Madeleine King

30 March 2021

MATERIALS AND METHODS

Background on TMPRSS2

TMPRSS2 was further researched using biological databases such as UniProt (O15393) and NCBI Gene (7113). UniProt displayed gene expression, the possible products of TMPRSS2 by alternative splicing, catalytic triad active site, and domains of the protease (The Uniprot Consortium, 2021). NCBI Gene also had the number of exons, protein products, domains, and function (Gene, 1988). The full genomic sequence was viewed using the "Genome Data Viewer" in the NCBI Gene database under the "Genome Browsers" header. Missense variations were viewed in the genome by selecting "Tracks", followed by "Configure Tracks" and by configuring the "Missense" option. Isoforms were also visualized by selecting "Genomic regions, transcripts, and products", then "Switch ON mode 'show All' for Gene tracks" in the toolbar.

Gene Map of TMPRSS2

work in progress

Exploratory Search for TMPRSS2 SNPs

Baughn (2020) mentions two common TMPRSS2 SNPS whose frequencies differ in ancestry as well as population. Therefore, these SNPs (rs12327690 and rs75603675) were used as a guide to find more SNPs that may be crucial to TMPRSS2 structure and function. PubMed database (NCBI Resource Coordinators, 2016) was used for discovering relevant papers by

search words such as: “rs12327690”, “rs75603675”, “TMPRSS2 SNPs” and “TMPRSS2 SNPs SARS-CoV-2”.

Several databases such as ClinVar, gnomAD, and dbSNP were also used to search for relevant TMPRSS2 SNPs. The search query “TMPRSS2[Gene]” was used on ClinVar to find SNPs that may have been already discovered in clinical studies related to human disease (Landrum, 2018). The most frequent SNPs were found using the gnomAD database (Karczewski, 2020). “TMPRSS2” was used as a search query, and the database contained all the known variants, and the option to sort them by various filters (ID: ENSG00000184012.7). To view possibly lethal SNPs, “Synonymous” and “Other” SNPs were deselected so only “Missense/Inframe indel” and “pLoF” remained. Furthermore, to view most common SNPs, the “Allele Frequency” filter was selected, which displayed them from highest to lowest frequency. Due to the rarity of nonsynonymous variants in TMPRSS2, SNPS with an allele frequency $> 10^{-6}$ were selected to be further studied (n=17). Each gnomAD variant’s referenced dbSNP number used to label each SNP due to homogeneity. Each variant’s dbSNP number was inputted into the dbSNP database to determine minor allele frequency (Sherry, 2001). Allele Frequency Aggregator (ALFA) was used to analyze population frequencies of TMPRSS2 because it is a modern project that aims increase the number of subjects with each quarterly release (Phan, 2020). The ALFA for each TMPRSS2 variant was obtained under the “Frequency” tab, which was broken down into populations.

All the total ALFA frequencies were visualized using a histogram to analyze TMPRSS2 common variants and their appearances in the population. Due to the relatively small values, the antilogarithm was taken of the frequencies for better visualization.

Predicting Severity of TMPRSS2 SNPs

Both PolyPhen-2 and SIFT were used to predict effects of TMPRSS2 SNPs. PolyPhen-2 predicts whether specific amino acid changes are benign or non-benign to the structure and function of a human protein (Adzhubei, 2010). Similarly, SIFT predicts effects of SNPs that are present in the NCBI dbSNP database (Sim, 2012). SNP ID for each variant (n=17) were inputted into the "Protein or SNP identifier" box on Poly-Phen2. TMPRSS2 isoform 1 (O15393) amino acid sequence was inputted into the " Protein Sequence in FASTA format" box, followed by the position, wild type amino acid, and the change. For SIFT, each of the dbSNP IDs were inputted at once and submitted.

A heat map using PredictProtein was used to visualize SNPS and determine hotspots of variants in the TMPRSS2 protein. FASTA format of TMPRSS2 isoform 1 (O15393) was inputted into the textbox. Heat map was visualized in the "Effect of Point Mutations" tab under the "Function Annotation" header. The nonsynonymous SNPs were located in the protein by using the slider.

Modeling TMPRSS2 using Protein Prediction Software

Since no crystal structure is available, several protein prediction servers: SWISS-MODEL, HHPred, RaptorX, I-TASSER, and PredMP were used to model the 3D structure of TMPRSS2. SWISS-MODEL uses homology modeling in order to predict tertiary structures of protein and predict functionality and various sequence features (Waterhouse, 2018). Isoform 1 of TMPRSS2 sequence was inputting into "Target Sequence" input box. Serine protease hepsin (1z8g.1) was used as a template, with 33.62% homology and 0.55 coverage. 3D Model was saved as a PDB file. Similar to SWISS-MODEL, HHPred uses sequence similarity and

homology-based modeling to visualize proteins (Zimmerman, 2018). FASTA sequence was inputted into “Input” field, followed by selecting hepsin as a template, and then “Create Model Using Selection” was clicked. The PIR file created was then inputted into “MODELLER” under the “3ary structure” tab. Structure was saved as a PDB file.

RaptorX is a structure prediction server that can predict secondary and tertiary protein structure, disordered regions, and binding sites (Källberg, 2012). The FASTA sequence of TMPRSS2 isoform 1 was inputted into the “Sequence” box. Five predicted models as well as a contact probability matrix was generated. The first 3D model was saved as a PDB file. I-TASSER is an award-winning software that uses threading to predict protein secondary and 3D models (Roy, 2010). Isoform 1 amino acid sequence of TMPRSS2 was inputted into text box with default parameters. The most accurate of the five final models (model 1) was saved a PDB file.

Molecular Docking of TMPRSS2 and SARS-CoV-2

Docking between TMPRSS2 and SARS-CoV-2 was performed by HADDOCK 2.4 (Van Zubert, 2016). For the TMPRSS2 structure, the PDB file generated by I-TASSER was inputted under “Choose File”. For type of molecule docking, “Protein or Protein-Ligand” was selected. SARS-CoV-2 S-trimer, open conformation (ID:7DK3) was used to dock against the protease. Chain A was selected as the tertiary structure of the spike protein is a homotrimer. Once again, “Protein or Protein-Ligand” was selected for the class of molecule. Active residues for TMPRSS2 were inputted as previously described (Hussain, 2020). The catalytic triad (His296, Asp345, Ser441) as well as the substrate binding sites (Asp435, Ser460, Gly462) were inputted as “Active residues” for TMPRSS2. The S2’ cleavage site (Arg815 and Ser816) was inputted as

"Active residues" for SARS-CoV-2 S protein. Default settings were used for remaining options, and cluster was saved as a PDB file.

TMPRSS2/SARS-CoV-2 was visualized using iCn3D (Wang, 2020) by opening the PDB file from HADDOCK 2.4. Interactions were viewed under the "Analysis" tab, followed by "Sequences and Annotations", and checking the "Interactions" box. Residues on the interactions track were recorded.

Visualizing Favorable Residues using Ramachandran Plots

Predicted protein structures attained from SWISS-MODEL, HHPred, RaptorX, and I-TASSER were assessed by MolProbity (Williams, 2018). PDB files obtained from each software were uploaded by selecting "Choose File". After the run finished, "Analyze geometry without all-atom contacts" was selected. Ramachandran plots were downloaded as PDF files, and summary statistics are shown in table.

Mapping SNPs on TMPRSS2/SARS-CoV-2 Complex

SNPs were visualized on TMPRSS2 interacting with SARS-CoV-2 using Chimera (Petterson, 2004). TMPPRS2/SARS-CoV-2 PDB file obtained from HADDOCK 2.4 was inputted into Chimera by opening "File" then "Open". Amino acid sequence was displayed by "Tools", "Sequence", "Sequence", then TMPRSS2 was selected (chain B in this case). The desired amino acid was selected by 'dragging and clicking' on the residue in the sequence. "Atoms/Bonds" was selected under the "Action" tab, followed by "Show" to further accentuate the residue as a stick. The residue was mutated to represent the SNP by "Tools", "Structure Editing" then "Rotamers". The highest probability rotamer was selected, the "Existing side chain(s)" was changed to "retain" so both the wild type and mutated residue can be seen

simultaneously. To differentiate, the wild-type residue's color was changed by clicking under the "Actions" tab.

Steric hindrance analysis were determined as previous described (NIAID Bioinformatics, 2013). Mutated residue was selected, then the up arrow was used to select entire residue. Under the "Tools" tab, "Surface Binding Analysis" was selected, then "Find Clashes and Contacts". For the external window that appeared, "Designate" was selected, then "All other atoms" were checked. The following boxes were checked under "Treatment of Clash/Contact Atoms": "Select", "Color", "Draw pseudo bonds of…", and "Write information to reply". To perform the analysis, select "Apply". Pseudo bonds between classing atoms were shown and highlighted in yellow.

Multiple Sequence Alignment of TMPRSS2 SNPs

[work in progress]

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

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

Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Gene ID: 7113, Homo sapiens transmembrane serine protease 2 (TMPRSS2), DNA; [cited 2021 Mar 25]. Available from: https://www.ncbi.nlm.nih.gov/gene/7113

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. https://doi.org/10.3934/microbiol.2020021

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., Cummings, B. B., Alföldi, J., Wang, Q., ... & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... & Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, *46*(D1), D1062-D1067. https://doi.org/10.1093/nar/gkx1153

NIAID Bioinformatics (2013 May, 3). *Mutating a Residue in UCSF Chimera (Part 1)* [Video]. Youtube. https://youtu.be/bcXMexN6hjY

NIAID Bioinformatics (2013 May, 3). *Mutating a Residue in UCSF Chimera (Part 2)* [Video]. Youtube. https://youtu.be/eJkrvr-xeXY

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/gkx1095

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

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, *29*(1), 308-311. https://dx.doi.org/10.1093%2Fnar%2F29.1.308

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. https://doi.org/10.1093/nar/gks539

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

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. https://doi.org/10.1016/j.jmb.2015.09.014

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

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. https://doi.org/10.1002/pro.3330

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