




RESEARCH NOTE

Assessment of risk conferred by coding and regulatory variations of *TMPRSS2* and *CD26* in susceptibility to SARS-CoV-2 infection in human

SABYASACHI SENAPATI^{1*} , SHASHANK KUMAR², ATUL K. SINGH², PRATIBHA BANERJEE¹ and SANDILYA BHAGAVATULA¹

¹Department of Human Genetics and Molecular Medicine, School of Health Sciences, Central University of Punjab, Bathinda 151 001, India

²Department of Biochemistry, School of Basic and Applied Sciences, Central University of Punjab, Bathinda 151 001, India

*For correspondence. E-mail: sabyasachi1012@gmail.com.

Received 13 April 2020; revised 10 May 2020; accepted 11 May 2020; published online 9 June 2020

Abstract. At present, more than 200 countries and territories are directly affected by the coronavirus disease-19 (COVID-19) pandemic. Incidence and case fatality rate are significantly higher among elderly individuals (age > 60 years), type 2 diabetes and hypertension patients. Cellular receptor ACE2, serine protease *TMPRSS2* and exopeptidase *CD26* (also known as DPP4) are the three membrane bound proteins potentially implicated in SARS-CoV-2 infection. We hypothesised that common variants from *TMPRSS2* and *CD26* may play critical role in infection susceptibility of predisposed population or group of individuals. Coding (missense) and regulatory variants from *TMPRSS2* and *CD26* were studied across 26 global populations. Two missense and five regulatory SNPs were identified to have differential allelic frequency. Significant linkage disequilibrium (LD) signature was observed in different populations. Modelled protein–protein interaction (PPI) predicted strong molecular interaction between these two receptors and SARS-CoV-2 spike protein (S1 domain). However, two missense SNPs, rs12329760 (*TMPRSS2*) and rs1129599 (*CD26*), were not found to be involved physically in the said interaction. Four regulatory variants (rs112657409, rs11910678, rs77675406 and rs713400) from *TMPRSS2* were found to influence the expression of *TMPRSS2* and pathologically relevant *MX1*. rs13015258 a 5' UTR variant from *CD26* have significant role in regulation of expression of key regulatory genes that could be involved in SARS-CoV-2 internalization. Overexpression of *CD26* through epigenetic modification at rs13015258-C allele was found critical and could explain the higher SARS-CoV-2 infected fatality rate among type 2 diabetes.

Keywords. coronavirus disease-19; severe acute respiratory syndrome coronavirus-2; dipeptidyl peptidase-4; transmembrane protease serine 2; spike protein; type-2 diabetes.

Introduction

Coronavirus disease-19 (COVID-19) outbreak is caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) at present has jolted the entire human race. COVID-19 pandemic has affected natives in more than 200 countries and territories located in different geographical locations and climatic conditions. According to World Health Organization (WHO), nearly 3.8 million individuals have already

Atul K. Singh, Pratibha Banerjee and Sandilya Bhagavatula contributed equally to this work. SS conceptualized and designed the study. AKS performed protein bioinformatics pipeline. PB and SB performed the genetics data mining. SS interpreted the genetic findings; SK interpreted the bioinformatics findings and contributed in writing. SS compiled the findings and wrote the manuscript. All the authors reviewed and approved the final manuscript.

Electronic supplementary material: The online version of this article (<https://doi.org/10.1007/s12041-020-01217-7>) contains supplementary material, which is available to authorized users.

been infected globally and more than 260 thousands have died till date (WHO COVID-19 Situation Report-109, 08/5/2020). Recent trend indicates much lower overall case fatality rate (CFR) in COVID-19 pandemic (4.58%) compared to outbreak of SARS (11%) and Middle East respiratory syndrome (MERS) (35%) (Peeri *et al.* 2020). The incidence and CFR varies drastically in different populations, where the present global CFR (6.9%) is significantly lower than Europe (9.2%) and considerably higher than Africa (3.4%), East Mediterranean and Southeast Asia (3.6%), Western Pacific (4.1%) and America (5.5%) (WHO COVID-19 Situation Report-109, 08/5/2020). Recent epidemiological data showed the association of hypertension, and type 2 diabetes with the incidence and case fatality of SARS-CoV-2 infection (Huang *et al.* 2020; Wu *et al.* 2020).

Spike protein (S-protein) of coronaviruses interacts with specific membrane receptors for their entry into the human cells. S-protein of SARS-CoV-2 has conserved tertiary folds that are involved in interaction with human receptors, but has limited sequence similarity with previously known coronaviruses (such as SARS-CoV, MERS-CoV) (Xu *et al.* 2020). Multiple host receptors, such as angiotensin converting enzyme 2 (ACE2), transmembrane serine protease 2 (TMPRSS2) and dipeptidyl peptidase-4, DPP4 (CD26) are known to help in priming of S protein and subsequent entry of virus particle in the host cell (Li *et al.* 2005; Hoffmann *et al.* 2016; Cascella *et al.* 2020; Ibrahim *et al.* 2020; Vankadari and Wilce 2020). However, specific host cell factor that facilitate SARS-CoV-2 entry into the human cell is still elusive. Structural or expression variations in these cellular receptors (host factor) could influence the cell adhesion, cellular entry and virulence of SARS-CoV-2. Population level differences in rate of incidence and CFR could be attributed by the functional genetic variations in these receptors. Therefore, detailed genetic analyses of these receptors are warranted for future epidemiological, molecular and pharmaceutical research to tackle SARS-CoV-2 or related coronavirus outbreak.

A recent comparative genetic analysis of ACE2 receptor was unable to confirm population specific resistance to SARS-CoV-2 infection, but showed promise for further investigation (Cao *et al.* 2020). Here for the first time we report a comparative genetic study on *TMPRSS2* and *CD26*, and their predicted molecular interaction with SARS-CoV-2 spike protein.

Methods and results

Common genetic variants localized in the coding (nonsynonymous) and regulatory regions of *TMPRSS2* and *CD26* were evaluated. Genotypes of these variants in 26 populations from five major global regions were obtained from 1000 Genomes database (<https://www.internationalgenome.org>) and used for comparative analysis. A common missense variant rs12329760: C>T (Chr21:41480570), was found to

have noticeable variations in allele frequency in different populations (table 1 in electronic supplementary material). Highest frequency of minor T allele was observed among Chinese (CHB = 0.41 and CHS = 0.38) and Japanese (JPT = 0.39) populations. Considerably lower allelic frequency was observed in other major populations (table 1 in electronic supplementary material). Regional linkage disequilibrium (LD) analyses showed the presence of very close and high range LD and haplotype patterns in CHS compared to other populations (figure 1, a–e in electronic supplementary material). Fourteen common (allelic freq > 0.01) regulatory variants are present at the regulatory regions of *TMPRSS2*. Four of these variants (rs112657409, rs11910678, rs77675406 and rs713400) are independent of the nonsynonymous SNP rs12329760 and have significant eQTL effects on *MX1* and *TMPRSS2* in different tissues (table 2 in electronic supplementary material). GTEx database was used to evaluate the eQTL (Aguet *et al.* 2019). *MX1* encodes a GTP metabolizing protein which participates in cellular antiviral response by antagonizing the replication of several RNA and DNA viruses (Jung *et al.* 2019). Region 5'-flanking of *TMPRSS2* harbours SNP rs713400 beyond a CpG island (CG:156) and was found to influence the expression of *TMPRSS2* (table 2 in electronic supplementary material). rs713400-T allelic frequency is considerably high among east Asians (avg_freq = 0.28) compared to Europeans (avg_freq = 0.11), South Asians (avg_freq = 0.13), Africans (avg_freq = 0.02) and Americans (avg_freq = 0.11) (table 2 in electronic supplementary material).

Another nonsynonymous variations rs1129599 in *CD26* with moderate effects on protein structure/function was found to be population specific (table 1 in electronic supplementary material). This relatively rare variation is only present among Africans (avg_freq = 0.04) and absolutely monomorphic in rest of the world. Like *TMPRSS2*, variants around this region of *CD26* have very high degree of short range LD among southern Han Chinese population (CHS), which is otherwise absent in other major populations (figure 1, f–j in electronic supplementary material).

Regulatory SNP rs13015258 (G>T) at the exon 1 start site (chr2:162930725) of *CD26* was found to have significant eQTL effect ($P = 2.50E-07$) on the expression of *CD26* in lung tissue (table 2 in electronic supplementary material). This site falls within a 646 nucleotide long CpG island (figure 2 in electronic supplementary material) and hypermethylation of C allele (complementary to G) of this SNP was identified to significantly ($P = 0.001$) repress the expression of *CD26* in human visceral adipose tissue (Turcot *et al.* 2011). Presence of C allele further promotes the binding of several transcription factors (table 2 in electronic supplementary material) and regulates the self-expression. *CD26* is a cell-surface protease expresses in variety of tissues including specific sets of T-cells, adipose tissues, endothelial and epithelial cells. Its soluble form is also present in plasma and body fluid. Its role in glucose metabolism is well established. Significantly higher expression of

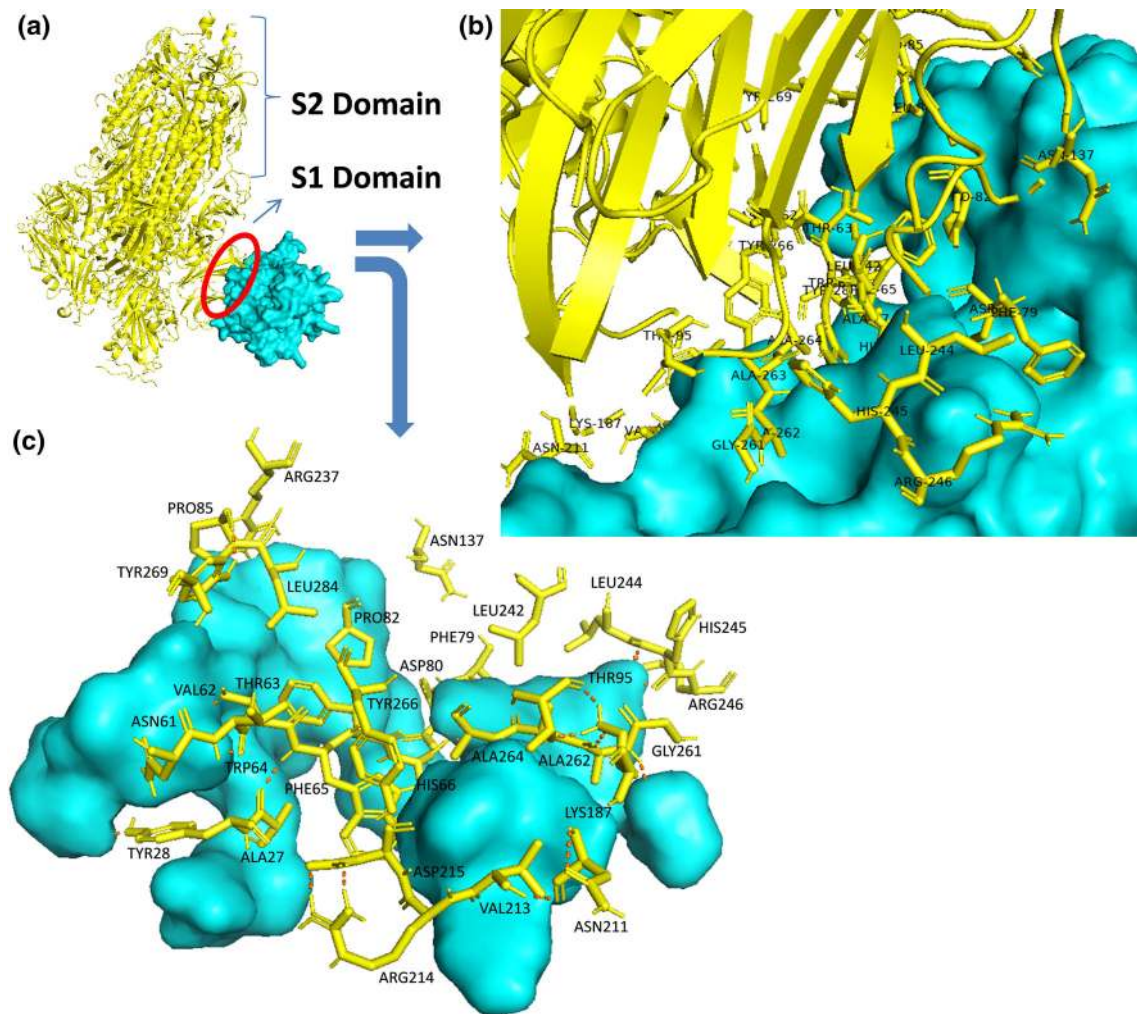


Figure 1. (a) Docking results showing protein–protein interaction between SARS-CoV-2 (yellow) (PDB:6VSB) and human TMPRSS2 modelled protein (cyan); (b) ribbon and surface structure diagram showing magnified protein–protein (SARS-CoV-2 and TMPRSS2) interaction region; (c) surface diagram showing magnified protein–protein (SARS-CoV-2 and TMPRSS2) interaction region with interacting amino acid residues. SARS-CoV-2 amino acid back bone is shown in yellow colour and TMPRSS2 surface structure is depicted in cyan colour.

CD26 was found associated with type 2 diabetes (Qiao *et al.* 2019). Frequency of G/T allele of rs13015258 varies considerably among different populations, where G is the major allele among Africans, Europeans and South Asians; but minor allele for Americans and East Asians (table 2 in electronic supplementary material).

Recently, Vankadari and Wilce (2020) predicted the homo-trimer structure of SARS-CoV-2 spike glycoprotein (modelled protein) and its interaction with human CD26 protein (crystalized). We hereby, for the first time report the computer-based interaction of crystalized SARS-CoV-2 spike glycoprotein with human CD26 protein. Moreover, the study also reports protein–protein interaction between SARS-CoV-2 spike glycoprotein and TMPRSS2 modelled protein.

PDB structure of SARS-CoV-2 spike glycoprotein (PDB:6VSB) of coronavirus and human CD26 receptor (PDB:4QZV) were retrieved from the Protein Data Bank

(PDB). Amino acid sequence of human TMPRSS2 protein was retrieved from Uniprot Database. Three dimensional structure of TMPRSS2 protein was modelled using Swiss-model structural bioinformatics server followed by refinement of the structure using ModRefiner sever (figure 3, a–c in electronic supplementary material). In human, the type-II transmembrane serine proteases (11 in number) are divided into four subfamilies. Hepsin (also known as TMPRSS1) and TMPRSS2 belong to the same subfamily (Hepsin/TMPRSS) and share similar type of extracellular carboxy-terminus protease domain (Sakai *et al.* 2014; Mukai *et al.* 2020). Thus, serine protease Hepsin (PDB:5CEL) was used as template for the TMPRSS2 protein homology modelling. Protein–protein docking of SARS-CoV-2 spike glycoprotein with human TMPRSS2 modelled and CD26 crystalized protein was performed using ClusPRO server (Kozakov *et al.* 2017). LigPlot⁺ v2.2 software was used to find the type of interaction among interacting proteins (Wallace *et al.*

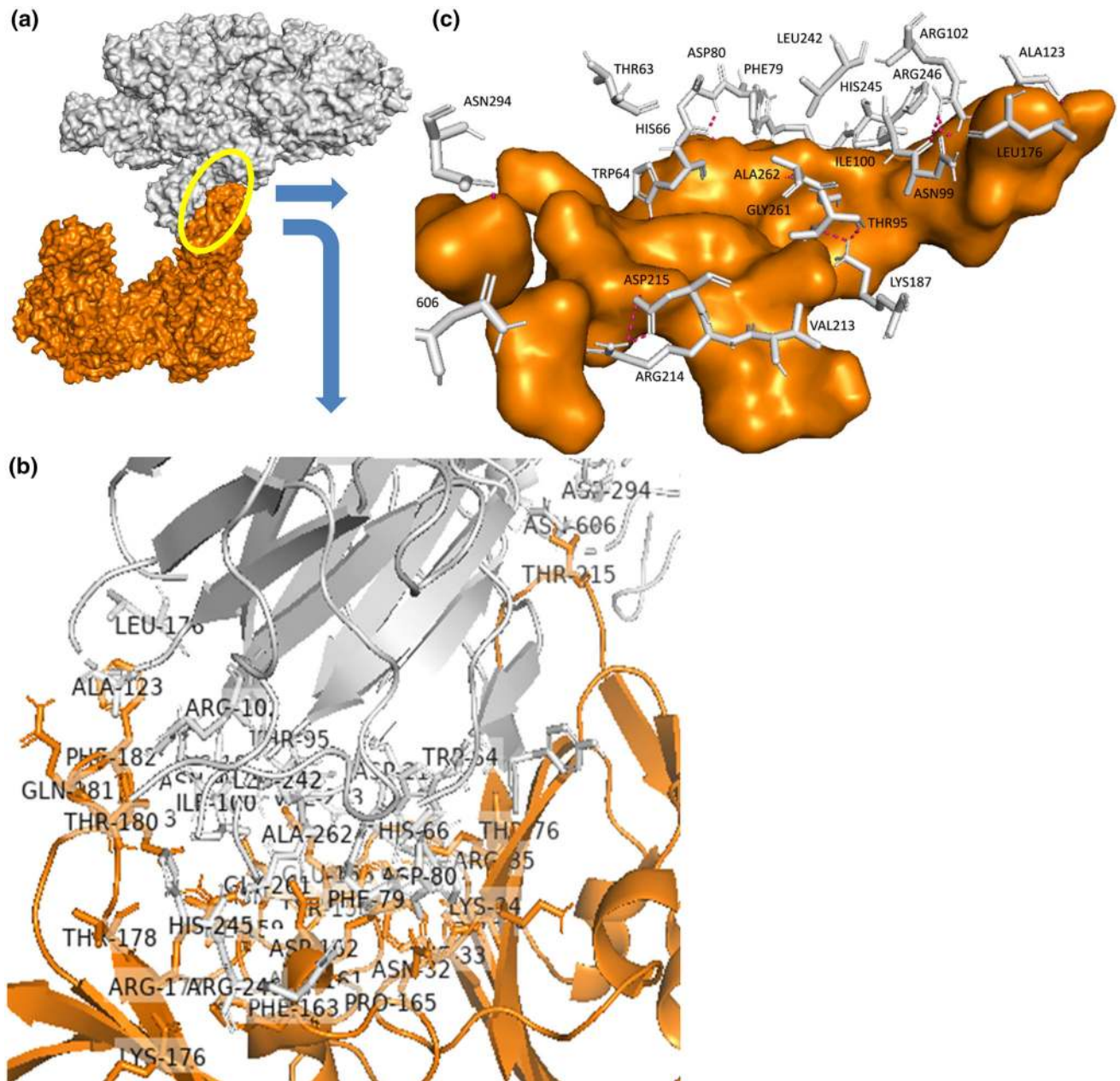


Figure 2. (a) Surface structure showing protein–protein interaction between SARS-CoV-2 (grey) (PDB: 6VSB) and human CD26 (orange) (PDB: 4QZV); (b) ribbon diagram showing magnified protein–protein (SARS-CoV-2 and CD26) interaction region; (c) surface diagram showing magnified protein–protein (SARS-CoV-2 and CD26) interaction region with interacting amino acid residues.

1995). Structure representation and visualization was performed by using PyMol software (DeLano 2002). Docking of SARS-CoV-2 spike glycoprotein with TMPRSS2 modelled protein (figure 1), and CD26 (figure 2) showed a large interface between the proteins. Amino acid residues involved in these interactions can be found in table 3 in electronic supplementary material. None of the missense variants, namely rs12329760 of *TMPPRSS2* and rs1129599 of *CD26* were found directly engaged in the protein–protein interaction with S1 domain of the viral spike protein. However, their indirect role in influencing this protein–protein interaction is beyond the scope of this study.

Discussion

Following the COVID-19 outbreak, this is the first report on the assessment of genetic susceptibility of *TMPPRSS2* and *CD26* (*DPP4*) for the SARS-CoV-2 infection. Based on the *in silico* prediction, in this study we also confirmed the molecular interactions between SARS-CoV-2 spike protein and human *TMPPRSS2* and *CD26/DPP4*. This study highlighted the differential allelic frequencies of two common missense variations from *TMPPRSS2* (rs12329760) and *CD26* (rs1129599) in different global populations. Noticeable LD differences in these loci probably indicated presence of

different haplotypes that could influence the overall receptor function. These two SNPs are not located within the receptor-ligand (S1 domain of SARS-CoV-2) binding site. Further study is warranted to find their effect on protein-protein interaction (PPI) dynamics, protein structure stability and turnover. Four regulatory SNPs from *TMPRSS2* (rs112657409, rs11910678, rs77675406 and rs713400) and one from *CD26* (rs13015258) have significant role in regulation of expression of key regulatory genes (*TMPRSS2*, *CD26* and *MXI*) that could be involved in SARS-CoV-2 infection. Epigenetic modification at rs13015258-C allele induces *CD26* overexpression which could explain the higher SARS-CoV-2 infected fatality rate among type 2 diabetes. Preliminary *in silico* predictions of interactions between *TMRSS2* and *CD26* with SARS-CoV-2 S-protein need to be confirmed by detailed molecular experiments. Findings from this study would guide further genetic epidemiological study and drug development or drug repurposing to tackle COVID-19.

Acknowledgement

We acknowledge the funding support from DST-SERB, New Delhi, Government of India (ECR/2016/001660), and UGC-BSR (F.30/2014-BSR) to SS. SK is supported by research funds from DST-SERB (EEQ/2016/000350), and UGC-BSR (F.30/372/2017-BSR). AKS (JRF) is supported by CSIR, India.

References

- Aguet F., Barbeira A. N., Bonazzola R., Brown A., Castel S. E., Jo B. *et al.* 2019 The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* (<https://doi.org/10.1101/787903>).
- Cao Y., Li L., Feng Z., Wan S., Huang P., Sun X. *et al.* 2020 Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, 1–4.
- Cascella M., Rajnik M., Cuomo A., Dulebohn S. C. and Di Napoli R. 2020 Features, evaluation and treatment coronavirus (COVID-19). *Statpearls* [internet], available at (<https://www.ncbi.nlm.nih.gov/books/NBK554776/>).
- DeLano W. L. 2002 PyMOL: an open-source molecular graphics tool. *Ccp4 Newsl. Protein Crystallogr.* **40**, 11.
- Hoffmann M., Krüger N., Zmora P., Wrensch F., Herrler, G. and Pöhlmann, S. 2016. The hemagglutinin of bat-associated influenza viruses is activated by *TMPRSS2* for pH-dependent entry into bat but not human cells. *PLoS one* **11**(3), e0152134. <https://doi.org/10.1371/journal.pone.0152134>.
- Huang C., Wang Y., Li X., Ren L., Zhao J., Hu Y. *et al.* 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506.
- Ibrahim I. M., Abdelmalek D. H., Elshahat M. E. and Elfiky A. A. 2020 COVID-19 spike-host cell receptor GRP78 binding site prediction. *J. Infection* (<https://doi.org/10.1016/j.jinf.2020.02.026>).
- Jung H. E., Oh J. E. and Lee H. K. 2019 Cell-penetrating Mx1 enhances anti-viral resistance against mucosal influenza viral infection. *Viruses* **11**, 109.
- Kozakov D., Hall D. R., Xia B., Porter K. A., Padhorny D., Yueh C. *et al.* 2017 The ClusPro web server for protein-protein docking. *Nat. Protoc.* **12**, 255.
- Li W., Zhang C., Sui J., Kuhn J. H., Moore M. J., Luo S. *et al.* 2005 Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* **24**, 1634–1643.
- Mukai S., Yamasaki K., Fujii M., Nagai T., Terada N., Kataoka H. *et al.* 2020 Dysregulation of Type II transmembrane serine protease and ligand-dependent activation of MET in urological cancers. *Int. J. Mol. Sci.* **21**, E2663.
- Peeri N. C., Shrestha N., Rahman M. S., Zaki R., Tan Z., Bibi S. *et al.* 2020 The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned?. *Int. J. Epidemiol.* (<https://doi.org/10.1093/ije/dyaa033>).
- Qiao J., Li L., Ma Y., Shi R. and Teng M. 2019 Biological function of dipeptidyl peptidase-4 on type 2 diabetes patients and diabetic mice. *Curr. Res. Transl. Med.* **67**, 89–92.
- Sakai K., Ami Y., Tahara M., Kubota T., Anraku M., Abe M. *et al.* 2014 The host protease *TMPRSS2* plays a major role in *in vivo* replication of emerging H7N9 and influenza viruses. *J. Virol.* **88**, 5608–5616.
- Turcot V., Bouchard L., Faucher G., Tchernof A., Deshaies Y., Pérusse L. *et al.* 2011 DPP4 gene DNA methylation in the omentum is associated with its gene expression and plasma lipid profile in severe obesity. *Obesity* **19**, 388–395.
- Vankadari N. and Wilce J. A. 2020 Emerging COVID-19 coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human *CD26*. *Emerg. Microbes Infect.* **9**, 601–604.
- Wallace A. C., Laskowski R. A. and Thornton J. N. 1995 LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134.
- Wu C., Chen X., Cai Y., Zhou X., Xu S., Huang H. *et al.* 2020 Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. *JAMA Intern. Med.* <https://doi.org/10.1001/jamainternmed.2020.0994>.
- Xu X., Chen P., Wang J., Feng J., Zhou H., Li X. *et al.* 2020 Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* **63**, 457–460.

Corresponding editor: MANOJ PRASAD