Discussion

Our analysis of TMPRSS2 revealed that there are SNPs located close to, and directly on, residues critical for the protease's binding interactions with the SARS-CoV-2 S protein. These SNPS may disrupt the viral entry of SARS-CoV-2 and therefore lower its ability to infect the host cell. Specifically, rs148125094, rs61735796, and rs768173297 are located close to residues important for binding and may have a damaging effect. Although they are predicted to be benign by PolyPhen-2 and SIFT, an analysis of their secondary and tertiary structures is needed to truly understand their implications on the binding ability of TMPRSS2. Rs142446494 is located directly on an interaction residue and is predicted to be damaging by SIFT, therefore it likely has destabilizing effects on TMPRSS2 and should also be further analyzed in its secondary and tertiary structure. While rs12329760 and rs150554820 are not located close to critical residues, they are both found in the conserved SRCR domain and are predicted to be damaging by both PolyPhen-2 and SIFT. Given that rs12329760 has already been cited in multiple studies for destabilizing TMPRSS2, it is likely that this SNP has potential to disrupt important binding interactions with the SARS-CoV-2 S protein (Vishnubhotla et al., 2020, Paniri et al., 2020). The potential that these SNPs have to alter the interaction between TMPRSS2 and the S protein are supported by previous studies that identify TMPRSS2 as a potential target against SARS-CoV-2 (Shen et al., 2020, Paniri et al., 2020).

Previous studies analyzing TMPRSS2 SNPs found that many of these variations were rare in the population, a limitation that was also present in this study (David et al., 2020). By analyzing SNPs that were above a designated frequency threshold, 0.000001, we attempted to identify variations that were more prevalent and therefore might have a greater impact on the general population in terms of differing susceptibility to SARS-CoV-2 infection. However, most

of the SNPs above the designated threshold were still rare in the overall population. The most frequent SNPS, rs75603675 and rs12329760, appeared in 30% and 22% of the global population, respectively, meaning that they could potentially underlie the global differences that have been observed in case severity of COVID-19. The rest of the SNPs analyzed fell below 1% frequency in the global population, with many of them falling orders of magnitude below that. Due to their rarity, it is unlikely that differences in binding interactions between TMPRSS2 and SARS-CoV-2, caused by these SNPs, are what is being observed through the different manifestations of clinical symptoms. Also important to note is that the sample sizes in which many of the SNP frequencies were taken from were relatively small, which may be contributing to the low frequencies reported. Without a larger sample size it is difficult to determine the true population frequencies of each SNP, however it is understandable that it is not practical to genotype large populations in a short amount of time. Still, the reported rarity of these SNPs make it unlikely that they play a leading role in symptom severity of COVID-19. They may still play an important role in the symptom severity in different hosts when combined with other factors.

Modelling docking interactions between TMPRSS2 and SARS-CoV-2 have previously been performed, and their importance in understanding how SNPs may affect critical binding interactions is underlined by this study (Hussain et al., 2020, Vishnubhotla et al., 2020). TMPRSS2 has not yet been crystallized, which may be due to its surface hydrophobicity, isoelectric point, and high percentage of coiled structure (Hussain et al., 2020). Therefore, predictive modeling using threading and homology is a useful tool in understanding the structure of TMPRSS2 and its likely interactions. This study's utilization of multiple softwares with different predictive methodologies allowed for analysis of possible TMPRSS2 structures and

ultimately the determination of which structure may be most accurate to perform docking analysis with. Homology modeling softwares, HHPred and Swiss-Model, utilized hepsin as a template for its model, as it has the closest homology to TMPRSS2 at 33.62%. However, this homology is relatively low and is likely why these softwares excluded large portions of TMPRSS2's sequence in their final structure. RaptorX is better suited for predicted protein structures with distant homology, however it's final structure also excluded portions of TMPRSS2's sequence. The structure of TMPRSS2 predicted by I-TASSER was chosen for analyzing docking interactions, as it modeled the complete sequence and is a highly cited software in protein prediction. While the ramachandran plot analysis of TMPRSS2 generated by I-TASSER had a higher percentage of amino acids present in unfavorable positions than those generated by other softwares, this is likely due to the fact that I-TASSER modelled a greater number of amino acids than the other softwares. Overall, the structure of TMPRSS2 generated was valid by ramachandran plot analysis and therefore offered understanding of potential docking interactions with the S protein.

The Docking interaction models showed that TMPRSS2 has 21 critical residues that are important for binding to SARS-CoV-2, however SNPs that are not located close to these residues may still have important implications for underlying disease severity. Specifically, the SNPs that were observed in conserved domains and have been predicted to be damaging by prediction softwares PolyPhen-2 and SIFT should also be analyzed for their structural and functional effects. Recent modelling of rs12329760 has identified that this V160M substitution results in structural deformation of TMPRSS2 because of the differences in topology and charge limit between the amino acids (Vishnubhotla et al., 2020). These findings support our hypothesis that the nonsynonymous SNPS may present structural challenges to TMPRSS2 that ultimately affect

its function and ability to bind to the S protein. Structural effects on TMPRSS2 caused by SNPs, like rs12329760, highlight that TMPRSS2 could be a useful therapeutic in treating or preventing SARS-CoV-2 infection and provide knowledge of how to develop such therapeutics. Therefore, the SNPs of interest should be further analyzed by use of computer modeling for their structural implications in future studies.

While this study focused on nonsynonymous, missense SNPS, there are other types of SNPs that could have important implications for TMPRSS2, namely intron or 3' UTR variants. Missense variants play an important role in TMPRSS2 stability, as the substitution of an amino acid on the primary sequence can have direct consequences of the folding of the secondary and tertiary structures that directly impact function, as was seen with rs12329760 (Vishnubhotla et al., 2020). Intron or 3' UTR variants may have the ability to alter the function of TMPRSS2 indirectly, by causing transcriptional changes that affect the level of TMPRSS2 expression. A previous study has identified SNPs that could potentially alter the post translational modification, spicing, and miRNA function of TMPRSS2 to ultimately affect its structure and function (Paniri et al., 2020). Such findings highlight the important role that SNPs in non-coding regions can play through epigenetic mechanisms, and should therefore be further investigated in TMPRSS2.

Overall, our findings contribute to the ongoing discussion of TMPRSS2s viability as a potential target against SARS-CoV-2. TMPRSS2 is known to play a crucial role in cleavage of the SARS-CoV-2 S protein to finalize its entry into the host cell (Hoffmann et al., 2020). The SNPs identified in this study have potential to decrease the binding ability of TMPRSS2 through amino acid substitutions that result in unstable protein folding. The SNPs identified in this study that occur on or close to a binding site have heightened potential to disrupt the binding interactions between SARS-CoV-2 and TMPRSS2 and should therefore be modeled and

analyzed for their structural interactions. An important limitation of these SNPs is that all but 2 occur rarely in the population, making it unlikely that they alone are responsible for the variation in disease severity and are more likely to, if at all, work with other factors to dictate these differences. Still, the implications SNPs have on TMPRSS2 structure and function are important in understanding SARS-CoV-2 infection and developing therapeutic interventions against COVID-19.

References

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.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N. H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*(2), 271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052

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.

Paniri, A., Hosseini, M. M., & Akhavan-Niaki, H. (2020). First comprehensive computational analysis of functional consequences of TMPRSS2 SNPs in susceptibility to SARS-CoV-2 among different populations. *Journal of Biomolecular Structure and Dynamics*, (just-accepted), 1-18.

Shen, L. W., Mao, H. J., Wu, Y. L., Tanaka, Y., & Zhang, W. (2017). TMPRSS2: A potential target for treatment of influenza virus and coronavirus infections. *Biochimie*, *142*, 1-10.

Vishnubhotla, R., Vankadari, N., Ketavarapu, V., Amanchy, R., Avanthi, S., Bale, G., ... & Sasikala, M. (2020). Genetic variants in TMPRSS2 and Structure of SARS-CoV-2 spike glycoprotein and TMPRSS2 complex. *BioRxiv*.