**Abstract: The Cleavage Potential of Cross-Species TMPRSS2 Variants toward SARS-CoV-2 S2’ Subunit**

**Tyshawn Ferrell1,2, Kelly Pierce1, Johnny Aldan1, Ethan Ho1, Dairian Balai 1**

1Texas Advance Computing Center**, University of Texas at Austin, Austin, TX, 78712-78758**

2Department of Biology, **Albany State University, Albany, GA 31705**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a positive-stranded RNA virus, has caused >100,000 infections and >50,000 deaths during the 2020 pandemic. SARS-CoV-2 successful infectivity is in part due to structural motifs and specific cell receptor targeting. SARS-CoV-2 spike (S) glycoprotein mediates cell attachment by recognizing the angiotensin converting enzyme II (ACE2) receptor, which initializes cleavage by host cell proteases to induce cell membrane fusion. In previous studies transmembrane protease/serine subfamily member 2 (TMPRSS2) have increased cleavage efficiency towards the S protein subunits. This study aims to characterize cross-species variants of TMPRSS2 and their potential to cleave S proteins subunits in silico. We hypothesize that TMPRSS2 aids in the spillover of SARS-CoV-2 to alternate host by having a conserved catalytic region. We investigated the conservation of TMPRSS2 sequences amongst several species and to the human homolog. We utilized protein databanks to form concise amino sequence alignments of six domestic animals for comparisons with the human variant. We characterized differences in catalytically important residues based off the position specific score and changes in amino acid properties. We found that important residues within TMPRSS2 are highly conserved, and found only substitutions likely to have minor effect on catalytic activity. This study has provided insight on possible spillover into alternate host with structurally similar TMPRSS2 variants compared to humans, posing an immense threat for epidemics in domesticated animal populations.

**Introduction**

**Proposed organization:**

1. **Viral entry: introduce ACE2, TMPRSS2, and other important host cell proteins as mediators of viral entry.**
2. **TMPRSS2: describe the salient details of this protein for SARS-CoV-2**
   1. **Receptor binding domain**
   2. **Catalytic triad**
   3. **Role of disulfide bonds**
3. **Viral host range as a property of host entry protein conservation: say a few words here about prior work looking at the conservation of ACE2 across species**
4. **Conclude with a statement on the aim of the research.**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel betacoronavirus, curtailed the world’s health system within a matter of months due to its high infectivity. As of July 20th, 2020, 14,348,858 laboratory-confirmed infections and 603,000 SARS-CoV-2 related deaths have been reported globally [12]. This is in part due to the high bonding affinity of the SARS-CoV-2 spike protein (S) for the angiotensin converting enzyme II (ACE2) receptor in airway epithelial cells. The S1 subunit of the S protein binds to ACE2 via its receptor binding domain (RBD), while the S2 subunit is used for virus-cell membrane fusion [1,3,6]. Though ACE2 is an essential component for attachment of the S protein, it is the host cell proteases that activates the S2 subunit that holds the transmembrane fusion machinery.

*The Role of TMPRSS2 in SARS-CoV-2 Infection*

Previous studies have shown that SARS-CoV has preference for cells with the transmembrane protease/serine subfamily member 2 (TMPRSS2) endoprotease, which increases the ability for virus-host membrane fusion [6,7,9]. TMPRSS2 is a member of the type II Transmembrane Serine Protease (TTSP), specifically belonging to the Hepsin/TMPRSS subfamily. TTSPs have four defining features: an N-terminal intracellular domain, transmembrane domain, “stem” domain, and a proteolytic domain. TMPRSS2 is synthesized as a zymogen that requires proteolytic processing to activate [2]. This activation results in cleavage of extracellular substrates initialized by a Serine residue at the catalytic site. The TMPRSS2s proteolytic domain has shown increased pathogenesis by cleaving monobasic sites of virus subunits within coronavirus and influenza family [3-7, 21]. The S protein in SARS-Cov-2 has an multibasic S1/S2’ cleavage site containing several arginine residues that are cleave by host cell proteases, which increases the efficacy of cell-cell spread [3]. However, in order for the virus-cell transmembrane fusion to occur, the monobasic S2’ cleavage site must be cleaved by TMPRSS2 [3-4]. TMPRSS2 proteolytic activity results from its catalytic domain, which consist of Ser-His-Asp, but Ser initializes the peptide hydrolysis by attacking the acyl compound of a lys or Arg residue in the S protein [2,4]. During this reaction two tetrahedral intermediates and one acylenzyme but finalizes with the protonation of an amine leaving group, carboxylic acid, water, and reformation of the catalytic triad [2]. Even though TMPRSS2 is a crucial part for virus-host transmembrane fusion, much of its functionality is still unclear. Since there are over 9000 sequences of TMPRSS2, a comparative analysis of its amino acid sequence can help predict its cleaving efficacy of SARS-CoV-2 spike protein.

*SARS-CoV-2 Host Range*

In silico studies have shown that SARS-CoV-2 ability to infect alternate host depends on ACE2 based on the conservation of the enzyme compared to the human ACE2 sequence [8].

With cases of SARS-CoV-2 in domesticated animals such as ferrets and cats, it is imperative to identify transmissible elements between humans and animals.

Additionally, ACE2 variants have been observed in multiple species, which may have lead SARS-CoV2 high infectivity in native and foreign host [1,8]. Exploring TMPRSS2 variants will yield valuable insight into SARS-CoV-2 zoonotic transmission potential to spill over in specie populations.

SARS-CoV-2, a zoonotic pathogen, has evolved to successfully infect a broad host range using common host cell machinery. With its high infectivity and unpredictable nature, SARS-CoV-2 is increasingly difficult to treat and contain. In previous studies, TMPRSS2 was shown to cleave SARS-CoV at a monobasic Arg site, causing cell-cell spread and increased severity of coronaviruses in infected lung cells of mice [20-21]. Additionally, TMPRSS2’s proteolytic activation takes place after SARS-CoV S protein undergoes receptor-induced conformational modifications [21]. However, TMPRSS2 deviations in other species with similar ACE2 receptors (compared to human strains) could lead to unfavorable conditions for SARS-CoV-2 activation. Therefore, we hypothesize that cross-species conservation of TMPRSS2 catalytic residues may play a role in alternate host spillover. By comparing residue alterations of TMPRSS2 multiple sequence alignments between various species and human, we wll be able to score the susceptibility/severity of infection for each species, in silica. Also, we’ll be able to categorize each substitution’s property delta, then visualize the structural context of that change in Pymol. By determining the molecular mechanisms leading to cell entry, we will gain insight on spillover potential of SARS-CoV-2 to alternate host. Which could harbor adaptive advantages for SARS-CoV-2 immune evasion and overall infectivity within a broad host range.

**Methods**

**Data Collection and Curation**

We retrieved 9,757 TMPRSS2 protein sequences from 120 species from EggN0G. [13]. Using Clustal Omega we generated the full set of each protein sequence alignment independently, which verified the authenticity of each alignment in the EggN0G dataset [14]. We utilized functions in Biopython [citation], a python biological computation package, to parse the lines within the alignment file to convert header and sequence pairs into iterables. We then filtered out target mammalian species (Homo sapiens, Mus musculus, Canis Lupus familiaris, Felis catus, Bos Taurus, Equus caballus, and Gallus gallus) from the multiple sequence alignment of all 120 species. From the target species list we extracted the H. sapien homolog sequence based upon UniProt’s TMPRSS2 homolog [17]. For each non-human species, we selected a sequence for comparison by choosing the sequence record with the greatest percent identity to human TMPRSS2. We used a position specific score matrix (PSSM) to find the percent probability of a particular residue at a specific location.

**Protein Sequence Analysis**

We identified 24 TMPRSS2 amino acid residues that were previously reported to be important for proteolytic cleavage and conformation [2,4,15,16,17]. These residues include the assumed monobasic recognition site (K223 and K224), disulfide bonds (C113, C120, C126, C133, C139, C148, C172, C185, C231, C241, C281, C297 C365, C410, C426, C437, and C465), catalytic triad (D345, H296, and S441), and the binding site (D435). We compared the residues of each mammalian species to the human TMPRSS2 sequence in search of differences in particular residues of interest. Each substitution at a residue of interest was categorized as a mutation, which was then analyzed for its property difference based on data from previous studies (tryna get the reference from a prof). We retrieved the predicted structure of TMPRSS2 from SWISS-MODEL [18]. We modified the structure of TMPRSS2 in Pymol to visualize sequence comparisons [19].

**Results**

**Conservation of Cross-species TMPRSS2 sequences**

We investigated the conservation of TMPRSS2 sequences across seven species at catalytic and structurally important positions. To accomplish this, we retrieved data from EggN0G and analyzed important positions within the multiple sequence alignment sequence using biopython. 20 out of 24 important residues are highly conserved, especially in the Serine domain where all species share the same catalytic triad sequence. Yet, there is slight variation with at monobasic recognition sites that have a deletion at 223 and a serine/arginine at 224 for select species. These positions reside in the Scavenger Receptor Cysteine Rich (SRCR) domain (residues 149-242), which functions in ligand-binding activity. This domain’s binding affinity for extracellular ligands is partially due to its pattern-recognition ability where clusters of basic residues form electrostatic bonds with select binding motifs [24]. The SRCR domain is normally near the surface of cells or proteins to act like phagocytic receptors [23]. The Ser224 residue in Cats and dogs are more likely to form intermolecular hydrogen bonds with adjacent amino acids, given that its accessible surface area (80Å3). However, with the deletion at 223 for both species it is uncertain if that will play a role in binding specificity for Ser224. Whereas the Arg224 Å residue in chickens would likely contribute to the basic cluster, still being efficient for electrostatic bonding with extracellular substrates.

In addition to differences in the SRCR domain there are also notable changes at positions 113 and 120 within the Low-Density Lipoprotein Receptor Class A (LDLRA) domain for cats and dogs. The LDLRA domain plays a role in zymogen activation by having a ligand-binding site that is specific Low-Density Lipoproteins and calcium ions [15] This activity stabilizes the orientation of the domain, which influences the ligand-binding potential in the SRCR domain [24]. However, mutations at positions 113 and 120 in cat and dog sequences would disrupt the disulfide bond that would form with residues 126 and 139, respectively. There are no other cysteines in that domain for those two species to form alternative disulfide bonds.

**Protein Visualization of TMPRSS2 residues**

We investigated the differences in TMPRSS2 by mutating select positions using the mutagenesis tool in pymol [18]. We retrieved the predicted TMPRSS2 structural model from SWISS-MODEL [19]. The structural model sequence began near the end of the LDLRA domain (145-492), so we only focused on the downstream monobasic recognition sites. Ser224 in cats and dogs may be more compacted within the TMPRSS2 no than Lys224 having a difference in its exposure to surface interaction with other molecules (percent buried= -72%∆ and surface accessible area= -87Å2∆). Additionally, Ser224 has only three rotamers where the OH group would likely interact with the O on Leu221 in the human sequence. However, due to the deletions at positions 221-223 in the cat and dog sequence, it was challenging to determine a predictable interaction between surrounding amino acids in pymol. Ser224 is also slightly more hydrophobic (-6∆) compared to Lys224, which is a notable factor in its ligand-binding specificity. At the same position in chickens lays an Arg residue that has similar physicochemical properties, but differs in hydrophobicity (-5∆) and accessible surface area (29∆). Position variability is most prominent at the monobasic recognition site, which results in a basic cluster in chicken sequence and a less predictable 3D structure for cat and dog sequences.

**Discussion**

SARS-CoV-2, the third member of the betacornarvirus subfamily known to cause epidemics or pandemics, employs host cell machinery for activation. TMPRSS2, a host cell protease, is required for cell entry of SARS-CoV and SARS-CoV-2 via cleavage of the S2 subunit [ 20,22,25]. TMPRSS2 is expressed in a broad range of tissues, including alveolar epithelium (lungs), upper airway epithelial cells, and prostate gland ductal and basal cells [20, 22]. In previous studies SARS-CoV-2 demonstrated a preference cell-surfaced TMPRSS2s over cathespins (endosomal pathway) and TMPRSS2+/+ cell-lines had a 10-fold increase in SARS-CoV-2 infected cells compared to TMPRSS-/- cells [6,9]. Which posed the question if other domesticated animals were susceptible to SARS-CoV-2, along with the pathway for infection within these species.

The aim of this study was to measure the conservation of important TMPRSS2 residues at select positions that aided in the cleavage ability of the protease domain towards the SARS-CoV-2 S protein S2 subunit. We observed that most species had sequences identical to the human homolog at positions of interest; however, there was slight variation for a few positions. Though the cat and dog sequence varied at four positions of interest compared to the human homolog, their sequences were identical for all important positions. It is known that the cat is highly susceptible to SARS-CoV-2 infection, however dogs have low susceptibility to SARS-CoV-2 [26]. This suggest that there is variation with the dog ACE2 receptor sequence, which in silica studies have characterized to have low binding affinity for the S protein [1].

Even though cats have a higher susceptibility to SARS-CoV-2 than dogs, the cleavage potential of feline TMPRSS2 may still be affected by changes within those positions. Because of the deletion at Lys223 and Ser224 replacing the Lys, there is no basic residue for electrostatic bond or basic cluster formation for monobasic recognition of the Arg residue in the S protein S2 cleavage site. This could lead to lack of ligand-binding specificity, likely decreasing the ability of TMPRSS2 to cleave the S2 subunit Arg at that specific site. On the opposite side the chicken sequence is right next to Lys223, which may form a basic cluster for the monobasic recognition. This basic cluster would likely have a minor effect on the S2 subunit cleavage site. Conformational changes may result in from the disruption of disulfide binding at positions 133 and 120 within the LDLRS domain between cat and dog sequences. The alteration of these two positions will calcium and LDL binding sites, which influence the monobasic recognition of extracellular substrates in the SRCR domain.

In conclusion the TMPRSS2 is highly conserved, especially within the catalytic triad, but there is slight variation within a few species at the monobasic recognition site and disulfide forming cysteines within the SRCR domain. Due to TMPRSS2s critical role in SARS-CoV-2 cell entry and conservation, it may serve as a target for serine protease inhibitors. In our future studies we plan to obtain the complete crystal structure of TMPRSS2 and compare it to the predicted model produced by SWISS-MODEL. By doing so, we’ll be able to verify the tertiary structure of TMPRSS2. Then we’ll investigate how the differences in specific residues affect the catalytic efficiency to cleave monobasic Arg residues in vivo. Finally, we will explore multiple sequence alignments for other species and analyze data for variation within each species at a nucleotide level. And after data analysis we’ll create a cross-species phylogenic tree of TMPRSS2 variants. This study has provided the foundation for cross-species TMPRSS2 variation among domesticated animals, which may be subjected to SARS-CoV-2 as an intermediate or native host.

**References**

1. Damas, J., Hughes, G. M., Keough, K. C., Painter, C. A., Persky, N. S., Corbo, M., Hiller, M., Koepfli, K. P., Pfenning, A. R., Zhao, H., Genereux, D. P., Swofford, R., Pollard, K. S., Ryder, O. A., Nweeia, M. T., Lindblad-Toh, K., Teeling, E. C., Karlsson, E. K., & Lewin, H. A. (2020). Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. bioRxiv : the preprint server for biology, 2020.04.16.045302. <https://doi.org/10.1101/2020.04.16.045302>
2. Hedstrom, L. (2002). Serine protease mechanism and specificity. *Chemical reviews*, *102*(12), 4501-4524.
3. Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell*.
4. Hoffmann, M., Hofmann-Winkler, H., & Pöhlmann, S. (2018). Priming time: How cellular proteases arm coronavirus spike proteins. In *Activation of Viruses by Host Proteases* (pp. 71-98). Springer, Cham.
5. Hofmann, H., & Pöhlmann, S. (2004). Cellular entry of the SARS coronavirus. Trends in microbiology, 12(10), 466-472.
6. Matsuyama, S., Nao, N., Shirato, K., Kawase, M., Saito, S., Takayama, I., ... & Sakata, M. (2020). Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proceedings of the National Academy of Sciences, 117(13), 7001-7003.
7. Meng, T., Cao, H., Zhang, H., Kang, Z., Xu, D., Gong, H., ... & Wei, H. (2020). The insert sequence in SARS-CoV-2 enhances spike protein cleavage by TMPRSS. BioRxiv.
8. 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.
9. Shirato, K., Kawase, M., & Matsuyama, S. (2018*). Wild-type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry*. Virology, 517, 9-15.
10. Shulla, A., Heald-Sargent, T., Subramanya, G., Zhao, J., Perlman, S., & Gallagher, T. (2011). *A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry*. Journal of virology, 85(2), 873-882.
11. 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.
12. WHO, 2020 Who, *Novel Coronavirus(2019-nCoV) Situation Report 182*

<https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200720-covid-19-sitrep-182.pdf?sfvrsn=60aabc5c_2>

1. *Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork*. [Nucleic Acids Res. 2019 Jan 8; 47(Database issue): D309–D314. doi: 10.1093/nar/gky1085](https://doi.org/10.1093/nar/gky1085)
2. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sciences. Protein Sci 27:135-145
3. Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., & Antonarakis, S. E. (1997). Cloning of the TMPRSS2 gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22. 3. Genomics, 44(3), 309-320.
4. Ohno, A., Maita, N., Tabata, T., Nagano, H., Arita, K., Ariyoshi, M., ... & Kishimoto, K. (2020). Cystal structure of inhibitor-bound human MSPL/TMPRSS13 that can activate high pathogenic avian influenza. bioRxiv.
5. The UniProt Consortium  
   **UniProt: a worldwide hub of protein knowledge**  
   [Nucleic Acids Res. 47: D506-515 (2019)](https://doi.org/10.1093/nar/gky1049)
6. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research **42**, W252–258, doi:10.1093/nar/gku340 (2014).
7. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. https://pymol.org/2/ (2020).
8. Iwata-Yoshikawa, N., Okamura, T., Shimizu, Y., Hasegawa, H., Takeda, M., & Nagata, N. (2019). TMPRSS2 contributes to virus spread and immunopathology in the airways of murine models after coronavirus infection. Journal of virology, 93(6).
9. Matsuyama, S., Nagata, N., Shirato, K., Kawase, M., Takeda, M., & Taguchi, F. (2010). Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. Journal of virology, 84(24), 12658-12664.
10. Kühn, N., Bergmann, S., Kösterke, N., Lambertz, R. L., Keppner, A., van den Brand, J. M., ... & Schughart, K. (2016). The proteolytic activation of (H3N2) influenza A virus hemagglutinin is facilitated by different type II transmembrane serine proteases. Journal of virology, 90(9), 4298-4307.
11. Areschoug, T., & Gordon, S. (2009). Scavenger receptors: role in innate immunity and microbial pathogenesis. Cellular microbiology, 11(8), 1160-1169.
12. Ojala, J. R., Pikkarainen, T., Tuuttila, A., Sandalova, T., & Tryggvason, K. (2007). Crystal structure of the cysteine-rich domain of scavenger receptor MARCO reveals the presence of a basic and an acidic cluster that both contribute to ligand recognition. *Journal of Biological Chemistry*, *282*(22), 16654-16666.
13. 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.
14. Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., ... & Zhao, Y. (2020). Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. Science, 368(6494), 1016-1020.