Genetic variations in the human severe acute respiratory syndrome coronavirus receptor *ACE2* and serine protease *TMPRSS2*

Kohei Fujikura [iD],[1] Kazuma Uesaka[2]

ABSTRACT

Aims The recent emergence of novel, pathogenic severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses a global health emergency. The coronaviral entry requires the spike (S)-protein for attachment to the host cell surface, and employs human angiotensin-converting enzyme 2 (hACE2) for entry and transmembrane protease serine 2 (TMPRSS2) for S-protein priming. Although coronaviruses undergo evolution by mutating themselves, it is also essential to know the host genetic factors. Here, we describe the single nucleotide variations (SNVs) in human *ACE2* and *TMPRSS2*.

Methods The genetic variants derived from five population-sequencing projects were classified by variant type, allele frequency (AF), ethnic group and estimated pathogenicity. The SNVs in SARS-CoV-2/hACE2 contact residues were investigated. The genetic variability was normalised using non-linear regression and the total number of SNVs was estimated by the derived formulas.

Results We detected 349 and 551 SNVs in *ACE2* and *TMPRSS2*, respectively, in a total of 156 513 individuals. The vast majority (>97%) of the SNVs were very rare (AF <0.1%) and population-specific, and were computationally estimated to be more frequently deleterious than the SNVs with high AF. These SNVs were distributed throughout the coding regions; some *ACE2* variants were located in the SARS-CoV-2/hACE2 contact residues, with a hemizygous state occurring in males. Using regression analysis, the total numbers of genetic variations in *ACE2* and *TMPRSS2* were 1.1×10^3 and 1.5×10^3, respectively, for a population of one million people.

Conclusion The majority of SNVs in *ACE2* and *TMPRSS2* are rare, population-specific and deleterious, and a multitude of very rare SNVs may explain different susceptibility to SARS-CoV-2.

INTRODUCTION

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is a novel coronavirus that was first detected in Wuhan, China, and causes COVID-19.[1–4] Since the initial detection of the virus, >10 million cases of COVID-19 have been confirmed worldwide as of 8 July 2020.[5] The COVID-19 outbreak has resulted in a social disorder and the collapse of the medical care system globally.[6–9]

Presently, some Western countries have recorded the highest rates of SARS-CoV-2 infection, the highest number of deaths and the highest mortality rates.[5] These data may be due to different factors, including: 1) the total number of tests performed, 2) the possible existence of more virulent strains, 3) the structure of population, 4) the percentage of chronic illness or smokers, 5) the occupational exposure to the virus as well as 6) the differences in environmental factors (eg, temperature, humidity and air pollution).[10–17] However, there may also exist some peculiar genetic characteristics of the populations that may affect susceptibility to viral infection, disease severity and the number of patients shedding huge amounts of the virus. In addition, it is suggested that COVID-19 is more likely to affect males than females, and can result in severe and even fatal respiratory diseases such as acute respiratory distress syndrome[12 18–25] and mild-to-moderate gastrointestinal symptoms, such as nausea, diarrhoea or abdominal pain.[26–30]

The spike (S)-protein of coronaviruses facilitates viral entry into target cells.[31 32] Entry depends on S-protein binding to a cellular receptor and S-protein priming by a cellular protease. The SARS-CoV S-protein engages human angiotensin-converting enzyme 2 (hACE2) as entry receptor[33–35] and employs the cellular serine protease, transmembrane protease serine 2 (hTMPRSS2), for S-protein priming.[36–38] The coronavirus S-protein/hACE2 interface has been elucidated and the efficiency of hACE2 usage has been shown to be a key determinant of SARS-CoV transmissibility. Importantly, a previous study has demonstrated that a number of *ACE2* variants could affect the association between ACE2 and S-protein in SARS-CoV or HCoV-NL63.[35] Recent reports suggested that SARS-CoV and SARS-CoV-2 share 73% amino acid identity[39] and the novel SARS-CoV-2 also uses the ACE2 and TMPRSS2 for entry into target cells.[40] Therefore, the genetic variation in these two genes in different populations might be also critical for the susceptibility, symptoms and outcome of SARS-CoV-2 infection. Yet, to date, a comprehensive overview of the genetic diversity of the two virus-entry-related genes is lacking.

Here, we provided the largest data set of *ACE2* and *TMPRSS2* gene polymorphisms from five extensive population-sequencing projects (total 156 513 individuals). The very rare SNVs we identified could contribute to a better understanding of gender differences and different susceptibilities or responses to SARS-CoV-2 in different human populations under similar conditions.

MATERIALS AND METHODS
Analysis of genetic variants

Data were collected from the genotyping pipelines of the 1000 Genomes (1000G) project (http://

www.ncbi.nlm.nih.gov/variation/tools/1000genomes/),[41] the US National Heart, Lung, and Blood Institute (NHLBI) (http://www.nhlbi.nih.gov/),[42] gnomAD (https://gnomad.broadinstitute.org/),[43] Tohoku Medical Megabank Organization (ToMMo) (https://www.megabank.tohoku.ac.jp/english/)[44 45] and UK10K (https://www.uk10k.org/),[46] which consisted of high-coverage whole-genome/whole-exome sequence data from various ethnic groups. The data set consisted of 156 513 individuals from various countries (online supplementary table S1; it should be noted that gnomAD includes the 1000G data set, but not other projects). The data set was then filtered using Variant Tools (http://varianttools.sourceforge.net/Annotation/HomePage) by variant type, allele frequency (AF), countries, ethnic/racial groups and pathogenicity. Information on variant types, positions and reference sequences were retrieved from NCBI dbSNP (http://www.nlm.nih.gov/SNP/).

Deleteriousness prediction methods
We comprehensively evaluated the predictive performance of 26 current deleteriousness-scoring methods, including 23 function prediction scores (SIFT, SIFT4G, PolyPhen-2-HDIV, PolyPhen-HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST4, MetaSVM, MetaLR, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, DEOGEN2, CADD, DANN, fathmm-MKL and GenoCnyon) and 3 conservation scores (GERP++, SiPhy and PhyloP). The scores were obtained from the dbNSFP database V.4.0.[47] It is noted that the prediction scores obtained from the dbNFSP database underwent transformation from the original prediction scores according to the threshold value (online supplementary table S2).

DOMAIN PREDICTION
Functional domains including transmembrane and signal peptide regions were predicted using InterPro (https://www.ebi.ac.uk/interpro/) with default options.

Non-linear regression model of *ACE2* and *TMPRSS2* genetic variation
The genetic variability of *ACE2* and *TMPRSS2* genes was normalised based on non-linear regression according to previous research.[48] Normalisation allows the estimation of populations with different accumulated sample sizes. The relationship between AF and genetic variation was determined using a scatter plot. This plot showed a pattern of exponential decay, and thus, a negative exponential model was fitted. The formula was then converted and plotted against the population size as follows:

$$N_{ACE-f} = 0.100\chi^{0.691} \ (R^2 = 0.984) \tag{1}$$

$$N_{ACE-m} = 0.0620\chi^{0.691} \ (R^2 = 0.984) \tag{2}$$

$$N_{TMPRSS2} = 0.569\chi^{0.569} \ (R^2 = 0.900) \tag{3}$$

where N_{ACE-m}, N_{ACE-f} and $N_{TMPRSS2}$ correspond to the estimated number of genetic variations, x refers to the population size and R^2 is the coefficient of determination. The total number of genetic variations in the two genes were estimated using formulas (1), (2) and (3). Since the *ACE2* gene is located on the X-chromosome, two different formulas, N_{ACE-m} (1) and N_{ACE-f} (2), were derived for males (46,XY) and females (46,XX), respectively.

STATISTICAL ANALYSIS
Statistical analysis was performed using the Mann-Whitney *U* test. A probability of p<0.05 was considered to be statistically significant. Statistical analyses were performed using JMP software (V.10.0; SAS Institute, Cary, North Carolina, USA).

RESULTS
Genetic variations in human *ACE2*
In order to obtain a comprehensive overview of the genetic diversity of *ACE2* and *TMPRSS2*, we collected SNV data sets from five extensive population-sequencing projects (1000G, NHLBI, gnomAD, ToMMo and UK10K). After removing the overlapping individuals, the data sets included genome/exome sequences of 156 513 individuals from diverse ethnic origins (online supplementary table S1).

The genetic diversity of *ACE2* and *TMPRSS2* is summarised in figures 1 and 2, respectively. *ACE2* is located on the X-chromosome, which raises the possibility that differences in sex chromosome dosage (46,XY vs 46,XX) could cause the phenotype to be always expressed in males. In *ACE2*, 349 SNVs were identified in the coding regions and splice sites, and were broadly distributed throughout the coding regions (figure 1A). A complete list of the variants can be found in online supplementary table S3. The most identified SNVs (n=247, 70.8%) were missense, while synonymous variants occurred in 26.9% of the SNVs (n=94) (figure 1B). The remaining 2% of SNVs were stop-gained (n=2), start-loss (n=1), splice site variants (n=2) or indels (n=3) (figure 1B). Per gene, this resulted in a non-synonymous/synonymous variant ratio (dN/dS) >1. The majority of these SNVs were rare (n=347, 99.4%) or quite rare (n=169, 48.4%) with AF <1% or AF <0.001%, respectively (figure 1C).

In silico prediction of the functional effects of SNVs in human *ACE2* gene
At present, the majority of the *ACE2* variants have not been functionally characterised. To obtain insights into the likelihood of functionally deleterious effects of the identified variants, we comprehensively evaluated the predictive performance of 26 current deleteriousness-scoring methods, including 23 function prediction scores (SIFT, SIFT4G, PolyPhen-2-HDIV, PolyPhen-HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST4, MetaSVM, MetaLR, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, DEOGEN2, CADD, DANN, fathmm-MKL and GenoCnyon) and 3 conservation scores (GERP++, SiPhy and PhyloP) (online supplementary table S4). Fraction deleteriousness ranged from 0% to 90.3%, possibly due to the differences in information and algorithms used for this prediction, but the median was 46.8% (calculated by LRT) (figure 1D). Importantly, rare SNVs were estimated to be more frequently deleterious than the SNVs with a high AF by some scoring methods (table 1).

Estimation of the genetic variability of *ACE2* in populations
To predict the number of overall *ACE2* genetic variants in populations, we calculated the normalised genetic variability using non-linear regression (figure 1E). The observed relationship between AF and genetic variations exhibited a pattern of exponential decay (figure 1E). A negative exponential model was fitted to these decay curves and then evaluated for unseen *ACE2* variations (figure 1E). Since allele count is different between males (46,XY) and females (46,XX), the per-individual genetic variability was estimated to be significantly higher in females than that in males regardless of population size (figure 1F). The total number of genetic variations in *ACE2* genes was expected to be approximately 4.2×10^3 for male populations and 6.8×10^3 for female populations (average, 5.5×10^3) for a sample size of

Figure 1 Genetic variation in *ACE2*. (A) The distribution of nucleotide polymorphisms along the full-length *ACE2* gene. The vertical bar indicates allele frequency (AF) (%). Single nucleotide variations (SNVs) are grouped by type: missense (blue), stop-gained (orange), start-lost (grey), indel (yellow) and splice site (green). The putative functional domains are depicted by coloured boxes. (B) Pie chart of 349 SNVs in *ACE2*. Each colour code corresponds to a different SNV type. (C) Relative abundance of SNVs plotted over their AFs. SNVs are grouped by type: missense (blue), synonymous (orange), other non-synonymous (grey) and indel (yellow). The inset demonstrates that the majority of SNVs in coding regions were quite rare (48.4%, AF <0.001%), while 90.3% were classified as rare (0.001%< AF <0.01%) and only 9.7% had low frequency or were common. (D) Percentage of fraction deleterious. Twenty-three software were employed to predict the pathogenicity of missense variants. (E) The non-linear regression fitting was performed based on the scatter plot showing the relationship between AF and variation of SNVs. (F) Relationship between total number of SNVs and population size. The *ACE2* variations are expected to rise as the population size increases. A higher number of rare SNVs could be detected in females compared with males. An enlarged view of the graph is also indicated in the upper panel. (G) Chow-Ruskey diagrams showing the number of shared and unique genetic variants for *ACE2* genes across four large-scale population studies. (H) Comparison of AFs across four large-scale population studies. ACE2, angiotensin-converting enzyme 2; HEMGH, metalloprotease zinc-binding site; NHLBI, National Heart, Lung, and Blood Institute; TM, transmembrane domain; ToMMo, Tohoku Medical Megabank Organization.
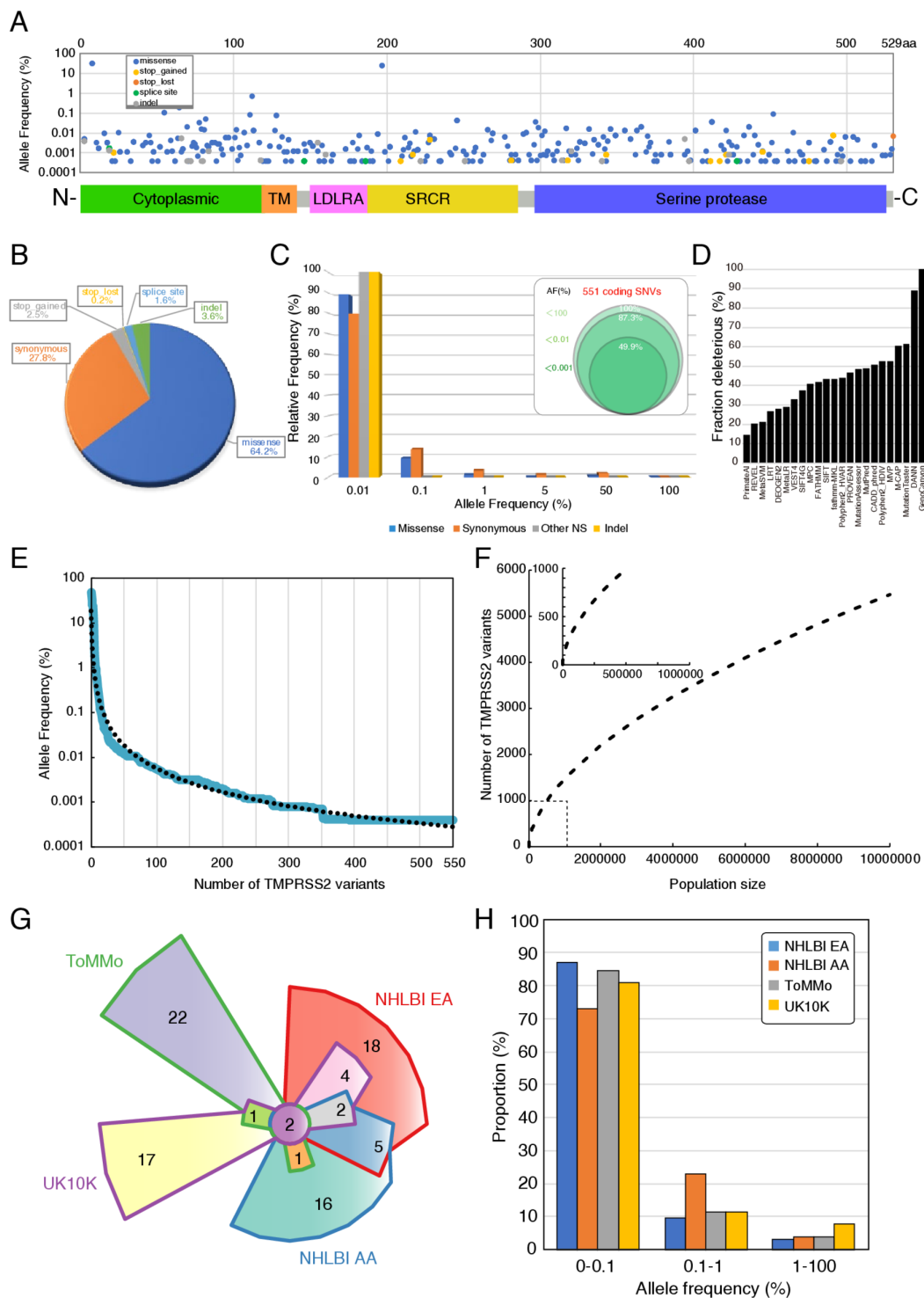
Figure 2 Genetic variation in *TMPRSS2*. (A) The distribution of nucleotide polymorphisms along the full-length *TMPRSS2* gene. The vertical bar indicates allele frequency (AF) (%). Single nucleotide variations (SNVs) are grouped by type: missense (blue), stop-gained (yellow), stop-lost (orange), splice site (green) and indel (grey). The putative functional domains are depicted by coloured boxes. (B) Pie chart of 551 SNVs in *TMPRSS2*. Each colour code corresponds to a different SNV type. (C) Relative abundance of SNVs plotted over their AFs. SNVs are grouped by type: missense (blue), synonymous (orange), other non-synonymous (grey) and indel (yellow). The inset demonstrates that the majority of SNVs in coding regions were quite rare (49.9%, AF <0.001%), while 87.3% were classified as rare (0.001%< AF <0.01%) and only 12.7% had low frequency or were common. (D) Percentage of fraction deleterious. Twenty-three software were employed to predict the pathogenicity of missense variants. (E) The non-linear regression fitting was performed based on the scatter plot showing the relationship between AF and variation of SNVs. (F) Relationship between total number of SNVs and population size. The *TMPRSS2* variations are expected to rise as the population size increases. An enlarged view of the graph is also indicated in the upper panel. (G) Chow-Ruskey diagrams showing the number of shared and unique genetic variants for *TMPRSS2* genes across four large-scale population studies. (H) Comparison of AFs across four large-scale population studies. ACE2, angiotensin-converting enzyme 2; LDLRA, low-density lipoprotein receptor A domain; NHLBI, National Heart, Lung, and Blood Institute; SRCR, scavenger receptor cysteine-rich domain; TM, transmembrane domain; TMPRSS2, transmembrane protease serine 2; ToMMo, Tohoku Medical Megabank Organization.

Table 1 Comparison of allele frequency between putative deleterious variants and putative tolerated variants

ACE2

	SIFT	SIFT4G	Polyphen2_HDIV	Polyphen2_HVAR	LRT	MutationTaster	MutationAssessor	FATHMM	PROVEAN	VEST4	MetaSVM
Deleterious	0.0008	0.0007	0.0011	0.0011	0.0011	0.0007	0.001	0.0011	0.001	0.0007	0.0011
Tolerated	0.0012	0.0011	0.0011	0.0011	0.0011	0.0011	0.0015	0.0011	0.0011	0.0011	0.0011
p-value	0.0661	0.1724	0.7992	0.2397	0.5946	0.1396	0.0617	0.0877	**0.043**	0.1569	0.9218

MetaLR	M-CAP	REVEL	MutPred	MVP	MPC	PrimateAI	DEOGEN2	CADD	DANN	fathmm-MKL	GenoCanyon
0.0012	0.001	0.0011	0.0006	0.0011	0.001	0.0006	NA	0.001	0.0011	0.001	0.0011
0.0011	0.0016	0.0011	0.001	0.001	0.0011	0.0011		0.0011	0.0011	0.0016	0.0013
0.2381	0.085	0.9882	**0.0365**	0.4816	0.333	0.1536		0.2478	0.86	**0.0497**	0.7635

TMPRSS2

	SIFT	SIFT4G	Polyphen2_HDIV	Polyphen2_HVAR	LRT	MutationTaster	MutationAssessor	FATHMM	PROVEAN	VEST4	MetaSVM
Deleterious	0.0011	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0004
Tolerated	0.0008	0.0008	0.0008	0.0008	0.0011	0.0012	0.0008	0.001	0.0011	0.0012	0.0032
p-value	0.9983	0.4043	0.3368	0.2947	**0.0026**	**0.0125**	0.3384	0.4184	0.1499	**0.0042**	**0.0008**

MetaLR	M-CAP	REVEL	MutPred	MVP	MPC	PrimateAI	DEOGEN2	CADD	DANN	fathmm-MKL	GenoCanyon
0.0007	0.006	0.0004	0.0008	0.0008	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008	NA
0.0012	0.0012	0.0012	0.0008	0.0012	0.0008	0.0011	0.0011	0.0012	0.0008	0.0012	
0.0071	0.082	**0.0004**	0.0766	**0.0238**	0.5259	**0.0179**	**0.0123**	**0.0271**	0.5279	**0.0388**	

The indicated values correspond to median allele frequency. Twenty-three deleteriousness-prediction software are employed in this study. P value is calculated using the Mann-Whitney U test. A probability of p<0.05 is considered statistically significant (values in bold).
ACE2, angiotensin-converting enzyme 2; NA, not available; TMPRSS2, transmembrane protease serine 2.

10 million individuals, indicating that variants described to date constitute only a small fraction of genetic variability in *ACE2* genes present on a population scale.

Genetic variants at the interface between the SARS-CoV-2 S-protein and human ACE2

A recent report suggested that while the sequence identity between the S-protein of SARS-CoV-2 and SARS-CoV is 73%, a significantly higher residue substitution rate was observed at the interaction interface with the ACE2 receptor.[49] Out of 29 interface residues, only 10 residues (34%) in SARS-CoV-2 are conserved with respect to SARS-CoV. Similarly, only 12 residues (40%) in the SARS-CoV are conserved with respect to SARS-CoV-2. Based on these recent data, we searched for genetic variations in the interface between S-protein and ACE2. A total of seven SNVs in *ACE2* were detected in the direct contact residues at the SARS-CoV-2 S-protein/hACE2 and SARS-CoV S-protein/hACE2 (table 2). There were also various SNVs in the neighbouring residues of the direct interface (online supplementary table S3).

Population-specific/population-enriched *ACE2* genetic variations

At the sequence level, genetic diversity is generated by de novo transmittable variants in one individual, which then undergo

natural selection and may spread through the population. Therefore, SNVs can become population-specific or population-enriched. Thus, we investigated the population-specific genetic variations in *hACE2* by analysing four large population sequencing projects (NHLBI EA (US European American, n=4300) and NHLBI AA (US African-American, n=2203), ToMMo (Japan, n=4773) and UK10K (UK, n=3781); total 15 057 individuals). Out of 52 non-synonymous variants, 44 (84.6%) were detected exclusively in any of the four populations (figure 1G) and no commonly shared variants were detected except synonymous variants (online supplementary table S3). The 1000G projects also revealed the country-specific/country-enriched SNVs, although further population size accumulation is necessary (online supplementary table S3). The fraction of very rare SNVs was lower in NHLBI AA and UK10K populations compared with NHLBI EA and ToMMo populations (figure 1H).

Genetic variations in human *TMPRSS2*

A total of 551 SNVs were identified in coding regions and splice sites and were broadly distributed throughout the coding regions (figure 2A). A complete list of the variants is summarised in online supplementary table S5. Most identified SNVs (n=247, 70.8%) were missense, while synonymous variants occurred in 26.9% of the SNVs (n=94) (figure 2B). The remaining 2% of SNVs were stop-gained (n=2), start-loss (n=1) or splice

Table 2 Genetic variations in hACE2 detected in the interface between SARS-CoV/SARS-CoV-2 S-protein and hACE2

SARS-CoV-2 S-protein	SARS-CoV S-protein	hACE2	Polymorphism in hACE2	GERP++_RS	phyloP100way_vertebrate	phastCons100way_vertebrate
A475, G476	P462	S19	p.Ser19Pro	−0.495	−0.801	0
F456, Y473, A475, Y489	L443, Y475	T27	p.Thr27Ala	1.85	0.345	0
Q493	–	E35	p.Glu35Lys	−0.741	−0.368	0
			p.Glu35Asp	0.813	−0.753	0.013
Y505	Y491	E37	p.Glu37Lys	5.81	2.967	0.995
F486	L472	M82	p.Met82Ile	−10.4	−1.463	0
–	R426	E329	p.Glu329Gly	2.8	1.025	0.001
T500, G502	T486, T487, G488	D355	p.Asp355Asn	5.34	7.905	1

hACE2, human angiotensin-converting enzyme 2; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Take home messages

- ► A total of 349 and 551 single nucleotide variations (SNVs) were detected in *ACE2* and *TMPRSS2*, respectively, in 156 513 individuals.
- ► The vast majority (>97%) of these SNVs were rare, population-specific and were computationally estimated to be deleterious.
- ► The SNVs in *ACE2* were distributed throughout the protein-coding regions and some were located in the severe acute respiratory syndrome coronavirus 2/human ACE2 contact residues.
- ► The total number of genetic variations in *ACE2* and *TMPRSS2* were estimated to be 1.1×10^3 and 1.5×10^3, respectively, for a population of 1 million people.

site variants (n=2) or indels (n=3) (figure 2B). Per gene, this resulted in a dN/dS >1. The majority of these SNVs were rare (n=347, 99.4%) or quite rare (n=169, 48.4%) with AF <1% or AF <0.001%, respectively (figure 2C). Fraction deleteriousness ranged from 14.4% to 100%, and the median was 43.5% (calculated by fathmm-MKL) (figure 2D, online supplementary table S6). Many scoring models revealed that rare SNVs were estimated to be more frequently deleterious than SNVs with a high AF (table 1).

On the basis of the regression analysis, the total number of genetic variations was expected to be approximately 1.5×10^3 and 5.5×10^3 at a sample size of 1 and 10 million individuals, respectively (figure 2E and F). Out of 88 non-synonymous variants identified in the four large population-sequencing projects (NHLBI EA, NHLBI AA, ToMMo and UK10K; total 15 057 individuals), 73 (83.0%) were detected exclusively in any of the 4 populations (figure 2G), whereas only 2 variants were shared among the 4 populations (online supplementary table S5). There was no large difference in the fraction of rare and common variants among the four different populations (figure 2H).

DISCUSSION

In the present study, we systematically investigated genetic variations in two coronavirus-related human genes. This large-scale effort yielded a huge global reservoir of genetic polymorphisms. The growing number of genetic variants in these genes, to which this publication contributes, will increase the potential to predict the different susceptibilities or responses to SARS-CoV-2 in different populations (including genders) under similar conditions. For example, the chemokine receptor protein encoded by the *CCR5* gene is best known to be the target for HIV infection, and genetic variations in *CCR5* can lead to differences in the rate of HIV infection, with the best known being the CCR5-Δ32 genetic polymorphism (rs333).[50 51] A previous study on SARS-CoV also has demonstrated that by altering the amino acid residues, ACE2 can be converted to an efficient/inefficient receptor for SARS-CoV and concluded that a number of *ACE2* variants could affect the association between ACE2 and S-protein in SARS-CoV.[35] Therefore, the present study could contribute to a better understanding of gender differences and different susceptibility and disease severity to viral infection in different ethnic groups. Furthermore, recent reports suggest that some candidate drugs do not target the coronavirus, but rather the hACE2 (eg, anti-ACE2 antibody[40]) or TMPRSS2 (eg, Nafamostat[52 53]). Therefore, the present study is also expected to be an essential step towards the translation of knowledge about

genetic polymorphisms into clinical strategies that benefit the patients.

ORCID iD
Kohei Fujikura http://orcid.org/0000-0002-0520-733X

REFERENCES

1 Zhu N, Zhang D, Wang W, *et al*. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;382:727–33.
2 Chen T, Wu D, Chen H, *et al*. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ* 2020:m1091.
3 Li Q, Guan X, Wu P, *et al*. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* 2020;382:1199–207.
4 Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *JAMA* 2020.
5 COVID-19 Dashboard by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU). Available: https://coronavirus.jhu.edu/map.html [Accessed 8 Jul 2020].
6 Wang C, Horby PW, Hayden FG, *et al*. A novel coronavirus outbreak of global health concern. *Lancet* 2020;395:470–3.
7 Prem K, Liu Y, Russell TW, *et al*. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* 2020;5:e261–70.
8 Armocida B, Formenti B, Ussai S, *et al*. The Italian health system and the COVID-19 challenge. *Lancet Public Health* 2020;5:e253.
9 Vigliar E, Iaccarino A, Bruzzese D, *et al*. Cytology in the time of coronavirus disease (covid-19): an Italian perspective. *J Clin Pathol* 2020.
10 Long C, Xu H, Shen Q, *et al*. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? *Eur J Radiol* 2020;126:108961.
11 Venter M, Richter K. Towards effective diagnostic assays for COVID-19: a review. *J Clin Pathol* 2020;73:370–7.
12 Zhang J-J, Dong X, Cao Y-Y, *et al*. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy* 2020. doi:10.1111/all.14238. [Epub ahead of print: 19 Feb 2020].
13 Andersen KG, Rambaut A, Lipkin WI, *et al*. The proximal origin of SARS-CoV-2. *Nat Med* 2020;26:450–2.
14 Liu W, Tao Z-W, Wang L, *et al*. Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease. *Chin Med J* 2020;133:1032–8.
15 Lan F-Y, Wei C-F, Hsu Y-T, *et al*. Work-Related COVID-19 transmission in six Asian countries/areas: a follow-up study. *PLoS One* 2020;15:e0233588.
16 Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *Lancet Digit Health* 2020;2:e201–8.
17 Livadiotis G. Statistical analysis of the impact of environmental temperature on the exponential growth rate of cases infected by COVID-19. *PLoS One* 2020;15:e0233875.
18 Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395:497–506.
19 Tian S, Hu W, Niu L, *et al*. Pulmonary pathology of early-phase 2019 novel coronavirus (COVID-19) pneumonia in two patients with lung cancer. *J Thorac Oncol* 2020;15:700–4.
20 Tian S, Xiong Y, Liu H, *et al*. Pathological study of the 2019 novel coronavirus disease (COVID-19) through postmortem core biopsies. *Mod Pathol* 2020;33:1007–14.

21 Zeng Z, Xu L, Xie X-Y, et al. Pulmonary pathology of early phase COVID-19 pneumonia in a patient with a benign lung lesion. *Histopathology* 2020. doi:10.1111/his.14138. [Epub ahead of print: 06 May 2020].

22 Hanley B, Lucas SB, Youd E, et al. Autopsy in suspected COVID-19 cases. *J Clin Pathol* 2020;73:239–42.

23 Fox SE, Akmatbekov A, Harbert JL, et al. Pulmonary and cardiac pathology in African American patients with COVID-19: an autopsy series from new Orleans. *Lancet Respir Med* 2020. doi:10.1016/S2213-2600(20)30243-5. [Epub ahead of print: 27 May 2020].

24 Menter T, Haslbauer JD, Nienhold R, et al. Post-Mortem examination of COVID19 patients reveals diffuse alveolar damage with severe capillary congestion and variegated findings of lungs and other organs suggesting vascular dysfunction. *Histopathology* 2020. doi:10.1111/his.14134. [Epub ahead of print: 04 May 2020].

25 Youd E, Moore L. COVID-19 autopsy in people who died in community settings: the first series. *J Clin Pathol* 2020. doi:10.1136/jclinpath-2020-206710. [Epub ahead of print: 30 Jun 2020].

26 Zhang H, Kang Z, Gong H, et al. Digestive system is a potential route of COVID-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process. *Gut* 2020;69:1010–8.

27 Jin X, Lian J-S, Hu J-H, et al. Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms. *Gut* 2020;69:1002–9.

28 Carvalho A, Alqusairi R, Adams A, et al. SARS-CoV-2 gastrointestinal infection causing hemorrhagic colitis: implications for detection and transmission of COVID-19 disease. *Am J Gastroenterol* 2020;115:942–6.

29 Lin L, Jiang X, Zhang Z, et al. Gastrointestinal symptoms of 95 cases with SARS-CoV-2 infection. *Gut* 2020;69:997–1001.

30 Hosoda T, Sakamoto M, Shimizu H, et al. SARS-CoV-2 enterocolitis with persisting to excrete the virus for approximately two weeks after recovering from diarrhea: a case report. *Infect Control Hosp Epidemiol* 2020;41:753–4.

31 Hofmann H, Pöhlmann S. Cellular entry of the SARS coronavirus. *Trends Microbiol* 2004;12:466–72.

32 Pillay TS. Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. *J Clin Pathol* 2020;73:366–9.

33 Li W, Moore MJ, Vasilieva N, et al. Angiotensin-Converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003;426:450–4.

34 Sims AC, Baric RS, Yount B, et al. Severe acute respiratory syndrome coronavirus infection of human ciliated airway epithelia: role of ciliated cells in viral spread in the conducting airways of the lungs. *J Virol* 2005;79:15511–24.

35 Li W, Zhang C, Sui J, et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *Embo J* 2005;24:1634–43.

36 Matsuyama S, Nagata N, Shirato K, et al. Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. *J Virol* 2010;84:12658–64.

37 Shirato K, Kawase M, Matsuyama S. Wild-Type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry. *Virology* 2018;517:9–15.

38 Glowacka I, Bertram S, Müller MA, et al. Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response. *J Virol* 2011;85:4122–34.

39 Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;395:565–74.

40 Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020;181:271–80.

41 Auton A, Brooks LD, Abecasis G. A global reference for human genetic variation. *Nature* 2015;526:68–74.

42 Fu W, O'Connor TD, Jun G, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013;493:216–20.

43 Wang Q, Pierce-Hoffman E, Cummings BB, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *bioRxiv* 2019:573378.

44 Nagasaki M, Yasuda J, Katsuoka F, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* 2015;6:8018.

45 Yasuda J, Katsuoka F, Danjoh I, et al. Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku medical Megabank project. *BMC Genomics* 2018;19:551.

46 UK10K Consortium, Walter K, Min JL, et al. The UK10K project identifies rare variants in health and disease. *Nature* 2015;526:82–90.

47 Liu X, Wu C, Li C, et al. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat* 2016;37:235–41.

48 Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family. *Pharmacogenet Genomics* 2015;25:584–94.

49 Brielle ES, Schneidman-Duhovny D, Linial M. The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor. *Viruses* 2020;12:986398.

50 Liu R, Paxton WA, Choe S, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* 1996;86:367–77.

51 Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722–5.

52 Yamamoto M, Matsuyama S, Li X, et al. Identification of nafamostat as a potent inhibitor of middle East respiratory syndrome coronavirus S protein-mediated membrane fusion using the Split-Protein-Based cell-cell fusion assay. *Antimicrob Agents Chemother* 2016;60:6532–9.

53 Yamamoto M, Kiso M, Sakai-Tagawa Y, et al. The anticoagulant nafamostat potently inhibits SARS-CoV-2 S protein-mediated fusion in a cell fusion assay system and viral infection in vitro in a Cell-Type-Dependent manner. *Viruses* 2020;12:E629.