Materials and Method

SNP Frequencies

Research on TMPRSS2 related SNPs and their frequencies were gathered from NCBI, ClinVar, and cited literature. The search bar was used to search TMPRSS2 on both NCBI dbSNP and ClinVar. Frequencies were obtained from dbSNP and sorted "by-ALFA" and "missense." We gathered information from the first 18 SNPs from this search query. Relevant information that was selected from the SNPs included the SNP rs number, sample sizes, and the population frequencies from the following groups: European, African, Asian, Latin American 1, Latin American 2, and the Total population.

SNP Predictions

PolyPhen-2 and SIFT were used to predict the outcomes of amino acid substitutions to the overall protein structure. For PolyPhen-2, TMPRSS2 SNPs of interest were submitted to the server. The rs number was used as the protein or SNP identifier, then the FASTA protein sequence was input, along with the position number, and the amino acid substitutions. There is an option to add a query description, and then the job was submitted by clicking "Submit Query" (Adzhubei et al., 2010). A score ranging from 0.1 was provided which rates the substitution as benign, possibly damaging, and probably damaging.

SIFT uses the dbSNP ID from NCBI to predict the potential tolerance of an amino acid substitution. All the SNPs of interest were entered one by one in the "Paste in rs id's" query and "Submit" (Vaser et al., 2016). Scores provided rate the SNP as either tolerated or deleterious.

Predict Protein provided a heatmap of the potential positive and negative effects of amino acid substitutions. The TMPRSS2 FASTA sequence was uploaded to the query and "PredictProtein" was selected to generate results (Bernhofer et al., 2021).

BioRender Gene Map (Update when complete)

A gene map of TMPRSS2 was created using BioRender.com. Features included in the gene map include the DNA exons and introns, regulatory regions, UTR sites, mRNA, isoforms, and domains: domain of unknown function (DUF3824), Low Density Lipoprotein Receptor Class A (LDLa), Scavenger Receptor Cysteine Rich Domain (SRCR_2) and Trypsin-like serine protease (Tyrp_SPc).

TMPRSS2 Structure Prediction

Four protein prediction softwares were used to create the structure of TMPRSS2. I-TASSER is a protein modeling resource generated by Zhang Lab that uses threading to predict secondary and 3D models of proteins. The FASTA sequence was inputted in the field "I-TASSER On-line Server" in the empty text box. Once added, we provided an email, password, and optional name, then clicked "Run I-TASSER". The program will predict the secondary structure, hydrophobic and hydrophilic residues and 3D models (Yang, Roy, Xu, & Zhang, 2015) (Roy, Kucukural, Zhang, 2010) (Zhang, 2008).

SwissModel uses homology modelling and already modeled TMPRSS2 using Uniprot:O15393 (Waterhouse et al., 2018). For the purposes of this research, the FASTA sequence for TMPRSS2 was pasted into the text box after clicking "Start Modelling." We provided a project title and email, then clicked "Build Model."

Raptor X was developed by the Xu lab and can predict secondary and tertiary protein structures, contacts, solvent accessibility, disordered regions, and binding sites. We registered with an email address, then from the main page clicked "Submit" under "RaptorX Software Prediction" then provided a job name and email address. The TMPRSS2 FASTA sequence was inputted and the job was submitted. Wait time was around 2-3 days (Wang, Li, Liu, & Xu, 2016).

HHpred has the ability to detect protein homology and predict protein secondary and tertiary structure. HHpred uses data from multiple databases such as PDB, SCOP, Pfam,

SMART, COGs, and CDD. To generate protein structures, the FASTA sequence for TMPRSS2 was inputted in the "Input" field, and then "Submit" was clicked. Once the results were generated, we selected the templates that will be used to visualize the protein and then clicked "Create Model Using Selection." A PIR file was created and then inputted into MODELLER software under "3ary structure." The MODELLER license key was inserted to "Custom Job ID" and then we clicked "Submit" to generate the results (Zimmerman et al., 2018) (Gabler et al., 2020).

Docking Model

To visualize the docking of TMPRSS2 and SARS-CoV-2, HADDOCK 2.4 was used. HADDOCK 2.4 is a server that was created by BonvinLab that models the interaction between two molecular structures and their fit. To begin, we registered and created an account with their server, then clicked "Submit a New Job," and completed their requirements, such as job name and type of structures, and then clicked "Next." The TMPRSS2 structure generated from I-TASSER in PDB format and the SARS-CoV-2 S protein PDB:7DK23 was used. "Protein-Protein Ligand Docking" was specified for both. In the "Active Residues" field, the active residues gathered from Hussain 2020 were added: His296, Asp345, Ser441, Asp435, Ser460, Gly642. A PDB file of the proteins docked is created as the result (Van Zundert et al., 2016).

View Interactions

Interactions were viewed and located using iCN3D. The PDB file generated from HADDOCK.24 was uploaded to the iCN3D software by clicking "File" "Select a File" and "PDB file." The interaction sites were located by "Analysis" "View Sequence" "Interactions" then in "Details" the relevant SNPs can be highlighted in the bottom bar. Additionally, through

"Analysis" then "H-bonds and interactions," a list of the interaction residues can be displayed in a table or viewed in a map format under "Interaction Network."

Assessment of TMPRSS2 Model

Ramachandran plots were generated by using MolProbity (Williams et al., 2018). Ramachandran plots analyze which secondary structure of proteins are favored sterically. It provides information about poor and favored rotamors, number of outliers, and an overall clash score. PDB files were uploaded from each modeling software by selecting "Choose File" and "Upload." The results are generated by selecting "Analyze the geometry without all-atom contacts."

Multiple Sequence Alignment

To generate a multiple sequence alignment for each of the SNPs, the FASTA file for TMPRSS2 was copied. This FASTA sequence was pasted into a .txt application such as TextEdit or Notepad. Each sequence was manipulated by hand by copying the original TMPRSS2 FASTA sequence and then deleting the original amino acid and replacing it with the mutation.

The completed text was then uploaded to Phylogeny.fr (Dereeper et al., 2008) under "One-Click," then the text was pasted in the box, and "Submit" was selected. Under "Alignment," and "Outputs" the alignment in Clustal format was obtained.

Visualizing SNP Mutations

SNPs were visualized on the structure of the docked model by uploading to Chimera (Pettersen et al., 2004). To mutate a residue, the chain that highlights TMPRSS2 was selected. Then under "Favorites" and "Command line," the specific amino acid residue was typed in the command line query as the command "select:123" where 123 is the amino acid number. The residue was turned into a stick model by selecting "Actions" "Atoms/Bonds" "Show" was selected. Then, mutation of the residue was performed through "Tools" "Structure Editing" and

“Rotamers” and in the selection box, the new amino acid was selected in the rotamer type field. Clicking “retain” kept both the original amino acid and the mutation on both the same residue.

Clashes and hindrance were observed by performing a steric analysis. This was done by clicking “control” and left-clicking on the residue. Then“Tools” “Surface Binding Analysis” and “Find Clashes and Contacts” were selected and in the pop-up menu “Desigmate” and “all other atoms” were selected, and no other default parameters were changed. “Treatment of Clash/Contact Atoms” was selected to ensure that “Select” “Color” “Draw pseudobonds of…” and “Write information to reply” are checked off, then results were displayed through “Apply.”

ConSurf (Update when complete)

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods, 7(4), 248-249.

Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., ... & Rost, B. (2021). PredictProtein-Predicting Protein Structure and Function for 29 Years. bioRxiv.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., ... & Gascuel, O. (2008). Phylogeny. fr: robust phylogenetic analysis for the non-specialist. Nucleic acids research, 36(suppl_2), W465-W469.

Gabler, F., Nam, S. Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., ... & Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Current Protocols in Bioinformatics, 72(1), e108.

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.

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.

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.

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.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. Nature protocols, 11(1), 1.

Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. Nucleic acids research, 44(W1), W430-W435.

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.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., ... & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all‐atom structure validation. Protein Science, 27(1), 293-315.

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.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC bioinformatics, 9(1), 1-8.

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.