Materials and Methods

TMPRSS2 Gene Map and Model

The NCBI Gene database and the NCBI Genome Viewer were used to gather known information about TMPRSS2 structure in order to create a gene map (NCBI, 2016). TMPRSS2 (Gene ID: 7113) was searched in NCBI Gene and the Genome Data Viewer was used to visualize the location of exons and important transcriptional elements of the gene. Missense variations in each exon were visualized by setting up "Tracks" and selecting "Configure Tracks" for "Missense" variants (NCBI, 2016). Isoforms for TMPRSS2 were also identified and visualized with the Genome Data Viewer, and information about their conserved domains was discovered by searching those isoforms as a query in NCBI Gene. "TMPRSS2" was also searched in UniProt (O153930), PubMed, and Google Scholar to gain more information about the general function and structure of the protease (Bateman et al., 2020). Biorender was then utilized to convert the structural information about TMPRSS2 into a visual gene map that includes the location of exons, the chromosome location, promoter regions, enhancer regions, and regulatory regions. **(still working on bio render so don't have complete methodology*)**

Since the crystallized structure of TMPRSS2 has not yet been discovered, multiple protein prediction softwares were utilized to create 3D structures of the protease to provide greater understanding of its potential binding interactions. The FASTA sequence of isoform 1 of TMPRSS2 was retrieved from Uniprot (O153930) and directly inputted into I-TASSER, Swiss-Model, RaptorX, and HHPred. I-TASSER was utilized because it uses a highly regarded threading methodology to model proteins, and it was the only software able to predict the structure of the entire sequence of TMPRSS2 (Yang et al., 2015). Swiss-Model and HHPred were utilized for their ability to perform homology modelling, in which hepsin (33.62%

homology) was used as the template sequence for both softwares. Upon uploading the FASTA sequence of TMPRSS2, HHPred required selection of a template protein for modelling, and the one with most homology (hepsin) was selected and TMPRSS2 was then modelled using MODELLER (Comparative, 2016). Given that the homology between TMPRSS2 and hepsin was relatively low, both Swiss-Moedel and HHPred were not able to model the entire protease using this template. Raptor X was utilized in efforts to solve this problem, as it specialized in modeling proteins with distant homology, yet it was not able to model the entire sequence of TMPRSS2 (Källberg et al., 2012).

To ensure that all models of TMPRSS2 generated were sterically favorable and to determine which model should be utilized for further structural analysis with SNPs, each model was subjected to a ramachandran analysis using MolProbity (Vincent et al., 2010). The PDB files of TMPRSS2 generated by each protein prediction software were uploaded onto MolProbity and the geometry was analyzed without all atom contacts. Using the default output actions, ramachandran plots were created for each model. The plots were analyzed for how many outliers, defined as residues that were modeled in an non-allowed steric position, each model contained to determine which model generated the most sterically favored structure. In order to correct for the differences in residues modelled, since not all softwares were able to model the complete sequence of TMPRSS2, the number of outliers was divided by number of amino acids modeled to generate a more accurate percentage of accuracy of each modelling software.

Docking of TMPRSS2 and SARS-CoV-2 Spike Protein

The model of TMPRSS2 generated by I-TASSER was selected to be used for analysis of the protein's docking interaction with the SARS-CoV-2 spike protein since it was the only software able to model the entire protease. The PDB file of TMPRSS2 created by I-TASSER was

uploaded onto HADDOCK 2.4 to generate a model of molecular docking (Van Zundert et al., 2016). The catalytic serine active sites of TMPRSS2 selected were His296, Asp345, and Ser441 (Hussain et al., 2020). The substrate binding sites selected were Asp435, Ser460, and Gly462. The SARS-CoV-2 S2' cleavage site used was R815-S816 (Hussain et al., 2020). The output generated from the software was a cluster PDB file that was uploaded onto iCn3D to visualize the docking interactions (Wang et al., 2020). Interactions were viewed by selecting "View Sequences and Annotations" and "H bonds and Interactions" under Analysis.

SNPs of Interest

TMPRSS2 was searched in the NCBI dbSNP database (Gene ID: 7113) and gnomAD (Gene ID: ENSG00000184012.7) browser to identify all missense variants (323) of the protease (NCBI, 2016; Karczewski et al., 2020). To narrow down missense SNPS to those that showed potential of clinical significance in regards to SARS-CoV-2, a literature search was conducted to discover SNPs that have been studied in regards to disease development. PubMed and Google Scholar were used to locate SNPs that had been cited in literature as having any relation to disease including SARS-CoV-2, other respiratory illnesses, or other diseases. "TMPRSS2" and "TMPRSS2 SNPs" were searched in these databases to retrieve literature about TMPRSS2 SNPs. Some well-studied SNPs were identified in various studies, including rs12327690 and rs75603675, and these SNP identifiers were also searched on PubMed and Google Scholar to retrieve more relevant SNPs (Baughn et al., 2020). The relevant SNPS were searched in Clinvar to further determine potential clinical significance.

Previous studies presented limitations in the generalizability of their findings in regards to SNPs due to their rare frequency in the population, therefore this study attempted to overcome this limitation by analyzing SNPs that were more prevalent in the population (David et al., 2020).

The frequencies of each SNP was retrieved from the NCBI dbSNP database by searching each SNP identifier and recording the frequency from the ALFA Allele Frequency database (NCBI, 2016). Both the total global population frequencies and specific ethnic group population frequencies were recorded in order to analyze whether SNPs had greater prevalence in certain populations that may explain why different populations suffered differing severity of COVID-19. Only missense SNPs that had a frequency above 0.00001 were considered for analysis in this study (n=17). A histogram of SNP frequencies was constructed to analyze the relative frequencies of SNPs by grouping SNPs together that were within the same exponential group (i.e. 10E-1, 10E-2, etc.).

Prediction of Effects of SNPs of interest

PolyPhen-2 and SIFT were used to predict the effects of SNPs of interest on the structure of TMPRSS2 (Klaasen et al., 2020). PolyPhen-2 predicts if amino acid substitutions will be damaging by scoring them on a scale of 0 to 1 (Adzhubei et al., 2013). Scores between 0 and 0.5 are scored as "Benign", meaning that the amino acid substitution will likely not have a damaging effect on protein structure or function. Scores between 0.5 and 0.9 are scored as "Possibly Damaging", meaning that the amino acid substitution has potential to damage the protein's structure or function. Scores between 0.9 and 1.0 are scored as "Probably Damaging", meaning that the amino acid substitution is likely damaging to the protein's structure or function (Adzhubei et al., 2013). The FASTA sequence of TMPRSS2 retrieved from Uniprot (O153930) was inputted into Poly-Phen2, and the residue substitution was manually inputted by selecting the original residue present at the specific site as well as the variant residue. SIFT predicts if amino acid substitutions will be damaging by scoring them on a scale to 0 to 1 (Sim et al., 2012). Scores below 0.05 are marked "Deleterious", meaning that the amino acid substitution is likely

damaging to the protein's structure or function. Scores above 0.05 are marked “Tolerated”, meaning that the amino acid substitution is not likely damaging to the protein’s structure or function (Sim et al., 2012). SNP identifiers were directly inputted into SIFT to retrieve scores and corresponding markings. These predictions were used in conjunction with 3D modeling of each SNP to analyze their potential to affect the binding interactions of TMPRSS2 with SARS-CoV-2.

Modeling and Analysis of TMPRSS2 SNPs

FASTA sequences of each SNP were generated by inputting the variant residue of each SNP onto the Uniprot FASTA sequence of TMPRSS2. Then, all SNP FASTA sequences were uploaded onto Phylogeny.fr to perform a multiple sequence alignment (MUSCLE format) which was then converted into CLUSTAL format (Dereeper et al., 2008). This sequence alignment allowed for identification of conserved regions among TMPRSS2 SNPS that may provide insight into which domains a SNP may be more damaging in. To generate visual models of each TMPRSS2 SNP for further structural and binding analysis, the PDB file of the docking interaction between TMPRSS2 and SARS-CoV-2 Spike protein generated by HADDOCK 2.4 was uploaded onto Chimera (Pettersen et al., 2004) . The TMPRSS2 strand was then mutated for each of the SNPs using Chimera. To model each SNP of TMPRSS2, “Favorites” then “Command Line” was first selected on the software. In the command line, the residue position of each SNP was typed in and selected on the model by selecting “Actions”, “Atoms/Bonds”, and “Show”. Then, to mutate the residue to the variant residue, “Tools”, “Structure Editing”, and “Rotamers” was selected and the variant residue was inputted (Pettersen et al., 2004) . The highest probability romater was chosen and it was selected to model both the new variant residue while retaining the original residue. Steric analysis was then performed on the modeled variant

residue to determine if there was clashing. The residue was selected and the “Find Clashes and Contacts” feature was used by designating to check the residue against “all other atoms” and using default parameters. Under “Treatment of Clash/Contact Atoms”, the boxes next to “Select”, “Color”, “Draw pseudo bonds of…”, and “Write information to reply..” were checked and the program was run to complete the steric analysis of each SNP (Pettersen et al., 2004). The results of the steric analysis were utilized to determine whether the variant residue presented clashes with other atoms on TMPRSS2 or the SARS-CoV-2 spike protein that could ultimately hinder the binding ability of TMPRSS2 to affect its function as a protease.

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.

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., ... & UniProt Consortium. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*.

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

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

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.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J. F., Guindon, S., Lefort, V., Lescot, M., Claverie, J. M., & Gascuel, O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, *36*(Web Server issue), W465–W469. https://doi.org/10.1093/nar/gkn180

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.

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

Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

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

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

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.

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.

Vincent B. Chen, W. Bryan Arendall III, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson and David C. Richardson (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica D66: 12-21.

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

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

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