Madeleine King

26 February 2021

Discussion:

The sequence-structure-function relationship of a protein is essential to understanding its role in the biological system as well as being a potential target for therapeutic use. As COVID-19 is a highly infectious disease with a range of cases from asymptomatic to severe, it is difficult to determine how to interfere with viral spread and pathogenicity. TMPRSS2 is the membrane protein that cleaves the S protein at its S2' site (Bestle, 2020) and therefore natural variants present in different populations could be a potential explanation for the range of cases. In this study, we asked whether nonsynonymous variants in human TMPRSS2 lead to differences in SARS-CoV-2 spread and severity.

A gene map was made of TMPRSS2 to learn more about the evolution of the protein and functions (Figure 1). Through research, it was found that it contains a conserved scavenger-receptor cysteine-rich domain (SRCR) and close by downstream is the trypsin-like serine protease domain. Even though the SRCR domain has not been well studied yet, its interest is growing exponentially. It has been discussed that it may be related to innate immunity, tumor development, and stem cell biology and that is a very ancient sequence found in only eukaryotes (Resnick, 1994). Likewise, the serine protease domain is superfamily domain that contains the Ser-His-Asp catalytic triad that form the active site of the protein. In particular, trypsin-like serine protease domains cleave a site following a positive amino acid like lysine or arginine. This is evident as TMPRSS2 cleaves the S2' site (Arg816) for exposure of the fusion peptide on the spike protein (Walls 2020, Bestle 2020). Therefore, we predicted that nonsynonymous variants in these regions may have an affect on protein structure and function.

Through the exploratory research, TMPRSS2 was discovered to have three different isoforms. Even though isoform 1 of TMPRSS2 is the unspliced sequence, isoform 2 is cited

more frequently and appears to have been more studied. This may be due to the fact isoform 2 is spread in lung-tissues: the reservoir for betacoronaviruses. Therefore, in this study the isoform 2 was used for protein modeling and docking. However, Zmora (2015) discovered that TMPRSS2 isoform 1 also activates respiratory viruses in lung-derived cells. Therefore, further research is needed to determine which isoform, or both, would be the best target for understanding COVIID-19 susceptibility and severity.

Common nonsynonymous variants of TMPRSS2 were chosen to be further studied and compared by their population frequencies (Table 1). Even though ALFA is a modern project aiming to obtain a large database collection of samples, their methods reveal flaws that need to be improved on. As seen in Table 1, the sample size is varied for each SNP, ranging from ~17,000 to 300,000. Because of this discrepancy, it is difficult to classify variants as common, rare, or neutral in the population due to the range of sampling in the different genome projects. Likewise, sample sizes between different populations may lead to inaccurate results. For example, even though the sample size for V160M is large, 90% of the samples were from European population category. Therefore, discrepancies in the sample sizes not only between SNPs but also populations indicate that statistical analysis for these results may be inaccurate.

SNPs that were chosen for analysis were plotted into an allele frequency graph to determine the distribution of frequencies in the TMPRSS2 gene (Figure 2). It is shown that TMPRSS2 ‘common’ SNPs are not very common at all. Only two variants (Gly8Arg, Val160Met) have a population frequency greater than 20%, and the rest equal or below 1% frequency. This agrees with previous studies, as the recommended frequency cutoff for benign SNPs in TMPRSS2 is 0.0001. Therefore, due to their rarity, most studies do not believe that coding variants in TMPRSS2 can lead to changes in pathogenicity or severity (Russo 2020). However, Baughn (2020) states that the two common SNPs (Gly8Arg and Val160Met) are important if TMPRSS2 was to be used as a potential candidate for therapeutic use, since they are fairly common in the population and vary between different ethnic groups. The Gly8Arg variant is only

found in isoform 1 and not belonging to a conserved domain that is known. Therefore, it is expected that this amino acid change is common in the population. On the other hand, Val160Met is located in the SRCR domain, so this amino acid may be considered lethal and disrupt the function of the protein. Val160Met has been previous determined to be associated with higher risk of prostate cancer (Mollica 2020), so it is reasonable to predict that it may interfere with the TMPRSS2-SARS-CoV-2 binding complex as well.

[discuss SIFT and Poly-Phen2 results]

The effects of point mutations in TMPRSS2 were visualized on a heatmap to determine hotspots (if any) of common nonsynonymous variants (Figures 3-11). It was predicted that more common frequency SNPs would be in less conserved domains since they would have a lower risk of affecting the functionality of the protein. By the same explanation, we predicted that there would be more lethal point mutations in the conserved domains of the protein. The expected correlation is found, excluding the Val160Met mutation which has a strong red signal (Figure 3). Even though valine and methionine are similar in size and polarity, any change at the 160th position appears to have a strong red signal (Figure 3). This indicates an important amino acid residue in the SRCR domain and may be the reason why Val160Met is so lethal. This agrees with previous studies which states that Val160 is a highly conserved residue across mammals. Fitzgerald (2008) may explain this possible residue, stating that it is in an exonic splicing enhancer site, therefore, changes to the residue are not well-tolerated. However, this contradicts the frequency in the population (25%) so more research is needed on the function of this residue.

Utilizing different protein modelling servers reveal similarities and differences in preciseness of 3D protein modelling (Figure 12-16). Visual differences, as well as the secondary structure differences were observed in the model. SWISS-MODEL and HHPred were the most similar structure and with the same amount of beta sheets (20) with 6 and 11 helices, respectively. Meanwhile, I-TASSER determined 17 beta sheets and RaptorX determined 15

beta sheets for their protein models. In addition, SWISS-MODEL and HHpred had a truncated amino acid sequence compared to the other servers (loss of ~100 amino acids). These results are not too surprising, as the prediction programs model a bit differently. For example, both HHPred and SWISS-MODEL use homology-based modelling to make 3D models of protein. Therefore, both servers only model the sequence that is matched to the target sequence chosen (hepsin). Since hepsin is only ~33% homology to TMPRSS2, the amino acids in the rest of the protein were not modelled. However, it is important to mention that TMPRSS2 shares 42.5% homology to kallikrein, yet only shares the trypsin serine protease domain. In comparison, hepsin shares only 33% homology to TMPRSS2 but contains both the SRCR and trypsin serine protease domain (Hussain 2020). Therefore, more manipulation in modelling the different domains in the protein may need to be carried out in future research. Despite this limitation, the conserved domains as well as the active site and substrate-binding sites were in the homologous part of the protein. Therefore, no significant data is lost when TMPRSS2 is modelled by SWISS-MODEL and HHPred. On the other hand, I-TASSER creates protein models from multiple threading and structural assembly simulations, which has led to the lab group winning several CASP experiments (Roy, 2010). The protein's function is then structurally matched using homology. Likewise, RaptorX uses threading and simulations, but also incorporates random coil fields to improve their Z score (Källberg, 2012).

[analyze Ramachandran plots & Z-scores]

Molecular docking of TMPRSS2 and SARS-CoV-2 Spike protein revealed interaction sites that may be important for binding (Figure 17). For multiple reasons, the structure chosen to continue further with the study was I-TASSER. Reasons include the simplicity of the server, the many citations and awards won by this protein server, previous research utilizing this server and lastly because it modelled the entire amino acid sequence. Likewise, the PDB ID: 7DK3 was chosen to use for the SARS-CoV-2 Spike Protein because this crystallization is present in the open conformation of the trimer. This conformation is necessary for the virus to recognize the

host receptor as well as get activated (Walls 2020). iCn3D structure viewer therefore was able to view the biomolecular complex and analyze the interactions present. Since there is only one predicted biomolecular complex of TMPRSS2 and SARS-CoV-2 Spike known to date (Hussain 2020), comparing to literature is limited in this regard. However, 12 of our 21 TMPRSS2 residues matched, including His296, Lys340, Lys392, Gln438, Lys467, Gly391, Gly462, Thr341, Glu229, Lys300, Val280, and Lys342. It is expected that not all of the interactions would line up, due to different methodology by combining templates from different prediction servers and imputing into MODELLER (Hussain 2020).

The effect of SNPs on TMPRSS2 binding to the spike protein can be determined by using the interaction sites found by I-TASSER (Table 3).

[compare SNPs to heatmap and poly-phen2/sift]

References

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

Bestle, D., Heindl, M. R., Limburg, H., Pilgram, O., Moulton, H., Stein, D. A., ... & Böttcher-Friebertshäuser, E. (2020). TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells. *Life Science Alliance, 3*(9). https://doi.org/10.26508/lsa.202000786

FitzGerald, L. M., Agalliu, I., Johnson, K., Miller, M. A., Kwon, E. M., Hurtado-Coll, A., ... & Huntsman, D. G. (2008). Association of TMPRSS2-ERG gene fusion with clinical characteristics and outcomes: results from a population-based study of prostate cancer. *BMC cancer*, *8*(1), 1-10. https://dx.doi.org/10.1186%2F1471-2407-8-230

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

Mollica, V., Rizzo, A., & Massari, F. (2020). The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer. *Future oncology (London, England)*, *16*(27), 2029–2033. https://doi.org/10.2217/fon-2020-0571

Resnick, D., Pearson, A., & Krieger, M. (1994). The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends in biochemical sciences*, *19*(1), 5-8. https://doi.org/10.1016/0968-0004(94)90165-1

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

Russo, R., Andolfo, I., Lasorsa, V. A., Iolascon, A., & Capasso, M. (2020). Genetic analysis of the coronavirus SARS-CoV-2 host protease TMPRSS2 in different populations. *Frontiers in genetics*, *11*, 872. https://dx.doi.org/10.3389%2Ffgene.2020.00872

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, *181*(2), 281-292. https://doi.org/10.1016/j.cell.2020.02.058

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