Prediction of the effects of the top 10 synonymous mutations from 26645 SARS-CoV-2 genomes of early pandemic phase [version 4; peer review: 1 approved, 3 approved with reservations, 2 not approved]

Previously titled: Prediction of the effects of the top 10 synonymous mutations from 26645 SARS-CoV-2 genomes

Wan Xin Boon, Boon Zhan Sia, Chong Han Ng

Faculty of Information Science and Technology, Multimedia University, Bukit Beruang, Melaka, 75450, Malaysia

Abstract

Background

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) had led to a global pandemic since December 2019. SARS-CoV-2 is a single-stranded RNA virus, which mutates at a higher rate. Multiple works had been done to study nonsynonymous mutations, which change protein sequences. However, there is little study on the effects of SARS-CoV-2 synonymous mutations, which may affect viral fitness. This study aims to predict the effect of synonymous mutations on the SARS-CoV-2 genome.

Methods

A total of 26645 SARS-CoV-2 genomic sequences retrieved from Global Initiative on Sharing all Influenza Data (GISAID) database were aligned using MAFFT. Then, the mutations and their respective frequency were identified. Multiple RNA secondary structures prediction tools, namely RNAdiff, IPKnot++ and Mfold2 were applied to predict the effect of the mutations on RNA secondary structure and their base

Open Peer Review

Approval Status

1 2 3 4 5 6

version 4
(revision)
18 Sep 2024

version 3
(revision)
29 Feb 2024

version 2
(revision)
05 Sep 2022

version 1
18 Oct 2021

1. Takahiko Koyama, IBM TJ Watson Research Center, Yorktown Heights, USA
2. Leyi Wang, University of Illinois, Urbana, USA
3. Chandran Nithin, International Institute of Molecular and Cell Biology in Warsaw,
pair probabilities was estimated using MutaRNA. Relative synonymous codon usage (RSCU) analysis was also performed to measure the codon usage bias (CUB) of SARS-CoV-2.

Results

A total of 150 synonymous mutations were identified. The synonymous mutation identified with the highest frequency is C3037U mutation in the nsp3 of ORF1a. Of these top 10 highest frequency synonymous mutations, C913U, C3037U, U16176C and C18877U mutants show pronounced changes between wild type and mutant in all 3 RNA secondary structure prediction tools, suggesting these mutations may have some biological impact on viral fitness. These four mutations show changes in base pair probabilities. All mutations except U16176C change the codon to a more preferred codon, which may result in higher translation efficiency.

Conclusion

Synonymous mutations in SARS-CoV-2 genome may affect RNA secondary structure, changing base pair probabilities and possibly resulting in a higher translation rate. However, lab experiments are required to validate the results obtained from prediction analysis.

Keywords
COVID-19, SARS-CoV-2, Synonymous mutations, RNA secondary structure

Warsaw, Poland
4. Roland Huber, Bioinformatics Institute, A*STAR, Singapore
5. Diego Forni, Scientific Institute IRCCS E Medea, Bosisio Parini, Italy
6. Tamar Schlick, New York University, New York, USA

Any reports and responses or comments on the article can be found at the end of the article.
Corresponding author: Chong Han Ng (chng@mmu.edu.my)

Author roles: Boon WX: Formal Analysis, Investigation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Sia BZ: Investigation, Methodology; Ng CH: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is supported by Multimedia University, Malaysia, IRFund 2.0 (grant number MMUI/210119 awarded to Chong Han, Ng). The funder has no role in study design, data analysis, decision to publish or manuscript preparation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Boon WX et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Boon WX, Sia BZ and Ng CH. Prediction of the effects of the top 10 synonymous mutations from 26645 SARS-CoV-2 genomes of early pandemic phase [version 4; peer review: 1 approved, 3 approved with reservations, 2 not approved] F1000Research 2024, 10:1053 https://doi.org/10.12688/f1000research.72896.4

First published: 18 Oct 2021, 10:1053 https://doi.org/10.12688/f1000research.72896.1
Introduction

In December 2019, coronavirus disease 2019 (COVID-19) cases first emerged from Wuhan, China\(^6\). Soon after, rapid spread of COVID-19 has resulted in a serious global outbreak. COVID-19 is an infectious and potentially lethal disease caused by a newly found coronavirus strain, known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus causes clinical manifestation ranging from asymptomatic to severe pneumonia and in the worst scenario, death\(^7\). SARS-CoV-2 seems to have a higher transmission rate\(^1\) but lower mortality rate\(^1\) in comparison to Middle East respiratory coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV).

SARS-CoV-2 is a single-stranded RNA virus with a genome size of 29,903 bases. In general, RNA viruses have a higher mutation rate than DNA viruses and this allows them to evolve rapidly, escaping the host immune defense response\(^1\). Different SARS-CoV-2 variants with multiple synonymous and nonsynonymous mutations have been reported since the beginning of the outbreak\(^1\). Some variants are classified as variants of concern (VOCs) since they are associated with the change in viral pathogenicity such as, higher disease severity, higher transmission rate, lower immunity response in the host as a consequence of the mutations\(^2\). However, it is expected that most of these mutations in SARS-CoV-2 genome are either neutral or mildly deleterious\(^1\). Numerous studies have been carried out to understand the molecular mechanisms of these nonsynonymous mutations on the functions of different SARS-CoV-2 proteins\(^3\). For example, the Alpha variant (B.1.1.7) of SARS-CoV-2, first identified in the UK in late 2020, is characterized by several mutations, including the D614G mutation in the spike protein, which enhances its binding affinity to the ACE2 receptor\(^4\). This variant exhibits increased transmissibility compared to the original virus, which led to its rapid spread globally\(^5\). However, there are only a few studies on the synonymous mutations of SARS-CoV-2 genome\(^6\).\(^7\).

Synonymous mutations are also known as silent mutations because the nucleotide mutations result in a change in the RNA sequence without altering the amino acid sequence\(^8\). Synonymous mutations have been suggested to have no functional consequence on the fitness of organisms and their evolution in long term\(^7\). However, numerous recent studies had showed that synonymous mutations may affect the folding and stability of RNA structures\(^14\). Interestingly a large scale study of synonymous mutations in multiple yeast genes has shown that most of synonymous mutations are not neutral, affecting the fitness of the cell\(^15\). For RNA viruses, even though synonymous mutations generally do not change their pathogenicity directly, some studies reveal that synonymous mutations may affect the RNA secondary structure of the virus\(^16\) and also change the codon usage bias of the genes in the virus\(^17\).\(^18\). The use of mRNA-based COVID-19 vaccines reduce the severity of the disease. However, mRNA molecule is susceptible to the degradation due to the presence of 2’ OH group in the ribose. To improve the stability of mRNA vaccine, Zhang et al. (2023) designed a novel algorithm, which optimizes the codon usage and RNA secondary structure by using synonymous codon\(^19\).

The synonymous mutations play some important biological roles, which may affect viral fitness and pathogenicity. However, the study of biological consequences of synonymous mutations have been largely overlooked. In this study, we identified synonymous mutations of SARS-CoV-2 genome from early pandemic phase. We predicted the effects of these synonymous mutations of the top 10 highest frequency on RNA secondary structures and codon usage bias of SARS-CoV-2 genome. These findings allow the researchers to prioritize these mutations for function analysis in the future.

Methods

Sequence retrieval

30,229 SARS-CoV-2 genomic sequences were downloaded from GISAID database (Global Initiative on Sharing All Influenza Data, RRID:SCR_018251)\(^10\) ranging from 31 December 2019 to 22 March 2021. SARS-CoV-2 genomic sequences were filtered by setting parameters to keep only sequences with complete genome and high coverage. The sequences were further filtered to remove those sequences with higher than 0.1% “N” unresolved nucleotides and ambiguous letters. A total of 3,584 sequences were removed by applying this filter. The reference sequence of SARS-CoV-2 genome (NC_045512.2)\(^10\) was retrieved in fasta format from NCBI database (NCBI, RRID:SCR_006472). It is a Wuhan isolate with a complete genome which comprises of 29,903 bases.

Multiple sequence alignment

The rapid calculation available in MAFFT online server (MAFFT, version 7.467, RRID:SCR_011811)\(^12\) was used to perform multiple sequence alignment (MSA) for 26,645 SARS-CoV-2 genomes. This option supports the alignment of more than 20,000 sequences with approximately 30,000 sites. The alignment length was kept, which means the insertions at the mutated sequences were removed, to keep the alignment length the same as the reference sequence. While other parameters were left as default.

Identification of mutations and their frequency in SARS-CoV-2 genomes

A simple Python script was written to identify the mutations in 26,645 SARS-CoV-2 genomes. To determine whether the identified mutations are synonymous or nonsynonymous, MEGA
X software, version 10.2.5 build 10210330 (MEGA Software, RRID:SCR_000667) was utilized to perform the translation for inspection purposes. The presence of amino acid changes was identified by referring to the genomic position of the nucleotide mutations. Synonymous mutations with the top 10 highest frequencies were generated.

SARS-CoV-2 RNA secondary structure prediction
The RNA secondary structure of wild type and mutant sequences were predicted using RNAfold program, version 2.4.18 (Vienna RNA, RRID:SCR_008550) with the incorporation of SHAPE reactivity data obtained from the study done by Manfredonia et al. (2020). The RNA secondary structure prediction was performed using a sequence length of 250 nucleotides upstream and downstream of the mutation site. Other than RNAfold, another two programs which are IPknot++ version 2.2.1 (SCR_022557) and MXFold2 (SCR_022558) were also used to perform the RNA secondary structure prediction of SARS-CoV-2 wild type and mutants.

Base pair probability estimation
To predict how the mutations affect RNA local folding, base pair probability was estimated by utilizing MutaRNA, version 1.3.0 (MutaRNA, RRID:SCR_021723). MutaRNA is a web-based tool that allows prediction and visualization of the structure changes induced by a single nucleotide polymorphism (SNP) in an RNA sequence. It includes the base pair probabilities within RNA molecule of both wild type and mutant. The parameters used in MutaRNA were set as default except the window size was changed to 501nt.

Relative Synonymous Codon Usage (RSCU)
Relative synonymous codon usage (RSCU) represents the ratio of the observed frequency of codons appearing in a gene to the expected frequency under equal codon usage. RSCU is calculated using the formula:

\[ RSCU_i = \frac{X_i}{\sum_{i=1}^{n} X_i}, \]

where \( X_i \) implies the number of occurrences of codon \( i \) and \( n \) stands for the number of synonymous codons encoded for that particular amino acid.

Results and discussion
A synonymous mutation is a change in the nucleotide that does not cause any changes in the encoded amino acid. Synonymous mutations were previously considered to be less important, but they are now proven to have some effects on RNA folding, RNA stability, miRNA binding and translational efficiency. Synonymous mutations may have significant effects on the adaptation, virulence, and evolution of RNA viruses. Another study done also indicated that synonymous mutations have association with more than 50 human diseases such as hemophilia B, tuberculosis (TB), cystic fibrosis (CF), Alzheimer, schizophrenia, chronic hepatitis C and so on. All these studies show that increasing importance has been associated with synonymous mutations over these years. Hence, it is necessary for us to study the effects of synonymous mutations of SARS-CoV-2 genome.

Identification of SARS-CoV-2 synonymous mutations
A total of 381 mutations were found in SARS-CoV-2 genomes by using python script, in which 150 of them are synonymous mutations. The distribution of these 150 synonymous mutations in 11 coding regions is shown in Figure 1. Among these mutations, ORF1a and ORF1b have a higher number of synonymous mutations at 76 and 33, respectively, which might be due to...
their longer sequence length. Besides that, our findings also show high C to U mutation rate in SARS-CoV-2 genome and this mutational skews are in line with multiple studies\(^{32-38}\). The high C to U mutation rate may be driven by host APOBEC-mediated RNA editing system and overexpression of APOBEC3 protein promotes viral replication and propagation in the human colon epithelial cell line\(^{39}\). These mutational skews are necessary to be considered when deducing the selection acting on synonymous variants in SARS-CoV-2 evolution\(^{40}\). Synonymous mutations are assumed subject to a lower selective pressure than nonsynonymous mutations, presumably the purifying selection force has stronger negative impact on the frequencies of nonsynonymous mutations. Interestingly there may be some selection force on synonymous mutations shown by a few studies, suggesting that these synonymous mutations are not random and neutral, may have some biological impact on viral fitness\(^{41-47}\).

The synonymous mutations in SARS-CoV-2 genomes with the top 10 highest frequency obtained from the analysis of 150 synonymous mutations were listed in Table 1. Our sequence samples are obtained from December 2019 to March 2021 and this period overlapped with the peak of Alpha variant (B.1.1.7) outbreak\(^1\). The defining synonymous SNPs of Alpha variant include C241T, C913T, C3037T, C5986T, C14676T, C15279T and T16176C\(^3\), and all except C241T are reported in our study as well. As shown in Table 1, synonymous mutations with the highest frequency identified from SARS-CoV-2 genomes is C3037U mutation located in nsp3 of ORF1a, followed by C313U mutation in nsp1 of ORF1a and C9286U mutation in nsp4 of ORF1a. Mutations with higher frequency are mostly found in ORF1a and ORF1b. Although there are some overlapping ORFs in the SARS-CoV-2 genome, such as ORF1a and ORF1b, ORF3a and ORF3c\(^4\), the top 10 highest frequency synonymous mutations are not located in these overlapping sites.

It is of great interest to find out the effect of these top 10 synonymous mutations on SARS-CoV-2 genome. However, it is important to take note that the high frequency of some mutations is not necessarily due to their positive effects. They may emerge during early stage of pandemic and are transmitted to all of their descendants, even though they have no or little effect on viral fitness\(^{48}\).

Similar to another companion paper, which focuses on the prediction analysis of nonsynonymous mutations of SARS-CoV-2 proteins\(^{49}\), the same SARS-CoV-2 virus genome data from GISAID database ranging from 1st January 20 to 22 March 21 were used in this study. The data collection time was overlapping with the period when the frequency of alpha variant reached the highest numbers around March–May 21\(^{34}\). There are seven synonymous mutations identified as the defining mutations in the alpha variant, of which all except C241T are also reported in our study. Due to the rapid evolution of SARS-CoV-2 genome, it is beyond the scope of our study to keep track SARS-CoV-2 mutational profile and to predict the consequences of these mutations. Two independent studies reported that alpha or alpha-like SARS-CoV-2 variants are circulating among wild deer population in North America in late 2021\(^{41,42}\). Although there is no reported case of viral spillback from deer to human transmission, we can’t simply rule out this possibility yet. Hence, our findings remain relevant despite of not using the latest genome dataset.

**Table 1. SARS-CoV-2 synonymous mutations with the top 10 highest frequency.**

| ORF | Position | Nucleotide Variation | Frequency |
|-----|----------|----------------------|-----------|
| 1a  | nsp1     | 313                  | C > U     | 8212      |
|     | nsp2     | 913                  | C > U     | 2820      |
|     | nsp3     | 3037                 | C > U     | 24651     |
|     |          | 5986                 | C > U     | 2860      |
|     | nsp4     | 9286                 | C > U     | 3880      |
| 1b  | nsp12    | 14676                | C > U     | 2859      |
|     |          | 15279                | C > U     | 2842      |
|     |          | 16176                | U > C     | 2838      |
|     | nsp14    | 18877                | C > U     | 2603      |
| M   |          | 26735                | C > U     | 2432      |
incorporating the nucleotide reactivity data derived from the experiments. If the reactivity value is high, the nucleotide is less likely to be paired, or vice versa. SARS-CoV-2 virus can form pseudoknot structures, which promote ribosomal frameshifting. However, many RNA secondary structure prediction programmes don’t predict pseudoknot structure since the calculation is computationally demanding. IPknot++ is one of the few programmes, which can predict pseudoknot structure. MXfold2 predicts RNA secondary structure using deep learning method with a large amount of training dataset.

Although different tools may produce similar results for identical RNA sequences, it’s important to note that there can be variations in prediction outputs due to differences in algorithm, thermodynamic parameter settings, inclusion of pseudoknot calculation, incorporation of experimental data and the assumptions of each tool. RNAfold, IPknot++ and MXfold2 apply the nearest neighbour model, using different thermodynamic parameters. In addition, MXfold2 implements deep learning models with max margin framework. Multiple experimental genome-wide mapping of RNA secondary structures studies showed that 5’ UTR of SARS-CoV-2 RNA genome forms 7 conserved stem-loop structures, and in some studies, 8 SLs, depending on the sequence length. To demonstrate the usability of prediction tools, we predicted RNA secondary structure of the sequence of 5’ UTR (1-480 nt) of SARS-CoV-2 (Extended data 1). The RNA secondary structures predicted by RNAfold with SHAPE data, IPknot++ and MXfold2 are similar, especially SL1, SL5-8 regions and they are comparable to most of the published experimental data. Both RNAfold with SHAPE data, and MXfold2 successfully predicted SL4, but IPknot++ predicted SL4 with pseudoknot structure, which has not been reported in other studies. Interestingly the result obtained from RNAfold without SHAPE data is quite different, possibly due to the missing experimental data. In addition, it has been shown that SARS-CoV-2 may adopt different RNA secondary structure conformations. Our study is aimed to predict if the sSNP may affect RNA secondary structure and the outcomes allow us to prioritize variants for the experiment functional studies in the future. Using multiple prediction tools may help to increase the accuracy and reliability of the prediction result. The prediction results for all 10 synonymous mutations using these 3 tools and the base pair probability estimation results are summarized in the Table 2 (✓ - changes, ✗ - no change). The results for all 10 synonymous mutations predicted with RNAfold, IPknot++ and MXfold2 are available in Extended data 2, 3 and 4, respectively. The base pair probabilities for all 10 synonymous mutations are shown as circular plots in Extended data 5. The darker the edge is, the more likely the two connected bases to form base pair. Of these 10 synonymous mutations, four mutants which are all located in ORF1ab, namely C913U, C3037U, U16176C and C18877U mutants show pronounced changes between wild types and mutants in all 3 prediction tools, suggesting these synonymous mutations may have some biological impact on viral fitness. Having say that, it is also possible that other mutants with only one or two changes predicted by these analyses, may also affect RNA secondary structures, having some impact on viral fitness. It has been shown that SARS-CoV-2 virus can form elaborated RNA secondary structures at 5’ and 3’ UTRs, and frameshifting element (FSE), located between the boundary of ORF1a and ORF1ab. The 5’ UTR of SARS-CoV-2 is important for viral mRNA stability and protein translation while the 3’ UTR may be involved in viral proliferation in the host cell. Interestingly it has been observed that base substitution type, transitions from C to U base occurred at higher frequency in the stem region of RNA secondary structure of 5’ and 3’ UTR of SARS-CoV-2 genome, possibly due to the less detrimental effect on the structure. The FSE can form pseudoknot structures, which regulate the relative protein expression of ORF1a and ORF1ab during viral infection.

Table 2. Summary of RNA secondary structure prediction and base pair probability estimation analysis of SARS-CoV-2 synonymous mutations.

| Mutation   | RNAfold (SHAPE) | IPknot++ | MXfold2 | MutaRNA |
|------------|-----------------|----------|---------|---------|
| C313U      | ✗               | ✗        | ✗       | ✗       |
| C913U      | ✓               | ✓        | ✓       | ✓       |
| C3037U     | ✓               | ✓        | ✓       | ✓       |
| C5986U     | ✓               | ✗        | ✓       | ✓       |
| C9286U     | ✗               | ✓        | ✗       | ✓       |
| C14676U    | ✓               | ✓        | ✗       | ✓       |
| C15279U    | ✓               | ✓        | ✗       | ✓       |
| U16176C    | ✓               | ✓        | ✓       | ✓       |
| C18877U    | ✓               | ✓        | ✓       | ✓       |
| C26735U    | ✓               | ✓        | ✗       | ✓       |

Other than 5’ and 3’ UTRs, Huston et al. (2021) found that ORF1ab region forms extensive RNA secondary structure network. Coincidentally all four mutations, C913U, C3037U, U16176C and C18877U reported in our study are located within ORF1ab. C913U mutation is found in the Nsp2, near the start codon (position 806) in ORF1a in SARS-CoV-2 genome. As shown in Figure 2, the wild type structures predicted by RNAfold and MXfold2 shares some degree of similarity around position 95–330 of 501 base long structure. C913U mutation has a pronounced effect on RNA secondary structure predicted by RNAfold. C913U mutation results in the appearance or disappearance of multiple loops, not only at the nearby mutated residue, but also at the sites further apart, suggesting this mutation may affect its long-range RNA interaction. While MXfold2 predicts that U913 mutant forms a shorter stem and a larger hairpin loop compared to C913 wild type. However, the structure predicted by IPknot++ is quite different from others, in which, C913U results in change of pseudoknot structure. Figure 2D shows that the base pair interactions of wild type RNA are changed substantially by
Figure 2. The effect of C913U mutation on RNA secondary structure of nsp2 in ORF1a. (A) RNA secondary structure of C913 wild type and U913 mutant predicted using RNAfold. (B) RNA secondary structure of C913 wild type and U913 mutant predicted using IPknot++. (C) RNA secondary structure of C913 wild type and U913 mutant predicted using MXfold2. (D) MutaRNA circular plots of base pairing probabilities of C913 wild type and U913 mutant. The black arrow indicates the position of WT and mutated nucleotides while the red arrow indicates the starting position of the query sequence.

U913 mutation. Previously it has been shown that Nsp2 protein suppresses host immune response by inhibiting the mRNA translation of interferon gene\(^5\). Although C913U mutation does not alter the amino acid residue of Nsp2 protein, it may be worthwhile to see if this C913U mutation plays a direct or indirect role in host immune response through Nsp2 protein. Since C913U is near to Nsp1 and Nsp2 protein boundaries, the altered RNA secondary structure may affect ribosome stalling, which, in turn, affect folding of nascent polypeptides and translation initiation. In addition, Nsp1 protein facilitates viral propagation by inhibiting host protein translation machinery\(^6\) and promoting host mRNA degradation\(^7\). It will be interesting to investigate if the C913U mutation affects these functions.

C3037U mutation is found in the Nsp3 in ORF1a. As shown in Figure 3, both IPknot++ and MXfold2 predict that U3037 mutant forms longer stem and smaller internal loop compared to wild type. On the contrary, RNAfold predicts that a small internal loop fuses into a bigger internal loop in U3037 mutant. MutaRNA circular plot shows that there is some minor difference in base pair probabilities between C3037 wild type and U3037 mutant. Nsp3 is a papain-like protease, which hydrolyzes several Nsp proteins, involved in viral replication\(^8\). Hence, we should investigate the effect of this mutation on its cleavage activity, probably through the change in transcription or translation level of Nsp3.

U16176C mutation is located in the Nsp12, close to the boundary of Nsp12 and Nsp13 genes in ORF1b. As shown in Figure 4, U16176C mutation results in a drastic change in RNA secondary structure predicted using RNAfold. IPknot++ predicts C16176 mutant forms new pseudoknot structures, which are absent in wild type U16176. On the other hand, MXfold2 predicts C16176 mutant forms a larger multi-branched loop and a shorter stem compared to wild
Figure 3. The effect of C3037U mutation on RNA secondary structure of nsp3 in ORF1a. (A) RNA secondary structure of C3037 wild type and U3037 mutant predicted using RNAfold. (B) RNA secondary structure of C3037 wild type and U3037 mutant predicted using IPknot++. (C) RNA secondary structure of C3037 wild type and U3037 mutant predicted using MXfold2. (D) MutaRNA circular plots of the base pairing probabilities of C3037 wild type and U3037 mutant. The black arrow indicates the position of WT and mutated nucleotides while the red arrow indicates the starting position of the query sequence.

C16176C mutation affects base pair potential at multiple sites. Nsp12 is one of the subunits of RNA-dependent RNA polymerase (RdRp), which is required for RNA synthesis. A study showed that a 1.4-kb-long SARS-CoV-2 RNA sequence (residues 15071–16451) located in the Nsp12 and Nsp13 regions is required to facilitate viral RNA packaging. Since U16176C mutation may affect RNA secondary structure, it will be interesting to see if it affects viral RNA packaging. U16176C together with C14676U and C15279U have very similar number of frequencies as shown in Table 1. Interestingly IPknot++ predicted all of them result in changes in pseudoknot structure as shown in Extended data 3. We speculated that these three sSNPs may be functionally related. These mutations are located downstream of the frameshifting element (residues 13405–13488) and this element forms a pseudoknot to promote ribosomal frameshifting during viral replication. It has been demonstrated that synonymous mutations affect both RNA secondary structure of the ribosomal frameshift signal and frameshifting efficiency in SARS-CoV virus. Another study had shown that this ribosomal frameshifting structure in SARS-CoV-2 virus involves long-range sequence interaction of 1.5 kb. It remains to be seen whether the long-range sequence interaction for ribosomal frameshifting can go beyond 1.5kb long.

C18877U mutation is located in Nsp14 in ORF1b. As shown in Figure 5, an additional internal loop is formed in U18877 mutant predicted by RNAfold. IPknot++ predicts U18877 mutant forms extra internal loops and longer hairpin near the mutated residue and it also affects the pseudoknot structure at 2 different sites further from the mutated residue. While MXfold2 predicts U18877 mutant forms one hairpin with multiple loops instead of one hairpin as seen in wild type. The changes at multiple base pairing sites due to the U18877
Figure 4. The effect of U16176C mutation on RNA secondary structure of nsp12 in ORF1b. (A) RNA secondary structure of U16176 wild type and C16176 mutant predicted using RNAfold. (B) RNA secondary structure of U16176 wild type and C16176 mutant predicted using IPknot++. (C) RNA secondary structure of U16176 wild type and C16176 mutant predicted using Mxfold2. (D) MutaRNA circular plots of the base pairing probabilities of U16176 wild type and C16176 mutant. The black arrow indicates the position of WT and mutated nucleotides while the red arrow indicates the starting position of the query sequence.

mutation is also observed in MutaRNA circular plot. Nsp14 is important to maintain high fidelity during viral RNA synthesis.63

RSCU analysis of SARS-CoV-2

Other than affecting RNA secondary structure, it has been shown that synonymous mutations may affect protein translation efficiency and accuracy through the formation of codon usage bias (CUB), which is non-random usage of synonymous codons, common in all species.64 It is a phenomenon where some codons are preferred over others for a specific amino acid. SARS-CoV-2 replicates using host cell’s machinery and synthesizes its protein by utilizing host cellular components. Hence, codon usage bias may affect the replication of viruses65.

Relative synonymous codon usage (RSCU) is a widely used statistical approach that can be used to measure codon usage bias in coding sequences. The RSCU values of SARS-CoV-2 are shown in Table 3 and the most preferred codons for each amino acid are marked in bold. Stop codons (UAA, UAG, UGA) and codons which code for an amino acid uniquely (AUG, UGG) are excluded from RSCU analysis.

Based on the RSCU values, the synonymous codons can be classified into five groups: i) codons with RSCU value equals to 1.0 are unbiased codons; ii) codons with RSCU value > 1.0 are codons preferred in a genome; iii) codons with RSCU value < 1.0 are codons less preferred in a genome; iv) codons with RSCU value > 1.6 are codons which are over-represented in a genome; v) codons with RSCU value < 1.6 are codons which are under-represented in a genome.65 There are 15 preferred codons (RSCU value > 1.0) and 11 over-represented codons (RSCU value > 1.6) in SARS-CoV-2 genome as shown in Table 3. The preferred codons in SARS-CoV-2 genome are GCA (Ala), CGU (Arg), AAU (Asn), GAU (Asp), UGU (Cys), CAA (Gln), GAA (Glu), CAU (His), AUU (Ile), UUG (Leu), AAA (Lys), UUU (Phe), CCA (Pro), AGU (Ser) and UAU (Tyr) while the over-represented codons are GCC (Ala), AGA (Arg), GGU (Gly), CUU (Leu), UUA (Leu), UUA (Leu), CCU (Pro),
Table 3. RSCU values of SARS-CoV-2 genome.

| Amino Acid | Synonymous Codons | RSCU |
|------------|------------------|------|
| Ala        | GCA              | 1.09 |
|            | GCC              | 0.58 |
|            | GCG              | 0.16 |
|            | GCU              | 2.17 |
| Arg        | AGA              | 2.67 |
|            | AGG              | 0.81 |
|            | CGA              | 0.29 |
|            | CGC              | 0.58 |
|            | CGG              | 0.19 |
|            | CGU              | 1.46 |
| Asn        | AAC              | 0.65 |
|            | AAU              | 1.35 |

A. RNAfold (SHAPE)

![RNAfold SHAPE](image1)

B. IPknot++

![IPknot++](image2)

C. Mfold2

![Mfold2](image3)

D. MutaRNA

![MutaRNA](image4)

Figure 5. The effect of C18877U mutation on RNA secondary structure of nsp14 in ORF1b. (A) RNA secondary structure of C18877 wild type and U18877 mutant predicted using RNAfold. (B) RNA secondary structure of C18877 wild type and U18877 mutant predicted using IPknot++. (C) RNA secondary structure of C18877 wild type and U18877 mutant predicted using Mfold2. (D) MutaRNA circular plots of the base pairing probabilities of C18877 wild type and U18877 mutant. The black arrow indicates the position of WT and mutated nucleotides while the red arrow indicates the starting position of the query sequence.
### Table 4. RSCU analysis of the top 10 synonymous mutations of SARS-CoV-2 genome.

| ORF | Mutation | Codon Change |
|-----|----------|--------------|
| 1a  | nsp1     | CUC -> CUU  |
|     | nsp2     | C913U        |
|     | nsp3     | C3037U       |
|     | nsp4     | C9286U       |
| 1b  | nsp12    | CCC -> CCU  |
|     | nsp14    | C18877U      |
| M   | C26735U  | UAC -> UUA  |

| Amino Acid | Synonymous Codons | RSCU |
|------------|-------------------|------|
| Ile        | AUU               | 1.53 |
| Leu        | CUA               | 0.66 |
|            | CUC               | 0.59 |
|            | CUG               | 0.30 |
|            | CUU               | 1.75 |
|            | UUA               | 1.63 |
|            | UUG               | 1.06 |
| Lys        | AAA               | 1.31 |
|            | AAG               | 0.69 |
| Phe        | UUC               | 0.59 |
|            | UUU               | 1.41 |
| Pro        | CCA               | 1.59 |
|            | CCC               | 0.29 |
|            | CCG               | 0.17 |
|            | CCU               | 1.94 |
| Ser        | AGC               | 0.36 |
|            | AGU               | 1.43 |
|            | UCA               | 1.67 |
|            | UCC               | 0.46 |
|            | UCG               | 0.11 |
|            | UCU               | 1.96 |
| Thr        | ACA               | 1.64 |
|            | ACC               | 0.38 |
|            | ACG               | 0.20 |
|            | ACU               | 1.78 |
| Tyr        | UAC               | 0.78 |
|            | UAU               | 1.22 |
| Val        | GUA               | 0.91 |
|            | GUC               | 0.56 |
|            | GUG               | 0.58 |
|            | GUU               | 1.95 |

UCA (Ser), UCU (Ser), ACA (Thr), ACU (Thr), and GUU (Val). The presence of the preferred and over-presented codons in a genome increases the protein synthesis rate.

Table 4 shows the RSCU analysis of the top 10 synonymous mutations. The codons in bold in the ‘codon change’ column are the codons with higher RSCU value, which means they are more preferred in SARS-CoV-2 genome. Most of the mutations change the codon to a more preferred codon as shown in Table 4. Nine of the ten synonymous mutations involve changes from C to U nucleotides and eight of them are located at the third position of codon, suggesting these changes are not random and possibly subjected to some selection pressure. In agreement with our study, the excessive changes of C to U nucleotides in SARS-CoV-2 genome has been reported in multiple studies. Since the preferred codons may have a better translation efficiency and accuracy compared to the nonpreferred codons, it is possible that most of these mutations may increase the viral fitness. While a study show that RNA secondary structures may be functionally linked to protein translation based on the evidence obtained from experimental work, it is difficult for us to establish the connection solely using in silico studies.

### Conclusions

The effects of SARS-CoV-2 synonymous mutations in various aspects such as RNA secondary structure and codon usage bias were studied, even though they do not cause changes in amino acid residue of the protein. C913U, C3037U, U16176C and C18877U mutants show pronounced changes between wild type and mutant predicted in all 3 RNA secondary structure prediction tools, suggesting these mutations may have some biological impact on viral fitness. In addition, these mutations showed changes in base pair potential estimated by MutaRNA. All mutations except U16176C change the codon to a more preferred codon, which may result in higher translation efficiency. Due to the shortcomings of prediction tools, experimental studies, such as protein translation assays, RNA packaging assays, are needed to give a more comprehensive understanding of the biological consequences of synonymous mutations on SARS-CoV-2 virus.

### Ethics and dissemination

No ethical approval is required for data analysis in this study (EA2702021).
Data and software availability

Underlying data
SARS-CoV-2 virus genome sequence data were obtained from the GISAID Database. The multiple alignment data can be assessed through FigShare.

Figshare: MSA (SARS-CoV-2). https://doi.org/10.6084/m9.figshare.20486178.v1

Extended data
Figshare: RNA secondary structure prediction and base pair probability estimation analysis
https://doi.org/10.6084/m9.figshare.20486166.v6

Extended data 1. Comparison of RNA secondary structure of SARS-CoV-2 5’ UTR (1-480 nt) predicted using RNAfold without SHAPE data, RNAfold with SHAPE data, IPknot++, MXfold2

Extended data 2. The RNA secondary structure of SARS-CoV-2 genome predicted using RNAfold.

Extended data 3. The RNA secondary structure of SARS-CoV-2 genome predicted using IPknot++.

Extended data 4. The RNA secondary structure of SARS-CoV-2 genome predicted using MXfold2.

Extended data 5. The base pair probabilities of SARS-CoV-2 genome estimated using MutRNA

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software
The python code for the identification of SARS-CoV-2 genome mutations can be assessed through GitHub.

Author contributions
CHN contributes to the concept, design, supervision of the project. WXB and SBZ contribute to the design, methodology, and data collection. WXB contributed to the analysis, and interpretation of data. All authors were involved in drafting and revising the manuscript and approved the final version.

References
1. Huang C, Wang Y, Li X, et al.: Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395(10223): 497-506. PubMed Abstract | Publisher Full Text | Free Full Text
2. Wu D, Wu T, Liu Q, et al.: The SARS-CoV-2 outbreak: what we know. Int J Infect Dis. 2020; 94: 44-48. PubMed Abstract | Publisher Full Text | Free Full Text
3. Sharma A, Tripathi S, Deb MK, et al.: Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. Int J Antimicrob Agents. 2020; 56(2): 106054. PubMed Abstract | Publisher Full Text | Free Full Text
4. Sanjuan R, Domingo-Calap P: Mechanisms of viral mutation. Cell Mol Life Sci. 2016; 73(23): 4433-4448. PubMed Abstract | Publisher Full Text | Free Full Text
5. Carabelli AM, Peacock TP, Thorne LG, et al.: SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat RevMicrobiol. 2023; 21(3): 162-177. PubMed Abstract | Publisher Full Text | Free Full Text
6. Oono S, Zhang Y, Ohe H, et al.: SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. Nat Commun. 2021; 12(1): 848. PubMed Abstract | Publisher Full Text | Free Full Text
7. Plante JA, Liu Y, Liu J, et al.: Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021; 592(7852): 116-121. PubMed Abstract | Publisher Full Text | Free Full Text
8. Zhu L, Wang Q, Zhang W, et al.: Evidence for selection on SARS-CoV-2 RNA translation revealed by the evolutionary dynamics of UTMs and CDSs. RNA Biol. 2022; 19(1): 866-876. PubMed Abstract | Publisher Full Text | Free Full Text
9. de Maio N, Walker CR, Turakhia Y, et al.: Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol Evol. 2021; 13(5): eva087. PubMed Abstract | Publisher Full Text | Free Full Text
10. Bin Y, Wang X, Zhao L, et al.: An analysis of mutational signatures of synonymous mutations across 15 cancer types. BMC Med Genet. 2019; 20(Suppl 2): 190. PubMed Abstract | Publisher Full Text | Free Full Text
11. Sharp PM, Averof M, Lloyd AT, et al.: DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci. 1999; 349(1329): 241-247. PubMed Abstract | Publisher Full Text
12. Charnay J, Parmley JL, Hurst LD: Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006; 7(2): 98-108. PubMed Abstract | Publisher Full Text
13. Shen X, Song S, Li C, et al.: Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature. 2022; 606(7915): 725-731. PubMed Abstract | Publisher Full Text
14. Muller S, Papamichael D, Coleman JR, et al.: Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virucity by lowering specific infectivity. J Virol. 2006; 80(19): 9687-9696. PubMed Abstract | Publisher Full Text | Free Full Text
15. Lauring AS, Acevedo A, Cooper SB, et al.: Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe. 2012; 12(5): 623-632. PubMed Abstract | Publisher Full Text | Free Full Text
16. Zhang H, Zhang L, Lin A, et al.: Algorithm for optimized mRNA design improves stability and immunogenicity. Nature. 2023; 621(7978): 396-403. PubMed Abstract | Publisher Full Text | Free Full Text
17. Mafft - multiple sequence alignment program. [Accessed: 23-Sep-2021]. Reference Source
18. Wu F, Zhao S, Yu B, et al.: A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579(7798): 265-269. PubMed Abstract | Publisher Full Text | Free Full Text
19. SHAPE directed RNA folding. Bioinformatics. 2016; 32(1): 145-7. PubMed Abstract | Publisher Full Text | Free Full Text
20. GISAID Initiative. [Accessed: 23-Sep-2021]. Reference Source
21. Kumar S, Stecher G, Li M, et al.: MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018; 35(6): 1547-1549. PubMed Abstract | Publisher Full Text | Free Full Text
22. Manfredonia I, Nithin C, Ponce-Salvatiera A, et al.: Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements.
28. Miladi M, Raden M, Diederichs S, et al.: Targeting stem-loop 1 of the SARS-CoV-2 5' UTR to suppress viral translation and Nsp1 evasion. Proc Natl Acad Sci U S A. 2022; 119(9): e21179619.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Yuan J, Feng Z, Wang Q, et al.: 3'UTR of SARS-CoV-2 spike gene hijacks host miR-296 or miR-520 to disturb cell proliferation and cytokine signaling. Front Immunol. 2022; 13: 924667.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Xu Z, Choi JH, Dai DL, et al.: SARS-CoV-2 Impairs interferon production via NSP2-induced repression of mRNA translation. Proc Natl Acad Sci U S A. 2022; 119(32): e2204539119.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Schubert K, Karoussis ED, Jomaa A, et al.: SARS-CoV-2 Nsp5 binds the ribosomal mRNA channel to inhibit translation. Nat Struct Mol Biol. 2020; 27(10): 959-966.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Burke JM, St Clair LA, Perera R, et al.: SARS-CoV-2 infection triggers widespread host mRNA decay leading to an mRNA export block. RNA. 2021; 27(11): 1318-1329.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Wolff G, Limpens RWAL, Zevenhoven-Dobbe JC, et al.: A molecular pore spans the double membrane of the coronavirus replication organelle. Science. 2020; 369(6509): 1395-1398.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Naydenova K, Muir KW, Wu LF, et al.: Structure of the SARS-CoV-2 RNA-dependent RNA polymerase in the presence of favipiravir-RTP. Proc Natl Acad Sci U S A. 2021; 118(7): e2021946118.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Terasaki K, Narayanan K, Makino S: Identification of a 1.4-kb-long sequence located in the nsp12 and nsp13 coding regions of SARS-CoV-2 genomic RNA that mediates efficient viral RNA packaging. J Virol. 2020; 94(1): 253.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Moeller NH, Shi K, Demir O, et al.: Structure and dynamics of SARS-CoV-2 proofreading exonuclease Eex07. Proc Natl Acad Sci U S A. 2020; 119(9): e2016379119.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Sun M, Zhang J, et al.: Preferred synonymous codons are translated more accurately: proteomic evidence, among-species variation, and mechanistic basis. Sci Adv. 2020; 6(27): eaaz9812.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Wong EH, Smith DK, Rababde R, et al.: Codon usage bias and the evolution of influenza a viruses. Codon usage biases of influenza virus. BMC Evol Biol. 2010; 10(1): 253.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Sharp PM, Tushy TM, Mosaorski KR: Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 1986; 14(13): 5125-5143.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Mäger DA, Cabral BJ, Premsankav V, et al.: mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A. 2019; 116(48): 24075-24083.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Boon WX, Ng CH: SARS (SARS-CoV-2), (accessed Aug. 17, 2022).
Reference Source
Open Peer Review

Current Peer Review Status:  ✗  ?  ✓  ✗  ?  ?

Version 4

Reviewer Report 27 September 2024

https://doi.org/10.5256/f1000research.171513.r324668

© 2024 Forni D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Diego Forni
Scientific Institute IRCCS E Medea, Bosisio Parini, Italy

I would like to thank the authors for answering my questions.
I still believe that it is not optimal to report counts without taking into account ORD length in figure 1.
I still also believe that differences among RNA secondary structures are relevant and I am not really sure that "C913U, C3037U, U16176C and C18877U mutants show pronounced changes between wild type and mutant" as stated in the conclusion.

Finally the two sentences starting with "Synonymous mutations are assumed subject to a lower selective pressure than nonsynonymous mutations, presumably.." in the introduction are not clear and need to be better explained.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: viral genomics, viral evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 3

Reviewer Report 11 September 2024

https://doi.org/10.5256/f1000research.162881.r304727
In this work, synonymous mutations of the SARS-CoV-2 genome are explored, with the rationale that these mutations impact function through altered RNA folding, despite unaltered protein products. Specifically, the researchers find 150 synonymous mutations after performing multiple sequence alignment and examining the protein products, and mapped the distributions in the coding regions. RNA secondary structure predictions and base pair probability calculations are then presented for 4 mutations C913U, C3037U, U16176C and C18877U that show pronounced changes between wildtype and mutant structures. Different prediction tools were used to confirm the structure predictions. The value of the study, while interesting, is limited because the long RNA lengths used make the disparate predicted structures unreliable. If the Information Classification: General authors examine some mutant systems experimentally with chemical reactivity like SHAPE or DMS, the results may be more meaningful.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
Diego Forni
Scientific Institute IRCCS E Medea, Bosisio Parini, Italy

In this manuscript, Dr Boon et al. analyzed the most common synonymous mutations of early pandemic SARS-CoV-2 genomes. They identified substitutions that influence the RNA secondary structure surrounding these positions. The authors have already modified and updated their manuscript based on previous comments, still there are few open questions in my opinion.

First of all, it is clear that the dataset is not up to date and it will not be changed, still this should be clearly stated, starting from the title, reporting or the time period or something like "at the beginning of the pandemic".

- Following this, also the introduction section should be focused on what we know about genomic variability and viral evolution in the time period being analyzed herein. If I understood well, most of these sequences belong to the alpha variant.
- Moreover, "variant of concern" (abbreviated VOC not VOI as stated in the text) refers to a specific lineage that carries several and distinct mutations. As it is stated in the text now it seems that a single mutation is a VOC. Please check it.
- It is expected that most mutations are found in ORF1ab, as this covers most of the genome, so counts should be normalized by ORF length (e.g. fig.1).
- Table 1 should report mutation frequency not count.
- Authors should discuss more why different predictors give very different results in terms of RNA secondary structure, and how this can affect their analysis.
- Why did the author select 250 bp flanking positions for their prediction? Which is the rationale? How does this influence the results?
- Are there any other mutations in the 501 nucleotide regions? Could mutations have an impact on predictions?
- How did the authors handle overlapping/internal ORFs? In these situations a syn mutations could also be non-synonymous for the other ORF. I do not think this is the case for the top 10 mutations, but still this should be explained.
- The CUB section seems to me not really linked to the rest of the analyses and could be expanded more.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** viral genomics, viral evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

---

Author Response 05 Sep 2024

**Chong Han Ng**

In this manuscript, Dr Boon et al. analyzed the most common synonymous mutations of early pandemic SARS-CoV-2 genomes. They identified substitutions that influence the RNA secondary structure surrounding these positions. The authors have already modified and updated their manuscript based on previous comments, still there are few open questions in my opinion.

First of all, it is clear that the dataset is not up to date and it will not be changed, still this should be clearly stated, starting from the title, reporting or the time period or something like "at the beginning of the pandemic".

**Author response:** The title has been revised to “Prediction of the effects of the top 10 synonymous mutations from 26645 SARS-CoV-2 genomes of early pandemic phase”.

**Reviewer comment:** Following this, also the introduction section should be focused on what we know about genomic variability and viral evolution in the time period being analyzed herein. If I understood well, most of these sequences belong to the alpha variant.

**Author response:** We added a few lines about Alpha variant in introduction, and result and discussion section (Identification of SARS-CoV-2 synonymous mutations).

**Reviewer comment:** Moreover, “variant of concern” (abbreviated VOC not VOI as stated in the text) refers to a specific lineage that carries several and distinct mutations. As it is stated in the text now it seems that a single mutation is a VOC. Please check it.

**Author response:** The abbreviation has been corrected from VOI to VOC. In the text, we mentioned that different SARS-CoV-2 variants with multiple synonymous and nonsynonymous mutations have been reported since the beginning of the outbreak. Therefore, it is clear that VOCs contain multiple mutations.

**Reviewer comment:** It is expected that most mutations are found in ORF1ab, as this covers most of the genome, so counts should be normalized by ORF length (e.g. fig.1).

**Author response:** Figure 1 show the distribution of 150 different types of synonymous
mutations identified from 26645 SARS-CoV-2 genomes. It is not about the count or the average of the mutations in each ORFs, thus it should not be normalized.

**Reviewer comment:** Table 1 should report mutation frequency not count.
**Author response:** Table 1 shows the mutation frequency as reported in this paper.

**Reviewer comment:** Authors should discuss more why different predictors give very different results in terms of RNA secondary structure, and how this can affect their analysis.
**Author response:** In the paper, there is about two paragraphs on comparison of these RNA secondary structure tools and we highlighted their similarities and differences of these tools. As mentioned in the discussion, we expected some differences in the outcomes obtained from different RNA secondary structure prediction tools due to the differences in algorithms, thermodynamic parameter settings, inclusion of pseudoknot calculation, incorporation of experimental data and the assumptions of each tool. Furthermore, SARS-CoV-2 may adopt different RNA secondary structure conformations. However, our major focus is to predict if the sSNP may affect/change RNA secondary structure, rather than the predicted structures per se. Importantly the outcomes allow us to prioritize variants for the experiment functional studies in the future. In addition, we demonstrated the usability of prediction tools by using 5' UTR (1-480 nt) of SARS-CoV-2, which is well-characterized in other studies in the Extended data 1. For the four sSNPs which showed pronounced changes, we also compared the differences of these structures predicted by different tools.

**Reviewer comment:** Why did the author select 250 bp flanking positions for their prediction? Which is the rationale? How does this influence the results?
**Author response:** Some sSNPs affect only RNA secondary structures of neighboring nucleotides while some sSNPs may have long-ranging effects, in addition to the changes in the neighboring nucleotides. The length of RNA sequences used should not be too short since we may miss out the long-ranging effects. On the other hand, we need to bear in mind that most of RNA secondary structure prediction tools are not suitable for long sequences analysis. We tried out different lengths of flanking sequences, including 100b, 250b, 500b and 1000b, and we found out that 250b flanking sequences produce the optimal results.

**Reviewer comment:** Are there any other mutations in the 501 nucleotide regions? Could mutations have an impact on predictions?
**Author response:** Yes, there may be other synonymous mutations within the 501 nucleotide regions, but they are not from the top 10 synonymous mutations. We are not doing epistasis analysis of different mutations since it is out of the scope of this study.

**Reviewer comment:** How did the authors handle overlapping/internal ORFs? In these situations a syn mutations could also be non-synonymous for the other ORF. I do not think this is the case for the top 10 mutations, but still this should be explained.
**Author response:** Yes, we don't find any synonymous mutation of one ORF, which is also a synonymous mutation for another ORF. We included one sentence about the overlapping ORFs in the result and discussion section in Identification of SARS-CoV-2 synonymous mutation. Since some mutations, including C913U, U16176C are located near the boundaries of different ORFs, we postulated their potential long ranging effects at the neighboring ORFs in the discussion section.
Reviewer comment: The CUB section seems to me not really linked to the rest of the analyses and could be expanded more.

Author response: We expanded CUB section, including additional relevant references, elaborating the results further to make it clearer to the reader.

Competing Interests: No competing interests were disclosed.

Version 2

Reviewer Report 18 October 2023

https://doi.org/10.5256/f1000research.137703.r208617

© 2023 Huber R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Roland Huber
Bioinformatics Institute, A*STAR, Singapore

I agree with previous reviewers that the analyzed sequences represent only a limited sample of variation in SARS-CoV-2, specifically from early in the pandemic. This might introduce unexpected biases in the analysis. E.g. it would be more likely to observe host adaption early on which would be consistent with more favourable codon usage.

With regard to the data that was analysed, the structure models obtained show limited consistency. The authors state that they do not expect the used tools to concur on the structures since they use different algorithms. This is concerning, as one would expect unambiguous structures to be consistent, even using different methodologies. Other tools, e.g. RNAstructure, also allow the inclusion of shape data and the prediction of pseudoknots. We are thus left with a series of diverging structure predictions and unsure what, if any, effect these specific mutations have. This is not helped by inconsistent presentation of the results. Figures 2-5 use different visualisations for the results of the 4 tools employed, which makes comparisons of the structures difficult for the reader.

The study unfortunately does nothing to associate the regions or structural elements with any type of functional or biological information. We are thus left with a study that analyses a limited set of synonymous mutations using inconsistent structure predictions and offers no additional biological insight.

Is the work clearly and accurately presented and does it cite the current literature?
Partly
Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational Biology, Structural Genomics, RNA biology, Virology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 06 September 2022
https://doi.org/10.5256/f1000research.137703.r149467

© 2022 Nithin C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chandran Nithin
Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

Authors have addressed the comments. However, the data set used in the study is outdated.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly
If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA structure prediction; Computational Structural Biology; RNA-protein complexes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Chandran Nithin
Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

The manuscript titled "Prediction of the Effects of Synonymous Variants on SARS-CoV-2 Genome" analyzes 30,299 SARS-CoV-2 genomes retrieved from GISAID for the period between 31st December 2019 and 22nd March 2021. The study identifies 381 mutations in the genome, of which 150 are synonymous. The study discusses two mutations, C913U and C26735U.

1. The dataset used in this study is outdated and represents only a short period between 31st December 2019 and 22nd March 2021. It will be necessary to update the data to include more sequences from a wider time frame.

   The authors had indicated (through the editorial manager) that they are unwilling to update the dataset; the authors need to indicate the same in the title and detail the limitations of the dataset in the introduction section.

2. The authors have used the RNAfold program to predict the secondary structure of wild-type and mutant sequences. The authors need to use additional prediction tools and check for
consensus between the predicted structure. Moreover, authors need to employ prediction tools to predict pseudoknots in the RNA secondary structure.

3. The secondary structure of SARS-CoV2 determined with the help of data from SHAPE-MaP, icSHAPE, and DMS-MaPseq experiments is available in the literature.\textsuperscript{1,2,3} The authors should utilize this data to augment this study.

4. The study requires the inclusion of appropriate controls which quantify the effect of mutations on the stability of the RNA structure. The authors state that "The minimum free energy value of the mutant (-146.90 kcal/mol) is slightly less negative than that of the wild type (-147.40 kcal/mol), which makes it a less thermodynamically stable structure compared to the wild type." This claim needs to be explored in detail and supported with evidence from experimental data and literature.

5. The authors state that C26735U mutation induces changes in the predicted RNA secondary structure by forming an extra multibranch loop at the mutation site. The authors should compare the same with secondary systems mentioned in point 3 (See Ref 1, 2, and 3). The authors need to perform a similar analysis for C913U.

6. The study identifies 381 mutations in the genome, of which 150 are synonymous; however, it discusses only two. The authors need to provide the data on the remaining 148 synonymous mutations and briefly discuss some of them in the manuscript.

7. The authors need to expand and include relevant literature in the introduction section.

8. The manuscript needs to be edited for language errors.

References
1. Lan T, Allan M, Malsick L, Woo J, et al.: Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nature Communications. 2022; 13 (1). Publisher Full Text
2. Huston N, Wan H, Strine M, de Cesaris Araujo Tavares R, et al.: Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Molecular Cell. 2021; 81 (3): 584-598.e5 Publisher Full Text
3. Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, et al.: Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Research. 2020; 48 (22): 12436-12452 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?  
No

Is the study design appropriate and is the work technically sound?  
Partly

Are sufficient details of methods and analysis provided to allow replication by others?  
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA structure prediction; Computational Structural Biology; RNA-protein complexes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 22 Aug 2022

Chong Han Ng

The manuscript titled "Prediction of the Effects of Synonymous Variants on SARS-CoV-2 Genome" analyzes 30,299 SARS-CoV-2 genomes retrieved from GISAID for the period between 31st December 2019 and 22nd March 2021. The study identifies 381 mutations in the genome, of which 150 are synonymous. The study discusses two mutations, C913U and C26735U.

1. The dataset used in this study is outdated and represents only a short period between 31st December 2019 and 22nd March 2021. It will be necessary to update the data to include more sequences from a wider time frame.

The authors had indicated (through the editorial manager) that they are unwilling to update the dataset; the authors need to indicate the same in the title and detail the limitations of the dataset in the introduction section.

Response: The landscape of mutational profile of SARS-CoV-2 genome is very dynamic, changing rapidly. As of 14 August 2022, there are more than 12 million SARS-CoV-2 virus genomes sequences deposited in GISAID database. Although it is useful to get more recent dataset, it is out of the scope of our study to update the data. Both identification of the SARS-CoV-2 mutations and the prediction analysis of the biological consequences of these mutations are very time-consuming processes. Based on the timelines of emergence of different variants, the cut-off date of our data collection overlaps with the period when the frequency of the alpha variant peaked. All the synonymous mutations except C241T reported in the alpha variant have been identified in our study as well. Although the alpha variant has been replaced by other variants in human population, it is detected from wildlife animals based on a few recent genomic surveillance analysis. It is possible that these alpha variant from wildlife animal may jump back to human, hence it remains relevant to predict the
biological consequences of these synonymous mutations. The title of the paper has been revised accordingly. The limitation of the dataset was included in the discussion.

2. The authors have used the RNAfold program to predict the secondary structure of wild-type and mutant sequences. The authors need to use additional prediction tools and check for consensus between the predicted structure. Moreover, authors need to employ prediction tools to predict pseudoknots in the RNA secondary structure.

Response: To improve the outcome of the RNA secondary structure analysis, 3 prediction tools, namely RNAfold, IPknot++ and MXfold2 are used in our study. Of these, IPknot++ predicted some synonymous mutations may affect the pseudoknot formation. The prediction results for all top 10 synonymous mutations using these 3 tools are summarized in the Table 2.

3. The secondary structure of SARS-CoV2 determined with the help of data from SHAPE-MaP, iCSHAPE, and DMS-MaPseq experiments is available in the literature. The authors should utilize this data to augment this study.

Response: To improve the outcome of the prediction analysis, the RNA secondary structure of wild type and mutant sequences were predicted using RNAfold program with the incorporation of SHAPE reactivity data obtained from the study done by Manfredonia et al. (2020).

4. The study requires the inclusion of appropriate controls which quantify the effect of mutations on the stability of the RNA structure. The authors state that "The minimum free energy value of the mutant (- 146.90 kcal/mol) is slightly less negative than that of the wild type (- 147.40 kcal/mol), which makes it a less thermodynamically stable structure compared to the wild type." This claim needs to be explored in detail and supported with evidence from experimental data and literature.

Response: Instead of comparing minimum free energy between the wild type and mutant, we applied three RNA secondary structure prediction tools to predict the effect of the mutation on RNA secondary structure.

5. The authors state that C26735U mutation induces changes in the predicted RNA secondary structure by forming an extra multibranch loop at the mutation site. The authors should compare the same with secondary systems mentioned in point 3 (See Ref 1, 2, and 3). The authors need to perform a similar analysis for C913U.

Response: The results of the 4 mutations showing changes in all 3 prediction tools are included in Figure 2-5.

6. The study identifies 381 mutations in the genome, of which 150 are synonymous; however, it discusses only two. The authors need to provide the data on the remaining 148 synonymous mutations and briefly discuss some of them in the
manuscript.

**Response:** The RNA secondary structures prediction results using 3 prediction tools for the top 10 synonymous mutations have been included in Extended data 1-3. There were 4 mutations showing changes in all 3 prediction tools and the results are discussed in the manuscript.

7. The authors need to expand and include relevant literature in the introduction section.

**Response:** The relevant literature in the introduction section has been updated.

8. The manuscript needs to be edited for language errors.

**Response:** The language errors have been edited accordingly.

**Competing Interests:** No competing interests were disclosed.
sequencing with complete genome and high coverage still contain bad sequences, how did the authors exclude those sequences?

In the abstract section, the results part only included those mutations with high frequency but did not contain results of other analyses. Instead, they included the results of RNA secondary structure analysis in the conclusion.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Clinical Virology Diagnosis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

**Author Response 22 Aug 2022**

**Chong Han Ng**

The manuscript entitled “Prediction of the Effects of Synonymous Variants on SARS-CoV-2 Genome” reports an analysis of the SARS-CoV-2 genome obtained in the GISAID database. The authors analyzed synonymous changes of the 30,229 SARS-CoV-2 sequences available online, and the effects of these changes on RNA secondary structures and base-pair probability. They identified 150 synonymous mutations in 11 coding regions. Mutations with higher frequency are mostly found in ORF1a and ORF1b and two mutations (C913U and C26735U) found with effect on the predicted RNA secondary structure are not those with higher frequency.

This study collected 30,229 sequences in the database in the period of 31 Dec 2019 to 22
March 2021. As of 17 Dec 2021, there are over 4.4 million complete genomes with high coverage. Will these sequences used in the study include all variants reported or what sequences of variants are included? A table to summarize all variants with these mutations or not is needed.

**Response:** We downloaded the raw data on 23rd March 2021 and completed the data analysis in late June 2021. We submitted the manuscript in late August 2021. The landscape of mutational profile of SARS-CoV-2 genome is very dynamic, changing rapidly. It is far beyond the scope of this study to include all variants reported. Based on the timelines of emergence of different variants, the cut-off date of our data collection overlaps with the period when the frequency of the alpha variant peaked. All the synonymous mutations except C241T reported in the alpha strain have been identified in our study as well. We added new paragraph in the discussion section to discuss the impact of our findings and to explain why the study of alpha variant remains relevant now. In addition, we used 3 RNA secondary prediction analysis tool, instead of one, to improve the outcome of the analysis. The prediction results for all top 10 synonymous mutations using these 3 tools and the base pair probability estimation results are summarized in the Table 2.

A major concern is that the sequences even filtered by setting parameters to keep only sequencing with complete genome and high coverage still contain bad sequences, how did the authors exclude those sequences?

**Response:** In our original method, we used the high coverage filter when downloading from GISAID database, in which only entries with less 1% N and 0.05% unique amino acids mutations are included. To further reduce bad sequences, we filtered to remove those sequences with higher than 0.1% N unresolved nucleotides and ambiguous letters. A total of 3584 sequences were removed by applying this filter. The list of the synonymous mutations remains unchanged despite of the revision of the mutation frequency.

In the abstract section, the results part only included those mutations with high frequency but did not contain results of other analyses. Instead, they included the results of RNA secondary structure analysis in the conclusion.

**Response:** The abstract has been revised to update the results and conclusion.

**Competing Interests:** No competing interests were disclosed.
IBM TJ Watson Research Center, Yorktown Heights, NY, USA

Boon et al. made an investigation of synonymous mutations of SARS-CoV-2 genomes. Authors first identified common synonymous mutations such as 3037C>T followed by secondary structure predictions using RNAfold.

First, the number of genomes authors used is too small and too old. Currently, nearly 5 million genomes are uploaded at GISAID; however, authors used only 30,229. The set does not represent the current mutational landscape of SARS-CoV-2. Authors need to update the data. I have 3 million genomes analyzed and 18108 distinct synonymous mutations identified. The list of highest frequency in Table 1 looks completely different with the latest dataset.

Secondly, authors employed RNAfold for RNA secondary structure prediction. Although RNAfold is a standard software for the task, the accuracy of the prediction is not impressive. Authors need to explore various other tools for the secondary structure or justify the usage of RNAfold.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome Analysis, SARS-CoV-2, Cancer, Immunology, Stem cell

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 04 Nov 2021

Chong Han Ng

Boon et al. made an investigation of synonymous mutations of SARS-CoV-2 genomes.
Authors first identified common synonymous mutations such as 3037C>T followed by secondary structure predictions using RNAfold.

First, the number of genomes authors used is too small and too old. Currently, nearly 5 million genomes are uploaded at GISAID; however, authors used only 30,229. The set does not represent the current mutational landscape of SARS-CoV-2. Authors need to update the data. I have 3 million genomes analyzed and 18108 distinct synonymous mutations identified. The list of highest frequency in Table 1 looks completely different with the latest dataset.

RE: Thank you for the comments and feedbacks. We downloaded SARS-CoV-2 virus genome data submitted from 31 December 19 to 22 March 21 on 23 March 21. There were about a total of 842,603 SARS-CoV-2 virus genomes sequences before applying filters. SARS-CoV-2 genomic sequences were filtered by setting parameters to keep only sequences with complete genome, high coverage, low coverage excluded and patient status options. The use of the low coverage excluded and patient status filters were unintentionally left out in the manuscript and they will be added later. In our study, the use of these filters allows us to analyze high quality data to identify SARS-CoV-2 mutations without the requirement of high computing power. Our data is representative of the mutational landscape of SARS-CoV-2 for this specified period. Same synonymous mutations, including C913T, C3037T, C5986T, C14676T, C15279T and T16176C were identified in alpha strain by other research groups (https://covariants.org/variants/20I.Alpha.V1) and (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf). The frequency of alpha strain peaked in April-May 21 and subsequently it has been quickly replaced by delta strain. Interestingly the defining synonymous mutations of delta strain (https://covariants.org/variants/21A.Delta) are very different from those of alpha strain. The mutational landscape of SARS-CoV-2 genome is very dynamic as shown by other research groups. Bear in mind that the main objective of our paper is not to monitor the mutational changes of SARS-CoV-2 genome. Although it is useful to get more recent dataset, it is out of the scope of our study to update the data. Both identification of the SARS-CoV-2 mutations and the prediction analysis of the biological consequences of these mutations are very time-consuming processes. The first draft of the manuscript was prepared in late June 21 and submitted in late August 21. Since different SARS-CoV-2 variants may have different sets of mutations, we will gain more insight to study the effects of different mutations in different SARS-CoV-2 variants. In our follow-up study, we plan to predict and compare the effect of the mutations of different variants. These analyses may provide some plausible explanation why some variants are replaced by other variants eventually.

Secondly, authors employed RNAfold for RNA secondary structure prediction. Although RNAfold is a standard software for the task, the accuracy of the prediction is not impressive. Authors need to explore various other tools for the secondary structure or justify the usage of RNAfold.

RE: We chose RNAfold because it is the most commonly used tools to predict the RNA secondary structure. Other than RNAfold, we also considered using Mfold to perform the RNA secondary structure prediction as well. However, the Mfold server was down during
the time of the study. RNAfold computes an optimal structure over the whole sequence length, which means its performance and accuracy can be affected as the sequences used get longer. However, in our case, the accuracy of the prediction is still acceptable since the sequence we used to do the structure prediction is relatively short. Besides, the prediction results we obtained from RNAfold is further supported with the prediction results obtained from MutaRNA, which predicts the structural changes induced by the mutation by estimating the base pairing probabilities. In our results, the circular plot from MutaRNA shows the changes in the base pairing probabilities near the mutation site correlates well with the RNA secondary structure predicted by RNAfold.

**Competing Interests:** No competing interests were disclosed.

---

**Author Response 21 Nov 2022**

**Chong Han Ng**

First, the number of genomes authors used is too small and too old. Currently, nearly 5 million genomes are uploaded at GISAID; however, authors used only 30,229. The set does not represent the current mutational landscape of SARS-CoV-2. Authors need to update the data. I have 3 million genomes analyzed and 18108 distinct synonymous mutations identified. The list of highest frequency in Table 1 looks completely different with the latest dataset.

**RE:** We submitted a revised version in Sep 2022. Although we didn't use the latest dataset, we provide some justification in the discussion to explain why our data may remain relevant. We hope you will consider to review our manuscript again. The landscape of mutational profile of SARS-CoV-2 genome is very dynamic, changing rapidly. It is far beyond the scope of this study to include all variants reported. Based on the timelines of emergence of different variants, the cut-off date of our data collection overlaps with the period when the frequency of the alpha variant peaked. All the synonymous mutations except C241T reported in the alpha strain have been identified in our study as well. Although the alpha variant has been replaced by other variants in human population, it is detected from wildlife animals based on a few recent genomic surveillance analysis. It is possible that these alpha variant from wildlife animal may jump back to human, hence it remains relevant to predict the biological consequences of these synonymous mutations. The title of the paper has been revised accordingly. The limitation of the dataset was included in the discussion.

Secondly, authors employed RNAfold for RNA secondary structure prediction. Although RNAfold is a standard software for the task, the accuracy of the prediction is not impressive. Authors need to explore various other tools for the secondary structure or justify the usage of RNAfold.

**RE:** In the revised version, we used 3 RNA secondary prediction analysis tool, namely RNAfold, IPknot++, to improve the outcome of the analysis. The prediction results for all top 10 synonymous mutations using these 3 tools and the base pair probability estimation results are summarized in the Table 2.
**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com