Analysis of C-terminal Domain ORF6 for Mutation Pattern SARS-Cov-2 using Slicing Index

Rini Arianty
Gunadarma University
Margonda Raya Street
Depok, Indonesia

Ety Sutanty
Gunadarma University
Margonda Raya Street
Depok, Indonesia

Esti Setiyaningsih
Gunadarma University
Margonda Raya Street
Depok, Indonesia

ABSTRACT
This study is to ascertain the extent to which ORF6 affects the rate at which the SARS-CoV-2 virus spreads as well as the variables that may contribute to the ORF6 protein's propensity to do so. If the virus with one or more mutations has a different phenotypic from the original virus, then the SARS-CoV-2 virus may have undergone a mutation. The genetic coding of the virus contains mutations. Compared to DNA viruses and bacteria, the SARS-CoV-2 virus is a single-stranded positive RNA virus with a fragile structure that is simple to modify. As a result, RNA viruses in vitro have a greater capacity for mutation. This research will analysis the ORF6 Protein SARS-CoV-2 dataset, which has mutations or not in the C-terminal section. Proteins that do not have their own mutations have a much higher replication capacity than those that have mutations. The present study employs Biopython and Slicing Index technique to investigate the ORF6 protein's interaction with cells and its impact on the immune system by means of visualizing the protein group data on string pieces by dividing the data based on the mutation index.

General Terms
Bio-informatics

Keywords
Biopython; ORF6; SARS-CoV-2; Slicing; Viral;

1. INTRODUCTION
The SARS-CoV-2 virus, better known as the COVID-19 virus [2], is a virus that spreads over the world and has a significant influence [3] on public health and the economy. The discovery of this virus highlighted the importance of understanding the virus mutation cycle [4] to be able to deal with it. Bioinformatics is one way that can play an important role in efforts to understand the SARS-CoV-2 virus. The emergence of the use of bioinformatics technology and tools encourages researchers to study the viral genome sequence and analyze the spread of the SARS-CoV-2 virus. [5]. The corona virus gene series consists of the ORF1ab gene, S gene, ORF3a gene, E gene, M gene, ORF6 gene, ORF7a gene, ORF 7b gene, ORF 8 gene, N gene, ORF10 gene. [6]. The addition of the SARS-CoV-2 variant increased along with the increase in gene mutations in the corona virus [7].

Several proteins in the SARS-CoV-2 virus play a role in causing Covid-19 in humans [8], such as sabotage (NSP1), decomposers (NSP3), proteins that have the ability to camouflage (NSP10) and several other functions of proteins as the cause of Covid-19 such as: triggering inflammation (ORF3a), blocking signals to the immune system (ORF6), triggering suicide self on infected cells (ORF7a) and many other different proteins that replicate and attack cells. One of the proteins that has become the object of attention because of its potential as a factor influencing the spread of the SARS-CoV-2 virus is the ORF6 protein in SARS-CoV-2.

Research conducted [9] on the influence of the virulence factor of ORF6 protein. Research [9] demonstrated that the ORF6 protein inhibited the nuclear localization of Signal Transducer and Activator of Transcription 1 (STAT1) when there was Interferon (IFN) stimulation. Type I IFN in humans [10] are a subgroup of interferon proteins that help regulate the activity of the immune system. STAT1 is a factor that influences the process of RNA synthesis by using one strand of the DNA molecule as a template (transcription) of the STAT protein family. [11]. These proteins respond to stimulation of IFN cell signaling (cytokines). Research result [9] exhibited a significant drop in response to IFN-y stimulation, indicating a decrease in STAT1 of the cell's biggest organ, the nucleus. The C-terminal of ORF6 aids in accelerating viral replication. [12].This is proven from the results of trials with viruses that have mutations in the C-terminal section. The test results show that viruses that do not have mutations in the C-terminal region have an increased ratio of the nucleus to all cells due to changes in the amino acid structure of the C-terminal region. Changes in the amino acid structure at the C-terminal and ORF-6 will form a bond directly with STAT1 through the C-terminal section. The C-terminal itself is the last Amino Acids in a Peptide. Amino Acids are compounds that combine to make proteins [13]. Research [14] utilized in dataset that was obtained from Hospices Civils de Lyon (HCL) hospital, as well as multiple hospitals located in the Lyon region and other regions. The present dataset reveals the identification of mutation variants of ORF6, characterized by the presence of deletions within the mutation sequences.. Deletion is [15] DNA's nitrogen sequence diminishing. For instance, if the DNA chain starts off as CCA-TTA-GCG and the cytosine at the start of the chain is gone, the DNA chain will then alter to CAT-AAG-CG. [16]. Research [14] Demonstrated evidence indicates that the mutation of ORF6, which transpired outside of the C-terminal region, did not have an impact on the rate of transmission of the SARS-CoV-2 virus.

This study will conduct an analysis of the ORF6 SARS-CoV-2 dataset, which has mutations or not in the C-terminal section using the Slicing Index technique. The present study employed the Slicing Index methodology to extract data from a dataset of ORF6 protein. This was achieved by accessing multiple elements in the code list of ORF6 protein. Proteins that do not have their own mutations have a much higher replication capacity than those that have mutations. The present study employs Biopython to investigate the ORF6 protein's

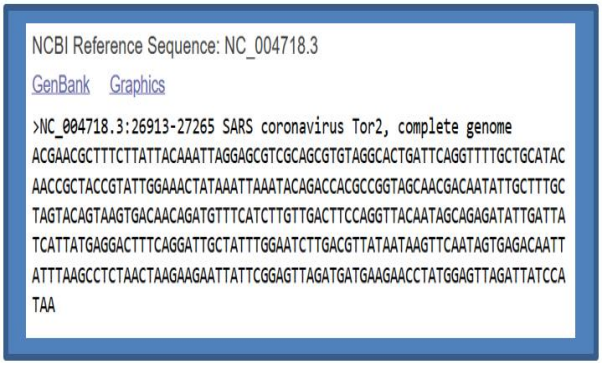
interaction with cells and its impact on the immune system by means of visualizing the C-terminal region of the ORF6 protein.

2. RESEARCH METHOD

The present research employs the ORF6 Sars-CoV-2 protein dataset, which is associated with the Gene ID: 1489673 [9] With fasta format. This study implements the use of the Biopython module to look at mutations in the C-terminal domain that occur in the ORF6 SARS-Cov-2 protein dataset. Cells infected with the SARS-COV-2 virus have their immune response suppressed by ORF6. It is also the most hazardous protein in SARS-COV-2, and the pathogenicity of the virus is linked to its functions. Coronavirus 2 that causes severe acute respiratory syndrome (SARS-COV-2) ORF6 blocks STAT1 nuclear localisation, acting as an antagonist of IFN mediated antiviral signalling.

2.1 ORF6 Protein Dataset

The present research employed the dataset of ORF6 protein in the FASTA format. FASTA files are a category of sequence file formats that contain protein sequences. These formats can be accessed at GenBank, which is managed by the National Center of Biotechnology Information (NCBI). The NCBI distributes sequences into four distinct FASTA file extensions. [17], fna is used for DNA sequences, .faa for protein coding sequence (CDS), .ffn for untranslated sequences for each CDS, and .fra for related RNA feature sequences. Sequencing produces a symbolic linear depiction of DNA which is the sequence of nitrogenous bases (Adenine, Guanine, Cytosine and Thymine) [18]. The dataset example could be seen on Fig. 1.


NCBI Reference Sequence: NC_004718.3

GenBank Graphics

>NC_004718.3:26913-27265 SARS coronavirus Tor2, complete genome

ACGAACGCTTTCTTATTACAAATTAGGAGCGTCGCAGCGTGTAGGCACTGATTCAGGTTTTGCTGCATAC
AACCGCTACCGTATTGGAAACTATAAATTAAATACAGACCACGCCGGTAGCAACGACAATATTGCTTTGC
TAGTACAGTAAGTGACAACAGATGTTTCATCTTGTTGACTTCCAGGTTACAATAGCAGAGATATTGATTA
TCATTATGAGGACTTTCAGGATTGCTATTTGGAATCTTGACGTTATAATAAGTTCAATAGTGAGACAATT
ATTTAAGCCTCTAACTAAGAAGAATTATTCGGAGTTAGATGATGAAGAACCTATGGAGTTAGATTATCCA
TAA

Fig 1: Example of Protein Sequence ORF6 [9]

The DNA base molecule consists of two groups [19] namely bases: (1) purines consisting of adenine (A) and guanine (G) and (2) pyrimidines consisting of: thymine (T) and cytosine (C). The DNA molecule is the genetic blueprint for each cell and the determining factor any particular characteristic of living organisms. A gene is given a symbol according to the order of its nucleotiode base pairs: A-T, G-C [20]. An example of the form of the sequence dataset in the research used can be seen in Fig. 2, consisting of 353 bp (base pair).

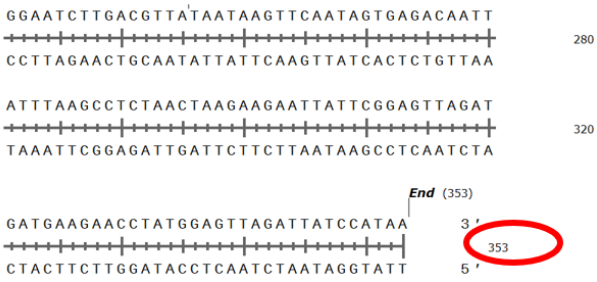


Fig 2: Protein ORF6353bp Example [9]

Position of the ORF6 Protein used in the Genomic Sequence dataset: NC_004718.3 [21] could be seen on Fig 3.

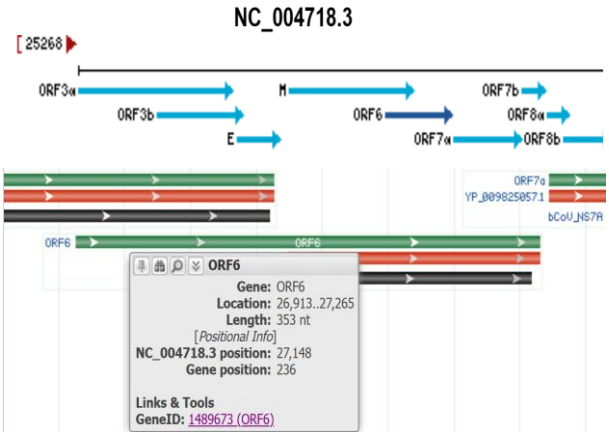


Fig 3:Position of Protein ORF6 DNA Sequence [9]

In the viral cluster genome, there are two external segments of the viral genome. The first segment is the ORF1a protein and the second segment is the C-terminal protein S, ORF3, E, M, ORF6, ORF7a, ORF8, N, and ORF10. Mutations occur in the SARS-CoV-2 virus [22] This is done by looking at the position in the genetic code of the virus [1].

2.2 ORF6 Protein in C-Terminal Domain Analysis

The spread of mutations in the ORF6 protein is carried out in the dataset by implementing the use of the Biopython module in the following stages:

1. Perform data cleaning to remove partial sequence data which is not possible for analysis because the protein sequence is incomplete and is only a fragment. The results of the sequence are saved into a new file using the pseudocode snippet as follows:

```
path='./dataset/ORF6/'
save_path='Unique Protein'
for cnt,I im enumerate(0s.listdir(path)):
  filename=i
  if filename!='Unique_Protein':
    record=SeqIO.read(path+filename,'fasta')
    if 'SARS-Cov-2' in record.description and 'partial' not in record.descript:
      if record.seq not in unique_value:
        unique_value.append(record.seq)
        with open(f'{path+save_path+filename}','w')as output_handle:
          SeqIO.write(record, output_handle, 'fasta')
```

2. Checking the C-terminal from the data that has been cleaned. This is done to determine whether there is a mutation in the protein group using the Slicing Index technique with the following pseudocode: :

```
path='./dataset/ORF6/Unique_Protein/'
unique_c_terminal=[]
mutated_c_terminal=0
normal_c_terminal=0
normal='DEEQPMEID'
diff=[]
for cnt,i in enumerate(os.listdir(path)):
  filename=i
  record=SeqIO.read(path+filename,'fasta')
  if record.seq[52:61].find(normal)==-1:
    mutated_c_terminal+=1
    diff.append(record.seq[52.61])
  else:
    normal_c_terminal+=1
  if(cnt=1):
    print(record.seq)
print(mutated_c_terminal,normal_c_terminal)
```

The Slicing Index technique is performed by accessing several elements in the list which contains the code for the C-terminal Domain. The protein code list contains an initial slicing index of 52 and a final limit of 61, then the index is separated by a "." and enclosed by "[" and "]" signs. Proteins that interact with the C-terminal (membrane protein) are less than Sec61 Amino Acids [23]. Research [23] demonstrated that Sec61 channels may participate in SARS-COV-2 replication and/or innate immunity responses [24].

3. Perform data visualization of the protein group data on string pieces by dividing the data based on the mutation index. The division is done using the Index Slicing Technique using the following pseudocode :

```
for i in range(53,62):
  data[f'position{i}']=difposition[i-53]
Xval=[f'indeks-{1}'for i in range(53,62)]
print(data)
Yval=[i for i in difposition]
```

3. RESULT AND DISCUSSION

Although genetically altered, the ORF6 protein in SARS-CoV-2 will retain its full ability to antagonize the innate immune response. [25]. The innate immune response is the first form of defense of the human body against all types of pathogens that enter the body.

3.1 Analysis Results based on ORF6 Protein Code Slicing Index

Experiments were carried out on the C-terminal Domain to determine the occurrence of mutations in the ORF6 protein group. As can be seen in Fig. 4, there are 2764 protein groups that have mutations at the C-terminal and 2663 that do not have mutations.

Fig 4:Mutation Analysis Visualization Results based on Slicing Index

Mutations occur in the following strings:

```
MFHLVDFQFTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEID
2764 2663
```

In the last few Amino Acids from the C-terminal protein ORF6, one of which is DEEQPMEID found in the SARS-CoV-2 virus, it is indispensable for the function of the ORF6 protein in blocking the activation of Interferon Regulatory Factor 3 (IRF3).[26] and STAT1. IRF3 contains several functional domains, one of which is the C-terminal IRF association domain [27]. IRF3 plays an important role in the response of the innate immune system to viral infections. IRF3 is directly responsible for the activation of IFNβ and IFNα4 production after viral infection, and once it is activated, it plays a critical role in the induction of IFN genes.

3.2 Mutation Visualization Results based on Sec61 Channel Index

In the visualization results of the C-terminal domain protein index mutations, the analysis was carried out at certain sequence indices as can be seen in Table I.

Table 1. Slicing Index Sec52 Until Sec61 C-Terminal Domain

Number	Slicing Indexing Result by Position	
	Position	***Sec***
1	1280	53
2	1176	54
3	1269	55
4	1264	56
5	1079	57
6	964	58
7	809	59
8	711	60
9	1400	61

For example, in line number 6 Slicing indexing in Sec 58 channel with position 964 Table I. The protein of the SARS-CoV-2 virus that plays a role in transmission is the spike (S) section which plays an important role in the ability of SARS-Cov-2 to reach innate immunity. The spike area is often reported to have mutations, in which the attachment process occurs between the Receptor Binding Domain (RBD) and ACE2 as a pathway for the virus to enter the host. [28] and has become a target in research on virus development and treatment of Covid-19[29]. Compared to variations in Spike region amino acid mutations between SARS-Like-CoV viruses, there are no mutations in the C-terminal domain, and there are deletions at positions 455–457, 463–464, and 485–497. Mutations occur mainly due to changes in amino acids so that the protein structure also changes [30]. The results of the visualization of Table I in graphical form can be seen in Fig. 5.

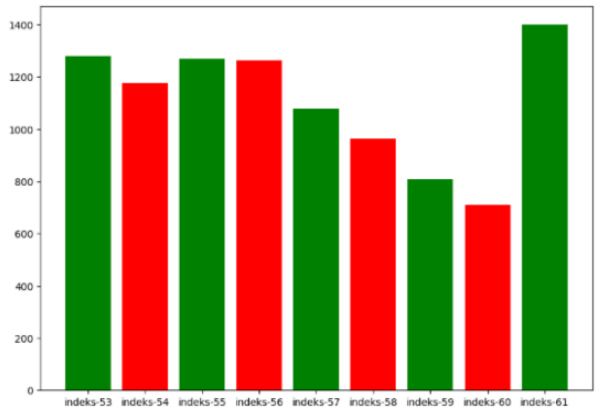


Fig 5. Mutation Analysis Visualization Results based on Slicing Index

4. CONCLUSION

From the results of the analysis of the trial results, it was found that the variations that had mutations in the c-terminus (which had a slower virus spread rate) numbered slightly more than the variations that did not have mutations in the c-terminus (which had a faster virus spread rate) and based on The results of the analysis also found that there were the most mutations at index 61 in the SARS-CoV-2 protein sequence in the C-terminus of the ORF6 protein, followed by index 53 and then index 55.

Further research can be carried out by conducting a more in-depth analysis of other proteins in DNA by implementing the use of Bioinformatics to understand how other proteins interact with cells and affect the immune system, so as to develop strategies for handling the Covid-19 virus and helping prevent its spread.

5. ACKNOWLEDGMENTS

Thank you for the support of the Gunadarma University Research Institute for the smooth implementation of this research.

6. REFERENCES

[1] E. Mohammadi, F. Shafiee, K. Shahzamani, and M. Mehdi, “Biomedicine & Pharmacotherapy Novel and emerging mutations of SARS-CoV-2 : Biomedical implications,” vol. 139, 2021.

[2] [D. N. Aisyah, C. A. Mayadewi, H. Diva, Z. Kozlakidis, Siswanto, and W. Adisasmito, “A spatial-temporal description of the SARSCoV-2 infections in Indonesia during the first six months of outbreak,” *PLoS One*, vol. 15, no. 12 December, pp. 1–14, 2020, doi: 10.1371/journal.pone.0243703.

[3] Y. Yang *et al.*, “SARS-CoV-2: characteristics and current advances in research,” *Virol. J.*, vol. 17, no. 1, pp. 1–17, 2020, doi: 10.1186/s12985-020-01369-z.

[4] [A. Rauf *et al.*, “COVID-19 pandemic: Epidemiology, etiology, conventional and non-conventional therapies,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 21, pp. 1–32, 2020, doi: 10.3390/ijerph17218155.

[5] S. M. Sadat, M. R. Aghadadeghi, M. Yousefi, A. Khodaei, M. Sadat Larijani, and G. Bahramali, “Bioinformatics Analysis of SARS-CoV-2 to Approach an Effective Vaccine Candidate Against COVID-19,” *Mol. Biotechnol.*, vol. 63, no. 5, pp. 389–409, 2021, doi: 10.1007/s12033-021-00303-0.

[6] R. A. Khailany, M. Safdar, and M. Ozaslan, “Genomic characterization of a novel SARS-CoV-2,” *Gene Reports*, vol. 19, no. March, 2020, doi: 10.1016/j.genrep.2020.100682.

[7] E. Volz *et al.*, “Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data,” *medRxiv*, vol. 4, no. 2, pp. 47–49, 2021, doi: 10.1101/2020.12.30.20249034.

[8] K. Hukma, A. Putri, K. P. Sari, and M. Y. Safitri, “Analisis Variasi Genetik Sekuen Gen Surface Glycoprotein (S) Pada SARS – CoV-2 Popset : 1843471817 Menggunakan RFLP Secara In- Sillico,” *Pros. Semin. Nas. Biol.*, vol. 1, no. 1, pp. 44–52, 2021, [Online]. Available: https://semnas.biologi.fmipa.unp.ac.id/index.php/prosiding/article/view/8

[9] Y. Miyamoto *et al.*, “SARS-CoV-2 ORF6 disrupts nucleocytoplasmic trafficking to advance viral replication,” *Commun. Biol.*, vol. 5, no. 1, pp. 1–15, 2022, doi: 10.1038/s42003-022-03427-4.

[10] [M. Sedegah *et al.*, “Cellular interferon-gamma and interleukin-2 responses to SARS-CoV-2 structural proteins are broader and higher in those vaccinated after SARS-CoV-2 infection compared to vaccinees without prior SARS-CoV-2 infection,” *PLoS One*, vol. 17, no. 10 October, pp. 1–23, 2022, doi: 10.1371/journal.pone.0276241.

[11] D. A. Jamison *et al.*, “A comprehensive SARS-CoV-2 and COVID-19 review, Part 1: Intracellular overdrive for SARS-CoV-2 infection,” *Eur. J. Hum. Genet.*, vol. 30, no. 8, pp. 889–898, 2022, doi: 10.1038/s41431-022-01108-8.

[12] R. Hall *et al.*, “SARS-CoV-2 ORF6 disrupts innate immune signalling by inhibiting cellular mRNA export,” *PLoS Pathog.*, vol. 18, no. 8, pp. 1–24, 2022, doi: 10.1371/journal.ppat.1010349.

[13] R. Zhou, R. Zeng, A. von Brunn, and J. Lei, “Structural characterization of the C-terminal domain of SARS-CoV-2 nucleocapsid protein,” *Mol. Biomed.*, vol. 1, no. 1, pp. 1–11, 2020, doi: 10.1186/s43556-020-00001-4.

[14] G. Quéromès *et al.*, “Characterization of SARS-CoV-2 ORF6 deletion variants detected in a nosocomial cluster during routine genomic surveillance, Lyon, France,” *Emerg. Microbes Infect.*, vol. 10, no. 1, pp. 167–177, 2021, doi: 10.1080/22221751.2021.1872351.

[15] P. V. Markov *et al.*, “The evolution of SARS-CoV-2,” *Nat. Rev. Microbiol.*, vol. 21, no. June, 2023, doi: 10.1038/s41579-023-00878-2.

[16] B. Cosar *et al.*, “SARS-CoV-2 Mutations and their Viral Variants,” *Cytokine Growth Factor Rev.*, vol. 63, no. July 2021, pp. 10–22, 2022, doi: 10.1016/j.cytogfr.2021.06.001.

[17] [M. Rashighi and J. E. Harris, *乳鼠心肌提取HHS Public Access*, vol. 176, no. 3. 2017. doi: 10.1053/j.gastro.2016.08.014.CagY.

[18] K. Kryukov, L. Jin, and S. Nakagawa, “Efficient compression of SARS-CoV-2 genome data using Nucleotide Archival Format,” *Patterns*, vol. 3, no. 9, p. 100562, 2022, doi: 10.1016/j.patter.2022.100562.

[19] D. M. Kristensen, Y. I. Wolf, and E. V Koonin, “ATGC database and ATGC-COGs : an updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation,” vol. 45, no.

October 2016, pp. 210–218, 2017, doi: 10.1093/nar/gkw934.

[20] A. Banerjee *et al.*, "Single-Molecule Analysis of DNA Base-Stacking Energetics Using Patterned DNA Nanostructures," 2022.

[21] NCBI, "No Title," *ORF6 Protein Dataset*, 2023. https://www.ncbi.nlm.nih.gov/gene/1489673

[22] R. Sanjua, "Mechanisms of viral mutation ′," pp. 4433–4448, 2016, doi: 10.1007/s00018-016-2299-6.

[23] M. Linxweiler, B. Schick, and R. Zimmermann, "Let ' s talk about Secs : Sec61 , Sec62 and Sec63 in signal transduction , oncology and personalized medicine," no. January, pp. 1–10, 2017, doi: 10.1038/sigtrans.2017.2.

[24] S. Sun, M. Mariappan, Y. W. Campus, and W. Haven, "C-terminal tail length guides insertion and assembly of membrane proteins," vol. 295, no. 21, pp. 15498–15510, 2020, doi: 10.1074/jbc.RA120.012992.

[25] S. E. Turvey, "NIH Public Access," no. June, 2018, doi: 10.1016/j.jaci.2009.07.016.

[26] M. Moustaqil *et al.*, "SARS-CoV-2 proteases cleave IRF3 and critical modulators of inflammatory pathways (NLRP12 and TAB1): implications for disease presentation across species and the search for reservoir hosts .," vol. 3, 2020.

[27] S. Fung, K. Siu, H. Lin, M. L. Yeung, and D. Jin, "SARS-CoV-2 main protease suppresses type I interferon production by preventing nuclear translocation of phosphorylated IRF3," vol. 17, 2021, doi: 10.7150/ijbs.59943.

[28] W. Dejnirattisai *et al.*, "The antigenic anatomy of SARS-CoV-2 receptor binding domain," *Cell*, vol. 184, no. 8, pp. 2183-2200.e22, 2021, doi: 10.1016/j.cell.2021.02.032.

[29] D. Ellis *et al.*, "Stabilization of the SARS-CoV-2 Spike Receptor-Binding Domain Using Deep Mutational Scanning and Structure-Based Design," *Front. Immunol.*, vol. 12, no. June, pp. 1–17, 2021, doi: 10.3389/fimmu.2021.710263.

[30] [L. Miorin, T. Kehrer, M. T. Sanchez-aparicio, K. Zhang, and P. Cohen, "SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling," 2020, doi: 10.1073/pnas.2016650117.