Final Project Report

Analysis of SARS-CoV-2 Spike protein.

Zunqiu Wang

12.19.2020

BI-GY 7663 by Mgavi Brathwaite

Outline

**1.** **Introduction**

**2.** **Graphics**

**3.** **Task checklist**

**4.** **Materials and Methodology**

**5.** **Analysis (screenshots with concise explanation)**

**6.** **Summary**

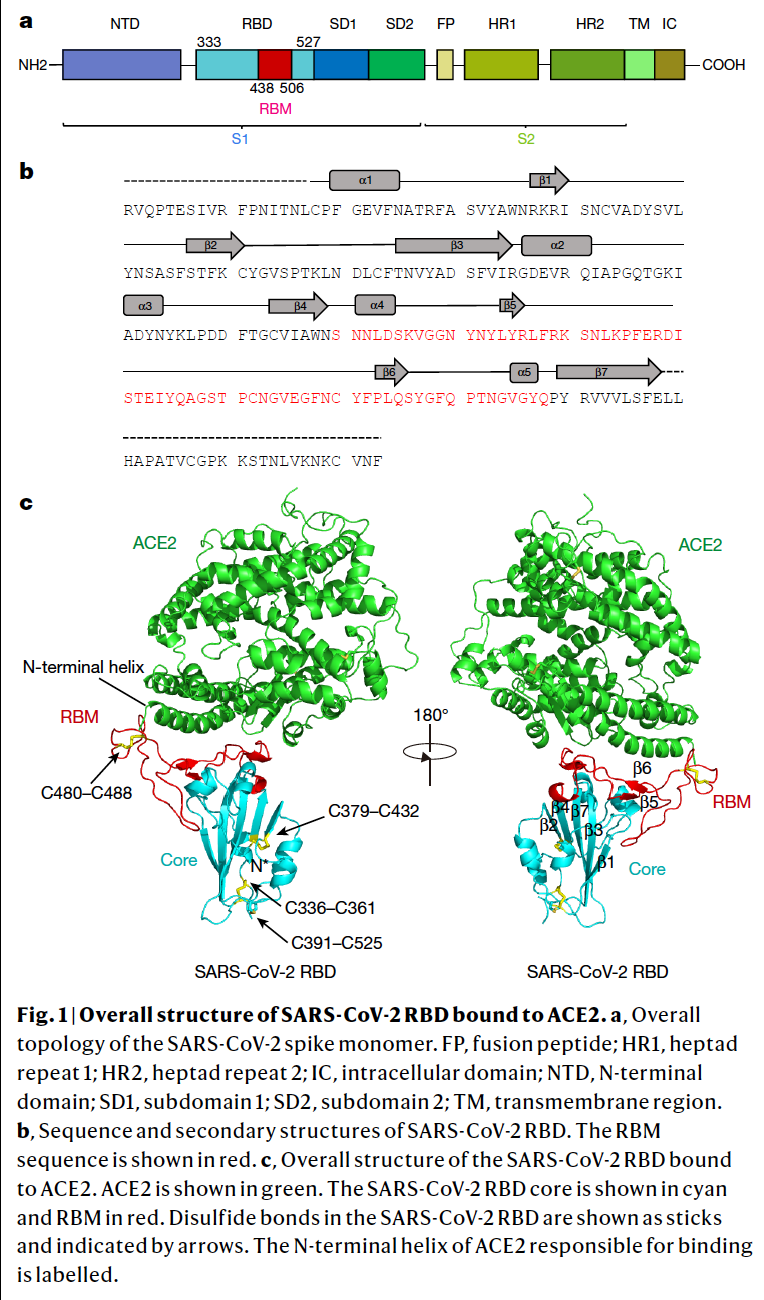
**7.** **Bibliography**

**8.** **Appendix with code, input & output files**

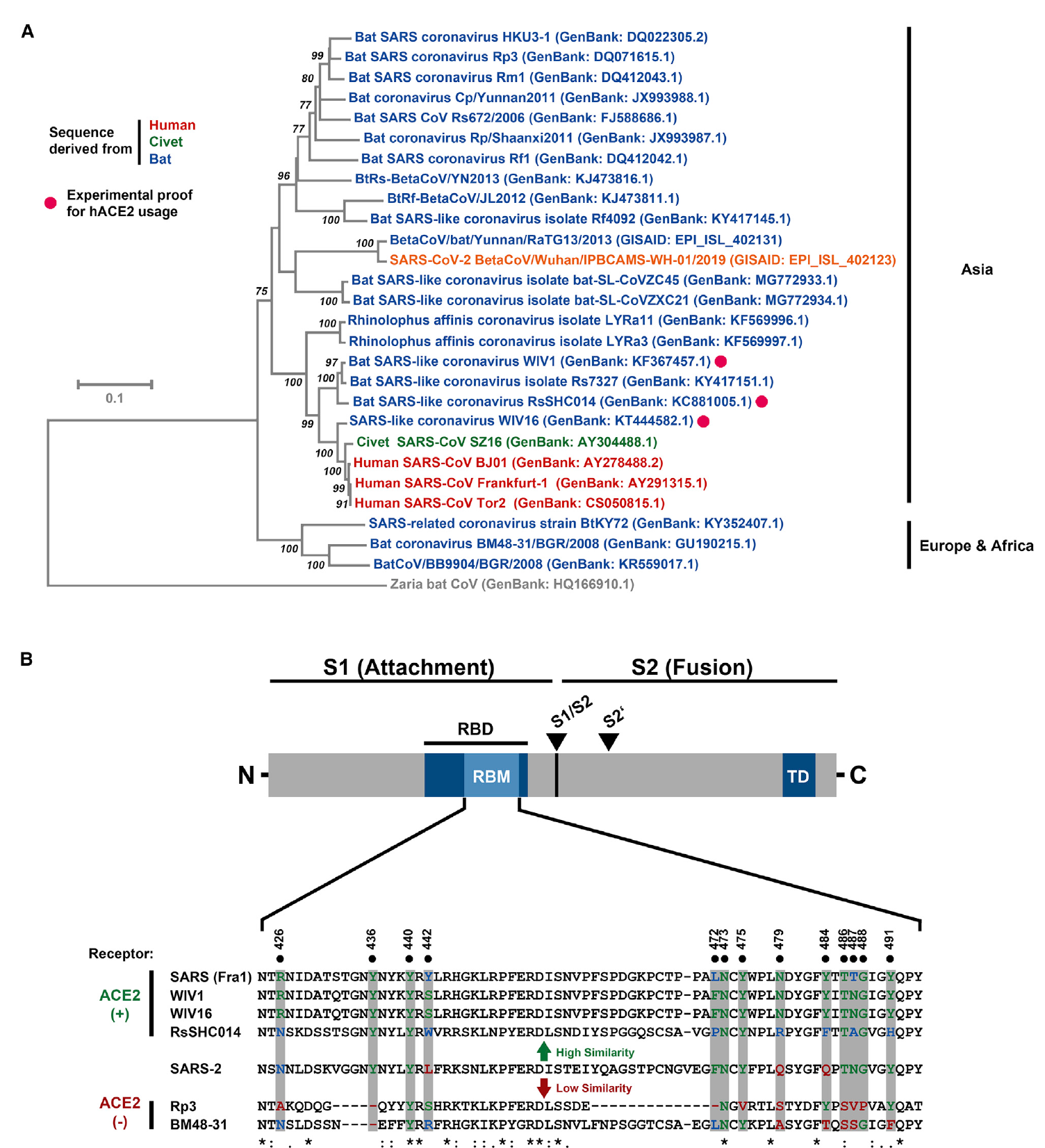
Introduction

The outbreak of epidemic severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has posed a huge threat to global public health and economic stagnance1. With the continuous spread of highly contagious virus, more than 80 countries have been impacted and thousands of patients have died after infection1. Previous research has reported that the spike glycoprotein on virus envelope, which has two functions: binding of host cell receptor and fusion of host cell membrane2. Particularly, virus binding to angiotensin-converting enzyme II (ACE2), which presents in many tissues, including lungs, is shown to be the most likely entry site after binding of receptor binding domain (RBD) of S glycoprotein on virus3. There are several approaches to curb the COVID-19 infection: blocking the entry, neutralizing antibodies, antiviral, and many others4-5. It is a pressing urgency to develop an effective therapeutic treatment for patients with severe complications. However, the mechanism of how spike protein is related to ACE2 levels, viral burden, and infection severity is still not comprehensively understood. With the advancement of technology as well as the need for systemic studies over multigenic disorders, a great number of computational methods have been generated and been used to interpret relevant information7. Over the years since the advent of bioinformatics, it helps scientists keep abreast of rapidly changing databases to establish new approaches to analyze high-throughput datasets8. The primary aim of this project is to develop a programming algorithm with a variety of computational tools to synthesize online database information to interrogate the roles of spike protein and ACE2 during the onset of SARS-CoV-2.

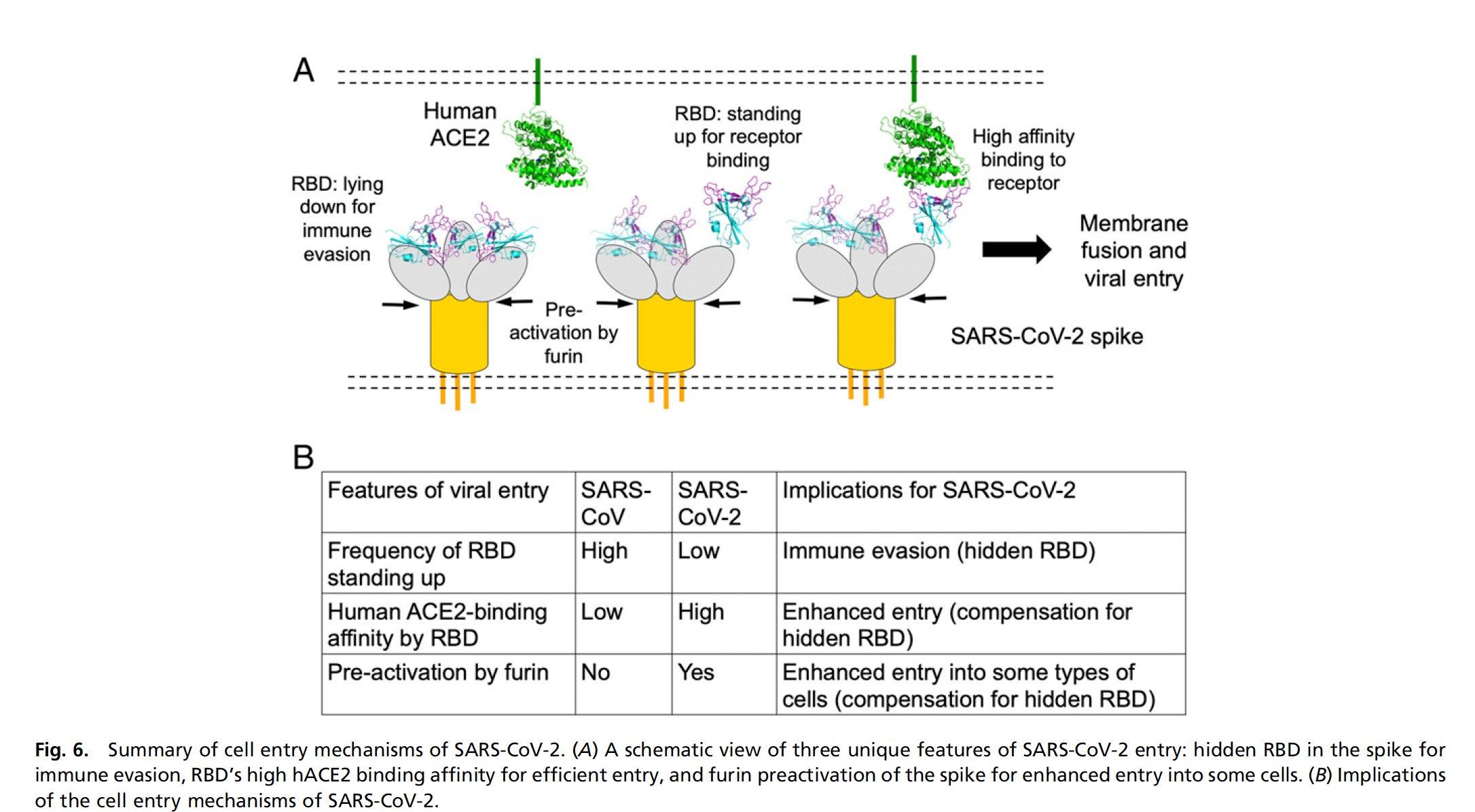
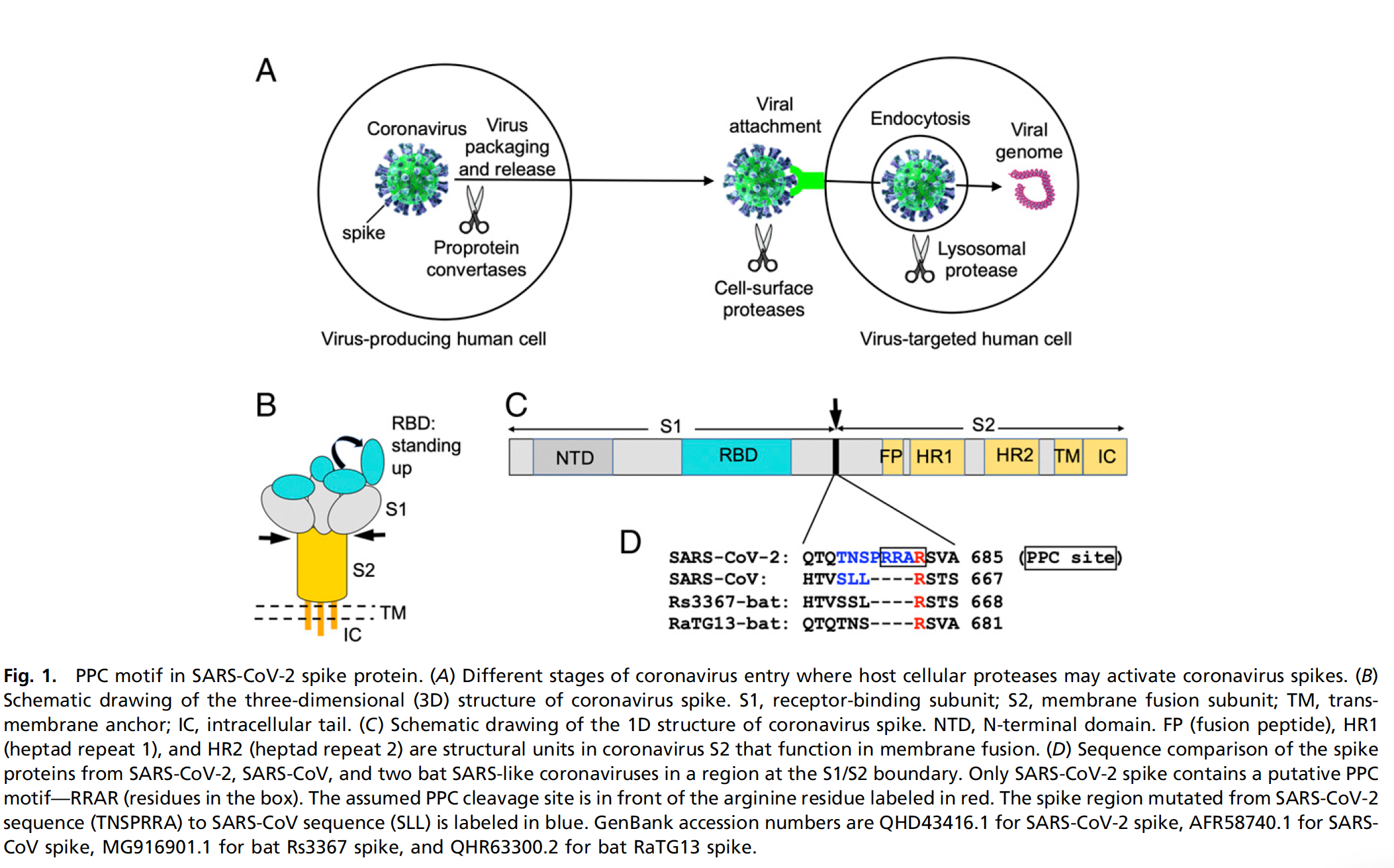
Graphics



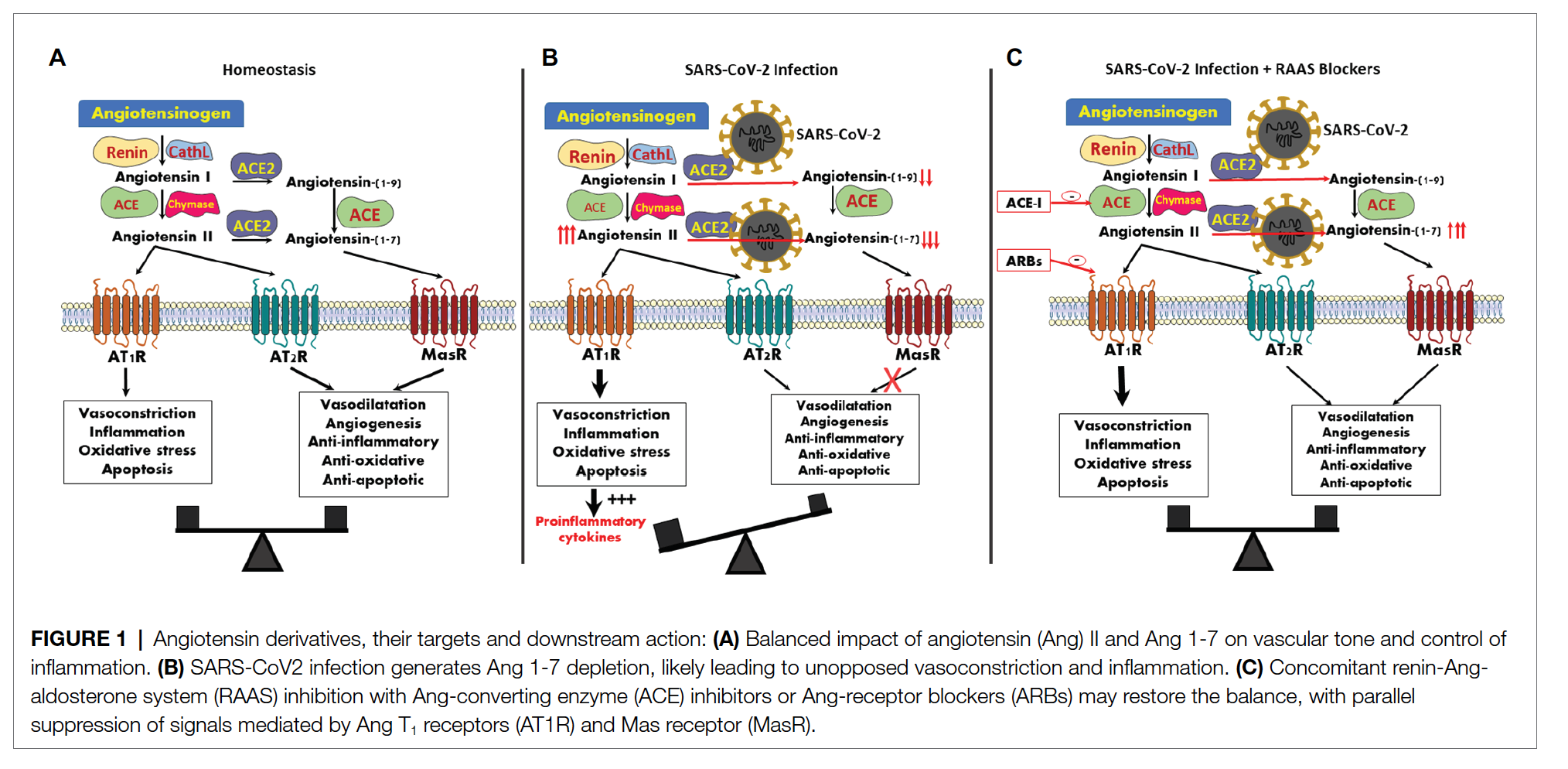
Adapted from (Lan et al., 2020)



Adapted from (Hoffmann et al., 2020)



Adapted from (Shang et al., 2020)



Adapted from (Abassi et al., 2020)

Task List

**Accessing Sequence Data and Performing an MSA**

#1 *Create a Biopython session and save a list of Spike protein fasta sequences. Run it to collect files containing fasta sequences. I obtained 1000 sequences with the 'retmax' parameter of the esearch method. Minimum retmax is 200 fasta sequences.*

*#2 Download Clustal Omega and use Python's OS module to run command line arguments to execute MSA on each fasta file. I aligned 1000 sequences in eight minutes. My machine has 32 gb of ram. If you have 16gb then it should align in under 20 minutes.*

*#3 Access the 'pubmed' database and return 10 articles about the spike protein. Reading the articles is*

*# up to you.*

*#4 Determine which protein family the Spike protein is a member of.*

*#5 Look for any variation between the fasta protein sequences.*

*#6* ***BONUS****: Determine gene ontology data via GO*

*#7 Determine the part of the spike that binds to the ACE2 receptor. Get a list of 1000 fasta sequences and run an msa.*

*#8 Access KEGG to get genome information on* 'SARS-CoV-2'. Use ID to get the COVID-19 description(COVID-19 [DS:H02398]). From here access all network data for Spike protein, 'S'.

#9 Use hsa(homo sapiens) # and kgml to get the pathway image file.

#10 Additional information such as pathways that 'SARS-CoV-2/COVID-19' participate in.

#11 What's the relationship between the ACE2 of Renin-angiotensin system and 'SARS-CoV-2/COVID-19'?

Methods and Materials

**#1 #2 #5**

I created a Biopython.Entrez session and searched and fetched 300 FASTA sequences data file based on the term “SARS-CoV-2 S protein”. Then, I ran ClustalOmega and created a command line, which was recognized by the OS system to locally perform Multiple Sequencing Alignment.

**#3**

I used Biopython.Entrez to retrieve pubmed on “SARS-CoV-2 Spike protein” and returned 10 articles.

**#7**

I firstly ran Biopython.Entrez and fetched 200 FASTA sequences data file based on

the term “SARS-CoV-2 S protein”. Next, I used SeqIO to find any descriptions with

“ACE2” and returned a file with corresponding sequences. I ran Biopython.Entrez

again with a narrowed term “spike receptor-binding domain ACE2” after determining

the binding site and ran ClustalOmega and created a command line, which was

recognized by the OS system to locally perform Multiple Sequencing Alignment.

**#4 #6**

To begin with, I used Uniprot search from bioservices based on term “SARS-CoV-2

Spike protein” and retrieved a table of all protein family associated with the disease.

After reading the returned file, I created a dictionary with the key as entry and value as

protein name and write to file only if “Spike protein” was in the value of the dictionary.

Next, I returned a UniprotKB ID from the entry whose value is located in the

dictionary. Using UniprotKB ID I accessed QuikGO to get annotations of that ID and

write to file. With the help of re, I located GO IDs and returned a set of GO IDs without

duplication. Last step is to access QuickGO again to get annotations of those GO IDs

and write to file.

**#8**

In KEGG session, I used term “SARS-CoV-2” in genome db to locate “gn” ID. Then,

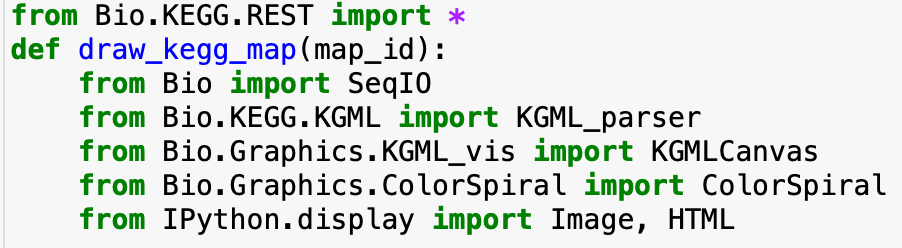
using this ID to locate disease ID with re module by getting “gn” ID annotation. From

disease ID annotation, I returned “hsa” ID using re based on pattern and I also wrote

disease ID annotation to file.

**#9**

Continuing to pass above functions and input of “hsa” ID, I created a function to

output a pathway image in PDF format. I imported the following packages. 

**#10**

Using #8 functions and a new one created with input of “hsa” ID and got its annotation

file with the help of Bio.KEGG REST. By reading the lines and searching for “GENE”,

“NETWORK”, and “ELEMENT” sections, write its corresponding content(gene

names, pathway names and their IDs, network IDs) to files. In addition, I used re

module to locate network IDs beginning with “N” and parse these ID annotations into

file. As a supplementary purpose, I used term “Renin-angiotensin system” for KEGG

and look for pathway based on term. Then, with re module I created a list based on

pattern and extract first column, entry ID, which were converted to “hsa” ID format.

Next, I got annotation file from ID and passed to KGML to get ACE2 pathway image.

Analysis

* *Create a Biopython session and save a list of Spike protein fasta sequences*
* *Download Clustal Omega and use Python's OS module to run command line arguments to execute MSA on each fasta file*
* *Look for any variation between the fasta protein sequences*

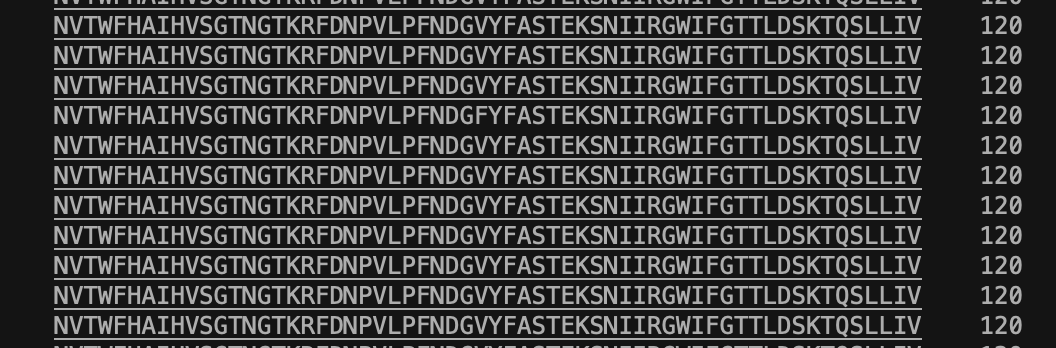
I tried several terms as well as restraining a limited number of returned ids to retrieve a FASTA file for a better looking output of a MSA file. In order to illustrate the representative look of output, I attached a screenshot of the MSA file. A good quality of alignment should be based on several criteria, including but not limited to low number of gaps and indicated annotations for quality assessment. The annotations consist of “\*” (perfect alignment), “:” (strong similarity), and “.” ( weak similarity)12. From MSA, upon inspection of variations(SNPs) among sequences, there are AA substitutions that are silenced mutation and probably will not alter a specific function. The characteristics of AA properties, such as polarity, charge, pKa value, will impact AA function. They may still change protein structurally and functionally but we need a more comprehensive modelling to predict how SNPs could lead to various outcomes of diagnostic, post infection, and effectiveness of a specific treatment. Aligned score shown on the far right. The “X” means undetermined AA; Gaps are not aligned.



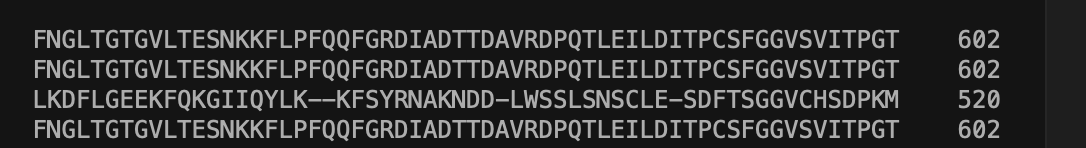
This portion of the sequence shares relatively high similarity, probably is the consensus region.

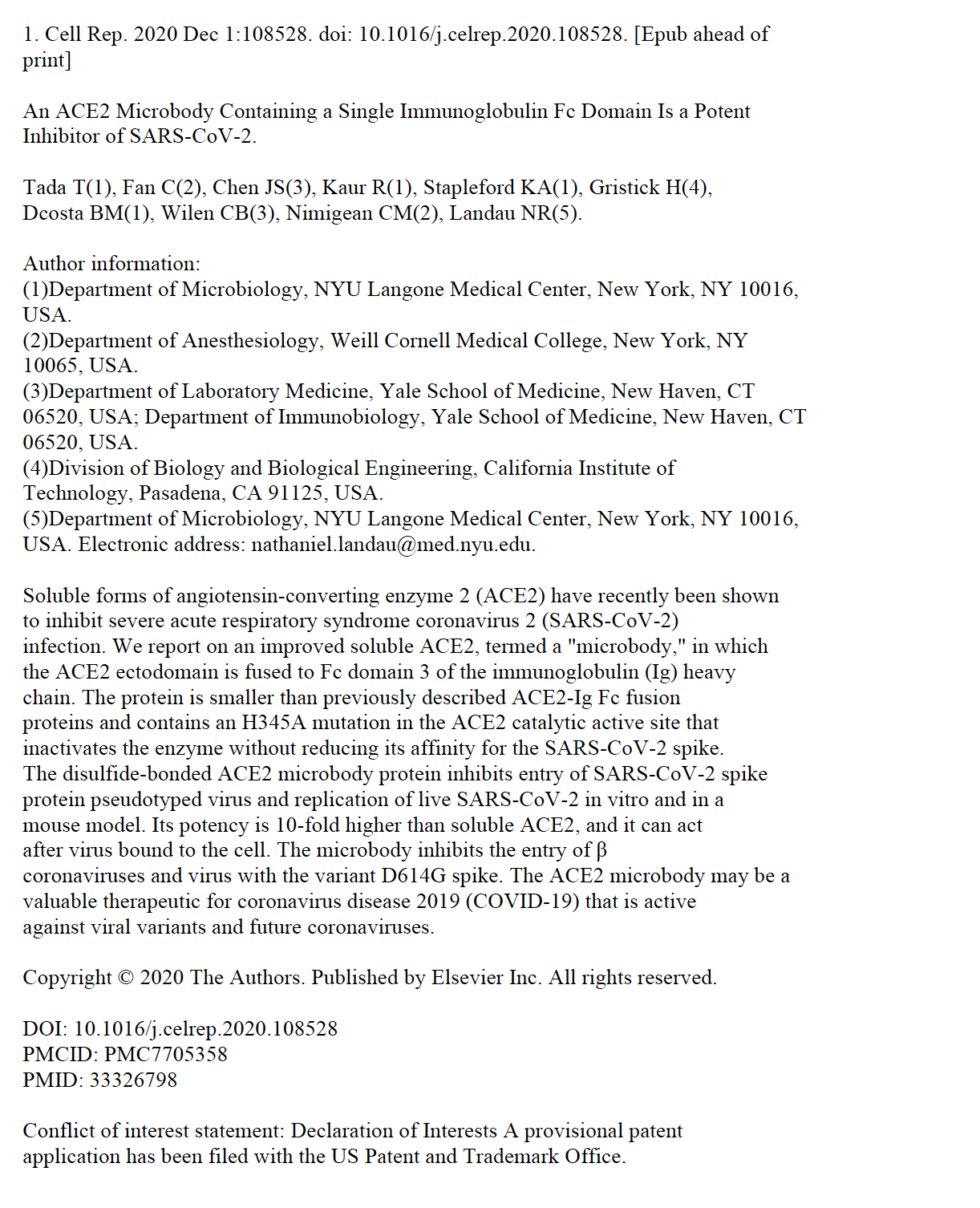


Here, V(non-ploar, aliphatic) is substituted by F(aromatic) and annotated by “.”, which shares least similarity.

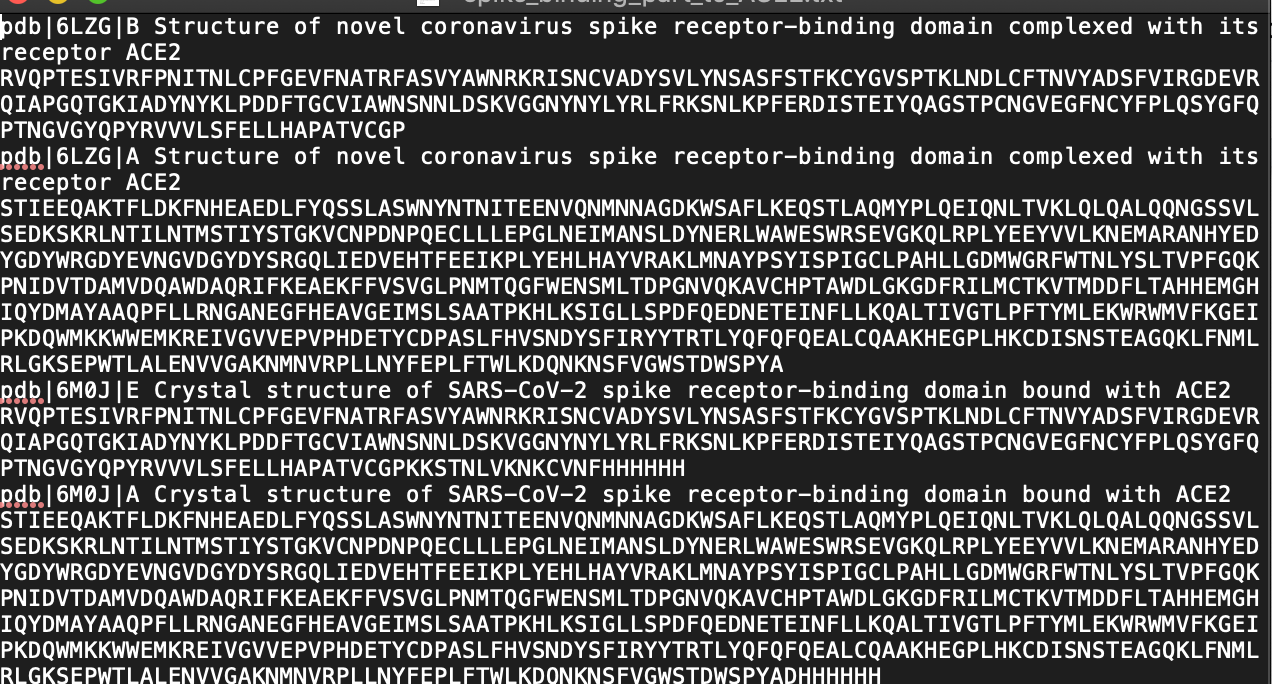
Here, Q is substituted by K and annotated by “:”, as they are both charged even though K is charged.

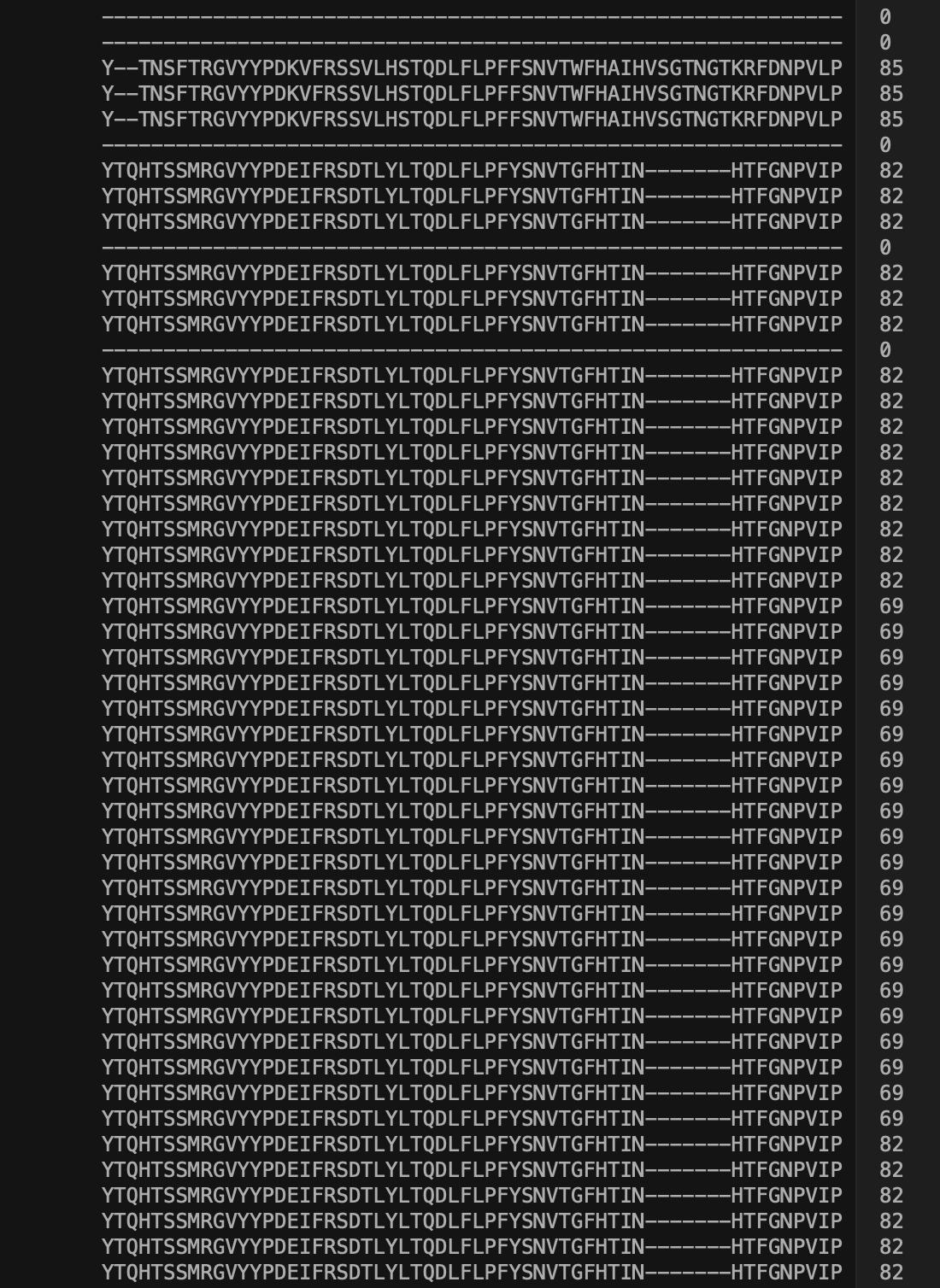
Here, D is substituted by E and annotated by “:”, as they are from the same amino acid group while they convert back and forth depending on environmental pH.



* *Access the 'pubmed' database and return 10 articles about the spike protein*
* *Determine the part of the spike that binds to the ACE2 receptor. Get a list of 200 fasta sequences and run an msa*

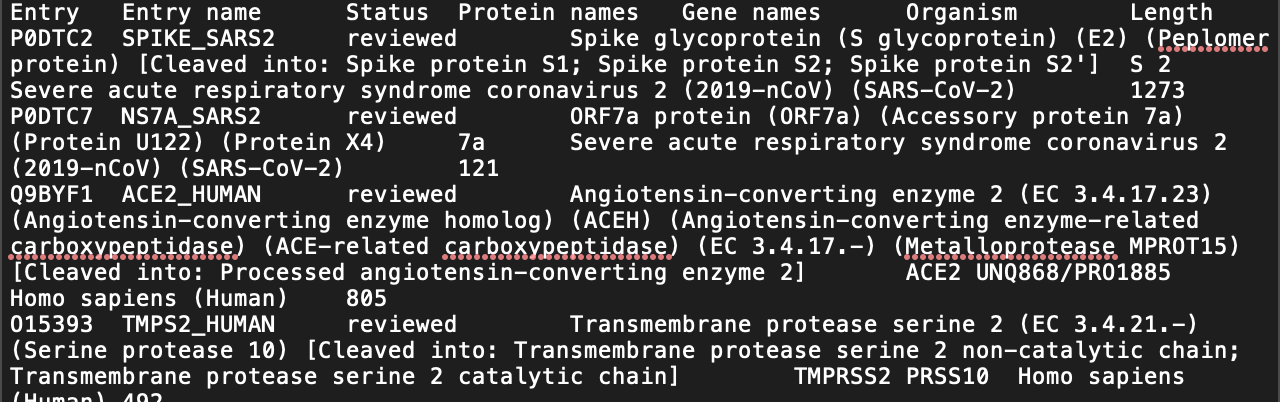
After getting the FASTA file from the first run, I determined the binding site is the spike receptor-binding domain with ACE2 so that I extracted only those sequences to demonstrate my understanding. Then, I used a different term for sequential FASTA file generation and MSA. With 200 sequences, MSA file does not look great as there are no annotations for quality of alignment.

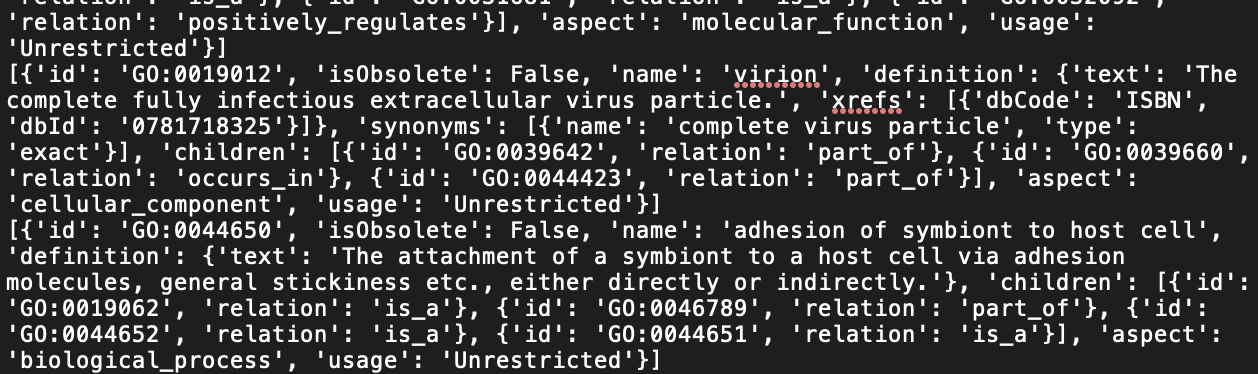
**

**

* *Determine which protein family the Spike protein is a member of*
* ***BONUS****: Determine gene ontology data via GO*

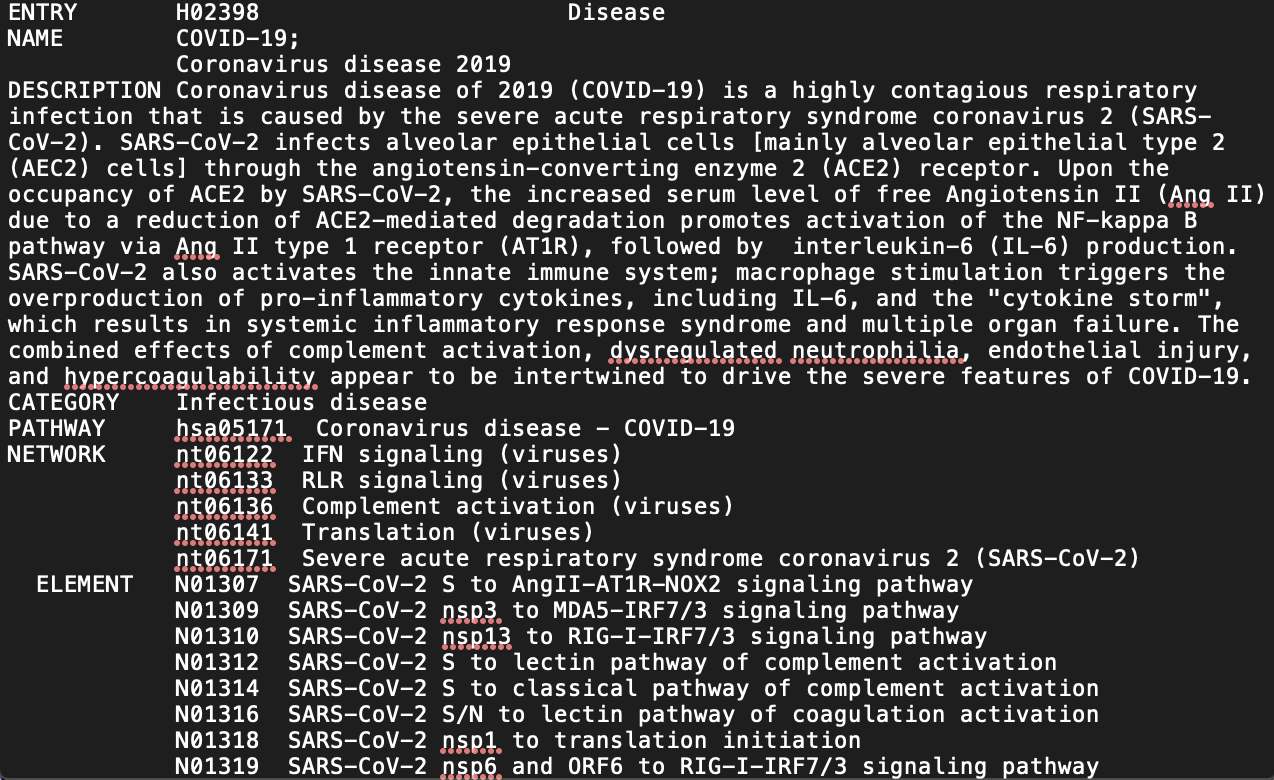
Upon examination, I determined that ‘PODTC2’ entry with the protein name “Spike glycoprotein(S glycoprotein) was the protein family spike protein belongs to. With that “PODTC2” as UniprotKB ID, I retrieved tons of annotations.

Using re, I could extract GO IDs based on the pattern provided in code. After getting GO data from QuickGO, it returned information of several aspects on how spike protein is involved, including but not limited to different pathway names and its relation, cellular components, biological processes, molecular function.



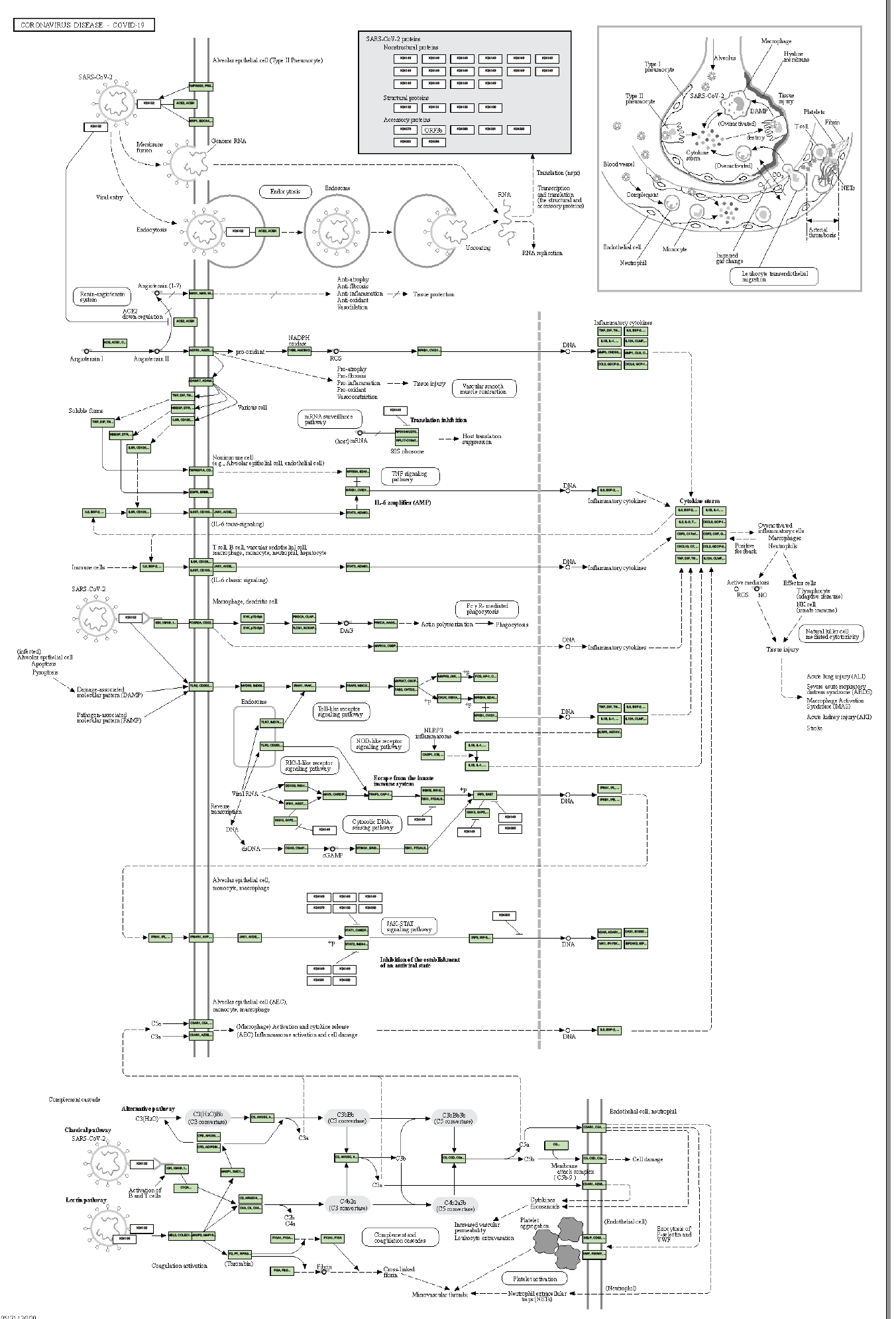
* *Access KEGG to get genome information on 'SARS-CoV-2'. Use ID to get the COVID-19 description(COVID-19 [DS:H02398]). From here access all network data for Spike protein, 'S'*

By using KEGG to get annotations and re module repeatedly, I got disease ID and its annotation.

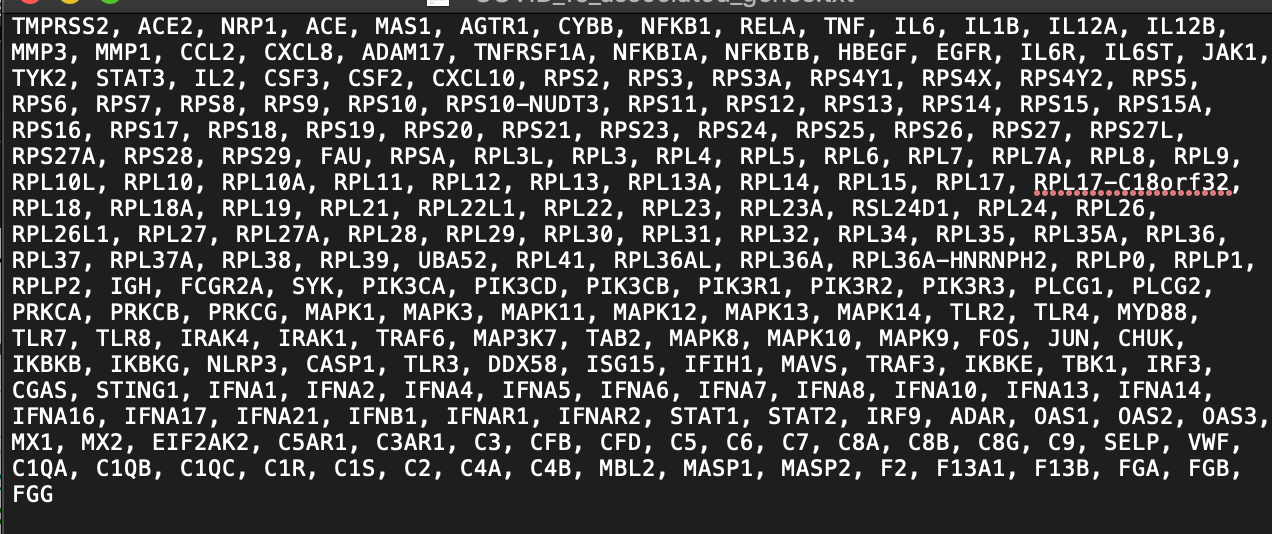


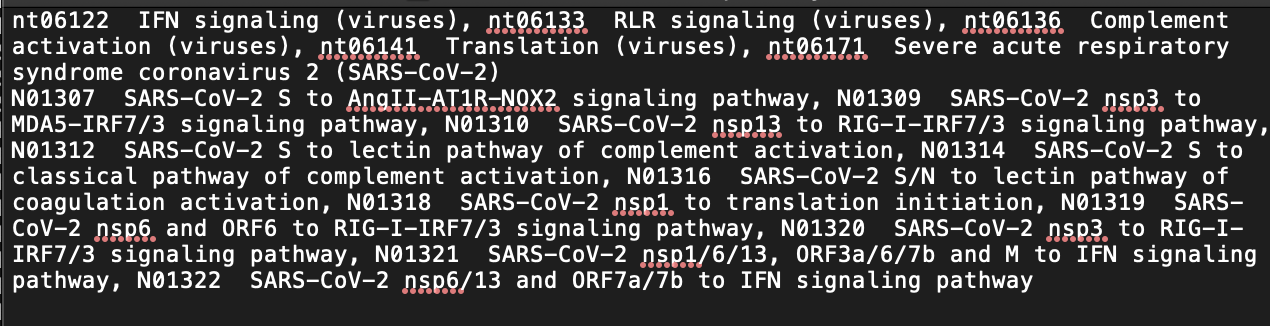
* *Use hsa(homo sapiens) # and kgml to get the pathway image file*

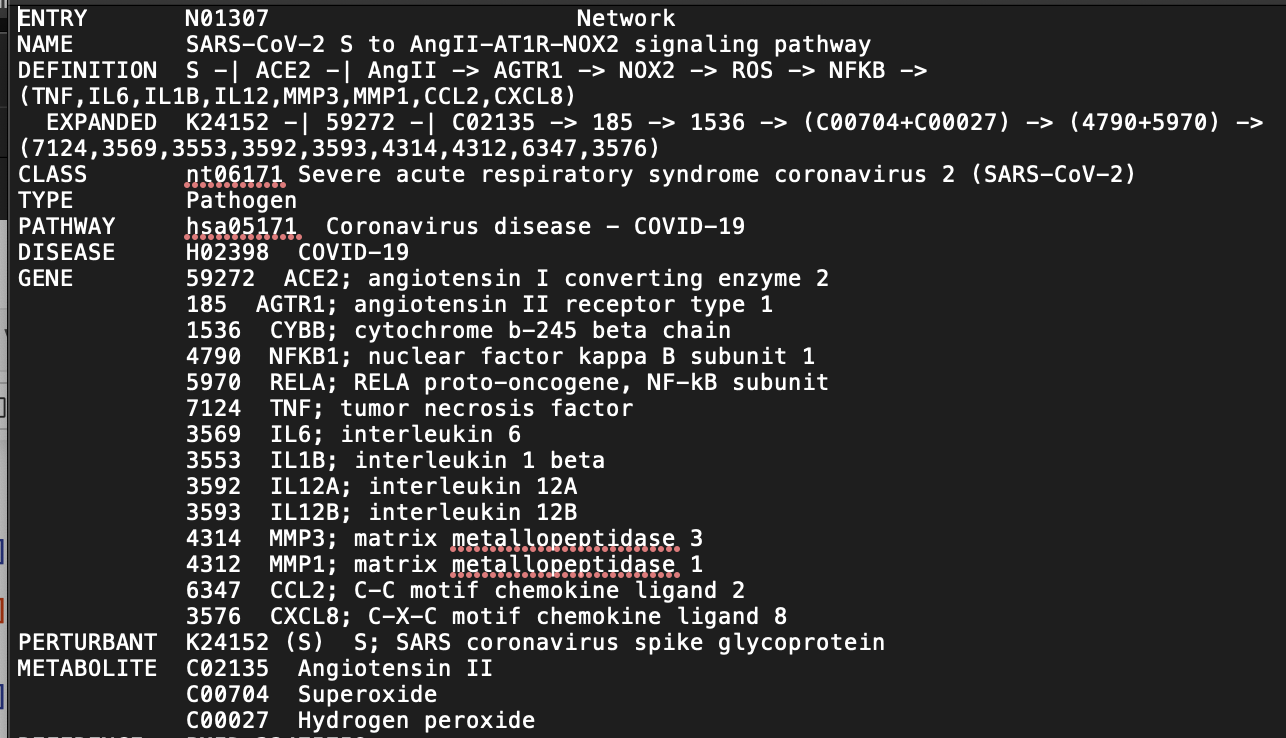
It seems a very complex network after infection of SARS-CoV-2 and there are a great number of pathways and associated receptors, cytokines, modulators that could be potential treatment targets.



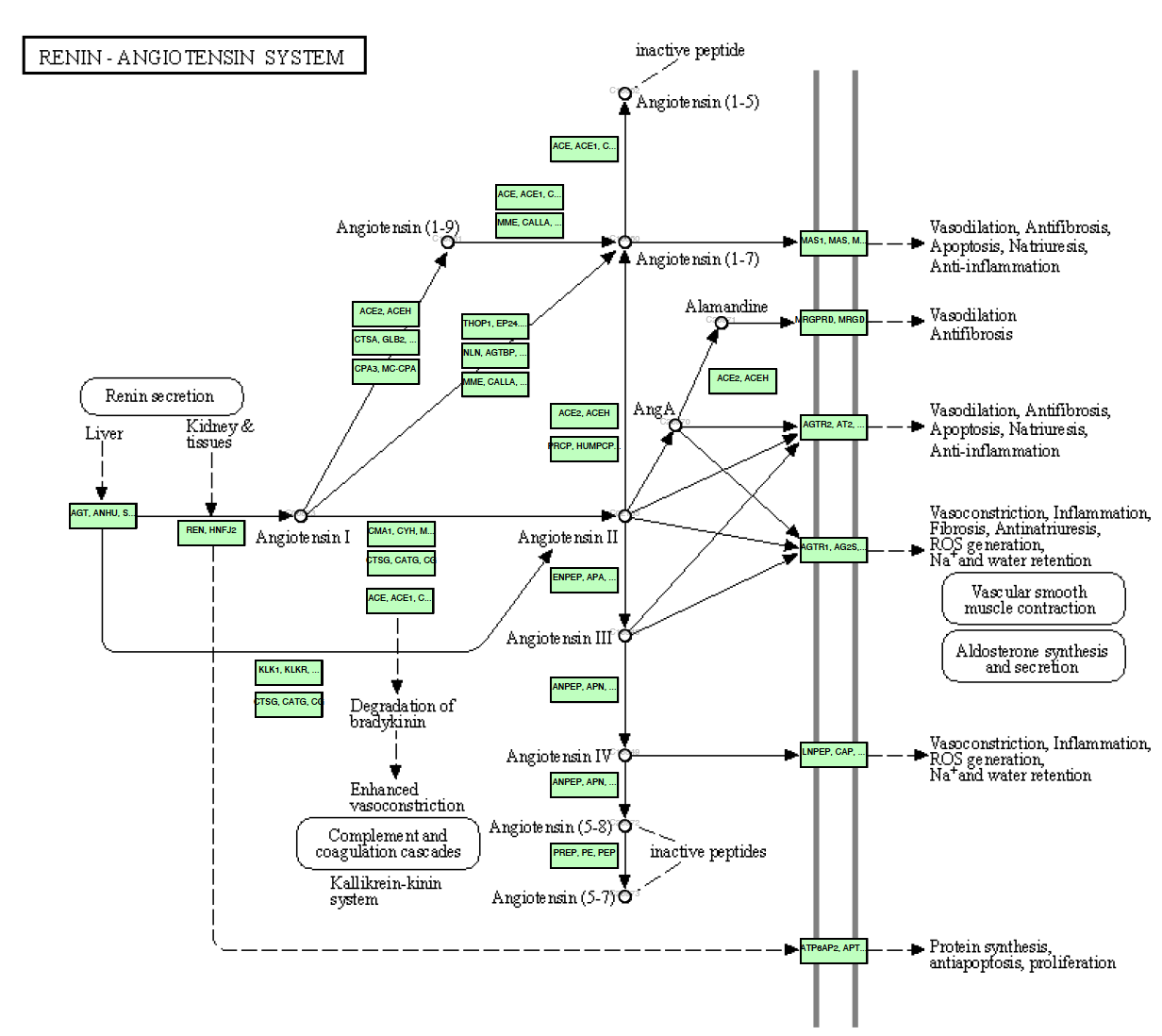
* *Additional information such as pathways that 'SARS-CoV-2/COVID-19' participate in*

Here are genes and pathways that are involved in COVID-19 infection. Each one plays a critical role in modulating disease severity. Many studies are focused on virus entry sites as well as receptors that are involved in ACE2 in the Renin-angiotensin system.

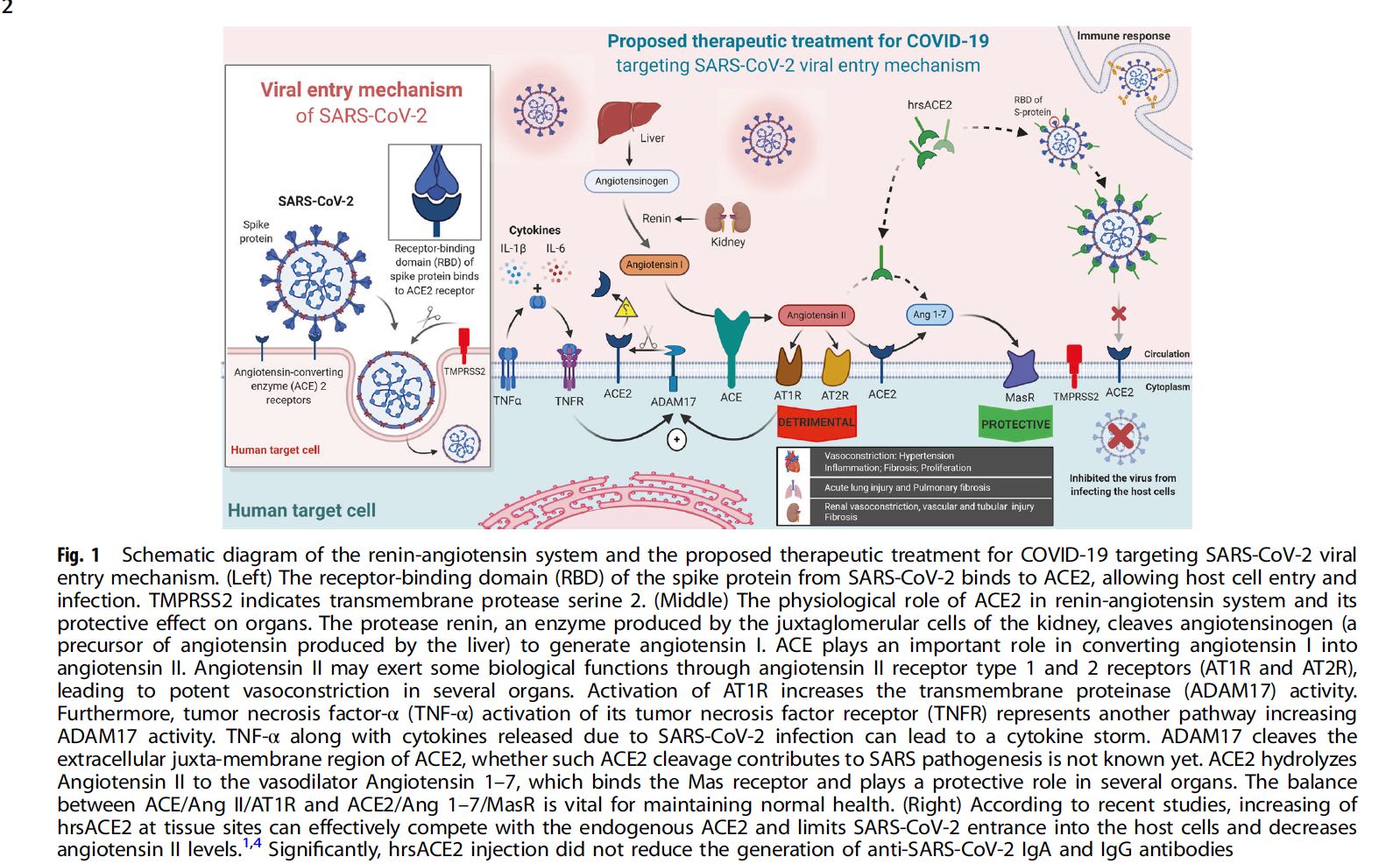


Here is a detailed look at the network of pathway annotations. Based on this, I get more information on what specifically each gene does that could affect infection outcomes.

The overview of ACE2 in the Renin-angiotensin system reveals a scheme at a cellular level how a single dysregulation of one pathway by SARS-Cov-2 could alter physiological homeostasis.



Summary

* *What's the relationship between the ACE2 of Renin-angiotensin system and 'SARS-CoV-2/COVID-19'?*

Adapted from (Abd El-Aziz et al., 2020)

In a nutshell, upon SARS-Cov-2 virus entry by binding of spike protein on virus envelope and ACE2 receptor of human lung cells, virus releases its viral genome and in the meantime upregulated free ACE2 causes immunological response by releasing inflammatory cytokines, which causes the “cytokine storm”13. Downstream of ACE2 induced pathways through AGTR1 drives vasoconstriction, pro-inflammation, and ultimately cellular damage in lung tissues14. Since there are variations of protein sequences of spike protein upon my analysis, it is challenging to elucidate the exact mechanism of infection from RBD binding of S protein to ACE2 to what damage it causes ranging from cellular level to organ level within a host. It is believed that bioinformatic tools will help validate structural and functional analyzes of current work and broaden the paths to pursue cutting-edge discoveries with newly developed programming algorithms.

Bibliography

1. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. <https://doi.org/10.1038/s41586-020-2180-5>
2. Ali, A. (2020). Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Scientific Reports*, 12.
3. Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., Gong, Y., Xiao, H., Fan, Z., Tan, S., Wu, G., Tan, W., Lu, X., Fan, C., Wang, Q., Liu, Y., … Liu, L. (2020). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science (New York, N.Y.)*, *368*(6496), 1274–1278. https://doi.org/10.1126/science.abc2241
4. Bioinformatics. (2020, December 10). Retrieved December 16, 2020, from <https://en.wikipedia.org/wiki/Bioinformatics>
5. Center for Drug Evaluation and Research. (2020, November 30). Coronavirus Treatment Acceleration Program (CTAP). Retrieved December 17, 2020, from https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap
6. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*(2), 271-280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>
7. Hogeweg P. (2011). The roots of bioinformatics in theoretical biology. *PLoS computational biology*, *7*(3), e1002021. https://doi.org/10.1371/journal.pcbi.1002021
8. Teufel, A., Krupp, M., Weinmann, A., & Galle, P.R. (2006). Current bioinformatics tools in genomic biomedical research (Review). International Journal of Molecular Medicine, 17, 967-973. <https://doi.org/10.3892/ijmm.17.6.967>
9. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, *117*(21), 11727–11734. <https://doi.org/10.1073/pnas.2003138117>
10. Abd El-Aziz, T. M., Al-Sabi, A., & Stockand, J. D. (2020). Human recombinant soluble ACE2 (hrsACE2) shows promise for treating severe COVID­19. *Signal Transduction and Targeted Therapy*, *5*(1), 258. <https://doi.org/10.1038/s41392-020-00374-6>
11. Abassi, Z., Higazi, A. A. R., Kinaneh, S., Armaly, Z., Skorecki, K., & Heyman, S. N. (2020). ACE2, COVID-19 Infection, Inflammation, and Coagulopathy: Missing Pieces in the Puzzle. *Frontiers in Physiology*, *11*, 574753. <https://doi.org/10.3389/fphys.2020.574753>
12. What are the meanings of the three symbols "\*", ".", and ":" in ClustalW? (2015, June 8). Retrieved December 18, 2020, from https://www.ddbj.nig.ac.jp/faq/en/explain-three-symbols-e.html
13. Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., & Manson, J. J. (2020). COVID-19: Consider cytokine storm syndromes and immunosuppression. *The Lancet*, *395*(10229), 1033–1034. <https://doi.org/10.1016/S0140-6736(20)30628-0>
14. Abassi, Z., Higazi, A. A. R., Kinaneh, S., Armaly, Z., Skorecki, K., & Heyman, S. N. (2020). ACE2, COVID-19 Infection, Inflammation, and Coagulopathy: Missing Pieces in the Puzzle. *Frontiers in Physiology*, *11*, 574753. <https://doi.org/10.3389/fphys.2020.574753>

Appendix