Applying bioinformatics to examine the structural and functional relationships between TMPRSS2 SNPs and SARS-CoV-2 and their potential implications for Covid-19 infection.

Annika Dinulos

Loyola Marymount University

Kam Dahlquist, Ph.D.

Annika Dinulos

Capstone Outline

Applying bioinformatics to examine the structural and functional relationships between TMPRSS2 SNPs and SARS-CoV-2 and their potential implications for Covid-19 infection.

Abstract

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.

Introduction

- SARS-CoV-2 is a virus that causes infection in humans called Covid-19, which may range from asymptomatic to severe. SARS-CoV-2 is part of the coronavirus family which utilizes a spike protein to bind to the human ACE-2 receptor and enter host cells (Hoffman et al., 2020).
    - Cell membrane fusion and viral entry is finalized through action of transmembrane protease serine 2 (TMPRSS2) proteases and FURIN, where TMPRSS2 cleaves the spike glycoprotein at the S1/S2 site.
    - TMPRSS2 plays a vital role in infection, and different single nucleotide polymorphisms may contribute to disease heterogeneity.
- The primary function of TMPRSS2 is unknown; however, functional research indicates that it is important in androgen regulation and increases the frequency of prostate cancer when fused with ERG (Mollica et al., 2020)
    - TMPRSS2 shows some activity with influenza and other respiratory diseases (Shen et al., 2020).
    - This could imply that varying TMPRSS2 expression may be connected to Covid-19 symptom severity.
- TMPRSS2 has 14 exons and has 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.
    - 18 SNPs were selected by highest frequency using the NCBI database and gnoMAD.

- It is important to determine if these SNPs have an effect on the protein, and if they are damaging. Software exists that score and categorize the variants as benign/damaging, and tolerated/deleterious (David et al., 2020)
- The aim of this research is to determine the structural and functional implications of these SNPs and how they may impact disease symptoms and progression.
    - Implications for this paper include providing more information on a relatively unknown protein, one that is critical to SARS-CoV-2 viral entry.
    - TMPRSS2 has no crystallized structure, so this research worked to develop structures using structure prediction programs that use homology modeling.
        - This includes: I-TASSER, RaptorX, SwissModel and HHPred
    - To examine interactions between TMPRSS2 and SARS-CoV-2, the predicted models of TMPRSS2 were docked in HADDOCK 2.4 from BonvinLab.
        - iC3nD was used to analyze the docked interactions.
- TMPRSS4 was also examined due to its possible connection to SARS-CoV-2; however, peer-reviewed data was sparse.

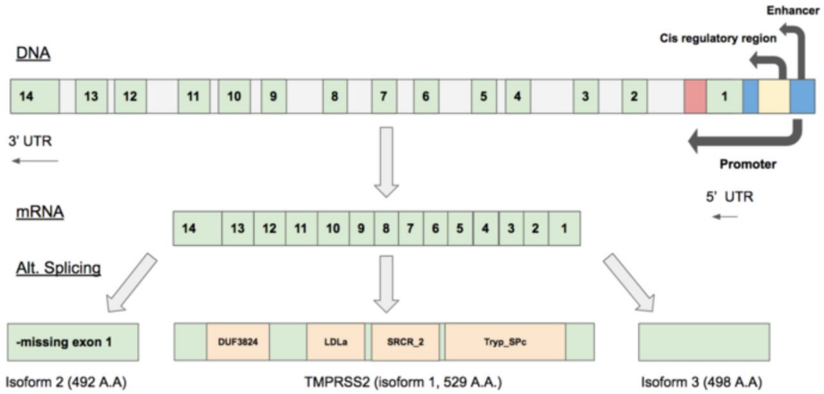
Methods

- Exploratory research on TMPRSS2 was conducted to obtain important information about the regulation, function and isoforms of the gene.
    - Gene and protein information was gathered from NCBI, Uniprot, and RCSB PDB
- A list of literature cited SNPs was compiled using information from NCBI dbSNP and ClinVar.
    - More SNPs were added using ALFA and gnoMad, and global frequency information was recorded.

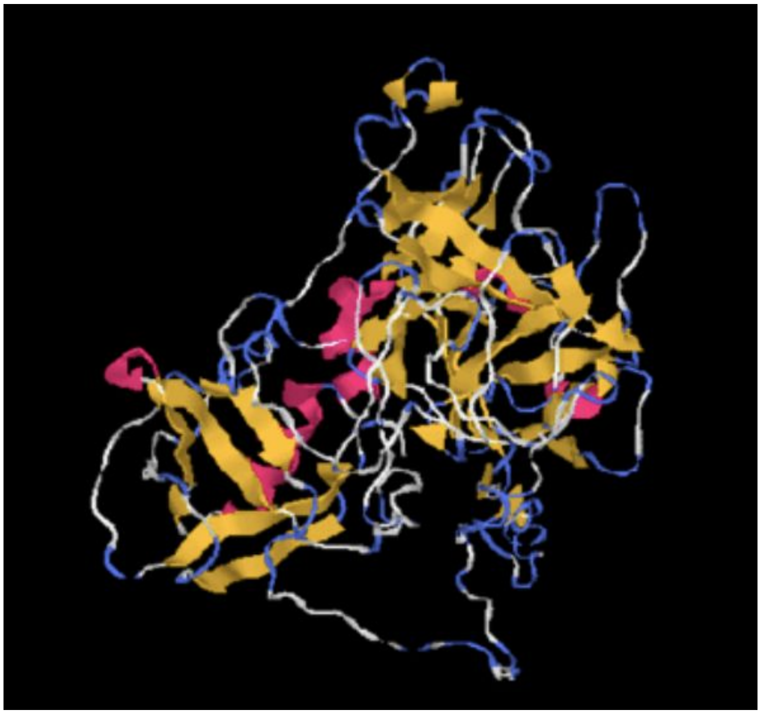
- We created a gene map of TMPRSS2 and its isoforms to visualize the exons.
- SNPs were analyzed using SIFT and PolyPhen2 to determine if any had any deleterious effects on the TMPRSS2 structure.
    - This can be done by entering the SNP name in the server
- Structure images of TMPRSS2 were generated using I-TASSER, HHPred, RaptorX, and SwissModel
    - FASTA sequences for TMPRSS2 were inputted into the software.
- TMPRSS2 and SARS-CoV-2 were docked using HADDOCK 2.4 to visualize the interaction between the two.
    - SARS-CoV-2 S protein PDB: 7dK3 chain A
    - TMPRSS2 PDB file generated from I-TASSER
    - Results create custom PDB file of docked molecules
- Docked PDB file was inputted into iC3nD to analyze the types of interactions between TMPRSS2 and SARS-CoV-2 S protein.
    - Steps: Upload, analysis, view sequence, interactions, details
        - Highlight SNPs on Interactions track
    - This generates lists and images of exact interactions between the two molecules in a form of a table and a network map respectively.
- 18 relevant SNPs were chosen using ALFA and gnoMad frequencies.
    - Individual FASTA sequences were made for each of the SNPs by manually changing the sequence in TextEditor
    - FASTA sequences were put into Phylogeny.fr to create a multiple sequence alignment, to determine the conservation of each of the SNPs.
        - Paste sequences, BLAST, Alignment, Clustal format

- Frequency vs. SNP graph was generated using Excel
- Heat map of spots where SNPs occurred on the protein will be generated.
- TMPRSS4 data was generated as well
    - Gene map and predicted structure through I-TASSER

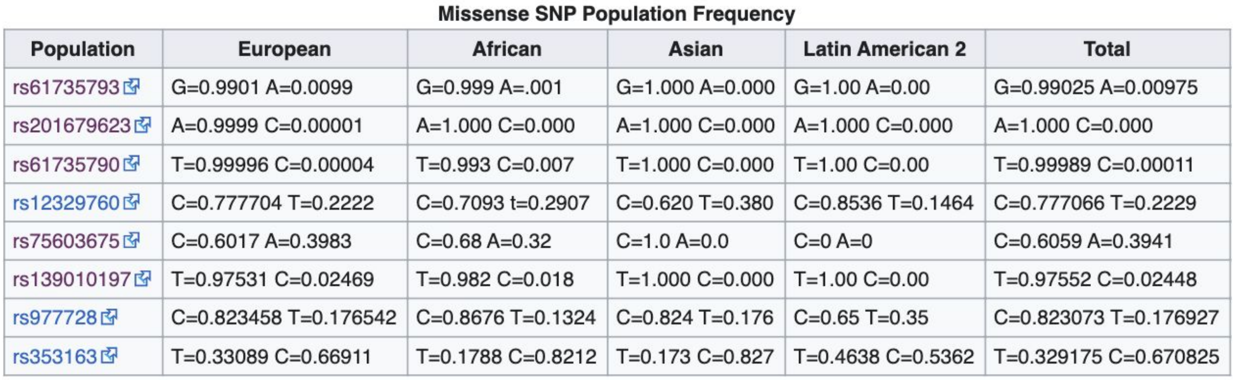
Results



- Gene map of TMPRSS2
    - Shows the regulatory regions, enhancer, promoter and UTRs
    - Alternative splicing isoforms are also indicated



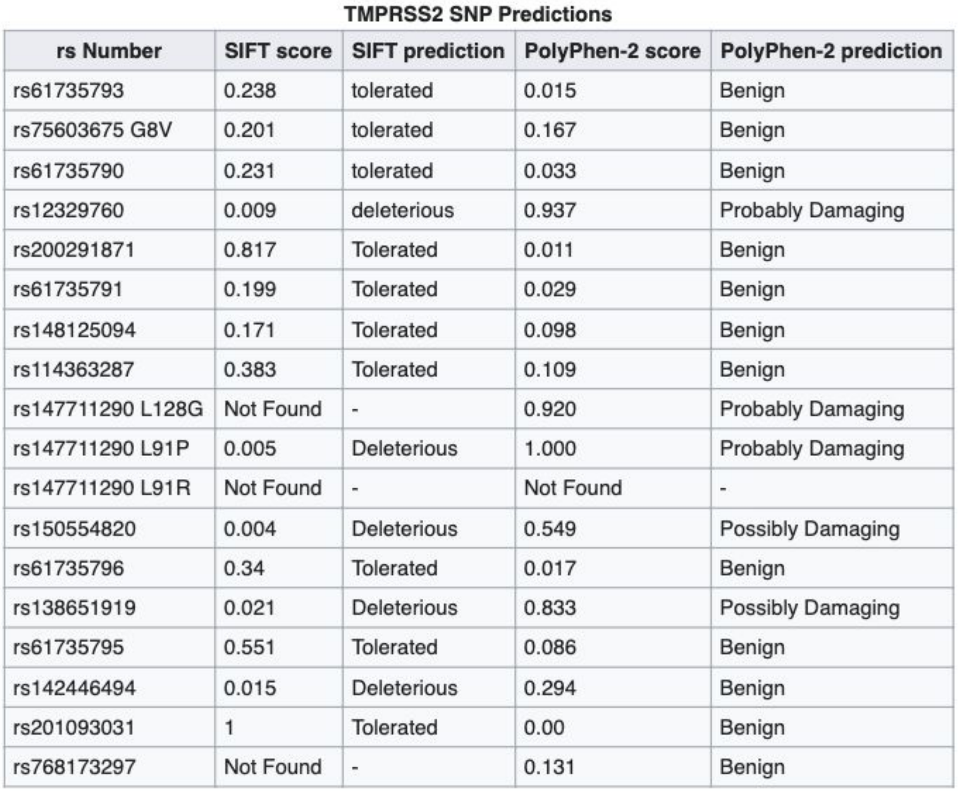
- Predicted structure of TMPRSS2 using I-TASSER, based on ~30% homology to hepsin



Missense SNP Population Frequency

Population	European	African	Asian	Latin American 2	Total
rs61735793	G=0.9901 A=0.0099	G=0.999 A=.001	G=1.000 A=0.000	G=1.00 A=0.00	G=0.99025 A=0.00975
rs201679623	A=0.9999 C=0.00001	A=1.000 C=0.000	A=1.000 C=0.000	A=1.000 C=0.000	A=1.000 C=0.000
rs61735790	T=0.99996 C=0.00004	T=0.993 C=0.007	T=1.000 C=0.000	T=1.00 C=0.00	T=0.99989 C=0.00011
rs12329760	C=0.777704 T=0.2222	C=0.7093 t=0.2907	C=0.620 T=0.380	C=0.8536 T=0.1464	C=0.777066 T=0.2229
rs75603675	C=0.6017 A=0.3983	C=0.68 A=0.32	C=1.0 A=0.0	C=0 A=0	C=0.6059 A=0.3941
rs139010197	T=0.97531 C=0.02469	T=0.982 C=0.018	T=1.000 C=0.000	T=1.00 C=0.00	T=0.97552 C=0.02448
rs977728	C=0.823458 T=0.176542	C=0.8676 T=0.1324	C=0.824 T=0.176	C=0.65 T=0.35	C=0.823073 T=0.176927
rs353163	T=0.33089 C=0.66911	T=0.1788 C=0.8212	T=0.173 C=0.827	T=0.4638 C=0.5362	T=0.329175 C=0.670825

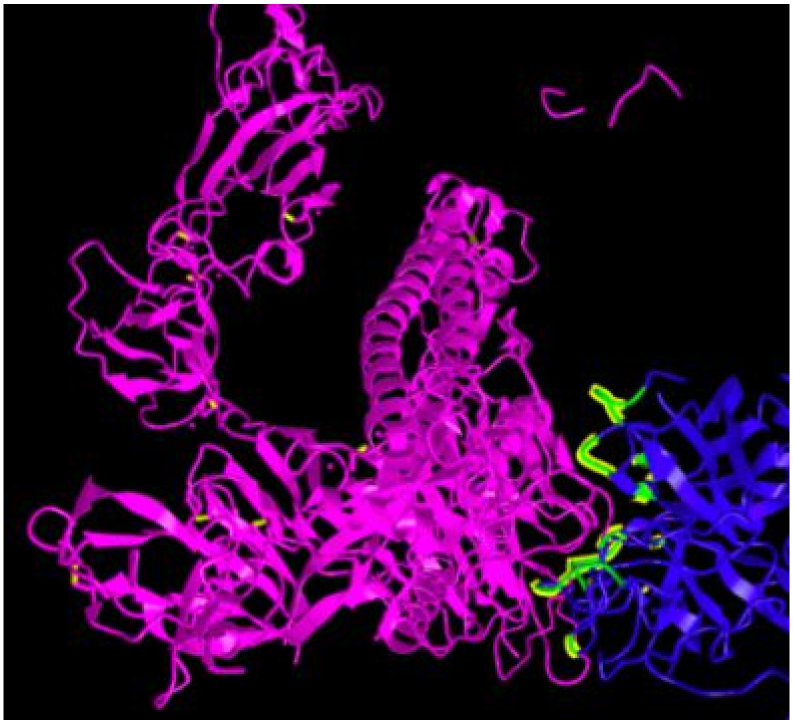
- Table displaying the frequencies of 8 TMPRSS2 SNPs according to NCBI
    - Will need to update this with the complete list of SNPs / ALFA frequencies
    - Most variant frequencies are small. The SNP with the greatest variation is rs75603675



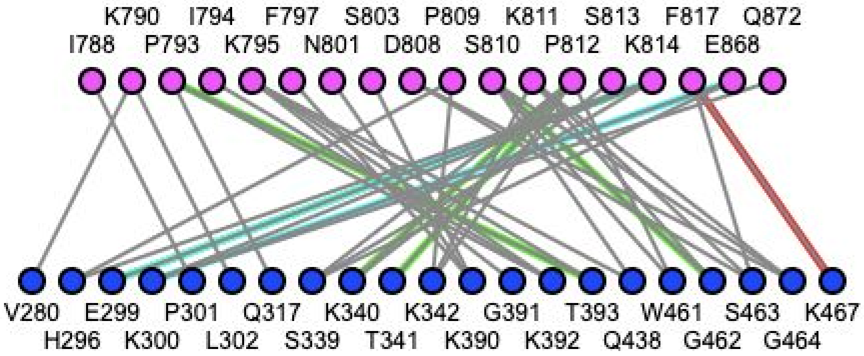
TMPRSS2 SNP Predictions

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

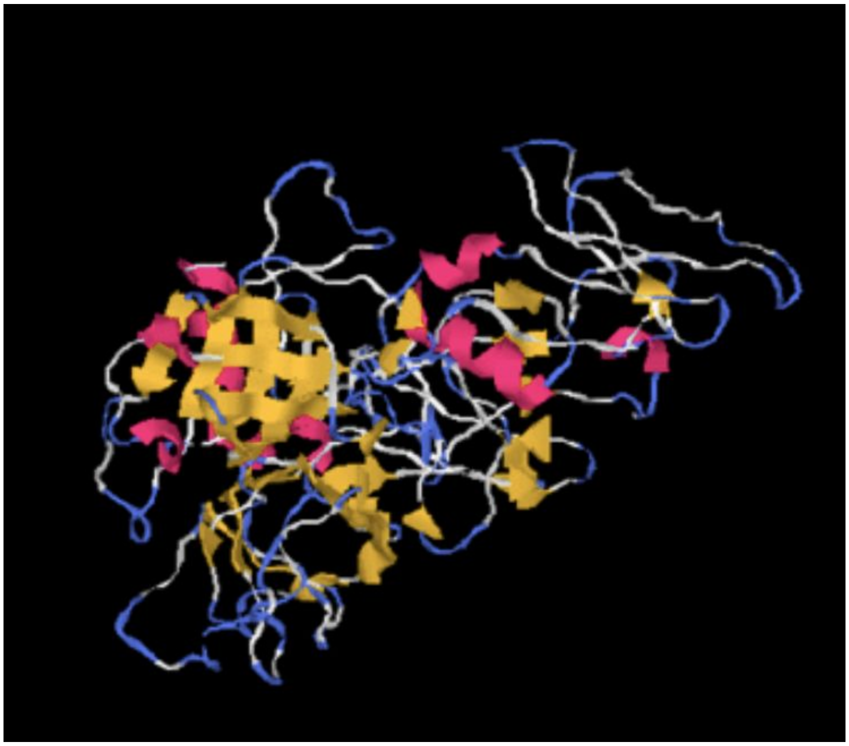
- TMPRSS2 SNP predictions using both SIFT and PolyPhen2
  - There are five SNPs that could be possibly damaging according to PolyPhen2
  - According to SIFT, there are six SNPs that are deleterious
    - The discrepancy between the two programs is rs1422446494



- TMPRSS2 and SARS-CoV-2 S docked molecules from HADDOCK 2.4 viewed on iC3nD.
- The green and yellow highlighted areas are sites of interaction between the two.


K790 I794 F797 S803 P809 K811 S813 F817 Q872
I788 P793 K795 N801 D808 S810 P812 K814 E868
V280 E299 P301 Q317 K340 K342 G391 T393 W461 S463 K467
H296 K300 L302 S339 T341 K390 K392 Q438 G462 G464

- TMPRSS2 and SARS-CoV-2 interaction map
  - Green: H-Bonds; Cyan: Salt Bridge/Ionic; Grey: contacts Magenta: Halogen Bonds; Red: π-Cation; Blue: π-Stacking



- TMPRSS4 predicted protein using I-TASSER
- Still needed:
  - Frequency vs SNP graph
  - Heat map of SNPS along protein

Discussion

- Five to six SNPs were determined to be deleterious according to SIFT and PolyPhen 2. These SNPS need to be visualized on the docked image derived from HADDOCK on iCn3D. We can discuss the actual amino acid change in relation to SIFT and PolyPhen2 data.
  - List out each SNP
  - If SNPs are not located at sites of interaction discuss their possible implications
- Compare protein servers between each other

    - Ex. HHPred generated incorrect proteins or proteins that could not be read by HADDOCK
    - I-TASSER was the most consistent and reliable predict protein server
- Important to note that most SNPs are not within the coding region, indicating that the frequencies of these SNPs will remain small, even as databases are updated with more information.
- Limitations to this study include choosing only one isoform to focus on. Other isoforms may show more SNPs with different effects on the protein, and thus their interaction with SARS-CoV-2. Additionally, we have narrowed down the SNPs to 18 by frequency; this does not indicate that these are the most damaging or important. Different criteria could be used. Protein structures were derived from predicted protein servers, as there is no crystallized structure available, it is possible that the actual protein structure looks different.
- Future directions of this research include gathering more information on more SNPs, there are 393 SNPs listed in the NCBI database, different criteria of exclusion could be used to provide a more full analysis of the impacts of polymorphisms

References

David, A., Khanna, T., Beykou, M., Hanna, G., & Sternberg, M. J. (2020). Structure, function and variants analysis of the androgen-regulated TMPRSS2, a drug target candidate for COVID-19 infection. *bioRxiv.*

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

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.