Ethnicity-dependent allele frequencies are correlated with COVID-19 case fatality rate

Sungwon Jeon[1], Asta Blazyte[1], Changhan Yoon[1], Hyojung Ryu[1], Yeonsu Jeon[1], Youngjune Bhak[1], Dan Bolser[2], Andrea Manica[3], Eun-Seok Shin[4], Yun Sung Cho[5], Byung Chul Kim[5], Namhee Ryoo[6], and Jong Bhak[1]

[1]UNIST
[2]Geromics Ltd.
[3]University of Cambridge
[4]Ulsan Medical Center
[5]Clinomics Inc.
[6]Keimyung University School of Medicine

October 16, 2020

Abstract

Coronavirus disease (COVID-19), caused by SARS-CoV-2, has a higher case fatality rate (CFR) in European ethnic groups than in others, especially East Asians. One explanation to this phenomenon might be TMPRSS2, a key processing enzyme essential for viral infection. Here, we analyzed the allele frequencies of two nonsynonymous variants rs12329760 (V197M) and rs75603675 (G8V) in the TMPRSS2 gene using over 200,000 present-day and ancient genomic samples. We found a significant association between the CFR of COVID-19 and the allele frequencies of the two variants. Interestingly, they had opposing effects on the CFR: inverse correlation by V197M, proportional correlation by G8V. East Asians have higher V197M and lower G8V allele frequencies than Europeans, possibly endowing resistance against SARS-CoV-2. Structural and energy calculation analysis of the V197M amino acid change showed that it destabilizes the TMPRSS2 protein, possibly affecting its ACE2 and viral spike protein processing negatively, ultimately resulting in reduced SARS-CoV-2 infection efficiency and CFR in East Asian ethnic groups.

Acknowledgements

This study was a part of Korean Genome Project (KGP) and was approved by the Institutional Review Board (IRB) of the Ulsan National Institute of Science and Technology (UNISTIRB-15-19-A, UNISTIRB-16-13-C). This work was supported by the Research Project Funded by Ulsan City Research Fund (1.200047.01) of UNIST. This work was also supported by the U-K BRAND Research Fund (1.200108.01) of UNIST(Ulsan National Institute of Science & Technology). We thank Dr. Seung Gu Park for advising the data visualization and Jasmin Junseo Lee for editing grammatical errors.

Competing Interests

The authors declare the following competing interests: Y.S.C is an employee and B.C.K. and J.B. are the CEOs of Clinomics Inc. Y.S.C., B.C.K., and J.B. have an equity interest in the company. D.B. is an employee and J.B. is the CEO of Geromics Ltd. The rest of the authors declare they have no competing interests.

Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Appearing first during late 2019 in Wuhan, China, COVID-19 has spread rapidly worldwide(1). As of May 23, 2020, SARS-CoV-2 has infected >5 million people in over 200 countries, killing more than 330,000 people(2). Europe has been particularly affected, with Spain and Italy each reaching over 200,000 cases of infection and more than 27,000 deaths, resulting in a maximum case fatality rate (CFR) of >10%(2). In contrast, East Asia did not experience such dire effects, with South Korea, for instance, reporting a peak CFR of 2.4%[(2)]. Multiple contributing factors could explain this difference, including timing and severity of lockdown measures[(3)], population age ratio[(4)], healthcare resource availability[(5)], smoking rate(6, 7), and early tuberculosis (Bacillus Calmette–Guérin) vaccination(8-10). In principle, genetic factors may also underpin differential susceptibility to SARS-CoV-2(11-13).

Genes encoding cellular serine protease (*TMPRSS2*), angiotensin-converting enzyme 2 (*ACE2*), cysteine proteases cathepsin B and cathepsin L (*CatB, CatL*), phosphatidylinositol 3-phosphate 5-kinase (*PIKfyve*), and two pore channel subtype 2 (*TPC2*) are notable for their critical roles in SARS-CoV-2 infection(14, 15). Particularly, the virus utilizes TMPRSS2 and CatB/L proteolytic activity for priming the viral spike protein, whereas ACE2 is the entry receptor for breaking into host cells(14, 15). A study has suggested TMPRSS2 inhibition as a clinical target because the priming step is a key factor determining successful entry into target cells(15). Most of the recent publications on the SARS-CoV-2 susceptibility so far focused on ACE2 and TMPRSS2 as possible genetic determinants by analyzing their associations with sex hormons, their gene expression in various tissues and cell lines, and interactions with spike protein or inhibitors at a gene level(15-19).

To understand the genetic background of complex phenotypes in human populations, researchers commonly assess correlations with allele frequency (AF)[(20, 21)]. This approach has identified a correlation between ancestral genetic composition and the CFR of COVID-19[(21)]. However, few have examined specific variants, their frequencies and individual contributions to SARS-CoV-2 susceptibility. Some reports are also based only on low-resolution intercontinental comparisons between Europeans and East Asians[(20-22)]. Based on these studies, not only do*TMPRSS2* variants appear to have wide population-specific variation(20), but, *TMPRSS2* also has low mutation burden in certain populations, a characteristic that could partially explain high*TMPRSS2* gene expression. Consequently, the latter is associated with a poor outcome in COVID-19[(20)]. Moreover, we know little about the evolutionary history of SARS-CoV-2 susceptibility-associated variants, including when they occurred or how their frequencies might have changed over time.

In this study, we investigated intercountry AF differences of*TMPRSS2* variants, estimated variant effects on TMPRSS2 protein structure stability, and linked them to the average of time-adjusted COVID-19 CFR (AT-CFR). We propose that the structural deviation causes TMPRSS2 to be less stable, resulting in a reduced overall infection rate that led to reduced CFR in East Asians. We collected and analyzed 221,498 genomes from public databases[(23-25)] and 2,262 whole genomes from the Korean Genome Project[(26)]. We also traced *TMPRSS2* AF distribution in ancient populations by region and time period. We aimed to increase the current understanding of the genetic variation underlying SARS-CoV-2 infections and explain the ethnic differences in CFR.

Results

Correlation of nonsynonymous *TMPRSS2* allele frequencies with COVID-19 AT-CFR

The AFs of two nonsynonymous *TMPRSS2* variants (G8V, rs75603675; V197M, rs12329760) were significantly correlated with COVID-19 AT-CFR (Spearman's correlation $\rho = 0.713$, $P = 0.00183$ for G8V and$\rho =$ -0.464, $P = 0.0157$ for V197M, Fig. 1a and b). The results are based on AF data from 17 and 27 countries, respectively (see Methods). However, the AFs of the two variants were not significantly correlated with total COVID-19 cases per million individuals (V197M:$P = 0.132$, G8V: $P = 0.165$; Supplementary Fig. 1). These two nonsynonymous variants were present among 20 *TMPRSS2* exonic variants with global AF of >1% from gnomAD(27). Thirteen of these were in 3' UTR and five were synonymous (Supplementary Fig. 2, Supplementary Data 1). G8V is located in a cytoplasmic domain with an undetermined 3D structure

(Supplementary Fig. 3). V197M is located in a stable beta-sheet of the scavenger receptor cysteine-rich (SRCR) domain (Supplementary Fig. 3).

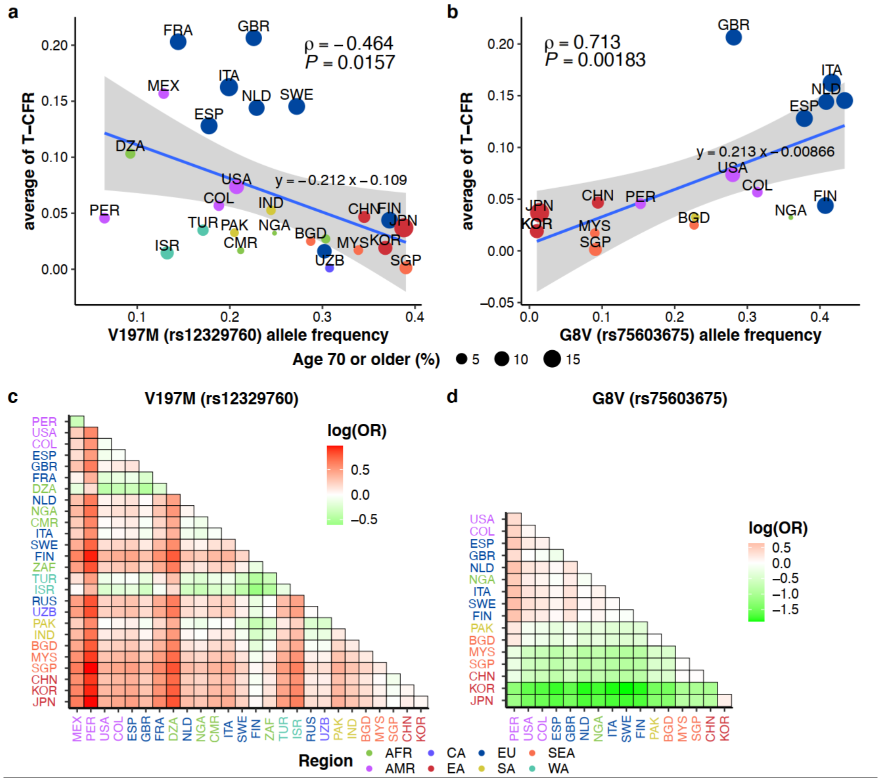


Fig. 1. Correlation plots of COVID-19 average of time-adjusted case fatality rate (AT-CFR) with allele frequencies of *TMPRSS2*variants. **a)** V197M (rs12329760) from 27 countries and**b)** G8V (rs75603675) from 17 countries. The size of dots indicates the proportion of people who are 70 or older in the countries. The correlations were estimated by Spearman's correlation test. Allelic odds ratios (ORs) (i.e., alternative/reference allele counts) of the Y-axis country to the X-axis country are presented in **c)** for V197M and **d)** for G8V. AFR: Africa, CA: Central Asia, EU: Europe, SEA: Southeast Asia, AMR: Americas, EA: East Asia, SA: South Asia, and WA: West Asia. Full country names and allele frequencies per country are in Supplementary Data 3.

Correlation between *TMPRSS2* V197M allele frequency and COVID-19 AT-CFR

The AF of V197M was negatively correlated with COVID-19 AT-CFR (Spearman's correlation coefficient, ρ = -0.464, P = 0.0157) (Fig. 1a). The AF distribution pattern was consistent with previous reports, with V197M AF being significantly lower in most Europeans than in East Asians(20) (Fig. 1a and c Supplementary Fig. 4a, Supplementary Data 2). In Chinese, Japanese, and Koreans, AF was 34.5%, 38.8%, and 36.8%, respectively (Supplementary Data 3). Among Europeans, the Finnish were a surprising outlier, with 37.3% AF (versus 19.9% in Italians, 17.8% in Spanish, and 22.6% in British) that corresponded to a low AT-CFR (Fig. 1a). Finnish AF differed only from the Chinese population among East Asians (P = 3.61×10^{-3}, Supplementary Fig. 4a, Supplementary Data 2). West Asians have AF that are similar to or lower than Europeans (Turkey 17.1%, Israel 13.2%). Latin Americans in general exhibited the lowest AFs, ranging from 18.8% in Columbia to 6.5% in Peru (Supplementary Data 3). Peruvian AF differed from all other countries except Mexico and Algeria (Supplementary Fig. 4a, Supplementary Data 2). We also found that V197M occurred in an extremely well-conserved position (phastCons17way_primate: 0.958, Supplementary Data 4) of the SRCR domain, suggesting that it is under purifying selection. Moreover, functional prediction tools SIFT[28] and PolyPhen2[29] regarded the variant as "deleterious" and "probably damaging", respectively (Supplementary Data 4).

TMPRSS2 V197M variant in ancient genomes

The V197M variant is absent in the great apes(30, 31) and in all sequenced archaic hominin genomes (Denisovan, Neanderthal). However, Tianyuan man's genotype showed that the variant was already present in humans 40,000 years ago in East Asia (Supplementary Data 5, Supplementary Data 6). We also found V197M in ancient genomes I7021 and I13180 from Mongolia, dated 5,211–5,000 BCE and 3,013–2,876 BCE, respectively (Supplementary Data 7). Starting from the pre-Ice Age (34,000–26,000 years ago), the variant was present in European inhabitants (37,250 BCE sample GoyetQ116-1 from Belgium(32)) and remained ever since (Supplementary Data 7, Supplementary Fig. 5). Although small sample sizes precluded statistical analysis, V197M AF appeared to be higher in ancient East Asian populations (33.3%) than in ancient Europeans (16.3%) (Supplementary Data 6, Supplementary Data 7, Supplementary Fig. 5).

Effect of V197M variant on TMPRSS2 protein structure

We used 3D protein models to investigate the effect of V197M on TMPRSS2. V197M increased energy score more than wild type (Table 1), suggesting reduced stability. Two programs (dDFIRE(33), nDOPE(34)) were used to measure the effect of V197M on the protein.

We used two homology modeling tools (Robetta[35], I-TASSER[36]) (Fig. 2) and transmembrane serine protease hepsin (PDB ID 1Z8G chain A)[37] as the template (Supplementary Fig. 6). The resultant model contains both SRCR and nearby peptidase S1 domains of TMPRSS2 (Fig. 2) because the former was too small for modeling. Despite only minor structural changes to the SRCR domain (Fig. 2), V197M had a consistently destabilizing effect in TMPRSS2 (Table 1). A further indication of reduced stability in mutants was a decrease in the favored region of the Ramachandran plot. Seven computational protein-stability prediction tools confirmed the V197M variant as destabilizing (Supplementary Data 8).

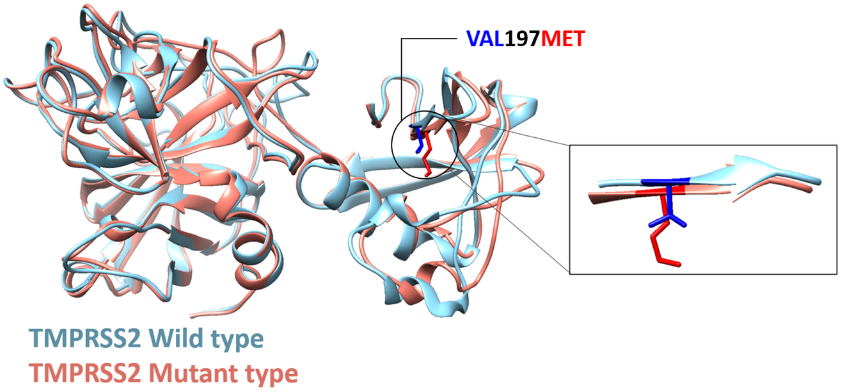


Fig. 2. TMPRSS2 protein structure of both wild type and mutant type (V197M), predicted with homology modeling using hepsin (1Z8G) template from the PDB database.

Correlation between *TMPRSS2* G8V variant AF and COVID-19 CFR

Unlike V197M, G8V was positively correlated with COVID-19 AT-CFR (Fig. 1b, Spearman's correlation coefficient, $\rho = 0.713$, $P = 0.00183$), indicating that it is a risk variant, rather than a protective one. Its frequency distribution formed Asian and non-Asian clusters (Fig. 1d). G8V frequencies were significantly higher in Europeans (Italians 41.6%, Spanish 31.1%, British 28.1%) than in East Asians (Chinese 9.4%, Singaporeans 9.1%, South Koreans 1.0%, Japanese 1.4%) or Latin Americans (Mexico 12.9%, Peru 6.5%) (Fig. 1d, Supplementary Fig. 4b, Supplementary Data 3, Supplementary Data 9). Koreans had the lowest G8V AF in the whole dataset, differing significantly from all 16 other countries spanning eight different regions (Fig. 1d). Nigerians had a variant frequency (36.0%) more similar to Europeans, while South Asian frequency (Pakistan 22.8%, Bangladesh 22.7%) fell between European and East Asian values (Supplementary Data 3). G8V occurred in a far less conserved region (phastCons17way_primate: 0.002, Supplementary Data 4) than V197M, and neither SIFT nor PolyPhen2 recognized it as functionally damaging (Supplementary Data 4). We also found G8V in both Denisovan and Neanderthal samples, but not in extant great apes (Supplementary Data 5, Supplementary Data 6). For modern humans, the earliest G8V occurrence was in a 45,000-year-old genome (Ust'Ishim) from Western Siberia(38) (Supplementary Data 10).

Discussion

This study has limitations. First, we only used public genome databases and variant frequency data that are not directly linked to COVID-19 patients and CFR. Furthermore, we could not completely normalize AT-CFR with relevant covariates, such as lockdown measures, mask availability, medical care standards, within-population or within-fatal-case age ratios, and SARS-CoV-2 test availability. However, we tested the Spearman's correlation between AT-CFR and thirteen socio-economic variables such as population density and Gross Domestic Product (GDP) per capita in a pairwise manner and found that only the proportion of the elderly (65 years and older) and the proportion of female smokers had significantly positive correlations (Supplementary Fig. 7). Another limitation is the lack of variant frequency data on chromosome X, absent from many public databases such as PGG.SNV, even though the X chromosome contains a key player, *ACE2* (14, 15). Notably, our protein structure modeling showed that TMPRSS2 and the template had a low sequence identity (32.49%). However, we confirmed that the V197M variant region of SRCR remained extremely consistent (Supplementary Fig. 6). Furthermore, ancient G8V data relied on sparse whole-genome-sequencing resources originating mainly from Europe and Russia, dated 2,000–1 BCE (Supplementary Data 9); these turned out too small to be conclusive. Finally, base-calling processing biases (e.g., haplodized ancient genome sequences) are a distinct possibility.

A previous report has noted that Europeans have significantly lower V197M AF than East Asians, a pattern speculated to be associated with COVID-19 CFR(20). Although we observed a significant correlation between the AFs of these two *TMPRSS2* variants and AT-CFR (Fig. 1), correlations between AFs and infection cases (per million) were non-significant (Spearman's correlation V197M: $P = 0.132$; G8V:$P = 0.165$) (Supplementary Fig. 1). One likely explanation is that infection cases are a more complex parameter than CFR. Factors such as high altitude had been reported to affect infection rate while not affecting CFR in COVID-19(39). Alternatively, CFR in infectious diseases reflects the importance of genetic factors more than infection rate(40). One example could be a study that evaluated incidence and CFR in sixteen yellow fever epidemics and found no significant differences between the infection rates of Caucasians and non-Caucasians while CFR differed significantly. Moreover, the study was unable to explain the differences observed by socioeconomic or demographic factors, or acquired immunity(41). To verify such trends in COVID-19, we require further studies investigating genomes, infection, treatment, and CFR data of COVID-19 patients.

Our evaluation of protein structural stability predicted that V197M destabilizes TMPRSS2 (Table 1, Supplementary Data 8). Unfortunately, we could not perform the same analysis on G8V because we lacked a homology modeling template. Our evidence (evolutionary conservation, protein domains) is insufficient to ascertain that G8V significantly affects TMPRSS2 protein structure and overall SARS-CoV-2 infection. However, one report has indicated that G8V affects residue torsion angles(42). The resultant flexibility reduction is more likely to affect TMPRSS2 interactions with ACE2 and the SARS-CoV-2 spike protein(42). We suspect V197M and G8V variants to be related to the overall *TMPRSS2* gene expression, however, we could not validate it.

In line with previous reports, we suggest that V197M acts to indirectly compromise the binding affinity of TMPRSS2 to SARS-CoV-2 spike protein and ACE2[35-37]. This implies a protective role of the V197M variant against SARS-CoV-2 infections, but neither we nor previous researchers(42-44) have uncovered any clear evidence or explanation for causation. Interestingly, the change from valine to methionine has a Grantham distance matrix value of only 22, the shortest distance from valine to any amino acid. Thus, V197M may lie on a thin boundary of extreme conservation versus functional benefit that may have arisen through viral invasion and polymorphisms in different ethnic groups that caused 3D structural deviation. We speculate that East Asians have already experienced similar viral infections in the past, leading to natural selection on V197M and G8V in *TMPRSS2* . Since V197M could have a synergistic or antagonistic effect with G8V and variants in other proteins, it perhaps accounts for only a fraction of resistance against SARS-CoV-2. Nevertheless, our CFR and genetic AF correlation study suggests that East Asian ethnic groups may have some genetic resistance that is reflected in the 3D structure of TMPRSS2 that negatively affects infection efficiency and hence the CFR of COVID-19.

Methods

Variant selection and data collection

Autosomal nonsynonymous variants located in *TMPRSS2* were extracted from Korea2K variome set (n = 2,262) from the Korean Genome Project[(26)], which turned out to contain 15 SNVs. Alternative AFs of other populations were obtained from the PGG.SNV database (GRCh38) (n = 220,147)(23), Italian Genome Reference Panel (IGRP1.0) (n = 926)(24), and Lithuanian high density SNP data (n = 425)(25). IGRP1.0 and Lithuanian genomes were lifted over to hg38 coordinates in Picard version 2.22.3(45), using LiftoverVcf with default options. The combined dataset included 223,760 samples from 4 variome databases with whole-genome sequencing, exome sequencing, or genotyping chip data (Supplementary Data 11). Allele counts were merged based on country of sample origin. Populations were excluded if they could not be assigned to any specific country, if fewer than 2,500 reported COVID-19 cases were present, or when CFR information was unavailable. Nonsynonymous variants were included only if they were present in >10 countries and had a global AF of >5%. The final dataset used to calculate AF and CFR correlations contained 69,168 samples (from 27 countries) for*TMPRSS2* V197M and 16,562 samples (from 17 countries) for*TMPRSS2* G8V.

Correlation with average of T-CFR

We downloaded COVID-19 data set on May 21, 2020 from Our World in Data (*https://github.com/owid/covid-19-data/tree/master/public/data*). We employed the equation from Daneshkhah et. al(46), to calculated average of T-CFR (Equation 1).

$$Average\ of\ T-CFR = \sum_{n=1}^{N} a_n \times T-CFR_n,\ a_n = c_n/\sum_{i=1}^{N} c_i \text{(Equation 1)}$$

Where N is the number of days which showed $<$ 2,500 confirmed cases on each country, a_n is a weight of T-CFR on day n , *T-CFR*$_n$ is T-CFR on day n ,c_i is the number of confirmed cases at day$i$.

Spearman's correlation test was conducted between AF and the average of T-CFR in R version 3.5.1.

Variant annotation

Variants were annotated in VEP version 99.2(47) with dbNSFP version 3.0(48) to evaluate deleteriousness and conservation. Additionally, phastCons scores were obtained for primates, mammals, and vertebrates to determine interspecific conservation of significant variant sites.

TMPRSS2 protein structure modelling and variant effects on the protein structure

We built a TMPRSS2 model using hepsin (1Z8G) as the template structure. The model was selected using PSI-BLAST sequence search(49), along with alignment from NCBI. Two sets of TMPRSS2 models were generated using the Robetta web server(35) and I-TASSER(36): a wild-type TMPRSS2 model based on 1Z8G and a V197M mutant model based on the wild-type one. Valine of residue 65 of 1Z8G was also substituted with methionine to generate mutant type. Protein energies of wild-type and variant models were compared in dDFIRE(33) and nDOPE(34) to determine structural stability (details in Supplementary Method). dDFIRE(33) scores have been extracted from the protein structure based on the distance between two atoms and the three angles involved in the dipole-dipole interaction. nDOPE(34) was used to measure protein energy as a statistical potential dependent on the calculated atomic distance in the protein structure.

Ramachandran favorable regions were measured through MolProbity(50). The following tools were used to predict variation in TMPRSS2 protein stability for both wild-type and mutant-type models: PoPMuSiC(51), CUPSAT(52), I-Mutant3(53), DUET(54), mCSM(55), SDM(56), MuPro(57). Visualizations were created in UCSF Chimera(58).

Ancient genome allele frequency analysis

Ancient genomes were downloaded from the David Reich Lab (*https://reich.hms.harvard.edu/datasets;* see Supplementary Data 7, Supplementary Data 10). Additional ancient European data for V197M (rs12329760) were obtained from the PGG.SNV database. Because the Reich Lab data did not cover G8V (rs75603675), only sample data from PGG.SNV was used for this variant. Data format conversion was handled using PLINK version 1.9(59). Presence of the two variants was verified and their frequencies calculated in different ancient populations (see Supplementary Data 7, Supplementary Data 10, Supplementary Data 12). Temporal variation in AF was visualized using the ggplot2 package in R.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. For the Korea2K variome set, the allele counts of the two *TMPRSS2* variants are available in supplementary data files. Detailed information about the Korea2K variome and request procedure can be found at*http://koreangenome.org.*

References

1. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents. 2020;55(3):105924.

2. ECDC. European Centre for Disease Prevention and Control,*https://www.ecdc.europa.eu/en/covid-19-pandemic*, access in 23 May 2020 2020 [Available from:*https://www.ecdc.europa.eu/en.*

3. Sonn JW, Kang M, Choi Y. Smart city technologies for pandemic control without lockdown. International Journal of Urban Sciences. 2020:1-3.

4. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proc Natl Acad Sci U S A. 2020;117(18):9696-8.

5. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. The Lancet Global Health. 2020;8(4).

6. Cai H. Sex difference and smoking predisposition in patients with COVID-19. The Lancet Respiratory Medicine. 2020;8(4).

7. Cai G. Bulk and single-cell transcriptomics identify tobacco-use disparity in lung gene expression of ACE2, the receptor of 2019-nCov. medRxiv. 2020.

8. Redelman-Sidi G. Could BCG be used to protect against COVID-19? Nat Rev Urol. 2020.

9. Miller A, Reandelar MJ, Fasciglione K, Roumenova V, Li Y, Otazu GH. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. medRxiv. 2020.

10. Hussein NR. Possible Factors Associated with Low Case Fatality Rate of COVID-19 in Kurdistan Region, Iraq. Journal of Kermanshah University of Medical Sciences. 2020;24(1).

11. Williams FMK, Freydin M, Mangino M, Couvreur S, Visconti A, Bowyer RCE, et al. Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. medRxiv. 2020.

12. Yuan FF, Velickovic Z, Ashton LJ, Dyer WB, Geczy AF, Dunckley H, et al. Influence of HLA gene polymorphisms on susceptibility and outcome post infection with the SARS-CoV virus. Virol Sin. 2014;29(2):128-30.

13. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov. 2020;6:11.

14. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun. 2020;11(1):1620.

15. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181(2):271-80 e8.

16. Matsuyama S, Nao N, Shirato K, Kawase M, Saito S, Takayama I, et al. Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proc Natl Acad Sci U S A. 2020;117(13):7001-3.

17. Song J, Li Y, Huang X, Chen Z, Li Y, Liu C, et al. Systematic analysis of ACE2 and TMPRSS2 expression in salivary glands reveals underlying transmission mechanism caused by SARS-CoV-2. J Med Virol. 2020.

18. Zhou L, Xu Z, Castiglione GM, Soiberman US, Eberhart CG, Duh EJ. ACE2 and TMPRSS2 are expressed on the human ocular surface, suggesting susceptibility to SARS-CoV-2 infection. Ocul Surf. 2020;18(4):537-44.

19. Mjaess G, Karam A, Aoun F, Albisinni S, Roumeguere T. COVID-19 and the male susceptibility: the role of ACE2, TMPRSS2 and the androgen receptor. Prog Urol. 2020;30(10):484-7.

20. Asselta R, Paraboschi EM, Mantovani A, Duga S. ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging (Albany NY). 2020;12(11):10087-98.

21. Das R, Ghate SD. Investigating the likely association between genetic ancestry and COVID-19 manifestation. medRxiv. 2020.

22. Kenyon C. Why Has COVID-19 Spread More Extensively in Europe than Asia? Preprints. 2020.

23. Zhang C, Gao Y, Ning Z, Lu Y, Zhang X, Liu J, et al. PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. Genome Biol. 2019;20(1):215.

24. Cocca M, Barbieri C, Concas MP, Robino A, Brumat M, Gandin I, et al. A bird's-eye view of Italian genomic variation through whole-genome sequencing. Eur J Hum Genet. 2020;28(4):435-44.

25. Urnikyte A, Flores-Bello A, Mondal M, Molyte A, Comas D, Calafell F, et al. Patterns of genetic structure and adaptive positive selection in the Lithuanian population from high-density SNP data. Sci Rep. 2019;9(1):9163.

26. Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. Science Advances. 2020;6(22).

27. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

28. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4.

29. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

30. Han S, Andres AM, Marques-Bonet T, Kuhlwilm M. Genetic Variation in Pan Species Is Shaped by Demographic History and Harbors Lineage-Specific Functions. Genome Biol Evol. 2019;11(4):1178-91.

31. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471-5.

32. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534(7606):200-5.

33. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins. 2008;72(2):793-803.

34. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006;15(11):2507-24.

35. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004;32(Web Server issue):W526-31.

36. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12(1):7-8.

37. Herter S, Piper DE, Aaron W, Gabriele T, Cutler G, Cao P, et al. Hepatocyte growth factor is a preferred in vitro substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. Biochem J. 2005;390(Pt 1):125-36.

38. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514(7523):445-9.

39. Segovia-Juarez J, Castagnetto JM, Gonzales GF. High altitude reduces infection rate of COVID-19 but not case-fatality rate. Respir Physiol Neurobiol. 2020;281:103494.

40. Petersen L, Andersen PK, Sorensen TI. Genetic influences on incidence and case-fatality of infectious disease. PLoS One. 2010;5(5):e10603.

41. Blake LE, Garcia-Blanco MA. Human genetic variation and yellow fever mortality during 19th century U.S. epidemics. mBio. 2014;5(3):e01253-14.

42. Sharma S, Singh I, Haider S, Malik MZ, Ponnusamy K, Rai E. ACE2 Homo-dimerization, Human Genomic variants and Interaction of Host Proteins Explain High Population Specific Differences in Outcomes of COVID19. bioRxiv. 2020.

43. Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, et al. Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. bioRxiv. 2020.

44. Paniri A, Hosseini MM, Akhavan-Niaki H. First comprehensive computational analysis of functional consequences of TMPRSS2 SNPs in susceptibility to SARS-CoV-2 among different populations. J Biomol Struct Dyn. 2020:1-18.

45. BroadInstitute. Picard toolkit. Broad Institute; 2018.

46. Ali Daneshkhah VA, Adam Eshein, Hariharan Subramanian, Hemant Kumar Roy, Vadim Backman. The Possible Role of Vitamin D in Suppressing Cytokine Storm and Associated Mortality in COVID-19 Patients. medRxiv. 2020.

47. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

48. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235-41.

49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

50. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 2018;27(1):293-315.

51. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011;12:151.

52. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. 2006;34(Web Server issue):W239-42.

53. Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics. 2008;9 Suppl 2:S6.

54. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;42(Web Server issue):W314-9.

55. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30(3):335-42.

56. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. 2017;45(W1):W229-W35.

57. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62(4):1125-32.

58. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

Author Contributions

S.J., C.Y., A.B., H.R. and J.B. designed and conceptualized this study. S.J., C.Y., H.R., A.B., and Y.B. conducted the data analysis and acquisition. S.J., C.Y., H.R., and A.B. performed the data visualization. A.B., S.J., C.Y., H.R., D.B., and J.B. wrote the manuscript. J.B., A.B., S.J., C.Y., H.R., D.B., Y.B., Y.J., A.M., E.-S.S., Y.S.C., N.R., and B.C.K. contributed to the manuscript editing process and critical revisions. All authors read and approved the finalized manuscript.

Figure Legends

Fig. 1. Correlation plots of COVID-19 average of time-adjusted case fatality rate (AT-CFR) with allele frequencies of *TMPRSS2*variants. **a)** V197M (rs12329760) from 27 countries and**b)** G8V (rs75603675) from 17 countries. Size of dots indicates the proportion of people who are 70 or older in the countries. The correlations were estimated by Spearman's correlation test. Allelic odds ratios (ORs) (i.e., alternative/reference allele counts) of the Y-axis country to the X-axis country are presented in **c)** for V197M and**d)** for G8V. AFR: Africa, CA: Central Asia, EU: Europe, SEA: Southeast Asia, AMR: Americas, EA: East Asia, SA: South Asia, and WA: West Asia. Full country names and allele frequencies per country are in Supplementary Data 3.

Fig. 2. TMPRSS2 protein structure of both wild type and mutant type (V197M), predicted with homology modeling using hepsin (1Z8G) template from the PDB database.

Fig. S1. Correlation plots of total COVID-19 cases per population of one million with allele frequencies of *TMPRSS2*variants: **a)** V197M (rs12329760) from 27 countries and**b)** G8V (rs75603675) from 17 countries.

AFR: Africa, CA: Central Asia, EU: Europe, SEA: Southeast Asia, AMR: American continents, EA: East Asia, SA: South Asia, and WA: West Asia.

Fig. S2. Exonic variants of *TMPRSS2* from gnomAD and Korea2K datasets with gnomAD global AF > 0.01.

Fig. S3 . Protein domain architecture of TMPRSS2. TM: transmembrane domain, LDLRA: LDL receptor class A domain, SRCR: scavenger receptor cysteine-rich domain. IBS version 1.0 was used to visualize domain architecture of TMPRSS2.

Fig. S4. Between-country differences in allele frequencies of**a)** V197M and **b)** G8V. Red box, allele frequencies of countries on the X and Y axes are significantly different ($P< 0.05$, Chi-squared test). White box, not significant. Gray box, small sample size precluded statistical analysis.

Fig. S5 . Allele frequency of V197M variant in ancient genomes. Dashed lines denote allele frequencies in present-day East Asians (red) and Europeans (blue), obtained from the gnomAD database. Allele frequencies in Europeans separated into Finnish (upper line) and non-Finnish (lower line). Numbers on the bar refer to sample size for the time frame on the X axis.

Fig. S6. Protein sequence alignment of SRCR and peptidase S1 domain from TMPRSS2 and hepsin. Darker and lighter colors represent matched and non-matched amino acids, respectively. Red box indicates amino acid position with the V197M variant.

Fig. S7. Correlations between average of T-CFR and factors in Our World in Data COVID-19 dataset. **a)** government response stringency index, **b)** population density, **c)** median age of the population, **d)** share of the population that is 65 years and older, **e)** share of the population that is 70 years and older, **f)** gross domestic product at purchasing power parity,**g)** share of the population living in extreme poverty,**h)** death rate from cardiovascular disease, **i)** diabetes prevalence, **j)** share of women who smoke, **k)** share of men who smoke, **l)** share of population with basic handwashing facilities on premises, and **m)** hospital beds per 100,000 people.

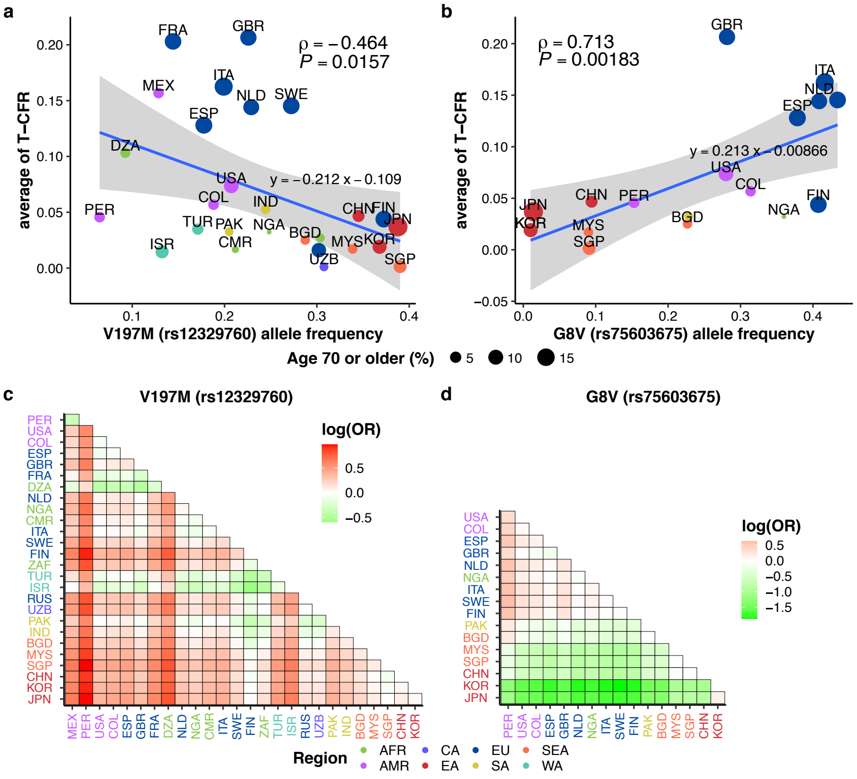
Table

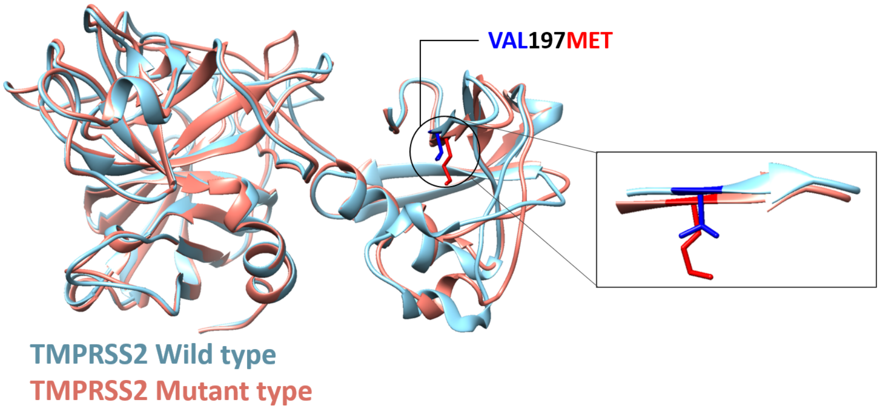
Table 1. Effect of V197M variant on structural features

Modeled structure	Modeled structure	Type of structures	dDFIRE	nDOPE	Ramachandran plot (Favored)(%)
Hepsin (1Z8G)	Hepsin (1Z8G)	WT[d]	-822.28	-1.586	97.53
		MT[e]	-812.80	-1.439	96.76
Robetta[a]	TMPRSS2(SRCR)[b]	WT	-183.43	-1.125	94.68
		MT	-176.48	-0.907	93.62
	TMPRSS2 (SRCR+Peptidase S1)[c]	WT	-730.79	-1.135	94.77
		MT	-725.40	-1.062	93.90
I-TASSER[a]	TMPRSS2(SRCR)	WT	-184.17	-0.909	91.67
		MT	-151.16	-0.156	86.17
	TMPRSS2 (SRCR+Peptidase S1)	WT	-700.23	-0.704	92.44
		MT	-615.98	-0.129	86.05

[a]Homology modeling tools, [b]SRCR domain separated from modeled TMPRSS2 structure,[c]Modeled TM-

PRSS2 structure, ^{d}Wild type structure, ^{e}Mutant type structure with V197M variant


a
average of T−CFR
ρ = −0.464
P = 0.0157
y = −0.212 x − 0.109
FRA
GBR
MEX
ITA
NLD
SWE
ESP
DZA
USA
COL
IND
PER
TUR
PAK
NGA
CMR
ISR
BGD
CHN
FIN
JPN
MYS
KOR
UZB
SGP
V197M (rs12329760) allele frequency
b
average of T−CFR
ρ = 0.713
P = 0.00183
y = 0.213 x − 0.00866
G8V (rs75603675) allele frequency
Age 70 or older (%)
c
V197M (rs12329760)
log(OR)
d
G8V (rs75603675)
log(OR)
Region
AFR
AMR
CA
EA
EU
SA
SEA
WA


VAL197MET
TMPRSS2 Wild type
TMPRSS2 Mutant type