutation and red means mutation to a stop codon.  $P_i^n$  and  $P_i^s$  denote the probabilities of N and S mutations given by a single point mutation from codon  $i$  respectively. (Mutations to stop codons are excluded from N mutations.)

Codon	aa	$c_i$	1U	1C	1A	1G	2U	2C	2A	2G	3U	3C	3A	3G	$P_i^n$	$P_i^s$	$c_i P_i^n$	$c_i P_i^s$
UUU	F	59		1.36	0.42	0.63		1.36	0.42	0.63		1.36	0.42	0.63	0.188	0.812	11.12	47.88
UUC		18		1.36	0.42	0.63		1.36	0.42	0.63	11.1		2	0.73	0.596	0.404	10.73	7.273
UUA		28		1.36	0.42	0.63		1.36	0.42	0.63		0.67	0.67	1.9	0.465	0.535	13.03	14.97
UUG		20		1.36	0.42	0.63		1.36	0.42	0.63	2.9	0.68	2.56		0.372	0.628	7.441	12.56
CUU	L	36	11.1		2	0.73		1.36	0.42	0.63		1.36	0.42	0.63	0.129	0.871	4.635	31.36
CUC		12	11.1		2	0.73		1.36	0.42	0.63	11.1		2	0.73	0.46	0.54	5.521	6.479
CUA		9	11.1		2	0.73		1.36	0.42	0.63	0.67	0.67	1.9		0.737	0.263	6.629	2.371
CUG		3	11.1		2	0.73		1.36	0.42	0.63	2.9	0.68	2.56		0.771	0.229	2.312	0.688
AUU	I	44	0.67	0.67		1.9		1.36	0.42	0.63		1.36	0.42	0.63	0.221	0.779	9.709	34.29
AUC		14	0.67	0.67		1.9		1.36	0.42	0.63	11.1		2	0.73	0.673	0.327	9.423	4.577
AUA		18	0.67	0.67		1.9		1.36	0.42	0.63	0.67	0.67		1.9	0.151	0.849	2.717	15.28
AUG		13	0.67	0.67		1.9		1.36	0.42	0.63	2.9	0.68	2.56		0	1	0	13
GUU	V	48	2.9	0.68	2.56			1.36	0.42	0.63		1.36	0.42	0.63	0.219	0.781	10.53	37.47
GUC		21	2.9	0.68	2.56			1.36	0.42	0.63	11.1		2	0.73	0.618	0.382	12.99	8.014
GUA		15	2.9	0.68	2.56			1.36	0.42	0.63	0.67	0.67	1.9		0.275	0.725	4.124	10.88
GUG		13	2.9	0.68	2.56			1.36	0.42	0.63	2.9	0.68	2.56		0.418	0.582	5.437	7.563
UCU	S	37		1.36	0.42	0.63	11.1		2	0.73		1.36	0.42	0.63	0.129	0.871	4.764	32.24
UCC		12		1.36	0.42	0.63	11.1		2	0.73	11.1		2	0.73	0.46	0.54	5.521	6.479
UCA		26		1.36	0.42	0.63	11.1		2	0.73	0.67	0.67	1.9		0.193	0.807	5.029	20.97
UCG		2		1.36	0.42	0.63	11.1		2	0.73	2.9	0.68	2.56		0.301	0.699	0.603	1.397
CCU	P	29	11.1		2	0.73	11.1		2	0.73		1.36	0.42	0.63	0.08	0.92	2.314	26.69
CCC		4	11.1		2	0.73	11.1		2	0.73	11.1		2	0.73	0.333	0.667	1.333	2.667
CCA		25	11.1		2	0.73	11.1		2	0.73	0.67	0.67	1.9		0.105	0.895	2.619	22.38
CCG		0	11.1		2	0.73	11.1		2	0.73	2.9	0.68	2.56		0.182	0.818	0	0
ACU	T	44	0.67	0.67		1.9	11.1		2	0.73		1.36	0.42	0.63	0.123	0.877	5.421	38.58
ACC		10	0.67	0.67		1.9	11.1		2	0.73	11.1		2	0.73	0.448	0.552	4.476	5.524
ACA		40	0.67	0.67		1.9	11.1		2	0.73	0.67	0.67	1.9		0.159	0.841	6.377	33.62
ACG		3	0.67	0.67		1.9	11.1		2	0.73	2.9	0.68	2.56		0.264	0.736	0.793	2.207
GCU	A	42	2.9	0.68	2.56		11.1		2	0.73		1.36	0.42	0.63	0.107	0.893	4.504	37.5
GCC		8	2.9	0.68	2.56		11.1		2	0.73	11.1		2	0.73	0.409	0.591	3.274	4.726
GCA		27	2.9	0.68	2.56		11.1		2	0.73	0.67	0.67	1.9		0.139	0.861	3.766	23.23
GCG		2	2.9	0.68	2.56		11.1		2	0.73	2.9	0.68	2.56		0.235	0.765	0.47	1.53
UAU	Y	40		1.36	0.42	0.63	0.67	0.67		1.9		1.36	0.42	0.63	0.194	0.806	7.757	32.24
UAC		14		1.36	0.42	0.63	0.67	0.67		1.9	11.1		2	0.73	0.663	0.337	9.286	4.714
UAA		0																
UAG		0																
CAU	H	13	11.1		2	0.73	0.67	0.67		1.9		1.36	0.42	0.63	0.07	0.93	0.905	12.09
CAC		4	11.1		2	0.73	0.67	0.67		1.9	11.1		2	0.73	0.359	0.641	1.437	2.563
CAA		46	11.1		2	0.73	0.67	0.67		1.9	0.67	0.67	1.9		0.206	0.794	9.484	36.52
CAG		16	11.1		2	0.73	0.67	0.67		1.9	2.9	0.68	2.56		0.211	0.789	3.38	12.62
AAU	N	54	0.67	0.67		1.9	0.67	0.67		1.9		1.36	0.42	0.63	0.153	0.847	8.251	45.75
AAC		34	0.67	0.67		1.9	0.67	0.67		1.9	11.1		2	0.73	0.547	0.453	18.59	15.41
AAA		38	0.67	0.67		1.9	0.67	0.67		1.9	0.67	0.67	1.9		0.21	0.79	7.975	30.02
AAG		23	0.67	0.67		1.9	0.67	0.67		1.9	2.9	0.68	2.56		0.214	0.786	4.925	18.07
GAU	D	43	2.9	0.68	2.56		0.67	0.67		1.9		1.36	0.42	0.63	0.115	0.885	4.952	38.05
GAC		19	2.9	0.68	2.56		0.67	0.67		1.9	11.1		2	0.73	0.478	0.522	9.09	9.91
GAA		34	2.9	0.68	2.56		0.67	0.67		1.9	0.67	0.67	1.9		0.195	0.805	6.642	27.36
GAG		14	2.9	0.68	2.56		0.67	0.67		1.9	2.9	0.68	2.56		0.203	0.797	2.838	11.16
UGU	C	28		1.36	0.42	0.63	2.9	0.68	2.56			1.36	0.42	0.63	0.129	0.871	3.609	24.39
UGC		12		1.36	0.42	0.63	2.9	0.68	2.56		11.1		2	0.73	0.545	0.455	6.541	5.459
UGA		0																
UGG		12		1.36	0.42	0.63	2.9	0.68	2.56		2.9	0.68	2.56		0	1	0	12
CGU	R	9	11.1		2	0.73	2.9	0.68	2.56			1.36	0.42	0.63	0.107	0.893	0.965	8.035
CGC		1	11.1		2	0.73	2.9	0.68	2.56		11.1		2	0.73	0.409	0.591	0.409	0.591
CGA		0	11.1		2	0.73	2.9	0.68	2.56		0.67	0.67	1.9		0.433	0.567	0	0
CGG		2	11.1		2	0.73	2.9	0.68	2.56		2.9	0.68	2.56		0.312	0.688	0.624	1.376
AGU	S	17	0.67	0.67		1.9	2.9	0.68	2.56			1.36	0.42	0.63	0.115	0.885	1.958	15.04
AGC		5	0.67	0.67		1.9	2.9	0.68	2.56		11.1		2	0.73	0.478	0.522	2.392	2.608
AGA		20	0.67	0.67		1.9	2.9	0.68	2.56		0.67	0.67	1.9		0.215	0.785	4.3	15.7
AGG		10	0.67	0.67		1.9	2.9	0.68	2.56		2.9	0.68	2.56		0.208	0.792	2.08	7.92
GGU	G	47	2.9	0.68	2.56		2.9	0.68	2.56			1.36	0.42	0.63	0.163	0.837	7.683	39.32
GGC		15	2.9	0.68	2.56		2.9	0.68	2.56		11.1		2	0.73	0.53	0.47	7.947	7.053
GGA		17	2.9	0.68	2.56		2.9	0.68	2.56		0.67	0.67	1.9		0.257	0.743	4.363	12.64
GGG		3	2.9	0.68	2.56		2.9	0.68	2.56		2.9	0.68	2.56		0.333	0.667	1	2
Total																	307	965
Rate																	0.241	0.759