SEQUENCE, STRUCTURE, AND FUNCTION OF TMPRSS2 MISSENSE MUTATIONS RELATING TO THE SEVERITY AND SUSCEPTIBILITY OF COVID-19 INFECTIONS

I. Abstract Draft (from conference entry)

SARS-CoV-2 is a highly infectious virus that is responsible for the COVID-19 global pandemic that swept the world in 2020. Disease outcomes range from asymptomatic to fatal. The virus initiates entry into host cells by the binding of its spike protein to the ACE2 receptor. Entry is finalized by the activation of spike glycoprotein by proteases including transmembrane protease, serine 2 (TMPRSS2) and FURIN which cleave the spike protein of the virus. Single nucleotide polymorphisms (SNPs) in TMPRSS2 may lead to functional changes which could underlie differences in disease severity. TMPRSS2 is also known to activate different respiratory illnesses including coronaviruses and influenza A (Shen et al., 2020). Previous studies have shown that knockout TMPRSS2 mice appeared healthy, experienced a decrease in viral spread within the respiratory system, and had a less severe immune response when infected with SARS-CoV and MERS-CoV (Baughn et al., 2020). Thus, we asked whether genetic variations in TMPRSS2 in humans lead to differences in infection rates or severity of disease symptoms of SARS-CoV-2. We examined the NCBI dbSNP database to identify SNPs in the TMPRSS2 gene. As of 10 December 2020, we found there were 11,023 intron variants, 393 missense variants, 186 synonymous variants, 3 in-frame insertion variants, 2 in-frame deletion variants, and 1 initiator codon variant reported. To narrow these down to 23 SNPs of interest, we first searched the ClinVar database to identify SNPs with general clinical significance, followed by searching the literature to determine SNPs specifically related to SARS-CoV-2 severity. One missense variant, rs12329760, results in an amino acid substitution, V160M, which has been predicted to alter TMPRSS2 function. A subset of these SNPs show differences in frequency in world populations, and we wondered if these SNPs had structural and functional consequences for the protein. A crystal structure of TMPRSS2 is not currently available. To visualize the structural consequences of amino acid substitutions, we performed homology modeling on TMPRSS2 (UniProt O15393) using the structure prediction software HHPred, RaptorX, and SwissModel based on the ~30% similarity to hepsin. The predicted structures of TMPRSS2 with various amino acid substitutions were then docked to the SARS-CoV-2 spike protein using I-TASSER and Haddock 2.4 to observe differences in binding interactions and therefore determine which sequence changes are predicted to alter binding interactions, potentially contributing to the wide variation of symptoms caused by COVID-19.

II. Introduction

a. Severe acute respiratory syndrome coronavirus, 2 (SARS-CoV-2) is a newly emerged virus that has caused a tragic pandemic.
    i. Recent human epidemics have been caused by coronaviruses:
        1. SARS-CoV: 2002

a. 8,098 infected, 774 deaths
2. Middle East respiratory syndrome coronavirus (MERS-CoV): 2012- present
    a. ~2,564 infected, 881 deaths (As of January 28, 2021)
3. SARS-CoV-2: 2019-present
    a. 101 million cases and 2.18 million deaths (As of January 28, 2021)
        i. Dong (2020)

ii. Unlike other viruses, SARS-CoV-2 cases range from asymptomatic to severe.
    1. The reasoning behind this is not completely clear.

b. The Spike (S) glycoprotein of SARS-CoV-2 is essential to virus attachment and entry. (Walls, 2020)
    i. It contains two subunits: S1 and S2 domain
        1. S1: identify and bind to host cell receptor.
        2. S2: fuses viral and host cell membrane together.
    ii. It been shown to use angiotensin-converting enzyme 2 (ACE2) using its S1 domain to attach onto the surface of a cell. (Hoffmann, 2020).
    iii. S glycoprotein is translated as precursor for entry.
        1. Like other coronaviruses, SARS-CoV-2 S protein must be cleaved at the S2' site for exposure of the fusion peptide.
        2. Unique of its kind, SARS-CoV-2 S protein has a furin-motif cleavage site at the S1/S2 site. (Walls, 2020)
        3. Host proteases like furin and transmembrane protease, serine 2 TMPRSS2 have both been seen to proteolytic cleave the S protein, activating it for entry. (Bestle, 2020)
            a. furin: S1/S2 site
            b. TMPRSS2: S2' site

c. TMPRSS2 is a safer approach as a target for COVID-19 understanding and therapeutics.
    i. ACE2 is part of the renal-angiotensin system, which is crucial for health.
        1. ACE2-knockout mice were observed to have severe cardiac defects and essential for heart function (Crackower, 2002).
    ii. TMPRSS2's physiological function is not completely understood.
        1. It's heightened expression in the prostate has been linked to development of prostate cancer tissue. (Mollica, 2020).
        2. TMPRSS2-knockout mice developed normally and survived to adulthood with no problems or noticeable differences. (Kim, 2006)
        3. Rats with knock-out TMPRSS2 had less severe immune response when infected with SARS-CoV and MERS-CoV and had less viral spread in the respiratory system. (Iwata-Yoshikawa, 2019)

iii. TMPRSS2 inhibitors is not unheard of in medicine.
1. Camostat mesylate, an inhibitor of TMPRSS2, has been shown in vitro to block SARS-CoV-2 entry and is a promising drug candidate, (Hoffmann 2020).
2. Bromhexine, is found in over-the-counter mucolytic cough suppressant and could be a safe treatment for coronaviruses.
a. It has been shown to inhibits TMPRSS2. (Lucas 2014)

d. Single nucleotide polymorphisms in coding regions of TMPRSS2 could affect interactions with S protein.
i. Single nucleotide polymorphisms (SNPs) are nucleotide changes that account for differences in the reference genome (99.99%) (Consortium, 2015)
ii. SNPs in the coding region can affect how a mRNA is transcribed and therefore how a protein is translated.
1. Missense mutations lead to an amino acid change which can be detrimental in terms of its structure and function.
a. Depends on gene locations & interactions.
i. Conserved domain, catalytic domain, etc.
iii. SNPs have been seen to possibly increase/decrease risk in infectivity.
1. Individuals with rs383510 and rs2070788 of the TMPRSS2 gene are more susceptible to get a severe form of influenza A and acute respiratory distress syndrome (Cheng, 2015).

e. TMPRSS4 may be involved with proteolytic cleavage of the SARS-CoV-2 Spike Protein (Zang, 2020).
i. High expression (>90%) in tissues such as the nasal cavity, esophagus, bronchus epithelium, colon, intestine, and oral cavity (UniProt - Q9NRS4)
ii. TMPRSS4 increased SARS-CoV-2 infectivity in intestinal epithelial cells.
iii. TMPRSS4 and ACE2 coexpressed in HEK293 cells lead to an increase in viral RNA levels.
iv. TMPRSS4 knockout led to a quadruple reduction in VSV-SARS-CoV-2 replication in human enteroid cells.
1. This was way larger than TMPRSS2 knockout cells.

f. **Research Question**: Can missense SNPs of TMPRSS2 and TMPRSS4 affect the function of the host protease and thus may explain the varied susceptibility and severity of COVID-19 cases?

III. Materials and Methods

a. Obtaining SNPs

i. Two missense variants with world frequency differences from Baughn 2020 were used as a baseline for significant SNPs/discovering new SNPs.
    1. rs75603675; c.23G>T p.Gly8Val
    2. rs12329760; c.589G>A p.Val160Met
ii. Each SNP was inputted onto PubMed to do a literature search.
iii. SNPs of TMPRSS2 that related to different diseases and conditions, including SARS-CoV-2 were seen as relevant.

b. TMPRSS2 background
    i. Since TMPRSS2 has not been studied extensively yet, TMPRSS2 was searched through different databases such as PubMed, ClinVar, UniProt, Google Scholar, and ScienceDirect in order to discover more about the gene/protein.
    ii. TMPRSS2 was searched into NCBI Gene and viewed with the Genome Browser.
        1. Missense variations can be seen by:
            a. Clicking ‘Tracks’ then ’Configure Tracks’ then search for variation. Select Missense and click configure to add the tracks to the gene.
        2. Isoforms can be seen by:
            a. Click “Genomic regions, transcripts, and products” section then “Switch ON mode ‘show All’ for Gene tracks” in the toolbar.
            b. Isoforms also were aligned using UniProt (O15393)
                i. Under “Sequences” click “Align” button

c. Using protein prediction programs
    i. Secondary structure prediction server was used to visualize point mutations in the amino acid sequence.
        1. PredictProtein
            a. FASTA format of TMPRSS2 was inputted into server.
    ii. TMPRSS2 has not been crystallized, therefore different homology protein predict servers were utilized to create 3D structures.
        1. SWISS Model
            a. FASTA format of isoform 1: TMPRSS2 (O15393) was inputted.
            b. 33.62% homology to hepsin
        2. HHPred
            a. FASTA format of isoform 1 was inputted into “Input” field and pressed submit.
            b. After run is finished, templates were selected to visualize and then select Create Model Using Selection.

c. PIR file was generated which was pasted into the MODDELLER software under 3ary structure.

3. RaptorX
    a. FASTA format of isoform 1 inputted into server.
    b. Contact probability matrix.
    c. 5 protein models
4. I-TASSER
    a. FASTA format of isoform 1 inputted
    b. Predicts secondary structures, solvent accessibility, C score, active sites, binding sites.
    c. Has 5 different 3D models.

iii. A de novo prediction server was used.

1. PredMP
    a. FASTA TMPRSS2 inputted into server.
    b. 5 predictive models

d. Molecular Docking of TMPRSS2 and SARS-CoV-2 Spike Protein

i. HADDOCK 2.4

1. PDB ID: 7DK3
2. PDB file obtained from I-TASSER was used for TMPRSS2.
    a. Catalytic Serine Active sites of TMPRSS2: His296, Asp345, Ser441 (Hussain 2020)
    b. Substrate binding sites: Asp435, Ser460, Gly462 (Hussain 2020)
    c. SARS-CoV-2 S2' Cleavage site: R815-S816
3. Clusters were saved as PDB files.

e. Visualizing & SNP Analysis

i. iCn3D Web-Based Structure Viewer

1. PDB file from HADDOCK (cluster 1) was inputted.
    a. File → Open File → PDB file
2. Interactions were viewed by:
    a. Under the Analysis tab, click "View Sequences and Annotations" then checking the Interactions box
    b. Under Analysis tab, click "H bonds and Interactions"
        i. Interaction map
        ii. Table
        iii. Force graph

ii. Multiple Sequence Alignment

1. Amino acid changes corresponding to each SNP were made and made into FASTA format.

2. All FASTA sequences were inputted onto Phylogeny.fr for a multiple sequence alignment (MUSCLE format)
    a. Alignment was put into CLUSTAL format by clicking the CLUTAL tab.

iii. Gene Heat Map with SNPs

iv. Allele frequency graph

1. A bar graph of allele frequency versus number of SNPS were graphed using Excel.
    a. Allele frequency for each SNP was obtained on the dbSNP webpage.
        i. ALFA frequency

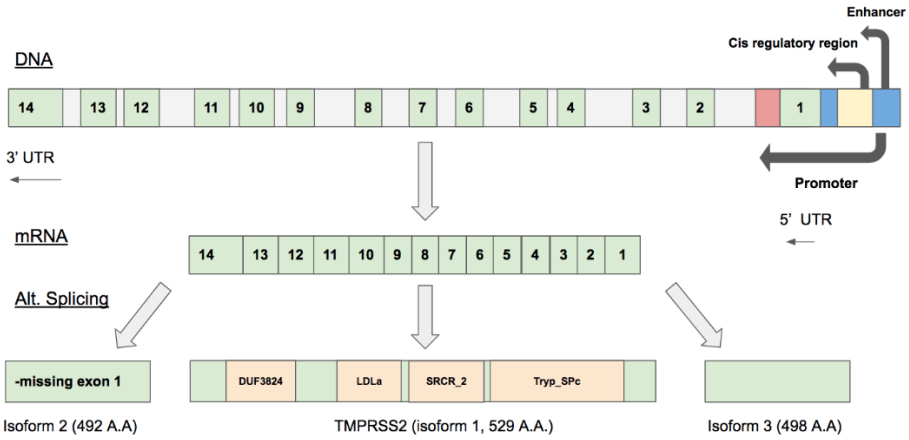
v. SNP prediction programs predicts how damaging a SNP would be.

1. Poly-Phen2
2. SIFT

IV. Results

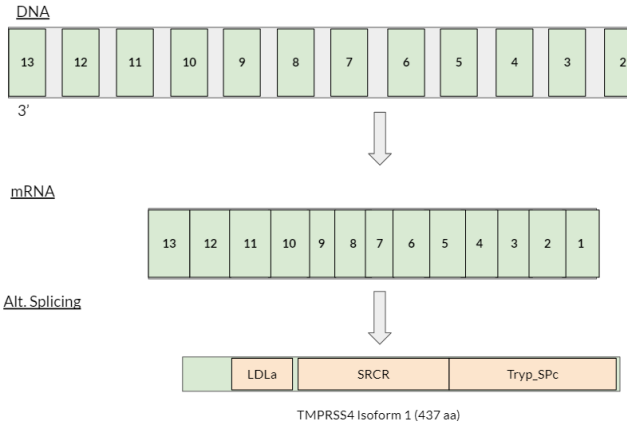
a. Gene Maps

i. TMPRSS2

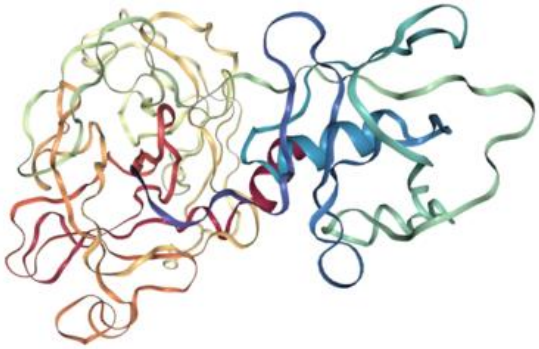
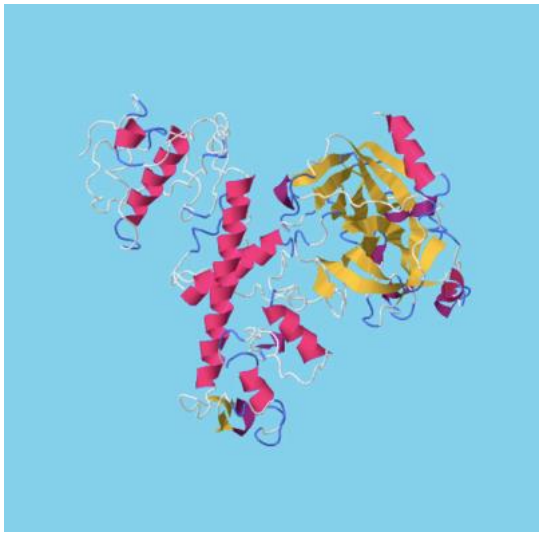


1.

ii. TMPRSS4

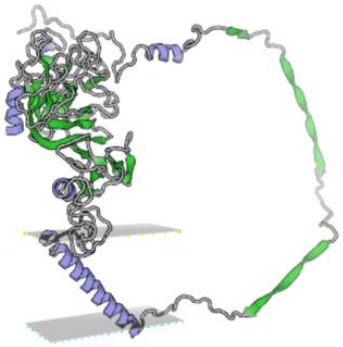


1.

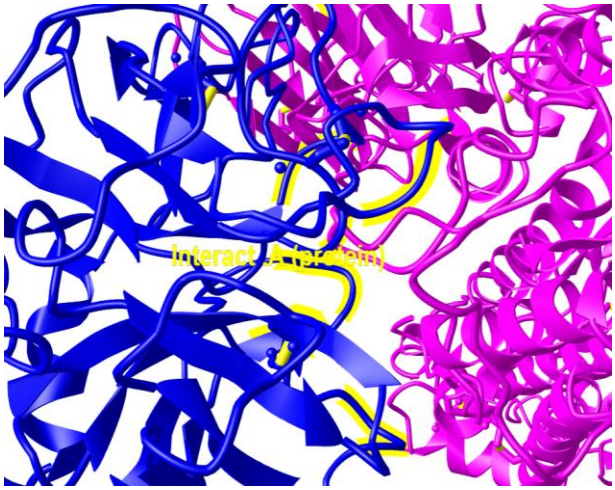
b. Protein Models
   i. SWISS-MODEL
      1. 
   ii. HHPred
      1. 
   iii. RaptorX
      1. 
   iv. I-TASSER

1. 

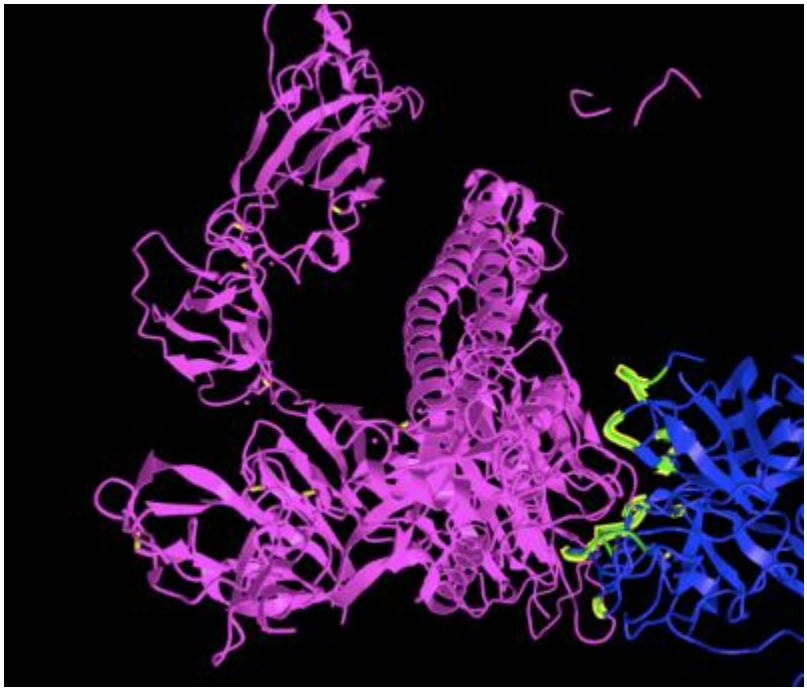
v. PredMP

1. 

c. Docking Model (Visualized using iCn3D)

i. 

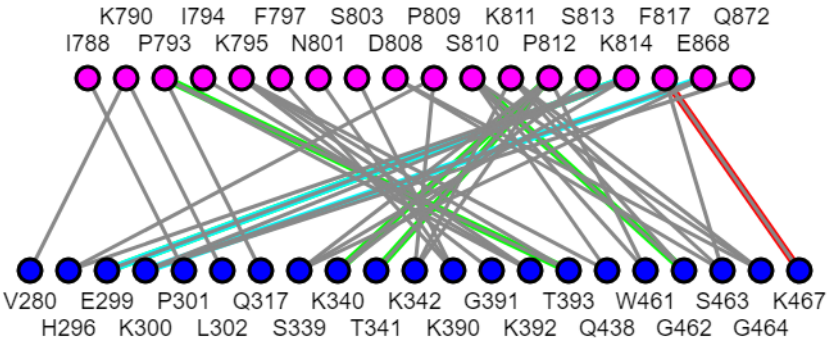
1. TMPRSS2: blue, SARS-CoV-2 S: pink

ii. 

1. Highlighted interactions in green

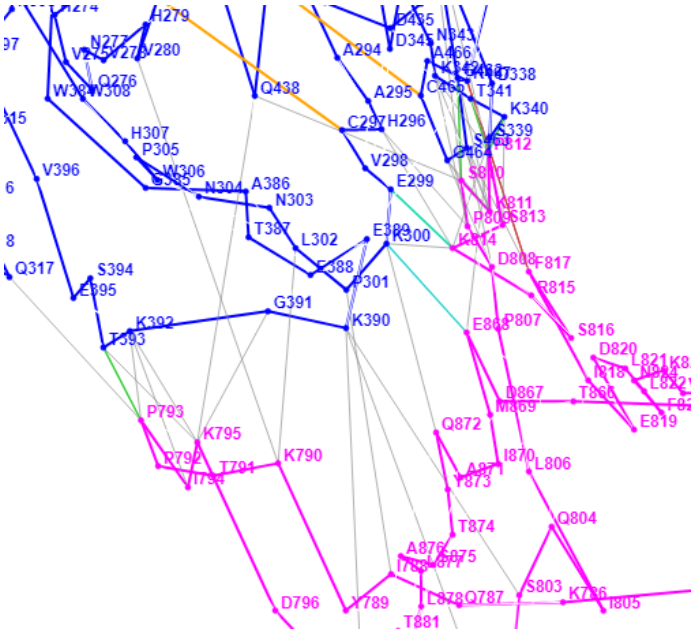
Hold Ctrl key to select multiple nodes/lines.
Green: H-Bonds; **Cyan**: Salt Bridge/Ionic; **Grey**: contacts
Magenta: Halogen Bonds; **Red**: π-Cation; **Blue**: π-Stacking

SVG PNG JSON **Scale**: 1



iii.

iv.



d. Heat Map

e. Allele Frequency Bar Graph

f. SIFT and Poly-Phen-2 Results

i.

TMPRSS2 SNP Predictions

rs Number	SIFT score	SIFT prediction	PolyPhen-2 score	PolyPhen-2 prediction
rs61735793	0.238	tolerated	0.015	Benign
rs75603675 G8V	0.201	tolerated	0.167	Benign
rs61735790	0.231	tolerated	0.033	Benign
rs12329760	0.009	deleterious	0.937	Probably Damaging
rs200291871	0.817	Tolerated	0.011	Benign
rs61735791	0.199	Tolerated	0.029	Benign
rs148125094	0.171	Tolerated	0.098	Benign
rs114363287	0.383	Tolerated	0.109	Benign
rs147711290 L128G	Not Found	-	0.920	Probably Damaging
rs147711290 L91P	0.005	Deleterious	1.000	Probably Damaging
rs147711290 L91R	Not Found	-	Not Found	-
rs150554820	0.004	Deleterious	0.549	Possibly Damaging
rs61735796	0.34	Tolerated	0.017	Benign
rs138651919	0.021	Deleterious	0.833	Possibly Damaging
rs61735795	0.551	Tolerated	0.086	Benign
rs142446494	0.015	Deleterious	0.294	Benign
rs201093031	1	Tolerated	0.00	Benign
rs768173297	Not Found	-	0.131	Benign

V. Discussion

a. Re-summarize introduction & research question.

b. Analyze SNPs in terms of frequency and location in the gene.

   i. Are there any hotspots?

      1. Are the frequencies related to hotspots?

c. Analyze SIFT/Poly-Phen Results and SNP Amino Acid changes location regarding to interactions.
    i. Interactions on iCn3D
    ii. Catalytic binding sites
    iii. Substrate-binding sites
    iv. Interaction near cleavage site on S protein
d. Comparing Protein Servers
    i. Which one was the most convenient?
    ii. Most accurate?
    iii. Easiest to use? User-friendly?
e. Compare results to literature.
    i. (Hussain 2020)
f. Future Research
    i. More SNPs
    ii. Utilizing Python or coding program to obtain a more accurate structure.
g. Limitations
    i. SNPs only available through available genomes/sequences
    ii. Lack of a crystallized structure for TMPRSS2
    iii. Lack of information regarding TMPRSS2's primary function.
h. Take home message.

References

Baughn, L. B., Sharma, N., Elhaik, E., Sekulic, A., Bryce, A. H., & Fonseca, R. (2020). Targeting TMPRSS2 in SARS-CoV-2 Infection. *Mayo Clinic proceedings*, 95(9), 1989–1999. https://doi.org/10.1016/j.mayocp.2020.06.018

Bestle, D., Heindl, M. R., Limburg, H., Pilgram, O., Moulton, H., Stein, D. A., ... & Böttcher-Friebertshäuser, E. (2020). TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells. *Life Science Alliance*, *3*(9). https://doi.org/10.26508/lsa.202000786

Cheng, Z., Zhou, J., To, K. K. W., Chu, H., Li, C., Wang, D., ... & Yuen, K. Y. (2015). Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic A (H1N1) influenza and A (H7N9) influenza. *The Journal of infectious diseases*, *212*(8), 1214-1221. https://doi.org/10.1093/infdis/jiv246

Consortium, G. P., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., & Kang, H. M. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74. https://dx.doi.org/10.1038%2Fnature15393

Crackower, M. A., Sarao, R., Oudit, G. Y., Yagil, C., Kozieradzki, I., Scanga, S. E., ... & Penninger, J. M. (2002). Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature*, *417*(6891), 822-828. https://doi.org/10.1038/nature00786

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533-534. https://doi.org/10.1016/S1473-3099(20)30120-1

Hussain, M., Jabeen, N., Amanullah, A., Baig, A. A., Aziz, B., Shabbir, S., ... & Uddin, N. (2020). Molecular docking between human TMPRSS2 and SARS-CoV-2 spike protein: conformation and intermolecular interactions. *AIMS microbiology*, *6*(3), 350. https://doi.org/10.3934/microbiol.2020021

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., ... & 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. https://doi.org/10.1016/j.cell.2020.02.052

Iwata-Yoshikawa, N., Okamura, T., Shimizu, Y., Hasegawa, H., Takeda, M., & Nagata, N. (2019). TMPRSS2 contributes to virus spread and immunopathology in the airways of murine models after coronavirus infection. *Journal of virology*, *93*(6). https://doi.org/10.1128/JVI.01815-18

Kim, T. S., Heinlein, C., Hackman, R. C., & Nelson, P. S. (2006). Phenotypic analysis of mice lacking the Tmprss2-encoded protease. *Molecular and cellular biology*, *26*(3), 965-975. https://doi.org/10.1128/MCB.26.3.965-975.2006

Lucas, J. M., Heinlein, C., Kim, T., Hernandez, S. A., Malik, M. S., True, L. D., ... & Nelson, P. S. (2014). The androgen-regulated protease TMPRSS2 activates a proteolytic cascade involving components of the tumor microenvironment and promotes prostate cancer metastasis. *Cancer discovery*, *4*(11), 1310-1325. https://doi.org/10.1158/2159-8290

Mollica, V., Rizzo, A., & Massari, F. (2020). The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer. *Future Medicine*, 16(27), 2029–2033. https://doi.org/10.2217/fon-2020-0571

Shen, L.W.; Mao, H.J., Wu, Y.L.; Tanaka,Y.; Zhang,W. (2017) TMPRSS2: A potential target for treatment of influenza virus and coronavirus infections, *Biochimie*, 142, 1-10. https://doi.org/10.1016/j.biochi.2017.07.016

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, *181*(2), 281-292. https://doi.org/10.1016/j.cell.2020.02.058

Zang, R., Castro, M. F. G., McCune, B. T., Zeng, Q., Rothlauf, P. W., Sonnek, N. M., ... & Ding, S. (2020). TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. *Science Immunology*, *5*(47), eabc3582. https://dx.doi.org/10.1126%2Fsciimmunol.abc3582