Results

TMPRSS2 Gene Map and Models

The NCBI Genome Viewer provided that TMPRSS2 (Gene ID: 7113) is located on 21q22.3 and contains 14 exons (NCBI, 2014). The gene's promoter is 1,393 nucleotides long (41,507,255-41,508,648), containing a cis regulatory region of 27 nucleotides (41,508,250-41,508,276), and is located next to an enhancer region of 1,008 nucleotides (41,508,128-41,509,135) (Figure 1). The 14 exons undergo alternative splicing to form 3 different isoforms. Isoform 3 is made from a 3200 bp sequence of mRNA and its final protein is 498 amino acids in length (Figure 1). This isoform is not well studied and therefore its function and structure is not well understood (NCBI, 2014). Isoform 2 is spliced from a 3450 base pair sequence of mRNA with a final protein length of 492 amino acids, and a literature review revealed that it plays a role in activating human respiratory viruses through proteolytic cleavage to activate glycoproteins (Figure 1, Thunders et al.,2020) . It is known to cleave the hemagglutinin glycoprotein necessary for influenza A activation as well as the spike protein that is crucial for cell entry of SARS-CoV and MERS-CoV (Zmora et al.,2015).

Isoform 1 is composed of 529 amino acids and is spliced from an mRNA sequence of 3250 base pairs. It is structurally similar to isoform 2 and they contain the same transmembrane and extracellular domains including the domain of unknown function (44 bp-91 bp), the low density lipoprotein receptor class A domain (150 bp-185 bp), the scavenger receptor cysteine-rich domain (190 bp-283 bp), and the trypsin-like serine protease domain (293 bp-524 bp) (Figure 1, Thunders et al.,2020). TMPRSS2 is understood to be activated by cleavage between its protease domain and the rest of the protein, which causes the protease domain to transform into the active state in a process known as autocatalytic activation (Zmora et al.,2015).

Isoform 1 and 2 undergo autocatalytic activation in the same way, however the specificity of the initial cleavage may differ between the two isoforms due to the additional 37 amino acids present in the N terminus of the cytoplasmic domain of isoform 1. The additional 37 amino acids present in isoform 1 are what mainly distinguish it from isoform 2, as after cleavage isoform 2 has only one N-terminus fragment whereas isoform 2 has two fragments (Zmora et al.,2015).

Isoform 1 has been observed to activate Influenza A, SARS-CoV, and MERS-CoV through proteolytic cleavage of the same glycoproteins that isoform 2 cleaves (Thunders et al., Zmora et al.,2015). Such findings may be largely due to the similarities that these isoforms share in structure. Given the potential role that isoform 1 and 2 play in activating viral expression, these isoforms were of interest in which relevant SNPs that could alter its structure and binding interactions were searched.

Models of TMPRSS2 generated by various modelling softwares provided insight into the general structure of the protein as a reference point for the subsequent models of TMPRSS2 SNPs to be compared to. There were slight variations in the presentations of the proteins in each software, specifically in the number of alpha helices and beta sheets. Swiss-Model provided a model that consisted of 11 helices and 20 beta sheets, while I-TASSER cited only 4 helices and 17 beta sheets in its model (Figure 2,3). HHPred's model was slightly more similar to that of Swiss-Model, as their model of TMPRSS2 presented 6 helices and 20 beta sheets (Figure 4). The model produced by Raptor-X was composed of 16 alpha helices and 15 beta sheets (Figure 5).

The differences in the structure of TMPRSS2 is likely due to the differences in the ways that each software constructed its model. Swiss-Model and HHPred both form models based on homology, using templates of proteins that are homologous to the target sequence. The serine protease Hepsin displayed 33.62% homology to TMPRSS2 and therefore was utilized as the

template sequence for the structural model by both softwares (Waterhouse et al., 2018). This likely explains why these models produced more similar models of TMPRSS2, specifically with the same number of beta sheets (Figure 2,3).

I-TASSEER develops structural models using standard threading, which utilizes known protein templates that display a similar fold to the target protein. It refines its generated model by creating the lowest free-energy conformation and compares the final model to functional templates of similar proteins (Yang et al., 2015). Raptor-X relies on a similar template-based modeling methodology, however it uses some unique threading techniques to better predict proteins with no close homologs. Firstly, the software generates a profile entropy score for each of the unique homologs that could serve as a template for the target protein in order to determine the quality of each sequence. They then use conditional random fields, which allow for biological signals not utilized by any other protein modelling software to impact their non-linear threading score (Källberg et al.,2020). Lastly, their multiple template threading procedure allows multiple models of the target protein to be generated from different template sequences. This procedure can improve the overall accuracy of the generated models by correcting errors in alignments (Källberg et al., 2020). While homology modeling is a standard approach to generating protein structures, the techniques utilized by Raptor-X can be more useful for proteins that do not have a closely-related homolog. Given that the closest homolog to TMPRSS2, hepsin, displayed relatively low homology, Raptor-X may provide a more accurate structural prediction using its distance-based protein folding.

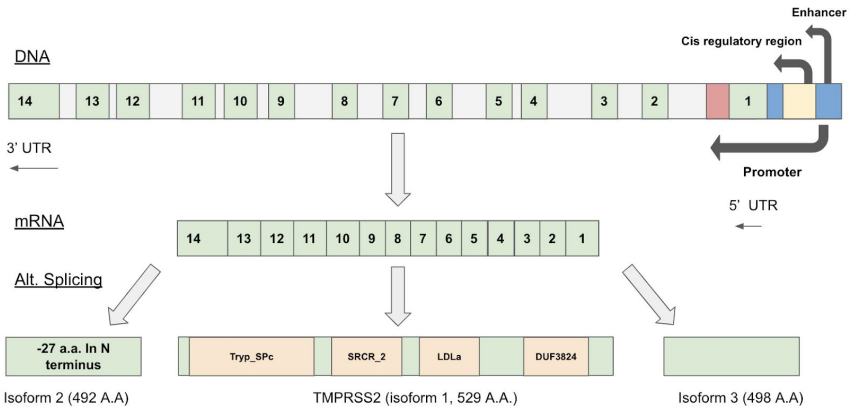
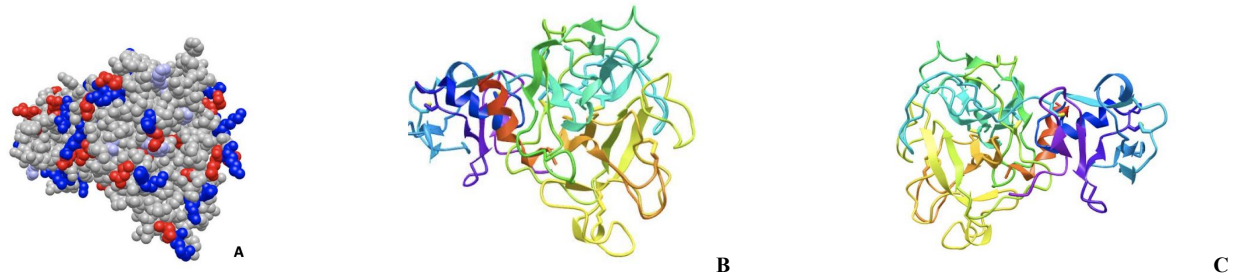


Figure 1. Gene Map of TMPRSS2 located on 21q22.3 from 41,464,305 bp - 41,508,158 bp with 14 labeled exons. On the DNA sequence, the gene's promoter is from 41,507,255 bp -41,508,648 bp. The cis regulatory region is from 41,508,250 bp - 41,508,276 bp. The enhancer region is from 41,508,128 bp - 41,509,135 bp. On Isoform 1 and 2, The Domain of Unknown Function (DUF3824) is 48 amino acids long at 44-91 bp.The Low Density Lipoprotein Receptor Class A (LDLa) domain is 36 amino acids long at 150 bp-185 bp. The Scavenger receptor cysteine-rich (SRCR_2) domain is 90 amino acids long at 190-283 bp. The Trypsin-like serine protease (Tryp_SPc) domain is 232 amino acids long at 293-524 bp. Isoform 2 is distinguished from isoform 1 by the 27 extra amino acids that are present in the N-terminus of isoform 1. Isoform 3 is not well studied and its domains have not been characterized.



c) d)

Figure 2. 3D models of TMPRSS2 using Swiss-Model (a) space filling model (b) ribbon model (c) ribbon model rotated 180°

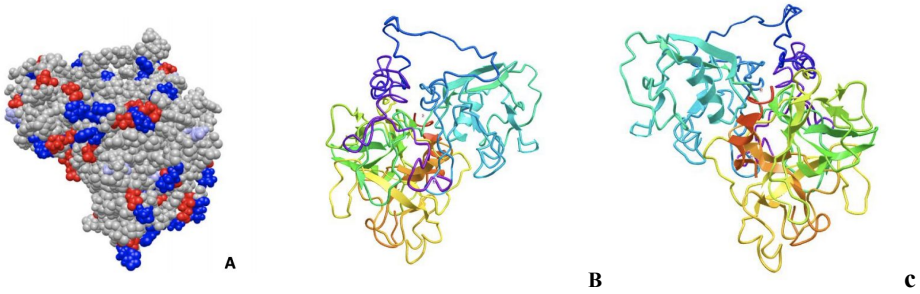


Figure 3. 3D models of TMPRSS2 using I-TASSER (a) space filling model (b) ribbon model (c) ribbon model rotated 180°

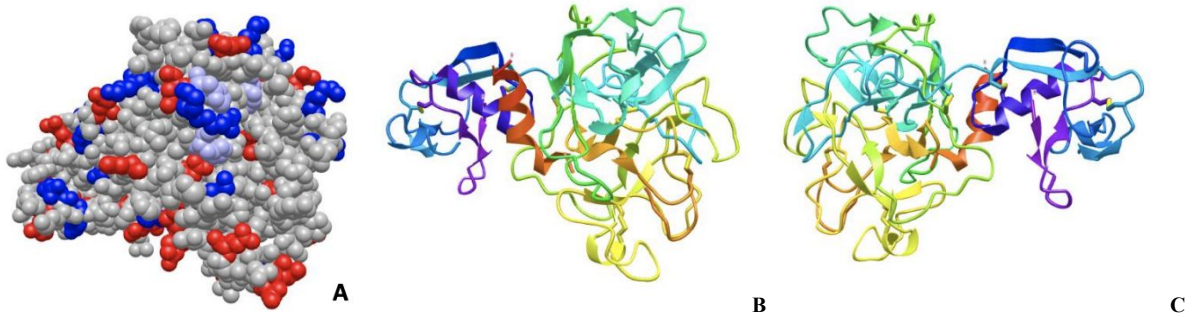


Figure 4. 3D models of TMPRSS2 using HHPred (a) space filling model (b) ribbon model (c) ribbon model rotated 180°

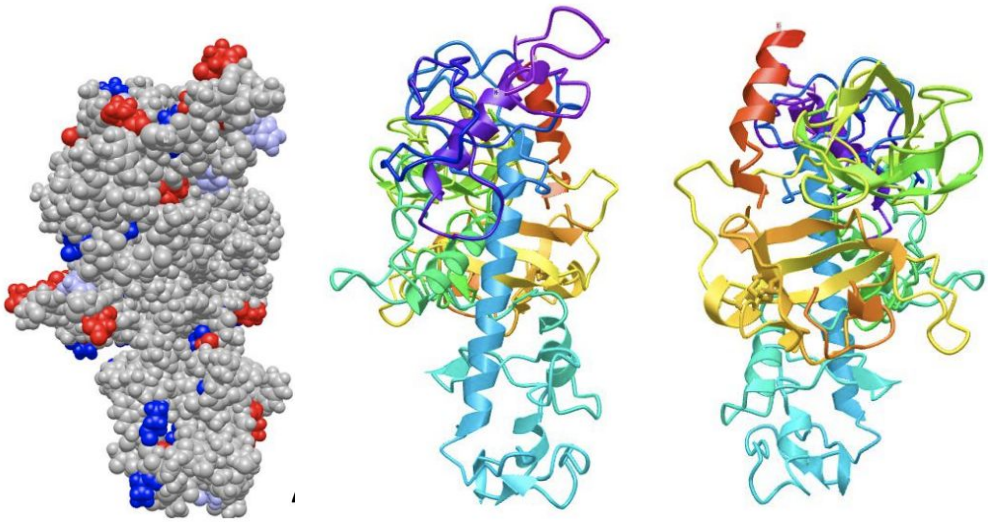


Figure 5. 3D models of TMPRSS2 using RaptorX (a) space filling model (b) ribbon model (c) ribbon model rotated 180°

Docking Interactions of TMPRSS2 and SARS-CoV-2

To understand how TMPRSS2 interacts with SARS-CoV-2, and which residues may be critical for their binding interaction, TMPRSS2 and the SARS-CoV-2 spike protein crystal structure (PDB: 7DK3) were docked using HADDOCK 2.4, and viewed with iCn3D (Figure 3a). The visualization revealed that there are 21 residues on TMPRSS2 and 18 residues on the spike protein that are critical for this binding interaction (Figure 3b). Each of these residues interact through specific interactions, including direct contact, hydrogen bonding, salt bridging, and π-cation. Given that nonsynonymous SNPs can alter the amino acid present at the specific location, a polar residue may be replaced with a non-polar residue which could affect the TMPRSS2's ability to form salt bridges with SARS-CoV-2. Alternate amino acids could also disrupt the network of hydrogen bonding and change the overall structure, affecting which residues come into contact during binding. Therefore, the residues that interact through hydrogen bonding, salt bridging, and direct contact are most likely to be altered by SNPs.

All of the residues of TMPRSS2 that are critical for binding to SARS-CoV-2 interact through direct contact with the spike protein, therefore they could all be affected by SNPs that result in changes to the structure of TMPRSS2 (Figure 3b). Two residues of TMPRSS2 interact with the spike protein through salt bridges (E299, K300), and four residues of TMPRSS2 interact with the spike protein through hydrogen bonding (K340,T341,T393, S463) (Figure 3b). These six residues of TMPRSS2 that have additional interactions with the spike protein besides direct contact may be at higher risk of being affected by a SNP.

The visualization of the docking interactions between TMPRSS2 and SARS-CoV-2 confirm that there are critical residues necessary for binding and, if altered by a SNP, may result in a disrupted binding interaction. These residues provide the specific framework for how TMPRSS2 and SARS-CoV-2 interact in a way that promotes viral entry into the host cell, providing insight into how these interactions could be altered to decrease viral entry. They also provided us with a reference point of which to compare our SNPs of interest, in order to discover if they alter the structure and subsequent function of one of these critical residues.

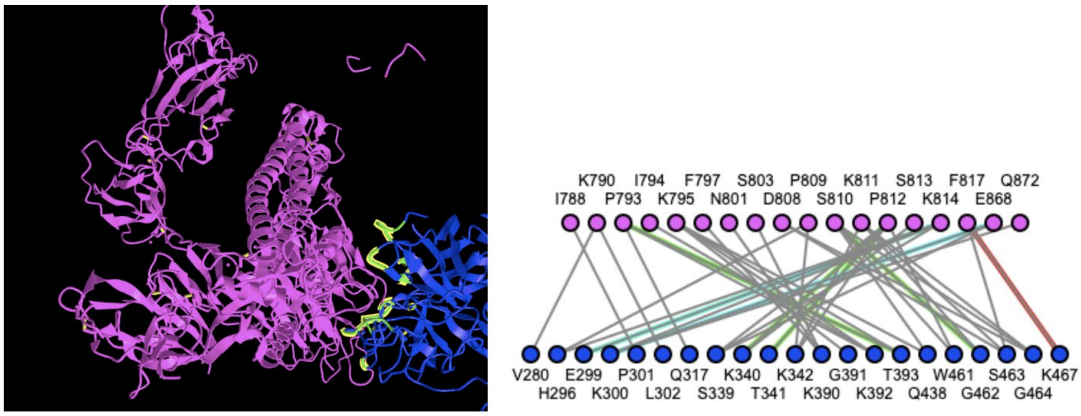


Figure 3. (a) Visualization of binding interactions between the SARS-CoV-2 spike protein and TMPRSS2 through iCn3D. The spike protein is pink, TMPRSS2 is blue, and the interacting residues between them are in green. (b) There are 21 amino acid residues of TMPRSS2 (blue dots) and 18 amino acid residues of the spike protein (pink dots) that are important for binding interactions. Interactions between each critical residue is denoted by a line connecting a spike protein residue to a TMPRSS2 residue. Lines are color coded based on their interactions. Grey

lines represent direct contact between residues, cyan lines represent salt bridges, green lines represent hydrogen bonding, and red lines represent π-cation.

figure b needs to be constructed into table

SNPs of Interest: Frequencies

We identified 18 non-synonymous SNPs of interest and their frequencies in global populations. Rs75603675 and rs12329760 displayed the highest total frequency, with an alternative allele frequency of 0.30337 and 0.224393, respectively (Table 1). Because these SNPs are more frequent, they possess potential to alter the binding interactions of TMPRSS2 in a manner that is more significant to the general population. Interestingly, Asians and Latin Americans did not present with the alternative allele of rs75603675 at all, while Europeans had the highest frequency of this SNP and African Americans had the second highest (Table 1). The differences in allele frequency among different cultural populations may underlie the variation in disease severity in the case of this SNP.

Each cultural group had similar frequencies of the alternative allele of rs12329760, with Asians having the highest frequency and Latin American 2 having the lowest frequency (Table 1). The remainder of SNPs of interest had total frequencies of the alternate allele between 0.01-0.00001, with the majority of the SNPs having frequencies between 0.0001 and 0.001 (Figure 4). Such frequencies denote that these SNPs are extremely rare in the overall population. Due to their rarity, it is difficult to conclude that these SNPs alone could be responsible for the variety in clinical presentations of SARS-CoV-2. The majority of these SNPs did not differ greatly from each cultural group, further signifying that these SNPs may not be directly responsible for the differences in severity of SARS-CoV-2 infections among different populations (Table 1). Another limitation noted is the vast differences in total sample size used to determine the

frequency of each SNP. The largest sample size contained almost 300,000 individuals (rs12329760) while the smallest sample size contained less than 18,000 (rs75603675) (Table 1). The discrepancies in the amount of individuals sampled for each allele are likely skewing the actual frequencies of each SNP. Moreover, even the highest sample size is largely under representative of the entire population which makes it difficult to conclude how prevalent these SNPs truly are. However, less common and even rare SNPs may still play an underlying role in the variation of disease symptoms when combined with other factors.

-Results from Heat Map will be added here-

Table 1. Allele Frequencies of 18 SNPs of interest, including their total frequencies as well as specific population frequencies. The first allele listed in each group is the reference allele, and the following allele is the alternative allele from the SNP. "African" includes all Africans, "Asian" include all Asians except South Asian, "Latin American 1" includes Latin Americans with Afro-Carribean ancestry, and "Latin American 2" includes Latin American individuals with mostly European and Native American Ancestry.

SNP ID	Amino Acid Change	Sample Size	European	African	Asian	Latin America 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 A=0.000412	G=0.9996 A=0.0004	G=0.9997 A=0.0003	G=1.0 A=0.0	G=1.0 A=0.0	G=0.999609 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
rs147711290	Leu81Arg	107770	A= 0.99997 C=0.0	A=0.9986 C=0.0	A=1.0 C=0.0	A=0.999 C=0.0	A=1.0 C=0.0	A= 0.999879 C= 0.000003978

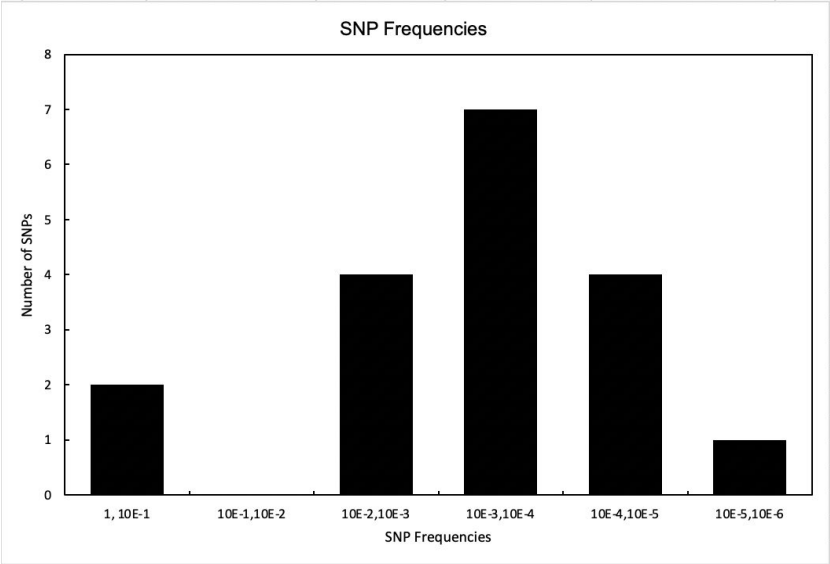


Figure 4. Relative allele frequencies of each SNP. The X-axis displays the frequency of the SNPs.. The Y-axis displays the total number of SNPs, out of the 18 SNPs of interest, that present the given frequency.

Figure 5. Heat map of allele frequencies (still needs to be constructed)

Prediction of Effects of SNPs

To predict the effects that each SNP of interest could potentially have on the stability of TMPRSS2 and subsequent binding interactions, two prediction softwares, PolyPhen-2 and SIFT, were utilized (Adzhubei et al.,2012, Sim et al., 2013). 4 SNPs (rs12329760, rs147711290 L91P, rs150554820, rs138651919) were predicted to be damaging by both softwares. 1 SNP (rs147711290 L128G) was predicted to be damaging by PolyPhen-2 but was not found in SIFT. 1 SNP (rs142446494) was predicted to be deleterious by SIFT but benign by Poly-Phen2. Each SNP predicted to be damaging by either software involved an amino acid substitution that replaced a nonpolar amino acid with a different nonpolar amino acid, so it is unlikely that polar interactions, such as salt bridging interactions, are what make these SNPs potentially damaging. Therefore, it is likely that these SNPS are potentially damaging to TMPRSS2 through changing

its overall structure and topology in ways that do not allow it to form intra-specific binding interactions that are necessary for stabilizing its structure. 12 SNPs were predicted to be benign by both softwares and 1 SNP was not found in either software (Table 2).

Table 2. Predictions of effects of SNPs on TMPRSS2 given by SIFT and PolyPhen-2. In SIFT, scores below 0.05 are marked "Deleterious" and scores above 0.05 are marked "Tolerated". Scores of 0-0.5 are marked "Benign", scores of 0.5-0.9 are marked "Possibly Damaging", and scores of 0.9-1.0 are marked "Probably Damaging" in PolyPhen-2.

SNP	Amino Acid Change	SIFT Score	SIFT Prediction	PolyPhen-2 Score	PolyPhen-2 Prediction
rs61735793	Thr112Ile	0.238	Tolerated	0.015	Benign
rs75603675	Gly8Val	0.201	Tolerated	0.167	Benign
rs61735790	His55Arg	0.231	Tolerated	0.033	Benign
rs12329760	Val197Met	0.009	Deleterious	0.937	Probably Damaging
rs200291871	Gly8Arg	0.817	Tolerated	0.011	Benign
rs61735791	Ala65Thr	0.199	Tolerated	0.029	Benign
rs148125094	Val452Ile	0.171	Tolerated	0.098	Benign
rs114363287	Gly111Arg	0.383	Tolerated	0.109	Benign
rs147711290	Leu128Gly	Not Found	-	0.92	Probably Damaging
rs147711290	Leu91Pro	0.005	Deleterious	1	Probably Damaging
rs147711290	Leu91Arg	Not Found	-	Not Found	-
rs150554820	Phe246Ile	0.004	Deleterious	0.549	Possibly Damaging
rs61735796	Glu297Lys	0.34	Tolerated	0.017	Benign
rs138651919	Pro78Leu	0.021	Deleterious	0.833	Possibly Damaging
rs61735795	Pro412Ser	0.551	Tolerated	0.086	Benign
rs142446494	Val317Met	0.015	Deleterious	0.294	Benign
rs201093031	Val70Ala	1	Tolerated	0	Benign
rs768173297	Thr346Met	Not Found	-	0.131	Benign

SNP Modelling and Docking Interactions

-results not yet constructed-

References

Adzhubei I, Jordan D.M., Sunyaev S.R. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20. Doi: 10.1002/0471142905.hg0720s76. PMID: 23315928; PMCID: PMC4480630.

Comparative Protein Structure Modeling Using MODELLER. Webb B, Sali A. Curr Protoc Protein Sci. 2016 Nov 1;86:2.9.1-2.9.37.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature protocols*, *7*(8), 1511–1522. https://doi.org/10.1038/nprot.2012.085

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, *44*(D1), D7–D19. https://doi.org/10.1093/nar/gkv1290

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, *40*(W1), W452-W457.

Thunders, M., & Delahunt, B. (2020). Gene of the month: TMPRSS2 (transmembrane serine protease 2). *Journal of clinical pathology*, *73*(12), 773-776.

Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., ... & Bonvin, A. M. J. J. (2016). The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, *428*(4), 720-725.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427

Wang, J., Youkharibache, P., Zhang, D., Lanczycki, C. J., Geer, R. C., Madej, T., Phan, L., Ward, M., Lu, S., Marchler, G. H., Wang, Y., Bryant, S. H., Geer, L. Y., & Marchler-Bauer, A. (2020). iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics (Oxford, England)*, *36*(1), 131–135. https://doi.org/10.1093/bioinformatics/btz502

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213

Zmora, P., Moldenhauer, A. S., Hofmann-Winkler, H., & Pöhlmann, S. (2015). TMPRSS2 Isoform 1 Activates Respiratory Viruses and Is Expressed in Viral Target Cells. *PloS one*, *10*(9), e0138380. https://doi.org/10.1371/journal.pone.0138380