

**Structure, function and variants analysis of the androgen-regulated *TMPRSS2*,  
a drug target candidate for COVID-19 infection**

Alessia David, Tarun Khanna, Melina Beykou, Gordon Hanna, Michael J E Sternberg

**Affiliation:**

Centre for Integrative Systems Biology and Bioinformatics, Department of Life  
Sciences, Imperial College London, London SW7 2AZ, UK

The Authors declare no competing interest

*Corresponding author:*

Dr Alessia David

Centre for Integrative Systems Biology and Bioinformatics

Department of Life Sciences

Imperial College London

London SW7 2AZ, UK

## Abstract

SARS-CoV-2 is a novel virus causing mainly respiratory, but also gastrointestinal symptoms. Elucidating the molecular processes underlying SARS-CoV-2 infection, and how the genetic background of an individual is responsible for the variability in clinical presentation and severity of COVID-19 is essential in understanding this disease.

Cell infection by the SARS-CoV-2 virus requires binding of its Spike (S) protein to the ACE2 cell surface protein and priming of the S by the serine protease TMPRSS2. One may expect that genetic variants leading to a defective TMPRSS2 protein can affect SARS-CoV-2 ability to infect cells. We used a range of bioinformatics methods to estimate the prevalence and pathogenicity of TMPRSS2 genetic variants in the human population, and assess whether TMPRSS2 and ACE2 are co-expressed in the intestine, similarly to what is observed in lungs.

We generated a 3D structural model of the TMPRSS2 extracellular domain using the prediction server Phyre and studied 378 naturally-occurring TMPRSS2 variants reported in the GnomAD database. One common variant, p.V160M (rs12329760), is predicted damaging by both SIFT and PolyPhen2 and has a MAF of 0.25. Valine 160 is a highly conserved residue within the SRCS domain. The SRCS is found in proteins involved in host defence, such as CD5 and CD6, but its role in TMPRSS2 remains unknown. 84 rare variants (53 missense and 31 leading to a prematurely truncated protein, cumulative minor allele frequency (MAF)  $7.34 \times 10^{-4}$ ) cause structural destabilization and possibly protein misfolding, and are also predicted damaging by SIFT and PolyPhen2 prediction tools. Moreover, we extracted gene expression data from the human protein atlas and showed that both ACE2 and TMPRSS2 are expressed in the small intestine, duodenum and colon, as well as the kidneys and gallbladder.

The implications of our study are that: i. TMPRSS2 variants, in particular p.V160M with a MAF of 0.25, should be investigated as a possible marker of disease severity and prognosis in COVID-19 and ii. in vitro validation of the co-expression of TMPRSS2 and ACE2 in gastro-intestinal is warranted.

**Keywords:** SARS-CoV-2 virus, Spike glycoprotein, TMPRSS2, ACE2

## INTRODUCTION

The transmembrane protease serine type 2 (TMPRSS2) protein has a key role in severe acute respiratory syndrome (SARS)-like coronavirus (SARS-CoV-2) infection, as it is required to prime the virus' spike (S) protein, thus facilitating its entry into target cells (Hoffmann et al., 2020), (Shang et al., 2020). TMPRSS2 is characterized by androgen receptor elements located upstream its transcription site (Lin et al., 1999). It is expressed in lung and bronchial cells (Lukassen et al., 2020), but also in the colon, stomach, pancreas, salivary glands, prostate and numerous other tissues (Vaarala et al., 2001). Moreover, TMPRSS2 has recently been shown to be co-expressed with the angiotensin-converting enzyme 2 (ACE2) – the SARS-CoV-2 cellular receptor (Zhou et al., 2020) (Hoffmann et al., 2020) – in bronchial and lung cells (Lukassen et al., 2020).

TMPRSS2 cleaves and activates the S protein of several coronaviruses, including the severe acute respiratory syndrome-related coronavirus (SARS-CoV) (Hoffmann et al., 2020) and the Middle East respiratory syndrome-related coronavirus (MERS-CoV) (Du et al., 2017), facilitating virus-cell membrane fusion and viral infection. TMPRSS2 is also crucial for the proteolytic cleavage and activation of hemagglutinin (HA) in the influenza A virus (Böttcher et al., 2006).

Studies in TMPRSS2-KO mice reported reduced SARS-CoV and MERS-CoV replication in lungs compared to wild-type mice, and a reduced proinflammatory viral response, particularly cytokine and chemokine response via the Toll-like receptor 3 pathway (Iwata-Yoshikawa et al., 2019). In vitro studies has shown that TMPRSS2 inhibitors prevent bronchial cell infection by SARS-CoV (Kawase et al., 2012). In animal studies, mice infected with SARS-CoV and treated with the serine protease inhibitor Camostat Mesilate showed a high survival rate (Zhou et al., 2015). Recently, Camostat Mesilate (which, in Japan, is already approved for use in patients with chronic pancreatitis and postoperative reflux esophagitis) was shown to block SARS-CoV-2 lung cell infection *in vitro* (Hoffmann et al., 2020). Based on these encouraging results, a clinical trial to study the effect of a TMPRSS2 inhibitor in COVID-19 patients is under way (<https://clinicaltrials.gov/ct2/show/NCT04321096>).

Since pharmacological block of TMPRSS2 prevents cell entry and infection by coronavirus, one could postulate that naturally occurring *TMPRSS2* genetic variants affecting the structure and function of the TMPRSS2 protein may confer some protection from SARS-CoV-2 infection. We performed a computational bioinformatics analysis on *TMPRSS2* variants reported in GnomAD, the database of population genetic variations, to identify variants that could result in TMPRSS2 loss of

structure/function, and assess their prevalence in the general population using data from the GnomAD database. Moreover, we explored the co-expression of *TMPRSS2* and *ACE2* in extrapulmonary tissues using data from the Human protein Atlas, focusing on the intestine in view of the high prevalence of diarrhoea and gastrointestinal symptoms in COVID-19 patients (Chen et al., 2020).

## METHODS

*TMPRSS2* variants, with their global and gender-specific minor allele frequency (MAF), were extracted from GnomAD (Karczewski et al., 2017). Cumulative MAF was calculated as the sum of the MAFs for each variant.

In order to assess the structural impact of *TMPRSS2* variants, the three-dimensional structure of the *TMPRSS2* was required. Currently, no experimental 3D structure for *TMPRSS2* is available. We, therefore, generated a 3D structural model using our in-house Phyre homology modelling algorithm (Kelley et al., 2015). The FASTA sequence of *TMPRSS2* was obtained from the UniProt protein knowledge database (The UniProt Consortium, 2017) (UniProt Id O15393, corresponding to 492 amino acid transcript Ensembl ID ENST00000332149.10). We deposited the 3D coordinates of the model in our PhyreRisk (Ofoegbu et al., 2019) database (<http://phyrerisk.bc.ic.ac.uk/search?action=fresh-search&searchTerm=TMPRSS2>).

For completeness, SWISS-MODEL (Waterhouse et al., 2018) was also implemented to model *TMPRSS2*.

Model quality was assessed using: i. VoroMQA (Voronoi tessellation-based Model Quality Assessment) (Olechnovič and Venclovas, 2017), a statistical-based tool based on inter-atomic and solvent contact areas, which returns both a local and global score between 0 and 1. A global score  $> 0.4$  indicates a good model and  $< 0.3$  a poor model, models with  $>0.3$  score  $<0.4$  cannot be correctly classified; ii. QMEANDisCO (Studer et al., 2020), which uses distance restraints calculated from experimental structures of proteins homologous to the query protein and returns a plot of local IDDT (distance difference comparison) scores for each residue between 0 and 1. A global IDTT  $>0.6$  indicates a good model structure. iii. ProSA (Wiederstein and Sippl, 2007). The latter returns: 1. the Z-score, which indicates the overall model quality of the target structure compared to the Z-score for all experimentally determined structures in PDB (X-rays and NMR); 2. the residue energy, which indicates the local model quality and is calculated over a 10- and 40-residue window. Negative energy values indicate good quality structure.

The impact of each variant on *TMPRSS2* protein structure was assessed by analysing the following 16 features, using our in house algorithm Missense3D (Ittisoponpisan et al., 2019): breakage of a disulfide bond, hydrogen bond or salt bridge, introduction of a buried proline, clash, introduction of hydrophilic residue, introduction of a buried charged residue, charge switch in a buried residue, alteration in secondary structure, replacement of charged with uncharged buried residue, introduction of a disallowed phi/psi region, replacement of a buried glycine with any

other residue, alteration in a cavity, replacement of cis proline, buried to exposed residue switch, replacement of a glycine located in a bend. In addition, we used the SIFT (Vaser et al., 2016) and Polyphen2 (Adzhubei et al., 2010) variant predictors, which mainly use evolutionary conservation to assess a variant's effect.

The effect of variant rs12329760 was further assessed using: i. CONDEL (González-Pérez and López-Bigas, 2011) which reports a weighted average of the scores from fatHMM and MutationAssessor, and ii. FoldX5 force field (Schymkowitz et al., 2005) which calculate the stability of a protein based on the estimation of its free energy. A  $\Delta\Delta G > 0$  kcal/mol (calculated as:  $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$ ) was predicted to have a destabilizing effect.

Expression data for *TMPRSS2* and *ACE2* were extracted from the Human Protein Atlas (HPA) (Uhlén et al., 2015): *TMPRSS2* at <https://www.proteinatlas.org/ENSG00000184012-TMPRSS2> and *ACE2* at <https://www.proteinatlas.org/ENSG00000130234-ACE2>. Three sources of RNA data are integrated in the HPA database: i. HPA-generated RNA sequencing data, ii. RNA-seq data from the Genotype-Tissue Expression Project (GTEx), which uses post-mortem tissue samples, and iii. RNA-seq values from the Functional Annotation of Mammalian Genomes 5 (FANTOM5) project. HPA provides normalized expression (NX) values. In brief, to take into account the removal of non-coding transcripts, the Transcripts Per Million (TPM) values within each sample were adjusted to sum to one million. Thereafter, library size and composition differences were accounted for by applying a trimmed mean of M-values (TMM) normalisation to the sample TPM values. Then Pareto scaling was applied at gene level within each of the data sources. Batch effects were removed when the tissue data from the three source datasets was integrated. Where a tissue has multiple sub-tissues, the reported NX value was the maximum NX value across all the sub-tissues. A consensus transcript expression level was produced for each gene and tissue, by selecting the maximum NX value for this combination across the three data sources. Further details are available from the HPA website <http://www.proteinatlas.org>.

## RESULTS

### ***TMPRSS2 protein***

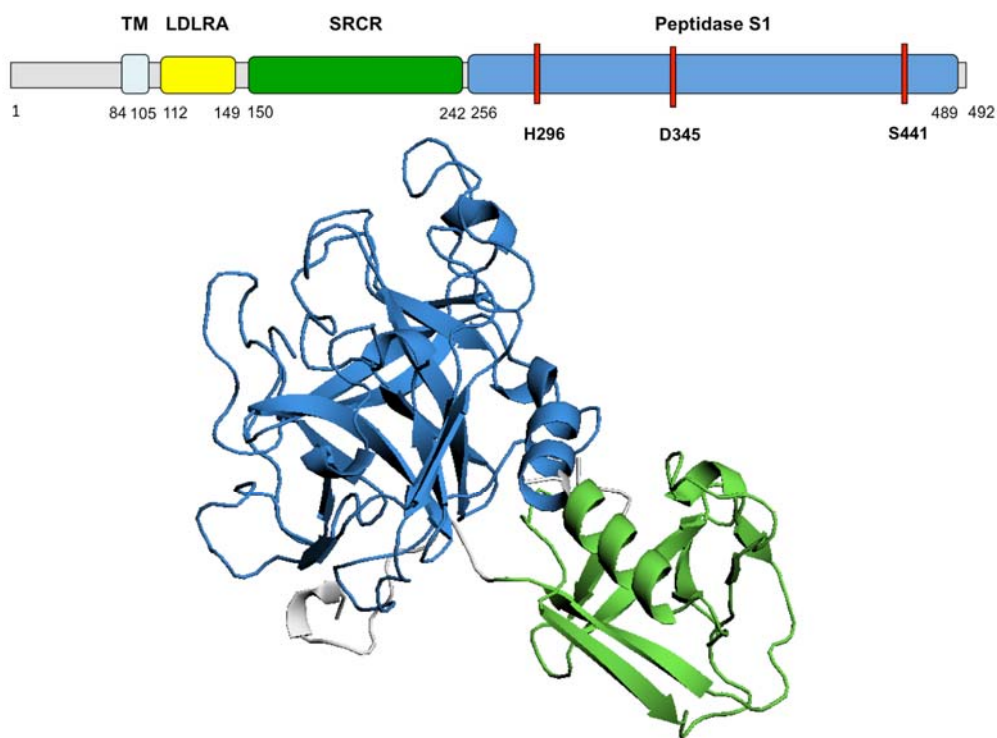
TMPRSS2 is composed of a cytoplasmic region (residues 1-84), a transmembrane region (amino acids 85-105) and an extracellular region (residues 106-492). The latter is composed of three domains: the LDLR class A (residues 112-149), the scavenger receptor cysteine-rich domain 2 (SRCR-2) (residues 150-242) and the Peptidase S1 (residues 256-489), which contains the protease active site: residues H296, D345 and S441. Two potential glycosylation sites are present in positions 213 and 249. A cleavage site at residues 255-256 has been shown to allow shedding of the extracellular region of TMPRSS2 (Afar et al., 2001).

No known experimental structure for TMPRSS2 is currently available. We therefore generated a 3D model of the extracellular region residues 145-491 corresponding to domains SRCR and Peptidase S1 (Figure 1 and Figure 2 and Table 1) using the transmembrane serine protease hepsin as a template (PDB: 1Z8G, chain A, X-ray structure with 1.55Å resolution. Model confidence 100%, sequence Id= 35% to target sequence; Model assessment: VoronMQA score = 0.462, QMEANDisCO score =0.67, and ProSA Z score = -8.96, also see Supplementary Figure S1. RMSD Phyre model vs SWISS-MODEL =0.30Å). A second model covering residues 111-150 (LDLRA domain) was generated using as a template PDB: 4U8U (giant haemoglobin), chain O, resolution 3.2Å (model confidence 97%, sequence Id= 30% to target sequence. (ProSA Z score= -3.08. VoronMQA score= 0.267, however this may not accurately reflect model quality since the software was trained on chains with more than 99 residues). Given the similarity of the Phyre and the SWISS-MODEL structures, the subsequent analysis should not be particularly sensitive to which model was selected.

**Table 1. 3D model of the SRCR and Peptidase S1 domain: template and quality assessment.**

Scores indicative of a good quality model are: VoronMQA >0.4, QMEANDisCO >0.6. For Prosa, the Z score ranges between -12 and -2.5 for X-ray structures of similar amino acid length (see suppl material).

	Template	Residues	Quality assessment		
			VoronMQA	QMEANDisCO	Prosa
<b>Phyre</b>	1Z8G	145-491	0.462	0.67	-8.96
<b>SWISS-MODEL</b>	1Z8G	146-489	0.499	0.68	-8.67

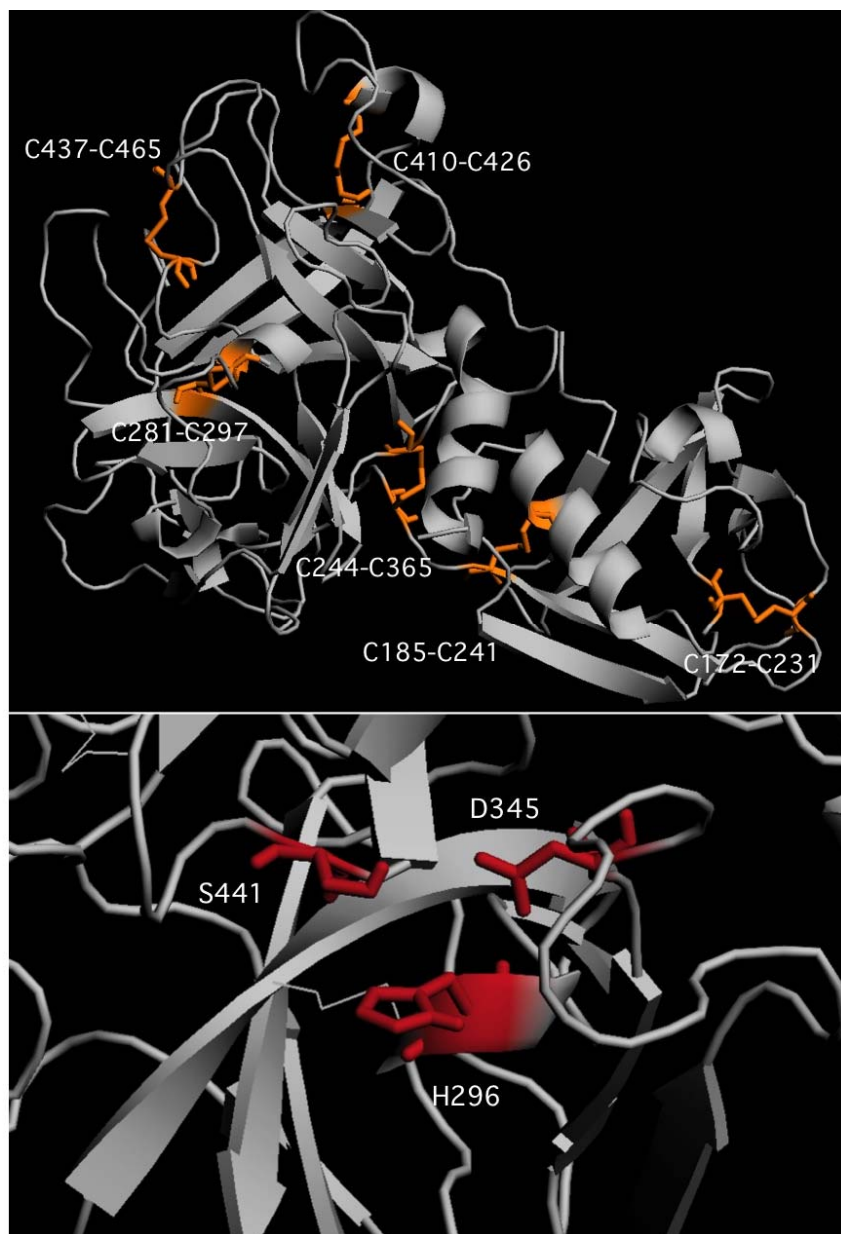


**Figure 1. Tmprss2 predicted 3D structure**

Diagram of Tmprss2 amino acid sequence and domains. The 3D model of Tmprss2 domains SRCR and Peptidase S1 is presented. The active site, residues H296, D345 and S441, is highlighted in red on the amino acid sequence.

TM, transmembrane domain; LDLRA, LDL-receptor class A; SRCR, scavenger receptor cysteine-rich domain 2; Peptidase S1, Serine peptidase.





**Figure 2. Tmprss2 structure**

Top panel: cysteine bonds in the SRCR and Peptidase S1 are presented in orange.  
Bottom panel: residues H296, D345 and S441 forming Tmprss2 active site are presented in red

### ***Tmprss2 variants***

We analysed 378 naturally occurring *Tmprss2* variants reported in the GnomAD database. One variant p.V160M (rs12329760) has a MAF of 0.248 in the population (0.2496 in males and 0.2488 in females), with 6.7% of individuals homozygotes for

this variant (9,587 homozygotes out of 141,456 individuals sequenced as part of the GnomAd project). This variant is predicted damaging by both SIFT (score=0.01) and PolyPhen2 (score=0.937). CONDEL also predicts this substitution to be damaging (score=0.788). Valine 160 is a highly conserved, small hydrophobic residue within the SRCS domain and may have a structural or functional role within TMPRSS2 (Figure 3). On manual inspection, substitution of valine with the large methionine may cause some degree of steric clash within the structure. However, this was not supported by the change in free energy ( $\Delta\Delta G_{\text{FoldX}}$  -0.47 kcal/mol).

The p.V160M substitution may affect TMPRSS2 function. The SRCS is a highly conserved domain, whose function is still not fully understood, although a role in ligand and/or protein interaction has been proposed (Aruffo et al., 1997). Interestingly, this domain is present in several proteins involved in host defence, such as CD5, CD6 and Complement factor I (Freeman et al., 1990), (Resnick et al., 1994).

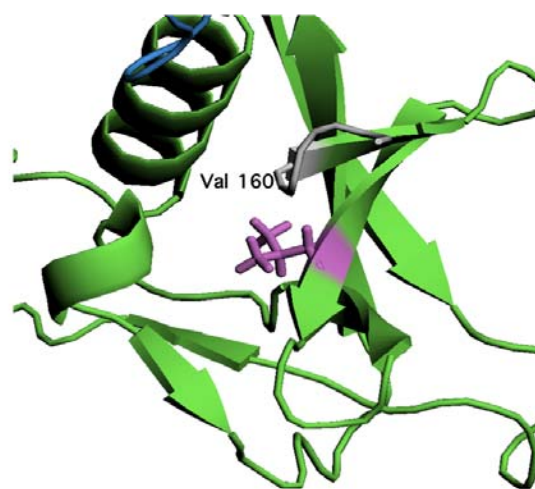
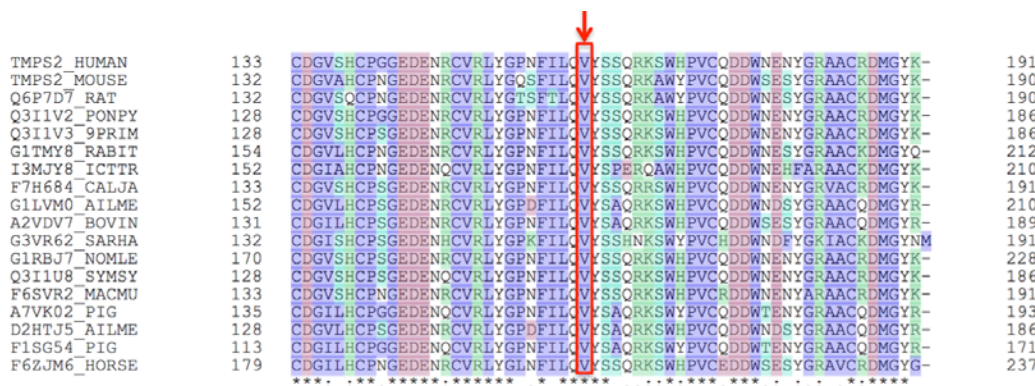


Figure 3. **V160M genetic variant (rs12329760) in TMPRSS2 SRCR domain.**

The position of Valine 160 is presented within a multi sequence alignment and highlighted in magenta on the 3D structure of the SRCR domain

31 variants lead to a prematurely truncated protein (17 frameshift variants causing premature termination and 14 stop gain variants). 304 out of 334 missense variants were mapped onto the 3D structure and 62 variants were predicted structurally damaging by Missense3D. Of these, 12 variants (p.C113R, p.C113Y, p.C139R, p.C185Y, p.C244F, p.231S, p.C244R, p.C281F, p.C281R, p.C297Y, p.C410Y, p.C465Y) result in substitution of the invariable cysteine forming disulfide bonds and are highly likely to cause destabilization and possibly misfolding of the TMPRSS2 protein structure. Further analysis showed that two variants disrupt TMPRSS2 function: p.R255S (rs769655195), which abolishes the TMPRSS2 cleavage site and p.S441G (rs1292701415), which abolishes TMPRSS2 active site. However, both variants are extremely rare in the population ( $MAF < 1 \times 10^{-5}$ ) and are unlikely to be useful as a marker of SARS-CoV-2 infection severity.

167 variants were predicted damaging by SIFT and 152 by Polyphen2. 137 variants were predicted damaging by both SIFT and PolyPhen2, and of these, 53 variants were also predicted to cause structural damage, thus reinforcing the damaging effect predicted for these variants.

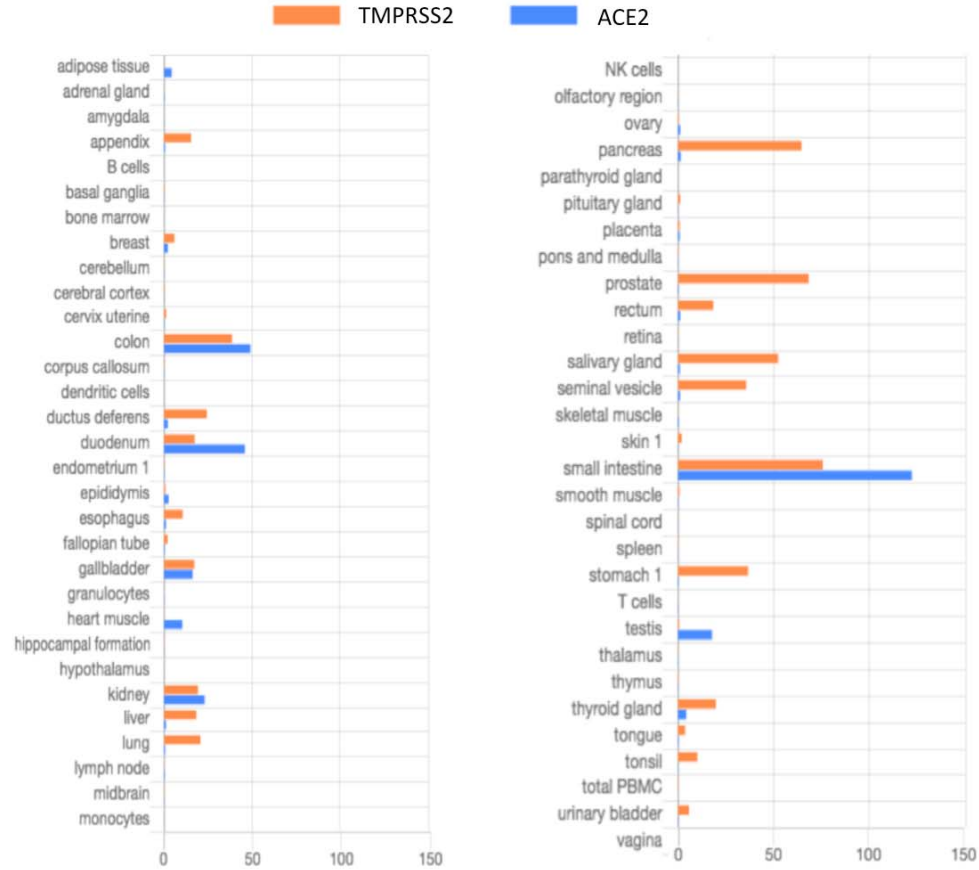
The 84 variants (53 missense and 31 leading to a prematurely truncated protein) that are predicted to cause loss of function with a high degree of confidence (predicted damaging by all three methods) are rare in the population. Their average MAF in the population is  $9.67 \times 10^{-6}$  and their cumulative MAF of  $7.34 \times 10^{-4}$ , therefore unlikely to be helpful as a marker of SARS-CoV-2 infection severity in the general population.

### **Both *TMPRSS2* and *ACE2* are expressed in the intestine**

We extracted data on *TMPRSS2* and *ACE2* tissue expression from the Human Protein Atlas. As shown in Figure 4, both proteins are expressed in the gut: colon, small intestine and duodenum. Other sites of expression are the kidneys and gallbladder. Interestingly, although *TMPRSS2* and *ACE2* have recently been shown to be co-expressed in the lung and bronchial cells, *ACE2* lung expression is almost negligible in the HPA, thus highlighting that specific *in vitro* experiments are necessary to study the co-expression of *TMPRSS2* and *ACE2* in extrapulmonary tissues. Moreover, although expression of *ACE2* arterial and venous endothelial cells and arterial smooth muscle cells was previously demonstrated (Hamming et al., 2004), it is not reported in the HPA.

**Figure 4. *TMPRSS2* and *ACE2* tissue expression**

Tissue Type	O15393 ( <i>TMPRSS2</i> )	Q9BYF1 ( <i>ACE2</i> )	Tissue Type	O15393 ( <i>TMPRSS2</i> )	Q9BYF1 ( <i>ACE2</i> )
adipose tissue	0.2	4.5	monocytes	0	0
adrenal gland	0.1	0.4	NK cells	0	0
amygdala	0.2	0.2	olfactory region	0.2	0.2
appendix	15.5	0.8	ovary	0.7	1.3
B cells	0	0	pancreas	64.5	1.6
basal ganglia	0.4	0.2	parathyroid gland	0	0
bone marrow	0	0	pituitary gland	1.3	0.2
breast	6.1	2.3	placenta	1.0	1.0
cerebellum	0.2	0.2	pons and medulla	0.4	0.2
cerebral cortex	0.5	0.2	prostate	68.2	0.5
cervix uterine	1.3	0.4	rectum	18.5	1.3
colon	38.7	49.1	retina	0.3	0.2
corpus callosum	0.4	0.2	salivary gland	52.3	1.1
dendritic cells	0	0	seminal vesicle	35.7	1.2
ductus deferens	24.4	2.3	skeletal muscle	0.2	0.7
duodenum	17.5	46.0	skin 1	2.1	0.2
endometrium 1	0.5	0.4	small intestine	75.6	122.0
epididymis	0.9	2.7	smooth muscle	0.9	0.3
esophagus	10.7	1.2	spinal cord	0.2	0.2
fallopian tube	2.2	0.6	spleen	0.3	0.2
gallbladder	17.3	16.4	stomach 1	36.7	0.5
granulocytes	0	0.2	T cells	0	0.3
heart muscle	0.2	10.5	testis	0.8	17.9
hippocampal formation	0.3	0.2	thalamus	0.3	0.2
hypothalamus	0.1	0.1	thymus	0.5	0.2
kidney	19.4	23.2	thyroid gland	19.9	4.5
liver	18.4	1.2	tongue	3.9	0.5
lung	20.7	0.8	tonsil	10.2	0.2
lymph node	0.2	0.6	total PBMC	0.4	0
midbrain	0.2	0.2	urinary bladder	5.9	0.4
vagina	4.3	0.9			



## DISCUSSION AND CONCLUSION

As of May 2020, SARS-CoV-2 has infected millions of people around the world and has caused over 300,000 deaths. However, the true number of infected individuals remains unknown, as studies have suggested that many remain asymptomatic or have a mild symptoms (Day, 2020), (Song et al.). Naturally occurring genetic variations that result in a defective *TMPRSS2* may explain why some individuals with COVID-19 develop mild disease. Our *in silico* analysis of *TMPRSS2* human variants shows that the predicted damaging substitution from valine to methionine at position 160 is a common genetic variant, present in almost 25% percent of the human population, with approximately 7% homozygotes according to GnomAd. *TMPRSS2* variants should be investigated further to understand the impact of a person's genetic background on their clinical presentation and prognosis when contracting SARS-CoV-2 .

Once SARS-CoV-2 infects an individual, it binds to ACE2 receptors on the surface of target cells. However, for the virus to enter the cells, double cleavage of the viral spike protein at the S1/S2 cleavage site and, subsequently, at the S2' site is required. This allows viral fusion with the cell membrane and internalization (Hoffmann et al., 2020). *TMPRSS2* is one the main cell surface proteases involved in the process of spike protein priming, although additional proteases, such as Furin and lysosomal cathepsin, are thought to be involved (Shang et al., 2020).

At present no known disease-association for *TMPRSS2* variants is known. A chromosomal aberration leading to gene fusion of *TMPRSS2* and its androgen promoter to *ERG*, known as the *TMPRSS2-ERG* fusion gene, has been identified in approximately one half of patients with prostate cancer (Haffner et al., 2010). Genetic variants leading to a defective and non-functioning *TMPRSS2* and *ACE2* are promising prognostic candidates for COVID-19 infection. Recent studies show that *ACE2* genetic variation is very rare in the population (Stawiski et al.) (MacGowan and Barton), thus making it an unlikely candidate to explain the wide range of symptoms (from asymptomatic to severely affected) observed in COVID-19 patients. We, therefore, focused on *TMPRSS2*, which together with *ACE2* plays an important role in SARS-CoV-2 infection. Although the majority of *TMPRSS2* variants are rare, the common variant V160M, which is predicted damaging, is an intriguing candidate for further study. Valine 160 is a highly conserved residue in the SRCS domain. It has been proposed that the latter may be involved in protein and/or ligand binding (Aruffo et al., 1997), however no known protein interaction partner is known to-date. The SRCR domain is often found in proteins involved in host defence (Freeman et al.,



1990), (Resnick et al., 1994), thus raising the possibility that the role of *TMPRSS2* in SARS-CoV-2 infection may extend beyond its peptidase activity related to viral protein priming.

Although respiratory tract infection is the most common manifestation of COVID-19, diarrhoea and gastrointestinal symptoms are also common. The co-expression of *ACE2* and *TMPRSS2* has been studied and experimentally ascertained in bronchial and lung cells, however co-expression of these proteins in the gut has not yet been investigated. We, therefore, used RNA expression data from the Human Protein Atlas (Uhlén et al., 2015) to show that *ACE2* and *TMPRSS2* are both expressed in the intestine. This suggests that, similarly to the lung, cells in the intestine may be susceptible to SARS-CoV-2 infection, resulting in gastro-intestinal symptoms in many patients, including children presenting with Kawasaki-like syndrome. This appealing pathogenetic hypothesis needs further *ad hoc in vitro* confirmation. It is, in fact, notable that the lung expression of *ACE2*, confirmed by dedicated *in vitro* studies (Lukassen et al., 2020), and its expression in arterial and venous endothelial cells and arterial smooth muscle cells (Hamming et al., 2004) are not reported in the HPA, thus suggesting that HPA data are incomplete.

In conclusion, SARS-CoV-2 is a new virus and still very little is known about the pathogenesis of COVID-19. Elucidating the molecular processes underlying SARS-CoV-2 infection and the factors responsible for the broad spectrum of disease severity observed in COVID-19 patients is crucial. Our data suggest that (i) *TMPRSS2* variants, particularly p.V160M which has a MAF of 0.25, should be investigated as a marker of disease prognosis and (ii) *in vitro* validation of the co-expression of *TMPRSS2* and *ACE2* in gastrointestinal cells is needed.

## REFERENCES

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249.
- Afar DE, Vivanco I, Hubert RS, Kuo J, Chen E, Saffran DC, Raitano AB, Jakobovits A. 2001. Catalytic cleavage of the androgen-regulated TMPRSS2 protease results in its secretion by prostate and prostate cancer epithelia. *Cancer Res* 61:1686–1692.
- Aruffo A, Bowen MA, Patel DD, Haynes BF, Starling GC, Gebe JA, Bajorath J. 1997. CD6-ligand interactions: a paradigm for SRCR domain function? *Immunol Today* 18:498–504.
- Böttcher E, Matrosovich T, Beyerle M, Klenk H-D, Garten W, Matrosovich M. 2006. Proteolytic activation of influenza viruses by serine proteases TMPRSS2 and HAT from human airway epithelium. *J Virol* 80:9896–9898.
- Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, et al. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395:507–513.
- Day M. 2020. Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* 369:m1375.
- Du L, Yang Y, Zhou Y, Lu L, Li F, Jiang S. 2017. MERS-CoV spike protein: a key target for antivirals. *Expert Opin Ther Targets* 21:131–143.
- Freeman M, Ashkenas J, Rees DJ, Kingsley DM, Copeland NG, Jenkins NA, Krieger M. 1990. An ancient, highly conserved family of cysteine-rich protein domains revealed by cloning type I and type II murine macrophage scavenger receptors. *Proc Natl Acad Sci USA* 87:8810–8814.
- González-Pérez A, López-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88:440–449.
- Haffner MC, Aryee MJ, Toubaji A, Esopi DM, Albadine R, Gurel B, Isaacs WB, Bova GS, Liu W, Xu J, Meeker AK, Netto G, et al. 2010. Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat Genet* 42:668–675.
- Hamming I, Timens W, Bulthuis MLC, Lely AT, Navis GJ, Goor H van. 2004. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J Pathol* 203:631–637.
- Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, et al. 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181:271-280.e8.
- Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. 2019. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol* 431:2197–2212.

- 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. *J Virol* 93:.
- Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D, The Exome Aggregation Consortium, et al. 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 45:D840–D845.
- Kawase M, Shirato K, Hoek L van der, Taguchi F, Matsuyama S. 2012. Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry. *J Virol* 86:6537–6545.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10:845–858.
- Lin B, Ferguson C, White JT, Wang S, Vessella R, True LD, Hood L, Nelson PS. 1999. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2. *Cancer Res* 59:4180–4184.
- Lukassen S, Chua RL, Trefzer T, Kahn NC, Schneider MA, Muley T, Winter H, Meister M, Veith C, Boots AW, Hennig BP, Kreuter M, et al. 2020. SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. *EMBO J* e105114.
- MacGowan S, Barton GJ. Missense variants in ACE2 are predicted to encourage and inhibit interaction with SARS-CoV-2 Spike and contribute to genetic risk in COVID-19.
- Ofoegbu TC, David A, Kelley LA, Mezulis S, Islam SA, Mersmann SF, Strömich L, Vakser IA, Houlston RS, Sternberg MJE. 2019. PhyreRisk: A Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants. *J Mol Biol* 431:2460–2466.
- Olechnovič K, Venclovas Č. 2017. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* 85:1131–1145.
- Resnick D, Pearson A, Krieger M. 1994. The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends Biochem Sci* 19:5–8.
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res* 33:W382-388.
- Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. 2020. Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci USA*.
- Song, Qui, Yin. A considerable proportion of individuals with asymptomatic SARS-CoV-2 infection in Tibetan population.
- Stawiski E, Diwanji D, Suryamohan K, Gupta R. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility.
- Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. 2020. QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics* 36:1765–1771.



The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419.

Vaarala MH, Porvari KS, Kellokumpu S, Kyllönen AP, Vihko PT. 2001. Expression of transmembrane serine protease TMPRSS2 in mouse and human tissues. *J Pathol* 193:134–140.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc* 11:1–9.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, Beer TAP de, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46:W296–W303.

Wiederstein M, Sippl MJ. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35:W407–410.

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579:270–273.

Zhou Y, Vedantham P, Lu K, Agudelo J, Carrion R, Nunneley JW, Barnard D, Pöhlmann S, McKerrow JH, Renslo AR, Simmons G. 2015. Protease inhibitors targeting coronavirus and filovirus entry. *Antiviral Res* 116:76–84.