Results

TMPRSS2 gene:

Databases such as NCBI Gene and UniProt (O15393) were utilized to discover background information of the TMPRSS2 protein. The molecular processing of the protein essential due to identification of coding regions, domains in the protein, and possible isoforms. After doing an exploratory search, TMPRSS2 is a 3250 bp gene discovered to be on the reverse strand of 21q22.3 in the human genome (41,464,305 - 41,508,158). The gene has 14 exons that will be transcribed into mRNA. Through alternative splicing, TMPRSS2 has three possible isoforms. The first two have been studied, but the third has not. Isoform 1 is 529 amino acids long, and isoform 2 is spliced after exon 1, leaving 492 amino acids. Both isoforms have been found to activate SARS-CoV and is expressed in viral cells (Zmora et al. 2015). TMPRSS2 has a domain with an unknown function (DUF3824, ~44 amino acids) and is followed by a cysteine-rich low density receptor lipoprotein class A domain (LDLa, 113-148 a.a.). In addition, TMPRSS2 has a highly conserved scavenger receptor cystine-rich domain (SRCR, 149-242 a.a.) that responds to foreign ligands, indicating an important role in molecular basis for disease (Sarrias et al. 2004). The last domain belongs to trypsin-like serine proteases (Tryp_SPc, 255-492 a.a.). Due to TMPRSS2 being a serine protease, the serine-histidine-aspartate active site is found in this domain. (H296, D345 and S441). Due to the conservation of the SRCR domain, and the catalytic triad in the Tryp_SPc domain, it was hypothesized that variants in this region could alter the structure of the protein and thus the function of TMPRSS2.

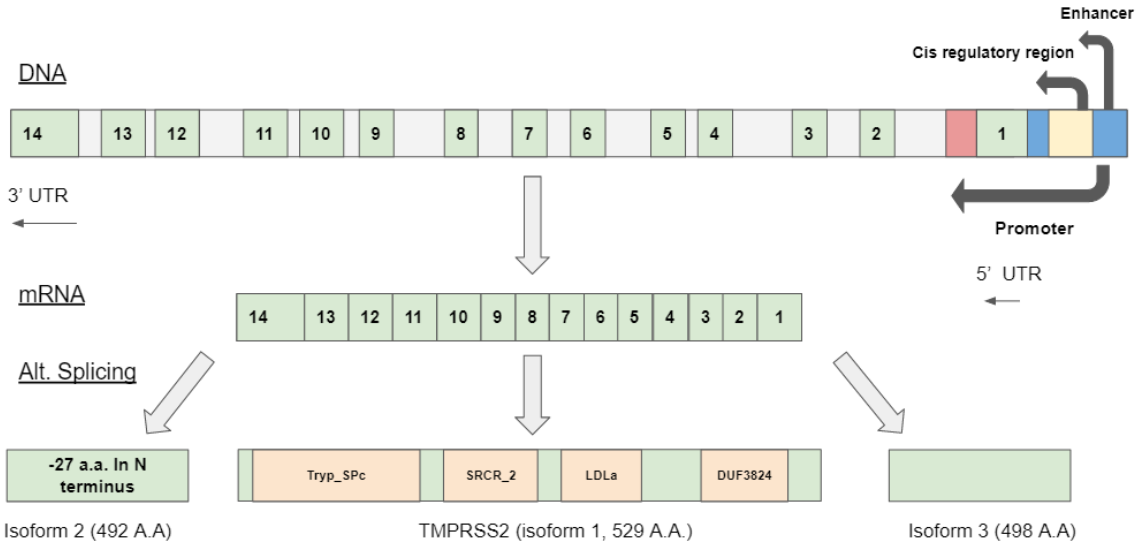


Figure 1. **TMPRSS2 gene map.** TMPRSS2 has fourteen exons, is transcribed by mRNA, and has three possible isoforms produced by alternative splicing. TMPRSS2 has four known domains with three with known functions.

Frequency data of TMPRSS2 variants:

Nonsynonymous coding region SNPs and their population frequencies were obtained from gnomAD (ENSG00000184012.7) and NCBI dbSNP, respectively. Polymorphisms in the coding region may alter the amino acid sequence of the protein, and thus alter the function of TMPRSS2. Nonsynonymous SNPs were sorted by clinical significance and allele frequency on gnomAD. Due to the rarity of nonsynonymous variants in TMPRSS2, SNPS with an allele frequency > 10^{-6} were selected to be further studied (n=17). Allele Frequency Aggregator (ALFA) was chosen to analyze population frequencies of TMPRSS2 because it is a modern project that aims increase the number of subjects with each quarterly release (Phan et al. 2020). Each SNP has a total frequency and is divided into different populations such as European, African, Asian, and Latin Americans (Table 1). The first allele listed is the reference allele, followed by the alternative allele (nucleotide change).

The frequencies of the 17 SNPs selected were visualized using a histogram (Figure 2). Most of the missense SNPs are considered to be rare, excluding two that are in more than 20% of the population. It is predictable that most of the nonsynonymous SNPs are rare in the population (less than 0.01%), or else it would result in a nonfunction proteinase.

Table 1. **Population frequency of nonsynonymous variants in TMPRSS2**

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=1.0 A=0.0	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	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=1.0 G=0.0	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=1.0 T=0.0	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=1.0 T=0.0	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=1.0 T=0.0	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	Lys309Met	44,404	G=0.99966 A=0.00034	G=1.0 A=0.0	G=1.0A=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=1.0 T=0.0	A=0.999 T=0.001	A=1.0 T=0.0	A=0.999879 T=0.000074
rs114363287	Gly111Arg	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

European= European; African= Africans and African Americans; Asian=All Asian individuals excluding South Asians; Latin American 1= Latin Americans with Afro-Caribbean ancestry; Latin American 2= Latin Americans with European and Native American Ancestry.

*SNPs only found in isoform 1.

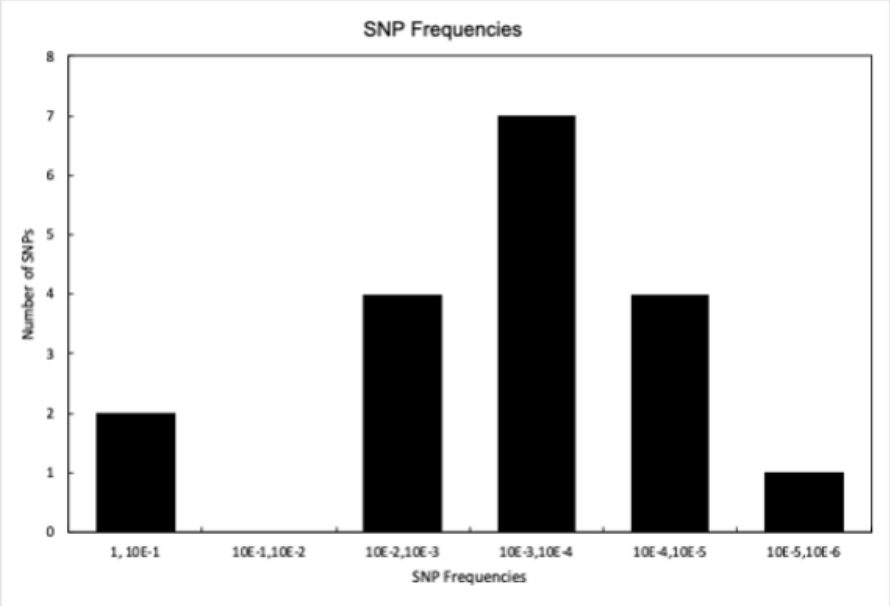


Figure 2. Minor allele frequency distribution of SNPs selected. Bins were manually made by using each power of 10 from the most to least frequent SNP (n=6 bins). The negative log of each allele frequency value was calculated to visualize better distribution.

Heat map of TMPRSS2 point mutations:

Even though the 17 SNPs were chosen by frequency and clinical significance, it is important to visualize them on a map and see their locations within the protein. Therefore, a heat map was used to determine variant hotspots within the protein, as well as to discover the location of the common nonsynonymous SNPs. PredictProtein was utilized to generalize this heat map by inputting the amino acid sequence of TMPRSS2 (Yachdav et al. 2014). Due to the more heavily studied isoform 2, the FASTA sequence for the shorter isoform was inputted into the prediction server. As seen in Figures 3-11, the overview of the gene is at top, with the highlighted yellow box being where a particular SNP is located. The heat map is scaled from -100 to 100 with the following score: vibrant red indicates a strong effect (score > 50), white indicates weak signals (-50 < score < 50), blue represents no effect (score < -50), and black squares corresponds to the wild-type residues. The effects of point mutations are predicted using SNAP2, which analyzes evolutionary information, secondary structure, and solvent

accessibility (Hecht et al. 2015). It was predicted that there will be more dark red columns towards the SRCS domain and Tryp_SPc domain due to the conservation and suggested importance of those regions. On the other hand, nonsynonymous SNPs with the least frequency should have a high score and vice versa. These heat maps are as predicted, with more frequent SNPs having little/no effect and less frequent SNPs having a stronger effect. It is also seen that in the overview heat map that hotspots are concentrated near the two conserved domains.

The only SNP that does not follow this trend is the change from valine to methionine at the 160th position (Figure 3). As seen in the heat map, any change at this amino acid position has a strong signal. Even though both amino acids are relative in terms of size and polarity, this residue is located in the conserved SRCR domain, and therefore could be potentially damaging to the function.

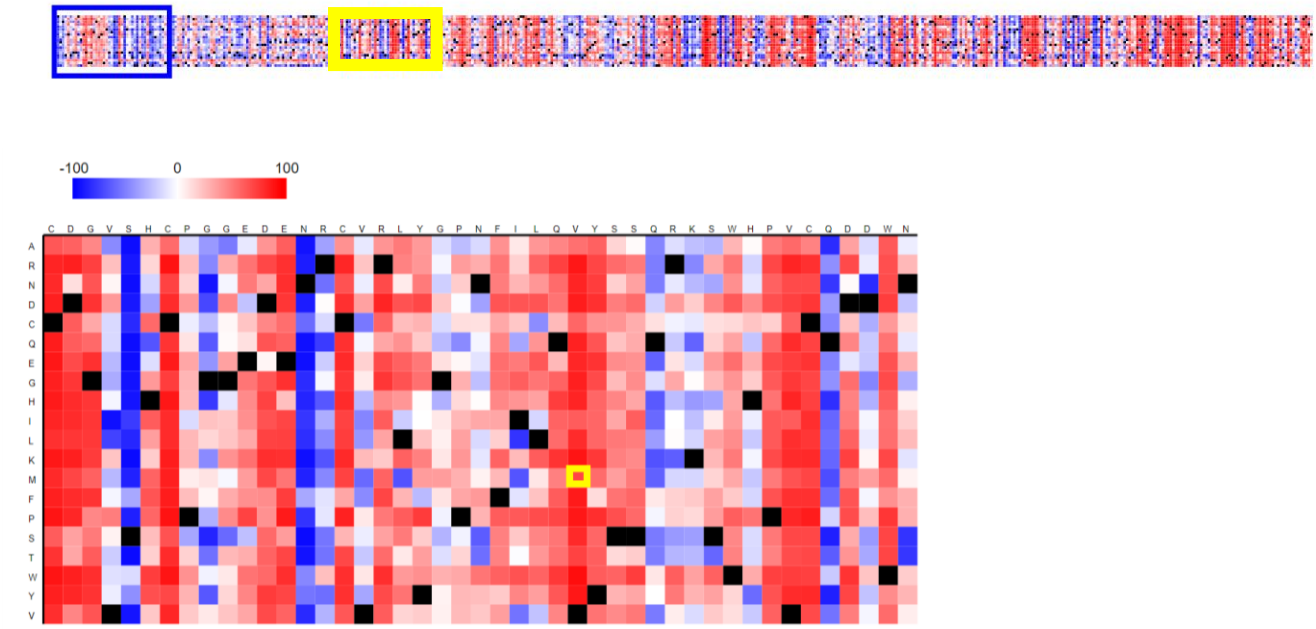


Figure 3. rs12329760 (V160M) visualized on TMPRSS2. This variant is shown to have a dark red signal, indicating that there is a strong signal for damage.

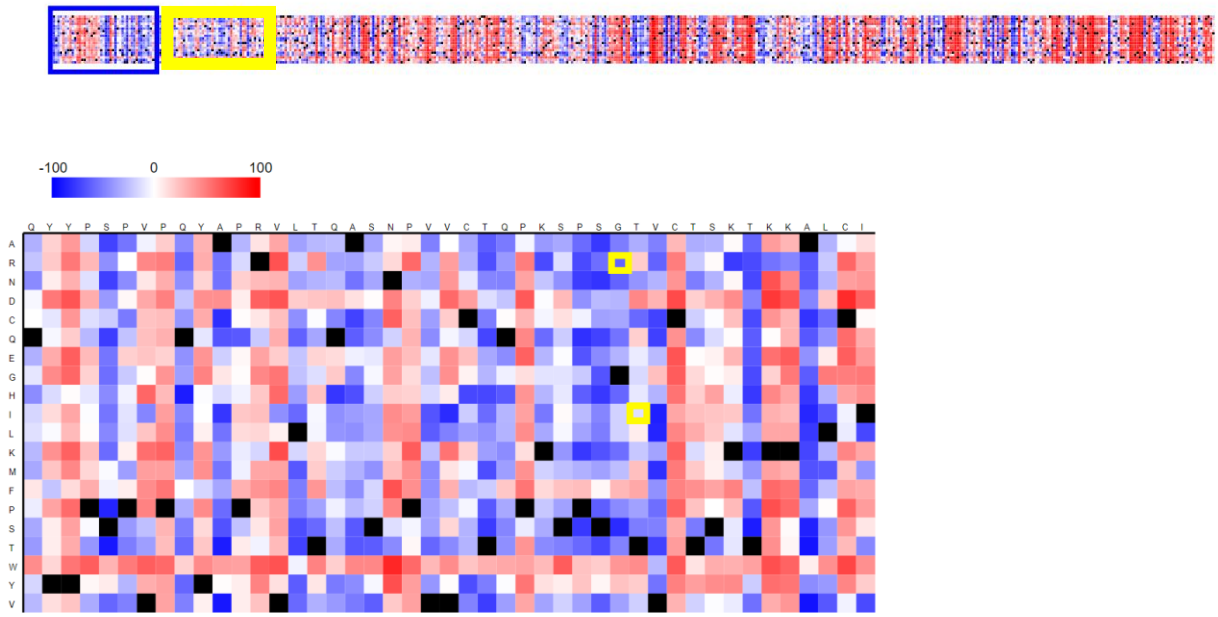


Figure 4. rs61735793 (Thr75Ile) and rs114363287 (Gly74Arg) visualized on TMPRSS2. Both variants have white/blue squares, indicating that these SNPs may not be important for TMPRSS2 function.

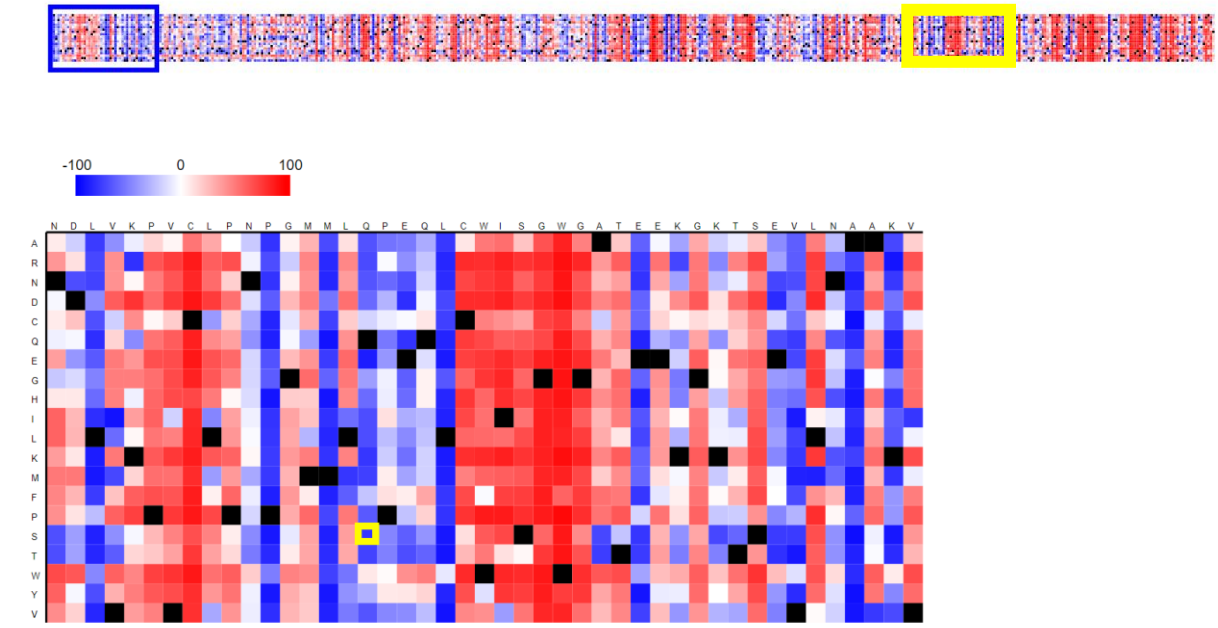


Figure 5. rs61735795 (Pro375Ser) visualized on TMPRSS2. This amino acid change has a dark blue square, indicating that it does not have a strong effect on TMPRSS2 function alone.

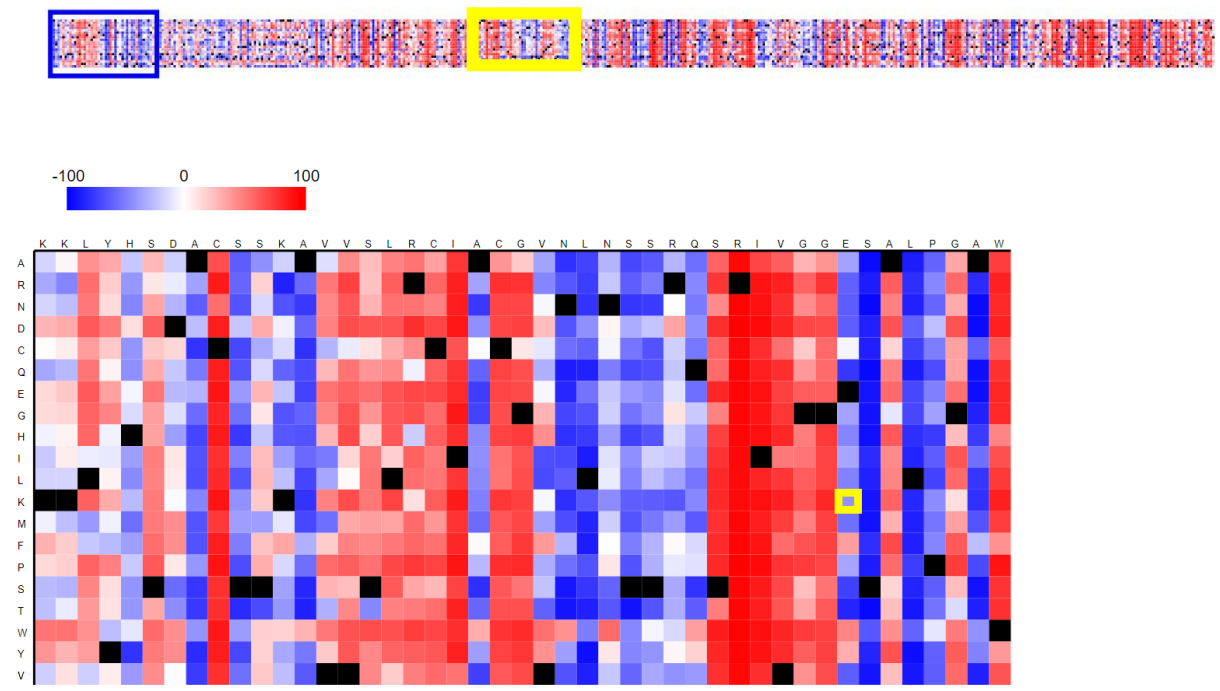


Figure 6. rs61735796 (Glu260Lys) visualized in TMPRSS2. This variant has a blue square and probably does not have an important effect on the protein.

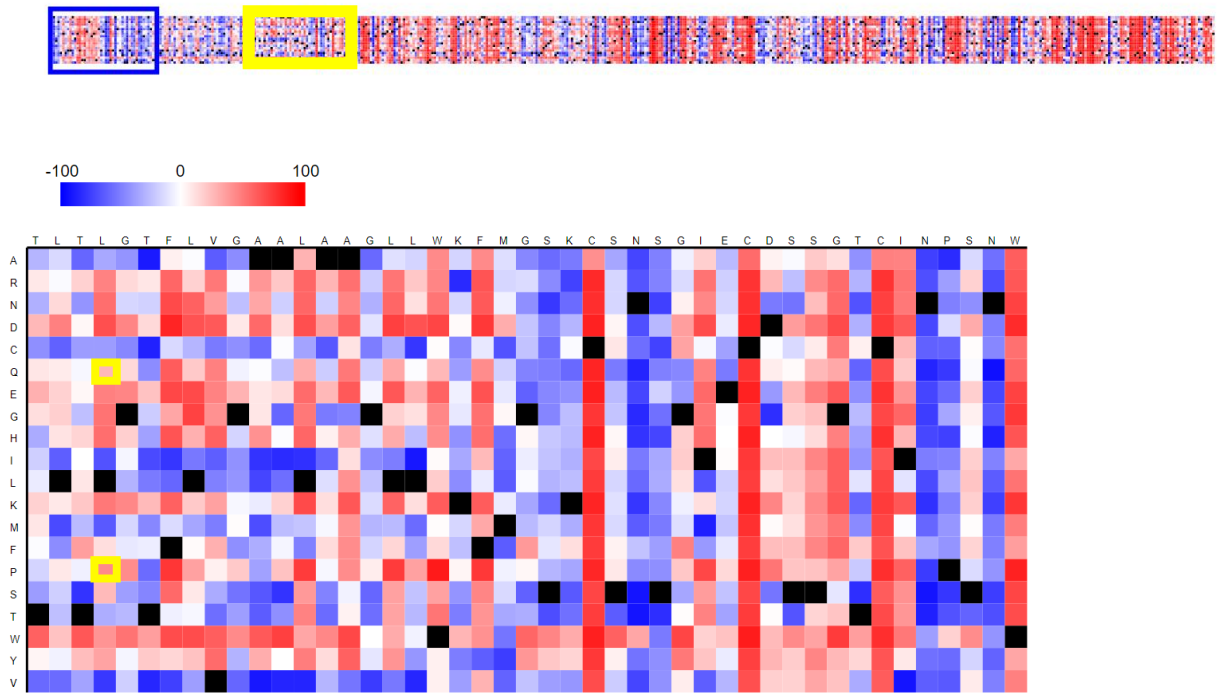
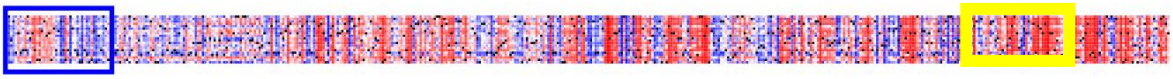


Figure 7: rs147711290 (Leu91Gln & Leu91Pro) visualized in TMPRSS2. Both of these amino acid changes have somewhat of a strong signal, with proline having a stronger signal.



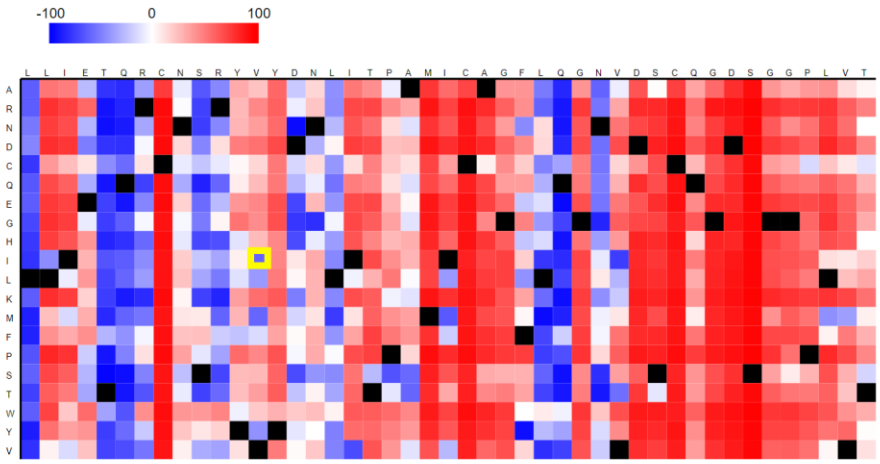


Figure 8: rs148125094 (Val415Ile) visualized in TMPRSS2. This variant has a blue signal indicating that this change may not be important.

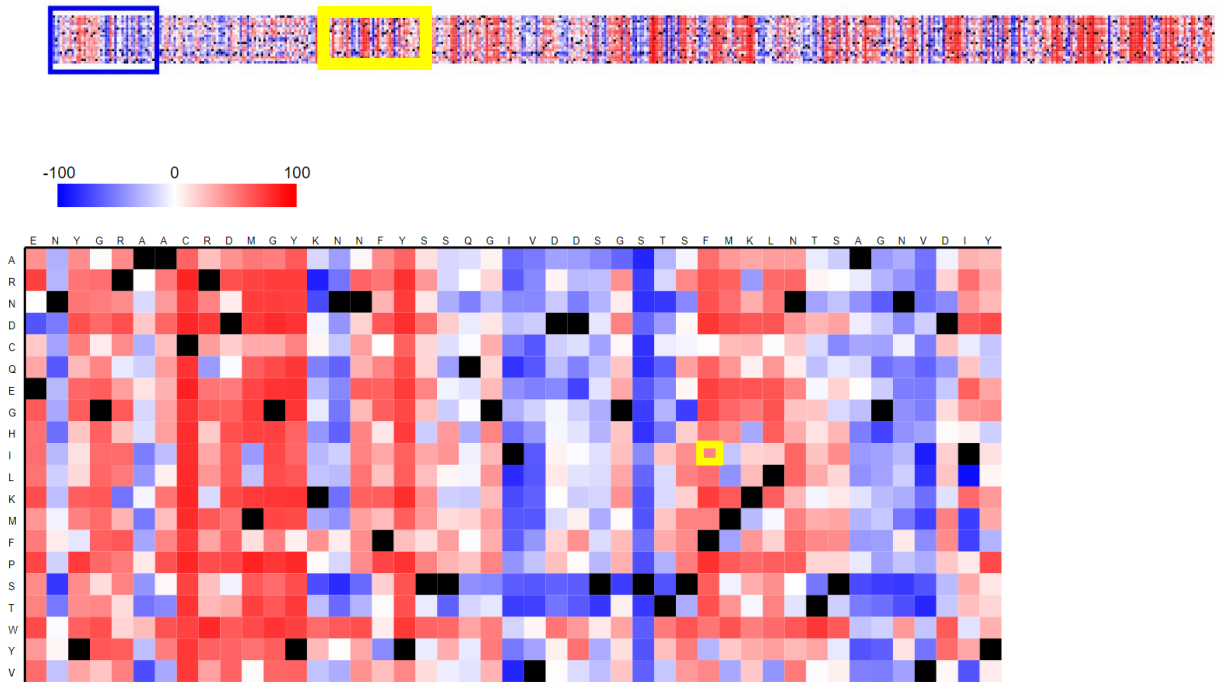


Figure 9: rs150554820 (Phe209Ile) visualized in TMPRSS2. This polymorphism has a red signal indicating that it may be lethal to the protein's function.

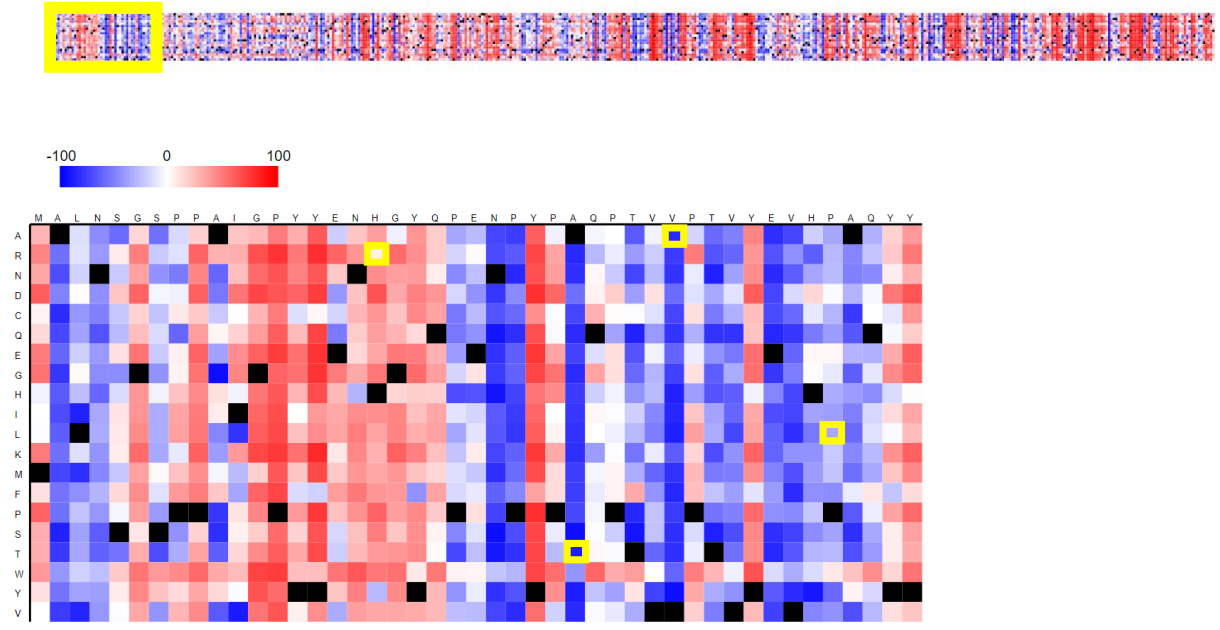


Figure 10: rs61735790 (His18Arg), rs61735791 (Ala28Thr), rs201093031 (Val33Ala), and rs138651919 (Pro41Leu) visualized in TMPRSS2. The change from histidine to arginine in the 18th amino acid is the only signal picked up as the other three are weak signals.

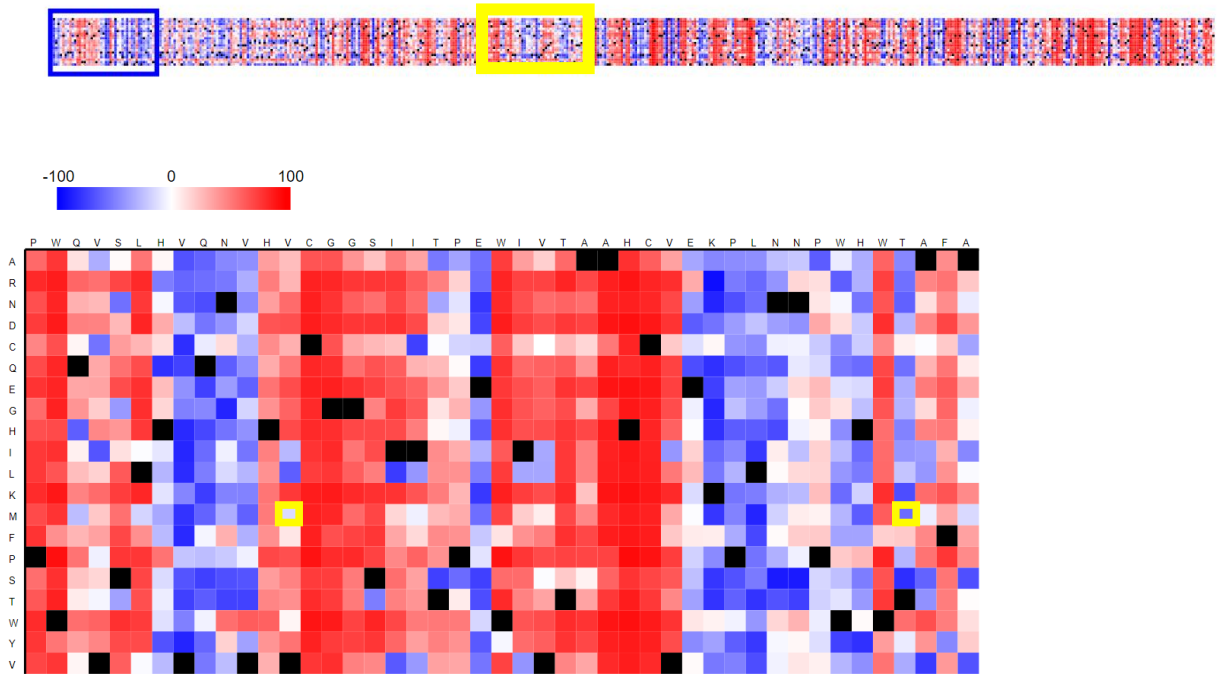


Figure 11: rs14244694 (Val280Met) and rs768173297 (Thr309Met) visualized in TMPRSS2. Both of these variants have weak signals and may not be important to TMPRSS2 function.

Molecular structures of TMPRSS2:

Currently, there is no crystallized structure of TMPRSS2 available. Therefore, multiple predict-protein servers such as HHPred (Zimmerman et al. 2018), SWISS-MODEL (Waterhouse et al. 2018), RaptorX (Kallberg et al. 2012), and I-TASSER (Roy et al. 2010) were used to visualize TMPRSS2. The structures were made by comparing the FASTA amino acid sequence to similar crystallized proteins, and therefore using it as a template to visualize TMPRSS2. For all the prediction software, hepsin (UniProtKB: P05981) was used as the foundation to model

TMPRSS2. Hepsin was most likely used as the template because they belong to the same family of peptidases (type II transmembrane serine proteases). Despite that, the sequence homology was only about ~30%. Even though this is relatively low, the SRCR conserved domain and the serine protease domain were still able to be visualized. Therefore, only the first third of the protein (domain with unknown function) was not able to be predicted.

Once the jobs were completed, the PDB files of the structures from each protein server were inputted into iCn3D web structure viewer (Wang et al. 2020). The structures generated from HHPred and SWISS-MODEL appear to have the most similarity with their globular-like shape (Figure 12,14). On the other hand, the TMPRSS2 structure made by RaptorX appears to be very elongated and thin (Figure 15). I-TASSER is globular similar to the prior two, yet it has more random coil formation (Figure 16).

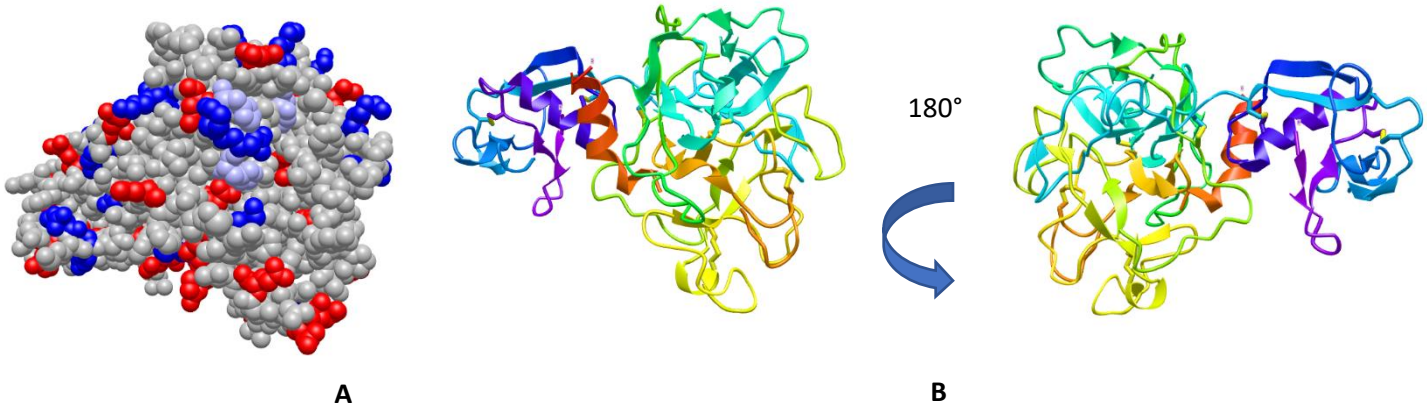


Figure 12. TMPRSS2 HHPred molecular structure. Structures were visualized in iCn3D. A) Space-filling model of TMPRSS2 colored by charge. B) Ribbon model of TMPRSS2 colored from N-terminus (violet) to C terminus(red), rotated 180 degrees on the y-axis.

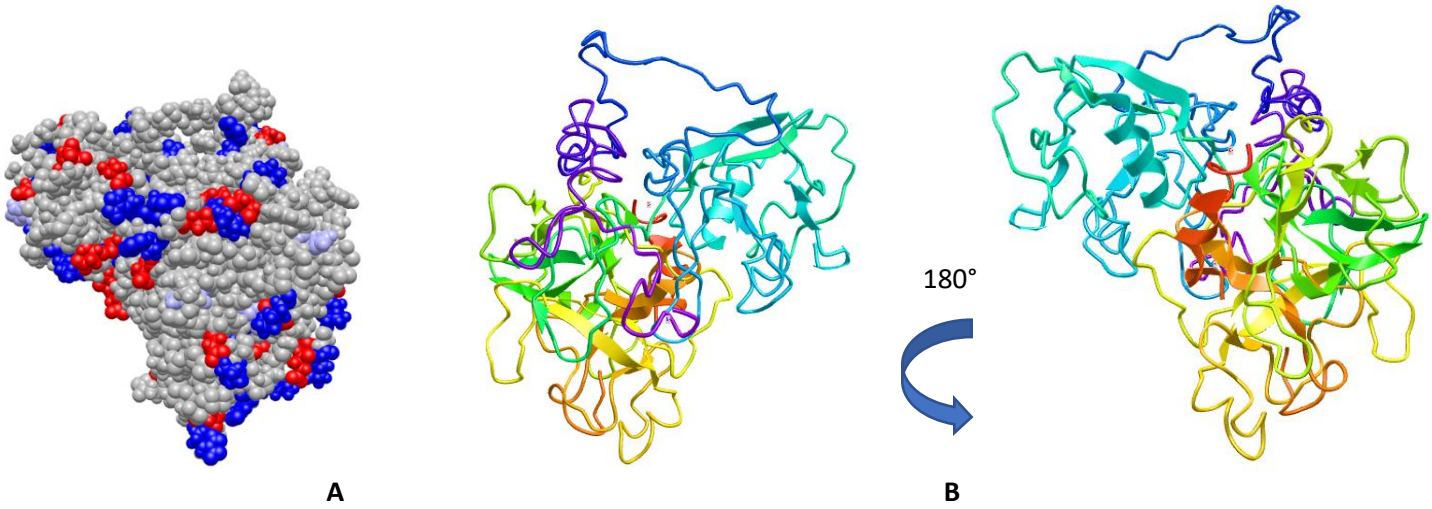


Figure 13. I-TASSER TMPRSS2 3D structure. Structures were visualized in iCn3D. A) Space-filling model of TMPRSS2 colored by charge. B) Ribbon model of TMPRSS2 colored from N-terminus (violet) to C terminus(red), rotated 180 degrees on the y-axis.

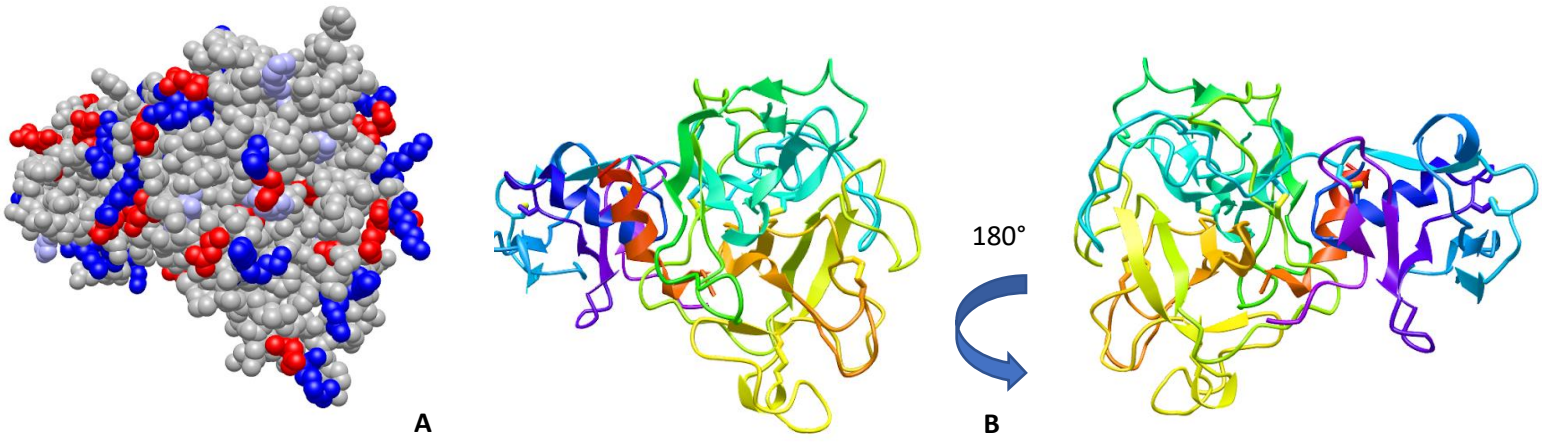


Figure 14. SWISS-MODEL structure of TMPRSS2. Structures were visualized in iCn3D. A) Space-filling model of TMPRSS2 colored by charge. B) Ribbon model of TMPRSS2 colored from N-terminus (violet) to C terminus(red), rotated 180 degrees.

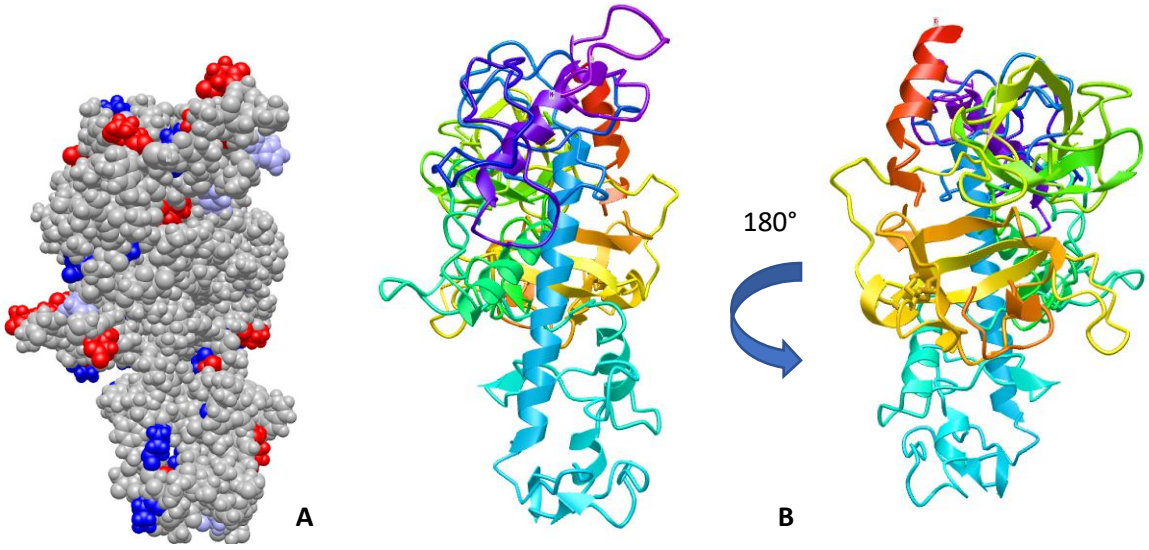


Figure 15. RaptorX generated 3D visualization of TMPRSS2. Structures were visualized in iCn3D. A) Space-filling model of TMPRSS2 colored by charge. B) Ribbon model of TMPRSS2 colored from N-terminus (violet) to C terminus(red), rotated 180 degrees.

[need to add]

Figure 16. Secondary structure sequence alignment of TMPRSS2. B= β sheet, H= α helix. The catalytic triad is shown in red.

Docking of TMPRSS2 and SARS-CoV-2 Spike Protein:

TMPRSS2 substrate and active sites were docked against the S2' cleavage site of the SARS-CoV-2 Spike Protein using HADDOCK 2.4. Viewing the complex in a structure viewer allowed for viewing of interactions between the two proteins. Amino acid changes in these interactions can therefore prevent TMPRSS2 from cleaving the Spike protein, and thus no

activation of the virus occurred. It was predicted that common nonsynonymous SNPs in these interaction regions, if any, could lead to unfavorable interactions and clashes, which will lead to less infectivity in host cells. If held true, this may explain the wide variety of severity in COVID-19 cases.

We found that 21 residues on TMPRSS2, and 18 residues on the Spike protein are important for cleavage (Table 2).

[need to add]

Figure 17. Molecular Docking of TMPRSS2 and SARS-CoV-2. PDB file obtained from HADDOCK 2.4 was visualized in iCn3D.

Table 2. TMPRSS2/SARS-CoV-2 complex predicted interactions.

[need to add]

References

Hecht, M., Bromberg, Y., & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC genomics*, *16*(8), 1-12. https://doi.org/10.1186/1471-2164-16-s8-s1

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

L. Phan, Y. Jin, H. Zhang, W. Qiang, E. Shekhtman, D. Shao, D. Revoe, R. Villamarin, E. Ivanchenko, M. Kimura, Z. Y. Wang, L. Hao, N. Sharopova, M. Bihan, A. Sturcke, M. Lee, N. Popova, W. Wu, C. Bastiani, M. Ward, J. B. Holmes, V. Lyoshin, K. Kaur, E. Moyer, M. Feolo, and B. L. Kattman. "ALFA: Allele Frequency Aggregator." *National Center for Biotechnology Information, U.S. National Library of Medicine,* 10 Mar. 2020, www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/

Protein [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. NP_001128571.1, Homo sapiens transmembrane protease serine 2 (TMPRSS2), isoform 1; [cited 2021 Feb 12]. https://www.ncbi.nlm.nih.gov/protein/NP_001128571.1?report=graph

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, *5*(4), 725–738. https://doi.org/10.1038/nprot.2010.5

Sarrias, M. R., Grønlund, J., Padilla, O., Madsen, J., Holmskov, U., & Lozano, F. (2004). The Scavenger Receptor Cysteine-Rich (SRCR) domain: an ancient and highly conserved protein module of the innate immune system. *Critical reviews in immunology*, *24*(1), 1–37. https://doi.org/10.1615/critrevimmunol.v24.i1.10

The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, D480-D489 https://doi.org/10.1093/nar/gkaa1100

Wang, J., Youkharibache, P., Zhang, D., Lanczycki, C. J., Geer, R. C., Madej, T., ... & Marchler-Bauer, A. (2020). iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics*, *36*(1), 131-135. https://doi.org/10.1093/bioinformatics/btz502

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, 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

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., ... & Rost, B. (2014). PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic acids research*, *42*(W1), W337-W343. https://dx.doi.org/10.1093%2Fnar%2Fgku366

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., ... & Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of molecular biology*, *430*(15), 2237-2243. https://doi.org/10.1016/j.jmb.2017.12.007

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