Molecules and Cells

Regional TMPRSS2 V197M Allele Frequencies Are Correlated with COVID-19 Case Fatality Rates

Sungwon Jeon[1,2,9], Asta Blazyte[1,2,9], Changhan Yoon[1,2,9], Hyojung Ryu[1,2], Yeonsu Jeon[1,2], Youngjune Bhak[1,2], Dan Bolser[3], Andrea Manica[4], Eun-Seok Shin[5,6], Yun Sung Cho[7], Byung Chul Kim[7], Namhee Ryoo[8], Hansol Choi[1,2], and Jong Bhak[1,2,3,6,7,*]

[1]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea, [2]Department of Biomedical Engineering, College of Information and Biotechnology, UNIST, Ulsan 44919, Korea, [3]Geromics, Ltd., Cambridge CB1 3NF, UK, [4]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK, [5]Division of Cardiology, Department of Internal Medicine, Ulsan Medical Center, Ulsan 44686, Korea, [6]Personal Genomics Institute (PGI), Genome Research Foundation (GRF), Cheongju 28160, Korea, [7]Clinomics, Inc., Ulsan 44919, Korea, [8]Department of Laboratory Medicine, Keimyung University School of Medicine, Daegu 42601, Korea, [9]These authors contributed equally to this work.
*Correspondence: jongbhak@genomics.org
https://doi.org/10.14348/molcells.2021.2249
www.molcells.org

Coronavirus disease, COVID-19 (coronavirus disease 2019), caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), has a higher case fatality rate in European countries than in others, especially East Asian ones. One potential explanation for this regional difference is the diversity of the viral infection efficiency. Here, we analyzed the allele frequencies of a nonsynonymous variant rs12329760 (V197M) in the *TMPRSS2* gene, a key enzyme essential for viral infection and found a significant association between the COVID-19 case fatality rate and the V197M allele frequencies, using over 200,000 present-day and ancient genomic samples. East Asian countries have higher V197M allele frequencies than other regions, including European countries which correlates to their lower case fatality rates. Structural and energy calculation analysis of the V197M amino acid change showed that it destabilizes the TMPRSS2 protein, possibly negatively affecting its ACE2 and viral spike protein processing.

Keywords: allele frequency, case fatality rate, COVID-19, SARS-CoV-2, TMPRSS2

INTRODUCTION

COVID-19 (coronavirus disease 2019) is an infectious disease caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). Appearing first during late 2019 in Wuhan, China, COVID-19 has spread rapidly worldwide (Lai et al., 2020). As of May 23, 2020, SARS-CoV-2 has infected >5 million people in over 200 countries, killing more than 330,000 people (European Centre for Disease Prevention and Control, 2020). Many European countries had been particularly affected, with Spain and Italy each having reached over 200,000 cases of infection and more than 27,000 deaths, reaching a maximum case fatality rate (CFR) of >10% (European Centre for Disease Prevention and Control, 2020). In contrast, many East Asian countries did not experience such dire effects, with South Korea, for instance, reporting a peak CFR of 2.4% (European Centre for Disease Prevention and Control, 2020). Multiple contributing factors could explain this difference, including timing and severity of lockdown measures (Sonn et al., 2020), population age ratio (Dowd et al., 2020), healthcare resource availability (Ji et al., 2020), smoking rate (Cai, 2020a; 2020b), and early tuberculosis (Ba-

Received 18 December, 2020; revised 14 June, 2021; accepted 10 July, 2021; published online 30 September, 2021

eISSN: 0219-1032

cillus Calmette–Guérin) vaccination (Hussein, 2020; Miller et al., 2020; Redelman-Sidi, 2020). In principle, genetic factors may also underpin differential susceptibility to SARS-CoV-2 (Cao et al., 2020; Williams et al., 2020; Yuan et al., 2014).

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 (Hoffmann et al., 2020; Ou et al., 2020). 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 (Hoffmann et al., 2020; Ou et al., 2020). A study has suggested TMPRSS2 inhibition as a clinical target because the priming step is a key factor determining successful entry into target cells (Hoffmann et al., 2020). 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 hormones, their gene expression in various tissues and cell lines, and interactions with spike protein or inhibitors at a gene level (Hoffmann et al., 2020; Matsuyama et al., 2020; Mjaess et al., 2020; Song et al., 2020; Zhou et al., 2020).

To understand the genetic background of complex phenotypes in human populations, researchers commonly assess correlations with allele frequency (AF) (Asselta et al., 2020; Das and Ghate, 2020). This approach has identified a correlation between ancestral genetic composition and the CFR of COVID-19 (Das and Ghate, 2020). 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 (Asselta et al., 2020; Das and Ghate, 2020; Kenyon, 2020). Based on these studies, not only do *TMPRSS2* variants appear to have wide population-specific variation (Asselta et al., 2020), 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 (Asselta et al., 2020). 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 structural stability, and linked them to the average of time-adjusted COVID-19 CFR (AT-CFR) using the method described in (Daneshkhah et al., 2020); and Materials and Methods section. We propose that the structural deviation causes TMPRSS2 to be less stable, resulting in a reduced overall infection rate that led to the reduced CFR in East Asians. We collected and analyzed 221,498 genomes from public databases (Cocca et al., 2020; Urnikyte et al., 2019; Zhang et al., 2019) and 2,262 whole genomes from the Korean Genome Project (Jeon et al., 2020). 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 regional differences in the CFR.

MATERIALS AND 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 (Jeon et al., 2020), which turned out to contain 15 single nucleotide variants (SNVs). Alternative AFs of other populations were obtained from the PGG.SNV database (GRCh38) (n = 220,147) (Zhang et al., 2019), Italian Genome Reference Panel (IGRP1.0) (n = 926) (Cocca et al., 2020), and Lithuanian high density SNV data (n = 425) (Urnikyte et al., 2019). IGRP1.0 and Lithuanian genomes were lifted over to hg38 coordinates in Picard version 2.22.3 (Broad Institute, 2020), 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 S1). Allele counts were merged based on country of sample origin.

Populations were excluded if they could not be assigned to any specific country, if had fewer than 2,500 reported COVID-19 cases, or when AF or CFR information was unavailable. Nonsynonymous variants were included only if they were present in >10 countries and had a global AF of >1%. The final dataset used to calculate AF and CFR correlations contained 72,907 samples (from 29 countries) for TMPRSS2 V197M.

Correlation with average case fatality rate (AT-CFR)

We downloaded COVID-19 data set on March 22, 2021 from Our World in Data (https://github.com/owid/covid-19-data/tree/master/public/data). We employed the equation from Daneshkhah et al. (2020), to calculate time-adjusted CFR (T-CFR) (Equation 1), which throughout this manuscript is averaged and referred to as AT-CFR.

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

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, $T\text{-}CFR_n$ 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 AT-CFR in R version 3.5.1.

Variant annotation

Variants were annotated in VEP version 99.2 (McLaren et al., 2016) with dbNSFP version 3.0 (Liu et al., 2016) to evaluate deleteriousness and conservation. Additionally, phastCons scores were obtained for primates, mammals, and vertebrates to determine interspecies 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 (Altschul et al., 1997), along with alignment from NCBI. Two sets of TMPRSS2 models were generated

using the Robetta web server (Kim et al., 2004) and I-TASSER (Yang et al., 2015): 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 (Yang and Zhou, 2008) and nDOPE (Shen and Sali, 2006) to determine structural stability (details in the Supplementary Methods). dDFIRE (Yang and Zhou, 2008) 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 (Shen and Sali, 2006) 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 (Williams et al., 2018). The following tools were used to predict variation in TMPRSS2 protein stability for both wild-type and mutant-type models: PoPMuSiC (Dehouck et al., 2011), CUPSAT (Parthiban et al., 2006), I-Mutant3 (Capriotti et al., 2008), DUET (Pires et al., 2014a), mCSM (Pires et al., 2014b), SDM (Pandurangan et al., 2017), MuPro (Cheng et al., 2006). Visualizations were created in UCSF Chimera (Pettersen et al., 2004).

Ancient genome allele frequency analysis

Ancient genomes were downloaded from the David Reich Lab (https://reich.hms.harvard.edu/datasets; Supplementary Data S2 and S3). Additional ancient European data for V197M (rs12329760) were obtained from the PGG.SNV database. Data format conversion was handled using PLINK version 1.9 (Chang et al., 2015). Presence of the two variants was verified and their frequencies calculated in different ancient populations (Supplementary Data S2-S4). Temporal variation in AF was visualized using the ggplot2 package in R.

RESULTS

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

We found two nonsynonymous exonic *TMPRSS2* variants (V197M, rs12329760; G8V, rs75603675) of which V197M AF was significantly correlated with COVID-19 AT-CFR (Spearman's correlation $\rho = -0.482$, $P = 0.00881$ for V197M) (Fig. 1B). Even though the COVID-19 AT-CFR has been stabilized in most of the countries (Fig. 1A; area indicated in pur-

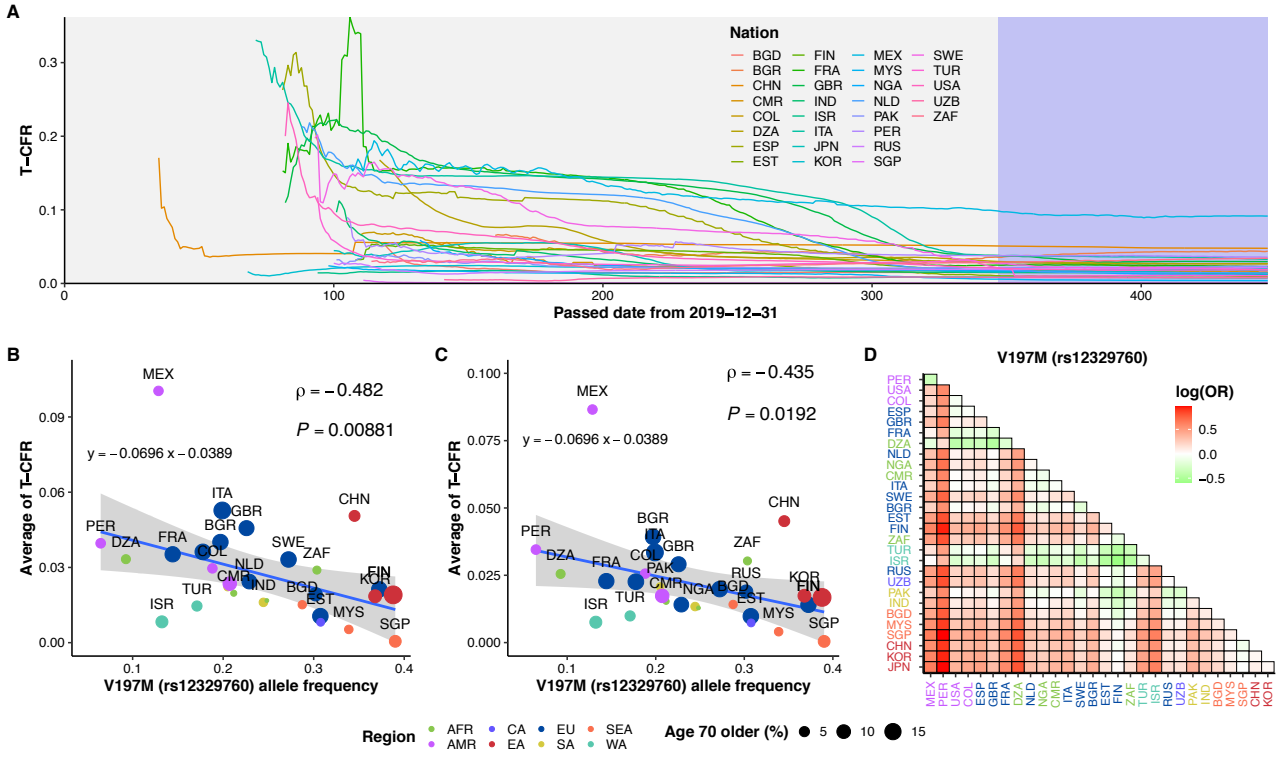

Fig. 1. COVID-19 AT-CFR correlation with allele frequencies of TMPRSS2 V197M (rs12329760) variant based on 29 countries. (A) COVID-19 AT-CFR dynamics in 29 countries from December 31, 2019 to March 22, 2021. (B) Correlation plot of TMPRSS2 V197M allele frequencies with COVID-19 AT-CFR. (C) Correlation plot of TMPRSS2 V197M allele frequencies with COVID-19 AT-CFR-100 (day 347-447, purple color in panel A). 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. (D) Allelic odds ratios (OR; 100 Genomes Project Consortium, 2010) (i.e., alternative/reference allele counts) of the Y-axis country to the X-axis country are presented in for V197M variant. Countries are color-coded by the region abbreviation. AFR, Africa; CA, Central Asia; EU, Europe; SEA, Southeast Asia; AMR, Americas; EA, East Asia; SA, South Asia; WA, West Asia. Full country names and allele frequencies per country are in Supplementary Data S5.

ple), we found a significant correlation between the regional V197M AF and the average of CFR for the last 100 days as well (AT-CFR-100; Spearman's correlation $\rho = -0.435$, $P = 0.0192$) (Fig. 1C). The V197M AF correlation with infection rate, which is presented as total COVID-19 cases per million individuals was less significant (infection rate, Spearman's correlation $\rho = -0.39$, $P = 0.0375$) (Supplementary Fig. S1). Contrary to the V197M, we could not find a significant correlation between G8V's regional AF and both AT-CFR and AT-CFR-100 (Spearman's correlation; AT-CFR: $\rho = 0.358$, $P = 0.133$ and AT-CFR-100: $\rho = 0.139$, $P = 0.57$) (Supplementary Fig. S2), therefore it was removed from the following analyses. These two variants were present among 20 *TMPRSS2* exonic variants with AF of >1% in gnomAD (Karczewski et al., 2020). Thirteen of these were in 3′ UTR, remaining five were synonymous (Supplementary Fig. S3, Supplementary Data S6). G8V is located in a cytoplasmic domain with an undetermined 3D structure (Supplementary Fig. S4). V197M is located in a stable beta-sheet of the scavenger receptor cysteine-rich (SRCR) domain (Supplementary Fig. S4, Fig. 2).

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

The AF of V197M was negatively correlated with COVID-19 AT-CFR and AT-CFR-100 (Figs. 1B and 1C). The AF distribution pattern was consistent with previous reports, with V197M AF being significantly lower in most Europeans than in East Asians (Asselta et al., 2020) (Fig. 1D, Supplementary Data S5 and S7). In Chinese, Japanese, and Koreans, AF was 34.5%, 38.8%, and 36.8%, respectively (Supplementary Data S5). Among Europeans, the Finnish were a surprising outlier, with 37.3% AF (vs 19.9% in Italians, 17.8% in Spanish, and 22.6% in British) that corresponded to a low AT-CFR (Supplementary Data S5). Finnish AF significantly differed only from the Chinese population among East Asians ($P = 3.61 \times 10^{-3}$) (Supplementary Fig. S5). 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 S5, Supplementary Fig. S5). Peruvian AF differed from all other countries except Mexico and Algeria (Supplementary Fig. S5, Supplementary Data S5).

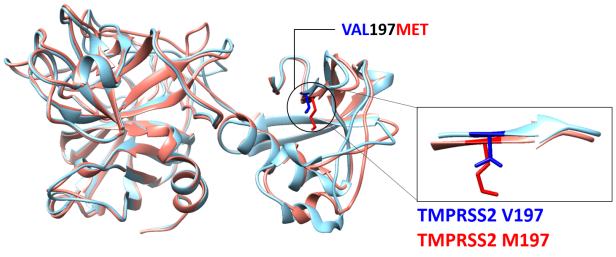

Fig. 2. TMPRSS2 protein structure of both wild type (V197) and mutant type (M197), predicted with homology modeling using hepsin (1Z8G) template from the PDB database.

TMPRSS2 V197M allele is archaic in humans

We also found that V197M occurred in an extremely well-conserved position (phastCons17way_primate: 0.958, Supplementary Data S8) of the SRCR domain, suggesting that it is under purifying selection. Moreover, functional prediction tools SIFT (Ng and Henikoff, 2003) and PolyPhen2 (Adzhubei et al., 2010) regarded the variant as "deleterious" and "probably damaging", respectively (Supplementary Data S8). This led us to investigate its selection in great apes and archaic hominin genomes.

Interestingly, the V197M variant is absent in the great apes (Han et al., 2019; Prado-Martinez et al., 2013) and in all sequenced archaic hominin genomes (Denisovan, Neanderthal) (Supplementary Data S9). We further investigated presence of V197M variant in ancient human genomes. Tianyuan man's genotype showed that the variant was already present in humans 40,000 years ago in East Asia (Supplementary Data S3). 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 S2). 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 [Fu et al., 2016]) and remained ever since (Supplementary Data S2 and S4). Although small sample sizes precluded statistical analysis,

Table 1. Effect of V197M variant on structural features

Modeled structure		Type of structures	dDFIRE	nDOPE	Ramachandran plot (favored) (%)
Hepsin (1Z8G)		V197[d]	-822.28	-1.586	97.53
		M197[e]	-812.80	-1.439	96.76
Robetta[a]	TMPRSS2 (SRCR)[b]	V197	-183.43	-1.125	94.68
		M197	-176.48	-0.907	93.62
	TMPRSS2 (SRCR + peptidase S1)[c]	V197	-730.79	-1.135	94.77
		M197	-725.40	-1.062	93.90
I-TASSER[a]	TMPRSS2 (SRCR)	V197	-184.17	-0.909	91.67
		M197	-151.16	-0.156	86.17
	TMPRSS2 (SRCR + peptidase S1)	V197	-700.23	-0.704	92.44
		M197	-615.98	-0.129	86.05

[a]Homology modeling tools, [b]SRCR domain separated from modeled TMPRSS2 structure, [c]Modeled TMPRSS2 structure, [d]Wild type (V197) structure, [e]Mutant type structure with M197 variant.

V197M AF appeared to be higher in ancient East Asian populations (33.3%) than in ancient Europeans (16.3%) (Supplementary Data S3 and S4, Supplementary Fig. S6).

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 [Yang and Zhou, 2008], nDOPE [Shen and Sali, 2006]) were used to measure the effect of V197M on the protein.

We used two homology modeling tools (Robetta [Kim et al., 2004], I-TASSER [Yang et al., 2015]) (Supplementary Methods) and transmembrane serine protease hepsin (PDB ID 1Z8G chain A) (Herter et al., 2005) as the template (Supplementary Fig. S7). 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 S10).

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 the CFR. However, a recent study conducted on COVID-19 patients in Italy confirmed V197M allele frequencies appear to be correlated to patients' clinical outcomes (Monticelli et al., 2021). 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 population density had significant positive correlations (Supplementary Figs. S8 and S9). 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* (Hoffmann et al., 2020; Ou et al., 2020). 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 (Fig. 2, Supplementary Fig. S7).

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 (Asselta et al., 2020). To confirm that this association in our study had not occurred by chance, we performed a V197M AF correlation test based on two different approaches (the correlation with AT-CFR and correlation with AT-CFR-100) applying multiple input filtering criteria (Supplementary Figs. S10-S12) as well as a series of linear regression analyses considering publicly available regional socio-economic and epidemiological variables (Supplementary Data S11-S16). Although we observed a very significant correlation between the AFs of the *TMPRSS2* V197M variant and AT-CFR (Fig. 1), correlation between AF and infection cases per million individuals was less-significant (29 countries, Spearman's correlation, $\rho = -0.39$, $P = 0.0375$) (Supplementary Fig. S1). 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 (Segovia-Juarez et al., 2020). Alternatively, CFR in infectious diseases reflects the importance of genetic factors more than infection rate (Petersen et al., 2010). One example, could be a study that evaluated the 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 (Blake and Garcia-Blanco, 2014). 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 S10). We suspect V197M variant 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 (Bhattacharyya et al., 2020; Petersen et al., 2010; Sharma et al., 2020). This implies a protective role of the V197M variant against SARS-CoV-2 infections, but neither we nor previous researchers (Bhattacharyya et al., 2020; Paniri et al., 2021; Sharma et al., 2020) 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 the 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 (Souilmi et al., 2020), leading to natural selection on V197M in *TMPRSS2*. However, our results do not disprove the alternative evolutionary mechanisms. The genetic drift due to population migrations and admixtures, and selection on *TMPRSS2* by unknown evolutionary drivers may result in the AF differences. Regardless of the underlying mechanisms that caused AF differences, our AT-CFR and the genetic AF correlation study suggests that East Asians 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.

Notably, the majority of the data in our study were obtained from PGG.SNV database (Zhang et al., 2019) (Supplementary Data S1), where we cannot check the duplicity as the genome data for each sample is not available, only allele

counts. This also complicates checking duplicity between the PGG.SNV (Zhang et al., 2019) and the external data used (such as Korean [Jeon et al., 2020], Lithuanian [Urnikyte et al., 2019], and Italian data [Cocca et al., 2020]) (Supplementary Data S1). Duplicity could have possibly occurred as the same individual could have been sequenced or genotyped for multiple projects either within PGG.SNV database, or for e.g., PGG.SNV and Korea1K. Still, the chances of this happening must be very low and even if it is the case, it would not affect the overall statistics significantly.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This research 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 Promotion of Innovative Businesses for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea)(P0016193)(2.210511.01). This work was also supported by the Establishment of Demonstration Infrastructure for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (P0016191)(2.210514.01). This work was also supported by the Research Project Funded by Ulsan City Research Fund (2.201052.01) 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.

AUTHOR CONTRIBUTIONS

S.J. and J.B. conceived the study design. S.J., C.Y., H.R., and A.B. performed analyses. A.B., C.Y., S.J., H.R., and H.C. wrote and edited the manuscript. J.B. secured funding. Y.J., Y.B., D.B., A.M., E.S.S., Y.S.C., N.R., and B.C.K. provided expertise and feedback. H.C. independently validated the results.

CONFLICT OF INTEREST

The authors declare the following financial interests/personal relationships which may be considered as potential 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 have no potential conflicts of interest to disclose.

ORCID

Sungwon Jeon https://orcid.org/0000-0002-2729-9087
Asta Blazyte https://orcid.org/0000-0001-7309-1482
Changhan Yoon https://orcid.org/0000-0003-0243-9853
Hyojung Ryu https://orcid.org/0000-0002-2276-850X
Yeonsu Jeon https://orcid.org/0000-0003-4560-4142
Youngjune Bhak https://orcid.org/0000-0002-9273-6984
Dan Bolser https://orcid.org/0000-0002-3991-0859
Andrea Manica https://orcid.org/0000-0003-1895-450X
Eun-Seok Shin https://orcid.org/0000-0002-9169-6968
Yun Sung Cho https://orcid.org/0000-0003-4490-8769
Byung Chul Kim https://orcid.org/0000-0002-4891-9679
Namhee Ryoo https://orcid.org/0000-0001-8383-709X
Hansol Choi https://orcid.org/0000-0002-3653-1474
Jong Bhak https://orcid.org/0000-0002-4228-1299

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248-249.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Asselta, R., Paraboschi, E.M., Mantovani, A., and Duga, S. (2020). ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging (Albany N.Y.) *12*, 10087-10098.

Bhattacharyya, C., Das, C., Ghosh, A., Singh, A.K., Mukherjee, S., Majumder, P.P., Basu, A., and Biswas, N.K. (2020). 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, https://doi.org/10.1101/2020.05.04.075911

Blake, L.E. and Garcia-Blanco, M.A. (2014). Human genetic variation and yellow fever mortality during 19th century U.S. epidemics. mBio *5*, e01253-14.

Broad Institute (2020). Picard toolkit. Retrieved May 7, 2020, from http://broadinstitute.github.io/picard/

Cai, G. (2020a). Bulk and single-cell transcriptomics identify tobacco-use disparity in lung gene expression of ACE2, the receptor of 2019-nCov. MedRxiv, https://doi.org/10.1101/2020.02.05.20020107

Cai, H. (2020b). Sex difference and smoking predisposition in patients with COVID-19. Lancet Respir. Med. *8*, e20.

Cao, Y., Li, L., Feng, Z., Wan, S., Huang, P., Sun, X., Wen, F., Huang, X., Ning, G., and Wang, W. (2020). Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov. *6*, 11.

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

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins *62*, 1125-1132.

Cocca, M., Barbieri, C., Concas, M.P., Robino, A., Brumat, M., Gandin, I., Trudu, M., Sala, C.F., Vuckovic, D., Girotto, G., et al. (2020). A bird's-eye view of Italian genomic variation through whole-genome sequencing. Eur. J. Hum. Genet. *28*, 435-444.

Daneshkhah, A., Eshein, A., Subramanian, H., Roy, H.K., and Backman, V. (2020). The possible role of vitamin D in suppressing cytokine storm and associated mortality in COVID-19 patients. MedRxiv, https://doi.org/10.1101/2020.04.08.20058578

Das, R. and Ghate, S.D. (2020). Investigating the likely association between genetic ancestry and COVID-19 manifestations. MedRxiv, https://doi.org/10.1101/2020.04.05.20054627

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

Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M.C. (2020). Demographic science aids in understanding the

spread and fatality rates of COVID-19. Proc. Natl. Acad. Sci. U. S. A. *117*, 9696-9698.

European Centre for Disease Prevention and Control (2020). COVID-19 statistics worldwide. Retrieved May 23, 2020, from https://www.ecdc.europa.eu/en/covid-19-pandemic

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwangler, A., Haak, W., Meyer, M., Mittnik, A., et al. (2016). The genetic history of Ice Age Europe. Nature *534*, 200-205.

Han, S., Andres, A.M., Marques-Bonet, T., and Kuhlwilm, M. (2019). Genetic variation in Pan species is shaped by demographic history and harbors lineage-specific functions. Genome Biol. Evol. *11*, 1178-1191.

Herter, S., Piper, D.E., Aaron, W., Gabriele, T., Cutler, G., Cao, P., Bhatt, A.S., Choe, Y., Craik, C.S., Walker, N., et al. (2005). 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. *390*, 125-136.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., 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.

Hussein, N.R. (2020). Possible factors associated with low case fatality rate of COVID-19 in Kurdistan Region, Iraq. J. Kermanshah Univ. Med. Sci. *24*, e103393.

Jeon, S., Bhak, Y., Choi, Y., Jeon, Y., Kim, S., Jang, J., Jang, J., Blazyte, A., Kim, C., Kim, Y., et al. (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. Sci. Adv. *6*, eaaz7835.

Ji, Y., Ma, Z., Peppelenbosch, M.P., and Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. Lancet Glob. Health *8*, e480.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443.

Kenyon, C. (2020). Why has COVID-19 spread more extensively in Europe than Asia? Preprints, https://doi.org/10.20944/preprints202005.0200.v1

Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. *32*(Web Server issue), W526-W531.

Lai, C.C., Shih, T.P., Ko, W.C., Tang, H.J., and Hsueh, P.R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. Int. J. Antimicrob. Agents *55*, 105924.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. *37*, 235-241.

Matsuyama, S., Nao, N., Shirato, K., Kawase, M., Saito, S., Takayama, I., Nagata, N., Sekizuka, T., Katoh, H., Kato, F., et al. (2020). Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proc. Natl. Acad. Sci. U. S. A. *117*, 7001-7003.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Miller, A., Reandelar, M.J., Fasciglione, K., Roumenova, V., Li, Y., and Otazu, G.H. (2020). Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19. MedRxiv, https://doi.org/10.1101/2020.03.24.20042937

Mjaess, G., Karam, A., Aoun, F., Albisinni, S., and Roumeguere, T. (2020). COVID-19 and the male susceptibility: the role of ACE2, TMPRSS2 and the androgen receptor. Prog. Urol. *30*, 484-487.

Monticelli, M., Hay Mele, B., Benetti, E., Fallerini, C., Baldassarri, M., Furini, S., Frullanti, E., Mari, F., Andreotti, G., Cubellis, M.V., et al. (2021). Protective role of a TMPRSS2 variant on severe COVID-19 outcome in young males and elderly women. Genes (Basel) *12*, 596.

Ng, P.C. and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812-3814.

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., et al. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat. Commun. *11*, 1620.

Pandurangan, A.P., Ochoa-Montano, B., Ascher, D.B., and Blundell, T.L. (2017). SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. *45*(W1), W229-W235.

Paniri, A., Hosseini, M.M., and Akhavan-Niaki, H. (2021). First comprehensive computational analysis of functional consequences of TMPRSS2 SNPs in susceptibility to SARS-CoV-2 among different populations. J. Biomol. Struct. Dyn. *39*, 3576-3593.

Parthiban, V., Gromiha, M.M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. *34*(Web Server issue), W239-W242.

Petersen, L., Andersen, P.K., and Sorensen, T.I. (2010). Genetic influences on incidence and case-fatality of infectious disease. PLoS One *5*, e10603.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605-1612.

Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014a). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. *42*(Web Server issue), W314-W319.

Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014b). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics *30*, 335-342.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471-475.

Redelman-Sidi, G. (2020). Could BCG be used to protect against COVID-19? Nat. Rev. Urol. *17*, 316-317.

Segovia-Juarez, J., Castagnetto, J.M., and Gonzales, G.F. (2020). High altitude reduces infection rate of COVID-19 but not case-fatality rate. Respir. Physiol. Neurobiol. *281*, 103494.

Sharma, S., Singh, I., Haider, S., Malik, M.Z., Ponnusamy, K., and Rai, E. (2020). ACE2 homo-dimerization, human genomic variants and interaction of host proteins explain high population specific differences in outcomes of COVID19. BioRxiv, https://doi.org/10.1101/2020.04.24.050534

Shen, M.Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci. *15*, 2507-2524.

Song, J., Li, Y., Huang, X., Chen, Z., Li, Y., Liu, C., Chen, Z., and Duan, X. (2020). Systematic analysis of ACE2 and TMPRSS2 expression in salivary glands reveals underlying transmission mechanism caused by SARS-CoV-2. J. Med. Virol. *92*, 2556–2566.

Sonn, J.W., Kang, M., and Choi, Y. (2020). Smart city technologies for pandemic control without lockdown. Int. J. Urban Sci. *24*, 149-151.

Souilmi, Y., Lauterbur, M.E., Tobler, R., Huber, C.D., Johar, A.S., and Enard, D. (2020). An ancient coronavirus-like epidemic drove adaptation in East Asians from 25,000 to 5,000 years ago. BioRxiv, https://doi.org/10.1101/2020.11.16.385401

Urnikyte, A., Flores-Bello, A., Mondal, M., Molyte, A., Comas, D., Calafell, F., Bosch, E., and Kucinskas, V. (2019). Patterns of genetic structure and adaptive positive selection in the Lithuanian population from high-density SNP data. Sci. Rep. *9*, 9163.

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018).

MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293-315.

Williams, F.M.K., Freidin, M.B., Mangino, M., Couvreur, S., Visconti, A., Bowyer, R.C.E., Le Roy, C.I., Falchi, M., Sudre, C., Davies, R., et al. (2020). Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. MedRxiv, https://doi.org/10.1101/2020.04.22.20072124

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

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

Yuan, F.F., Velickovic, Z., Ashton, L.J., Dyer, W.B., Geczy, A.F., Dunckley, H., Lynch, G.W., and Sullivan, J.S. (2014). Influence of HLA gene polymorphisms on susceptibility and outcome post infection with the SARS-CoV virus. Virol. Sin. *29*, 128-130.

Zhang, C., Gao, Y., Ning, Z., Lu, Y., Zhang, X., Liu, J., Xie, B., Xue, Z., Wang, X., Yuan, K., et al. (2019). PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. Genome Biol. *20*, 215.

Zhou, L., Xu, Z., Castiglione, G.M., Soiberman, U.S., Eberhart, C.G., and Duh, E.J. (2020). ACE2 and TMPRSS2 are expressed on the human ocular surface, suggesting susceptibility to SARS-CoV-2 infection. Ocul. Surf. *18*, 537-544.