Results

TMPRSS2 has 14 exons and three isoforms (Figure 1). The promoter extends past the first exon. Isoform 1 has 529 amino acids, isoform 2 has 492 amino acids, and isoform 3 has 498 amino acids. Specifically, isoform 2 excludes exon 1. Regions that have been located include the promoter, enhancer, and the cis regulatory region. No additional information about isoform 3 was discovered.

Only nonsynonymous SNPs of TMPRSS2 were examined due to the possibility of structural changes in the shape of the protein. Of the 18 most common SNPs according to ALFA frequencies on dbSNP, the most common SNP is rs75603675, with a total frequency of 0.30337 on the alternate allele A (Table 1). This SNP occurs in exon one and is not observed in isoform 2. Only one other SNP is not observed in isoform 2, rs200291871. It was discovered that the occurrence of nonsynonymous SNPs are rare. Only 2 SNPs (rs75603675 and rs12329760) occur in frequencies larger than 0.1 (Table 1) (Figure 2). The other 16 SNP frequencies range from 10e^-5 to 10e^-3 (Table 1) (Figure 2). Global frequencies indicate that in European and African populations (Table 1). Only a small number of SNPs were found to be potentially damaging or deleterious to the function of the protein according to SIFT and PolyPhen-2. There are five SNPs that could be possibly damaging according to PolyPhen2. While SIFT determined 6 deleterious variations. The discrepancy between the two servers was rs1422446494 (Table 2). However, PredictProtein determined that rs142244694 is probably not very damaging to the protein, as it did not meet their criteria for possible damage. The effects of point mutations at the specific amino acids for these SNPs are visualized in Figure 3.

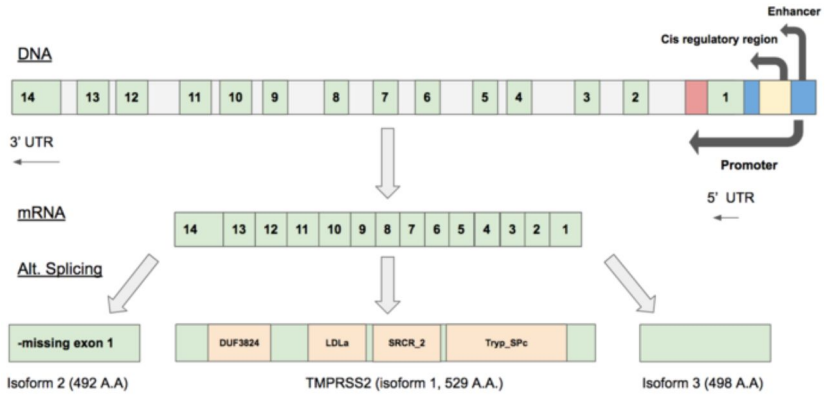

Figure 1. Gene Map displaying regulatory regions of TMPRSS2 and known isoforms. Isoform 2 excludes exon 1.

Table 1. Global frequency data obtained from NCBI dbSNP. Amino acid changes, sample size, and frequencies from European, African, Asian, Latin America, are listed. Total frequencies are listed in the rightmost column. Frequency data was compiled by NCBI from multiple databases including: ALFA, ExAC, gnomAD, GO exome sequencing project, and the PAGE study. Latin American 1 indicates those with Afro-Caribbean ancestry, while Latin American 2 indicated those with native ancestry. Asterisks indicate SNPs not found in isoform 2.

SNP ID	Amino Acid Change	Sample Size	European	African	Asian	Latin American 1	Latin American 2	Total Frequency
rs75603675*	Gly8Arg	17,922	C=0.65871 A=0.34129	C=0.9705 A=0.0295	C=1.0 A=0.0	C=1.00 A=0.00	C=1.0 A=0.0	C=0.69663 A=0.30337
rs12329760	Val160Met	295,780	C=0.779795 T=0.220205	C=0.70975 T=0.29025	C=0.6092 T=0.3908	C=0.7617 T=0.2383	C=0.8537 T=0.1463	C=0.775607 T=0.224393
rs61735793	Thr75Ile	191,500	G=0.989476 A=0.010524	G=0.9994 A=0.0006	G=1.0 A=0.0	G=0.998 A=0.002, C=0.000	G=0.9962 A=0.0038	G=0.990381 A=0.009619
rs200291871*	Gly8Arg	18,890	C=0.9887 G=0.013	C=0.9976 G=0.0024	C=1.0 G=0.0	C=1.000 G=0.000	C=1.0 G=0.0	C=0.99105 G=0.00895
rs61735791	Ala28Thr	203,412	C=0.996771 T=0.003229	C=0.9996 T=0.0004	C=0.9992 T=0.0008	C=1.000 T=0.000	C=0.999 T=0.001	C=0.996952 T=0.003048
rs148125094	Val415Ile	203,772	C=0.998865 T=0.001135	C=1.0 T=0.0	C=1.0 T=0.0	C=1.000 T=0.000	C=1.0 T=0.0	C=0.998955 T=0.001045
rs142446494	Val280Met	44,790	C=0.99939 T=0.00061	C=1.0 T=0.0	C=1.0 T=0.0	C=1.000 T=0.000	C=1.0 T=0.0	C=0.99929 T=0.00071
rs61735796	Glu260Lys	49,254	C=0.99919 T=0.00081	C=1.0 T=0.0	C=1.0 T=0.0	C=1.0 T=0.0	C=1.0 T=0.0	C=0.99933 T=0.00067
rs150554820	Phe209Ile	49,260	A=0.99928 T=0.00072	A=0.9992 T=0.0008	A=1.0 T=0.0	A=1.0 T=0.0	A=1.0 T=0.0	A=0.99933 T=0.00067
rs138651919	Pro41Leu	199,290	G=0.999588	G=0.9996	G=0.9997	G=1.0	G=1.0	G=0.999609

			A=0.000412	A=0.0004	A=0.0003	A=0.0	A=0.0	A=0.000391
rs61735790	His18Arg	199,596	T=0.999965 C=0.000035	T=0.9890 C=0.0110	T=1.0 C=0.0	T=0.991 C=0.009	T=1.0 C=0.0	T=0.999614 C=0.000386
rs768173297	Thr309Met	44,404	G=0.99966 A=0.00034	G=1.0 A=0.0	G=1.0 A=0.0	G=1.0 A=0.0	G=1.0 A=0.0	G=0.99973 A=0.00027
rs61735795	Pro375Ser	78,726	G=1.0 A=0.0	G=0.9963 A=0.0037	G=1.000 A=0.000	G=1.0 A=0.0	G=1.0 A=0.0	G=0.99982 A=0.00018
rs201093031	Val33Ala	58,202	A=0.99992 G=0.00008	A=1.0 G=0.0	A=0.994 G=0.006	A=1.0 G=0.0	A=1.0 G=0.0	A=0.99991 G=0.00009
rs147711290	Leu91Gln	107770	A= 0.99997 T=0.00000	A=0.9986 T=0.0012	A=1.000 T=0.000	A=0.999 T=0.001	A=1.000 T=0.000	A=0.999879 T=0.000074
rs114363287	Gly74Arg	199,516	C=0.999994 T=0.000006	C=0.9982 T=0.0018	C=1.0 T=0.0	C=0.998 T=0.002	C=1.0 T=0.0	C=0.999930 T=0.000070
rs147711290	Leu91Pro	107770	A= 0.99997 G=0.00003	A=0.9986 G=0.0002	A=1.0 G=0.0	A=0.999 G=0.0	A=1.0 G=0.0	A= 0.999879 G=0.000046

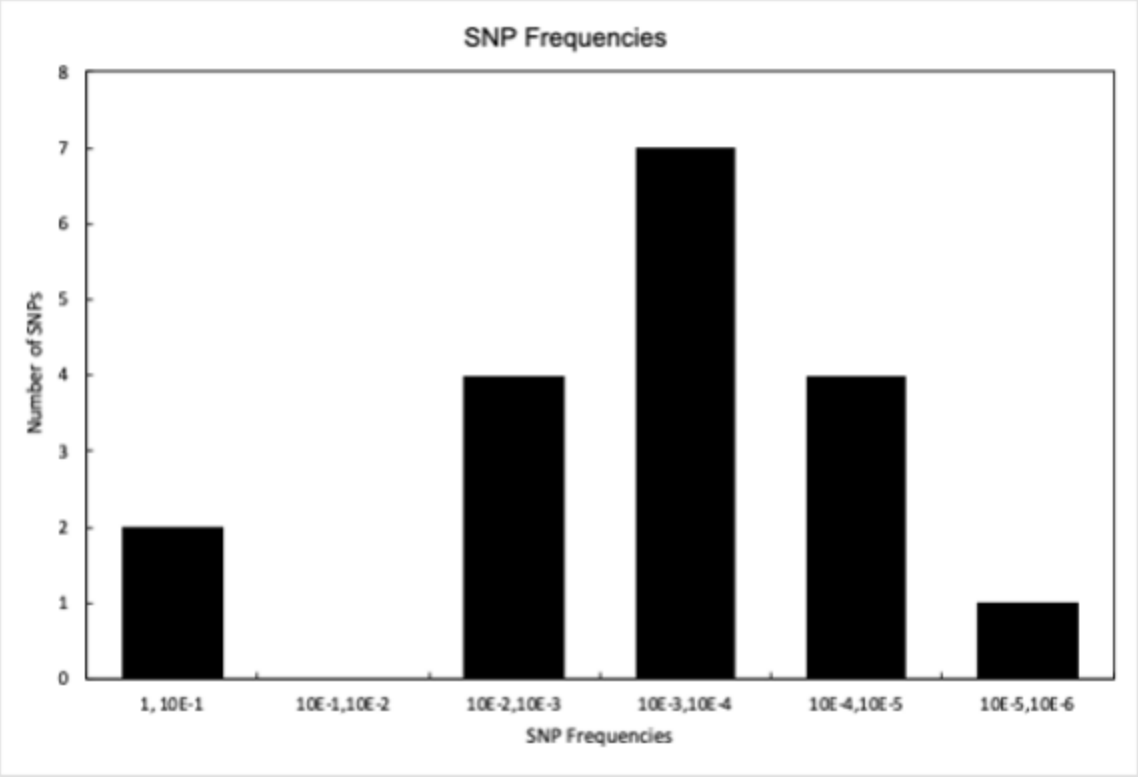

Figure 2. Histogram displaying the number of SNPs per set of frequencies. This will change to display the log bases from 10e^-6 to 1. Frequencies were obtained from dbSNP ALFA.

Table 2. SIFT and Polyphen-2 scores and predictions for chosen SNPs. Include criteria for scoring and predictions here.

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

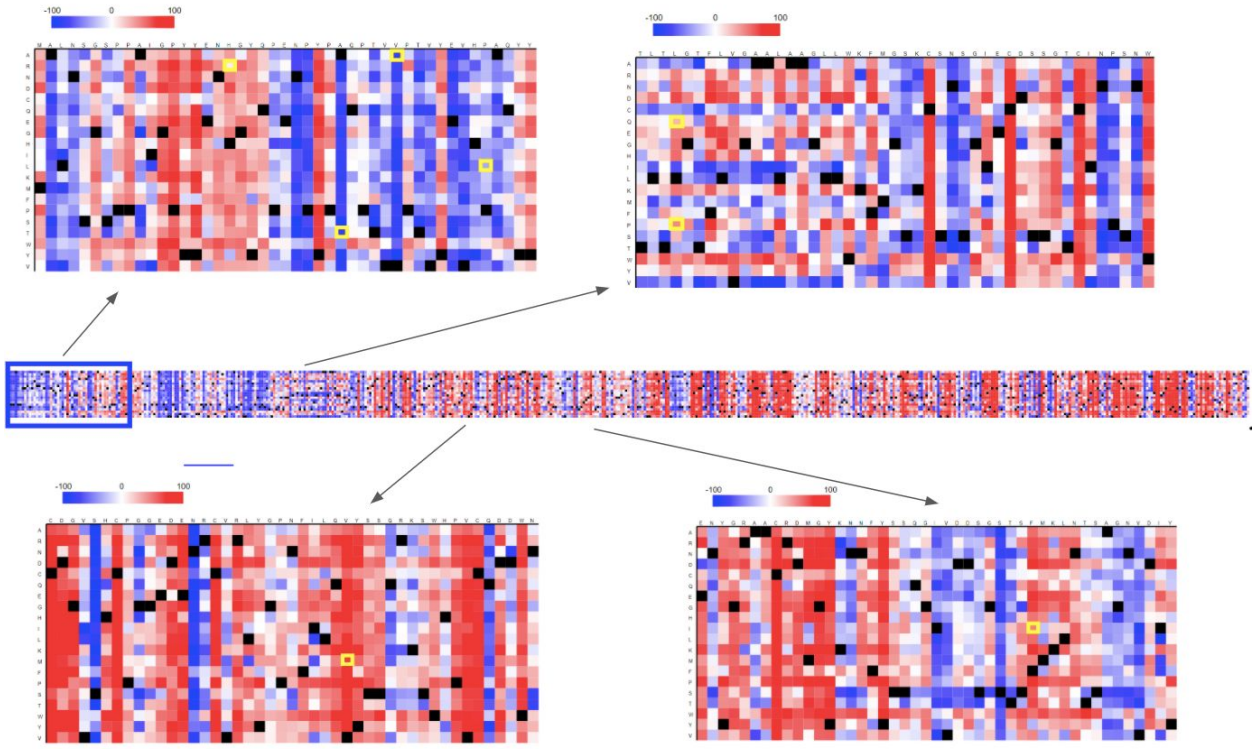

Figure 3. Heat map of the effect of point mutations in TMPRSS2. All 16 SNPs on isoform 2 were highlighted in yellow boxes. Only four panels are shown here. The four panels indicate damaging SNPs and do not account for most of the non-damaging and benign effects. Red shades indicate damaging effects, while blue shades indicate benign effects. Intensity of the colors indicates the severity of these mutations.

Due to the more extensive research done on isoform 2, the isoform 2 sequence was used to generate a predicted model for TMPRSS2, since no crystal structure has been confirmed. Several programs were used to generate models and compared through iCn3D including: SwissModel, HHPred, RaptorX, and I-TASSER.

The overall shape of the protein is globular, and appears to have more beta sheets compared to alpha helices. The numbers of each vary between each protein server. Each server matched TMPRSS2 with hepsin to build the general structures shown. Two programs, SwissModel (Figure 4) and HHPred (Figures 5) provided faulty structures that omitted parts of the sequence from their model of TMPRSS2. RaptorX (Figure 6) and I-TASSER (Figure 7)

generated complete structures. The RaptorX model appeared longer, compared to the wider, shorter structures generated by the other servers. Since I-TASSER is highly researched and included in many primary research articles, and showed the most consistent results of the four protein servers, we decided upon I-TASSER as the best model to dock TMPRSS2 with SARS-CoV-2.

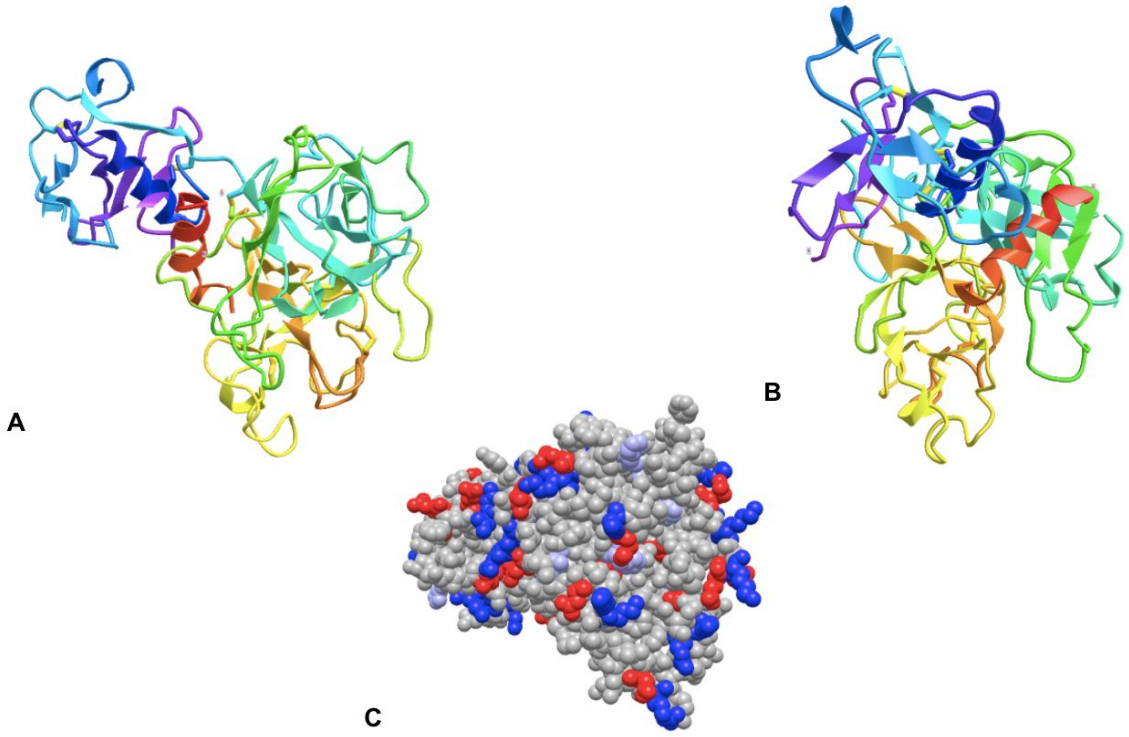

Figure 4. SwissModel structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 11 alpha helix sets and 20 beta sheets were predicted by SwissModel.

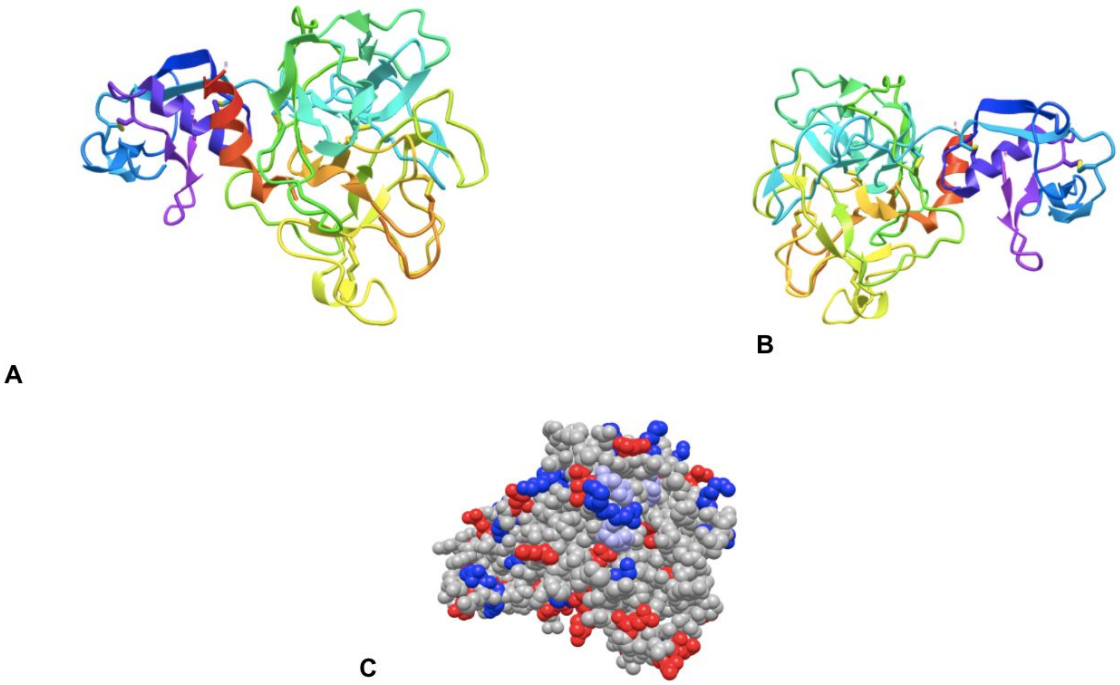

Figure 5. HHPred structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 6 alpha helix sets and 20 beta sheets were predicted by HHPred.

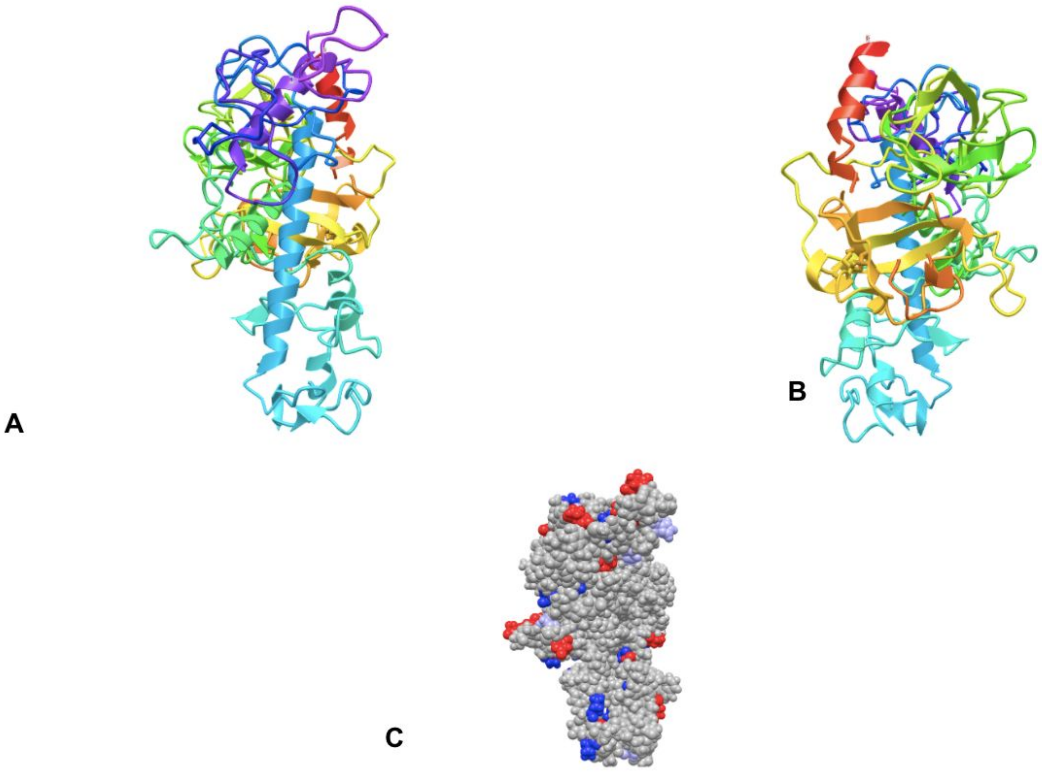

Figure 6. RaptorX structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 16 alpha helix sets and 15 beta sheets were predicted by RaptorX.

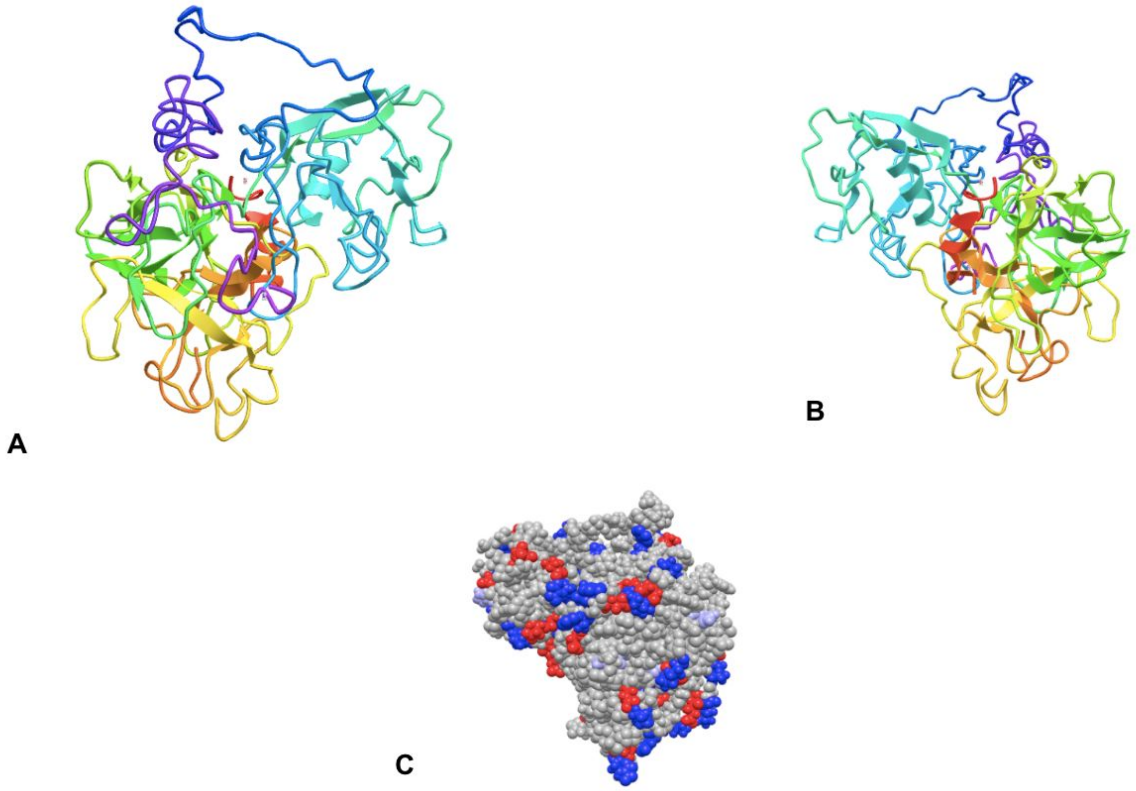

Figure 7. I-TASSER structure for TMPRSS2. The original view (A), 180 degrees rotated around the Y axis (B) and space-filling model (C) are shown here. Coloration of (A) and (B) is spectrum gradient based on proximity to n or c terminii. Coloration of the space-filling model (C) is done using charge. 4 alpha helix sets and 17 beta sheets were predicted by I-TASSER.

The predicted model of TMPRSS2 created by I-TASSER was docked with the known structure of SARS-CoV-2 S protein (PDB:7DK3) (Figure 8). 18 interaction sites were found on SARS-CoV-2 and 21 interaction sites were found on TMPRSS2 (Missing table). Two SNPs were found to be either on or close to the interaction sites. TMPRSS2 V280 interacts with SARS-CoV-2 S protein; SNP rs142446494 changes valine 280 to methionine. Four TMPRSS2 interaction sites exist from 300-317 amino acids; rs768173297 is located at 309 aa and changes threonine to methionine. It is interesting to note that the effect of point mutations for these SNPs were found to be not very damaging.

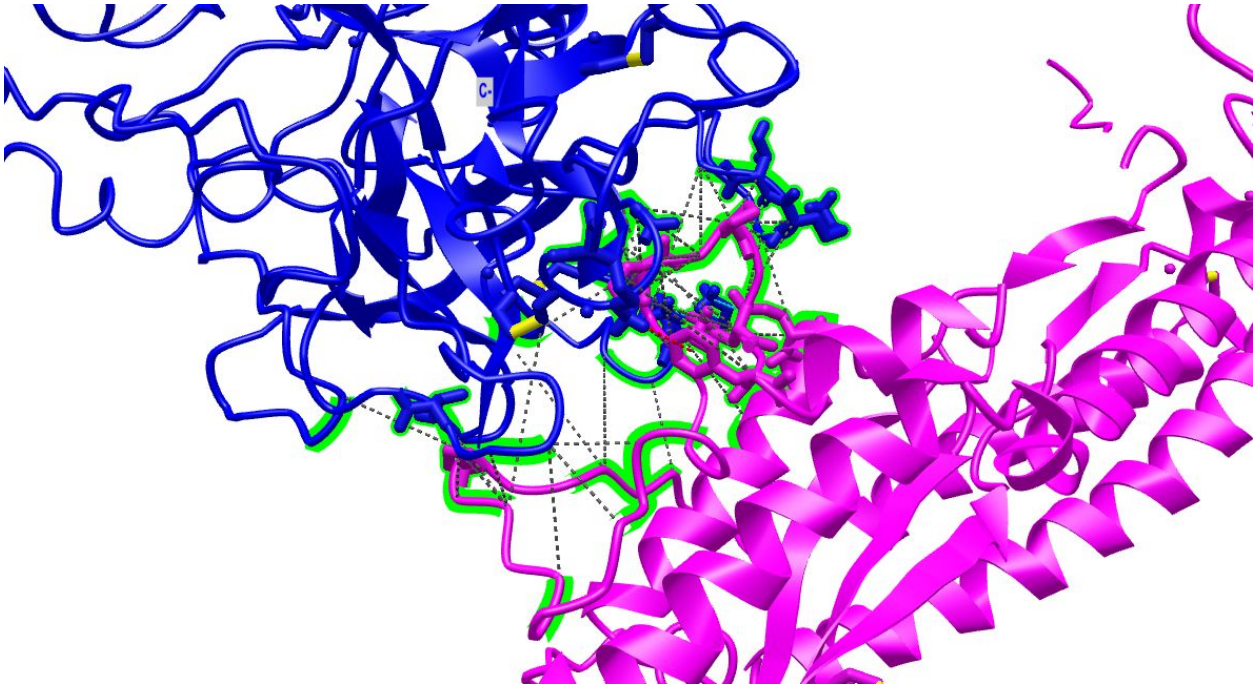

Figure 8. Interaction between TMPRSS2 and SARS-CoV-2. Green highlighted areas indicate interaction sites between the two molecules.

Discussion

TMPRSS2 is located on the 21st chromosome and has 14 exons. It's promoter region 41,507,255 - 41,508,648. The regulatory region is 41,508,250 - 41,508,276, and the enhancer is located at 41,508,128-41,509,135. Active sites of TMPRSS2 include His296, Asp345, Ser411 and substrate binding sites include Asp 435, Ser460 and Gly462 (Hussain et al., 2020). TMPRSS2 is necessary for spike protein priming and TMPRSS2 usage is necessary for SARS-CoV-2 infection of lung cells and it is possible that camostat mesylate treatment can inhibit TMPRSS2 and block SARS-CoV-2 infection (Hoffmann et. al, 2020).

Data on SNPs in global populations was not robust. Two nonsynonymous SNPs were identified as somewhat populous compared to other SNPs on TMPRSS2 (rsrs12329760 and rs rs756036750). These two SNPs occur in about 30-20% of global sample populations. The other SNPs range from values from 10e^-6 to 10e^-2 Additionally, it was discovered that only a few variations were considered to be deleterious or damaging according to PolyPhen-2 and SIFT,

which are servers that score and categorize the potential effects of variations (David et al., 2020). These were identified in small sample sizes in the study (Table 1). PolyPhen-2 predicts whether amino acid substitutions will be damaging or non-damaging to the protein. PolyPhen-2 determines that a substitution is benign if the score is less than 0.05, possibly damaging between 0.5-0.9 and probably damaging if the score is between 0.9 to 1. SIFT provides scores and predictions based on the substitution it is examining. The amino acid is considered damaging if the score is less than 0.05. Tolerated means that the score is greater than 0.05. These ALFA frequencies obtained from several studies provided information on selected ancestries and certainly do not give an accurate depiction of world distributions of variations. More data should be obtained from other populations with larger sample sizes to be able to generalize the findings. It is important to note that most SNPs in the population do not exist within the coding region, indicating that the frequencies of these nonsynonymous SNPs will remain small, even as databases are updated with more information. Additionally, these frequencies did not contain samples of people with Covid-19, so no direct conclusions can be drawn from the frequency of SNPs. Future research gathered on TMPRSS2 samples with positive Covid-19 diagnoses is important to advance these findings.

Evaluation of Structure Programs

No crystal structure for TMPRSS2 currently exists, so examination of structure was completed with the use of different structure prediction programs. The four programs used were HHPred, Raptor X, SwissModel, and I-TASSER. Raptor X is a structure prediction server developed by the Xu group. It predicts secondary and tertiary protein structures, contacts, solvent accessibility, disordered regions, and binding sites. HHPred can detect remote protein homology and structure prediction, including secondary and tertiary structure. Swiss Model also uses structure

homology-modelling and already had a model for TMPRSS2. I-TASSER was identified as a highly published and widely used structure program. It was created by Zhang Lab and it uses threading to predict protein secondary structures and 3d models. All models were visualized through iCn3D. User friendliness of each of the programs was about equal; however, both SwissModel and HHPred omitted areas of the sequence from their structure. Assessment of the TMPRSS2 models were also done using MolProbity. MolProbity generates ramachandran plots for each of the models. **Include analysis of plots when they are fixed here (z-score and number of outliers).** The I-TASSER model included the entire sequence so the structure generated by I-TASSER was used to dock with SARS-CoV-2 S protein. It should also be noted that a de novo program was also used to generate a predicted protein structure; however, the resulting model was not usable.

Docking was performed using HADDOCK 2.4 from BonvinLab. It models the interaction between two molecular structures and their fit. The docked structure was also visualized using iCn3D. This docking revealed 21 interaction residues on TMPRSS2 and 18 on SARS-CoV-2 S protein. Of these residues we discovered that a few SNPs are located on or nearby the interaction sites. SNP rs142446494 is an interaction site, changing valine 280 to methionine. It is located on a beta sheet and interacts with K790, which immediately follows the fusion peptide. Both valine and methionine are nonpolar amino acids, however, the addition of a sulfur atom could unfavorably bind to other nearby amino acids. Information about the SNP is limited and it has not been identified by ClinVar. Evaluation of this SNP was determined to be deleterious by SIFT but benign according to Poly-Phen. Rs148125094 changes valine 415 to isoleucine. It is located near interaction sites in the serine protease domain. PredictProtein did not determine this substitution to have negative effects. Rs61735796 changes glutamate 260 to lysine, and is somewhat close to interaction sites on the spike protein. It also exists in the serine

protease domain; SIFT and PolyPhen-2 have determined that the variation is tolerated. Rs61735796 changes threonine 309 to methionine and is also nearby interaction sites. It is located in a beta sheet and PredictProtein did not predict it to have a negative effect on the protein structure. PolyPhen and SIFT data have not been found for this SNP. Additionally, there are other SNPs that do not exist on the interaction site but are found in conserved regions. These include rs12329760 which changes valine 160 to methionine. It is found in the scavenger receptor cysteine rich domain (SRCR) conserved domain and PredictProtein, PolyPhen-2, and SIFT all determined the variation to be damaging. Rs150554820 changes phenylalanine 209 to isoleucine. This SNP is not located near interaction sites but it is in the SRCR conserved domain and was predicted to be deleterious and damaging according to PredictProtein, PolyPhen-2 and SIFT.

Research on TMPRSS2 is slim and until the structure is crystallized it cannot be determined that these variations have an absolute effect on Covid-19 disease pathogenicity. The data we have presented is predictive and hypothesized that these interaction sites would be critical for the binding of TMPRSS2 to SARS-CoV-2.

Limitations to this study include only examining isoform 2. Other isoforms may show more SNPs with different effects on the protein (for example, SNP rs75603675 is a SNP located on isoform 1 that exists in 30% of sequence populations). 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 as exclusion factors to examine more variations within TMPRSS2. Protein structures were derived from predicted protein serves, as there is no crystallized structure available. It is possible that the actual structure is very different from the predicted homology models.

Future directions of this research include gathering a more exhaustive list of SNPs to analyze. There are 393 SNPs listed in the NCBI database, so different criteria of exclusion could be used to provide a more full analysis of the impacts of polymorphism. Additionally, examination of other genes in the TMPRSS family, such as TMPRSS11 and TMPRSS4, could provide more information on SARS-CoV-2 disease. Klassen et. al (2020) noted that there could be a possible connection between TMPRSS2 and TMPRSS11 function that may impact priming of SARS-CoV-2. Other research indicates that TMPRSS4 is more expressed in intestinal enterocytes and is relevant to influenza A infection, which may increase susceptibility for flu-like responses to SARS-CoV-2 (Zang. et al, 2020).

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

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.
Hussain et. al provide very useful information on TMPRSS2 and SARS-CoV-2 spike protein interactions. They provide docking and conformational models that visualize several possible interactions. Several different polymorphisms are provided in their results section that should be examined.

Klaassen, K., Stankovic, B., Zukic, B., Kotur, N., Gasic, V., Pavlovic, S., & Stojiljkovic, M. (2020). Functional prediction and comparative population analysis of variants in genes for proteases and innate immunity related to SARS-CoV-2 infection. bioRxiv.

Zang, R., Gomez Castro, M. F., McCune, B. T., Zeng, Q., Rothlauf, P. W., Sonnek, N. M., Liu, Z., Brulois, K. F., Wang, X., Greenberg, H. B., Diamond, M. S., Ciorba, M. A., Whelan, S., & Ding, S. (2020). TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of

human small intestinal enterocytes. *Science Immunology, 5*(47), eabc3582. https://doi.org/10.1126/sciimmunol.abc3582