Research Paper

ACE2 and *TMPRSS2* variants and expression as candidates to sex and country differences in COVID-19 severity in Italy

Rosanna Asselta[1,2,*], Elvezia Maria Paraboschi[1,2,*], Alberto Mantovani[1,2,3], Stefano Duga[1,2]

[1]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan 20090, Italy
[2]Humanitas Clinical and Research Center, IRCCS, Rozzano, Milan 20089, Italy
[3]The William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK
*Equal contribution

Correspondence to: Stefano Duga; **email:** stefano.duga@hunimed.eu
Keywords: SARS-CoV-2, COVID-19, ACE2, TMPRSS2, genetic variants
Received: April 16, 2020 **Accepted:** May 25, 2020 **Published:** June 5, 2020

ABSTRACT

As the outbreak of coronavirus disease 2019 (COVID-19) progresses, prognostic markers for early identification of high-risk individuals are an urgent medical need. Italy has one of the highest numbers of SARS-CoV-2-related deaths and one of the highest mortality rates. Worldwide, a more severe course of COVID-19 is associated with older age, comorbidities, and male sex. Hence, we searched for possible genetic components of COVID-19 severity among Italians by looking at expression levels and variants in *ACE2* and *TMPRSS2* genes, crucial for viral infection. Exome and SNP-array data from a large Italian cohort were used to compare the rare-variants burden and polymorphisms frequency with Europeans and East Asians. Moreover, we looked into gene expression databases to check for sex-unbalanced expression.
While we found no significant evidence that *ACE2* is associated with disease severity/sex bias, *TMPRSS2* levels and genetic variants proved to be possible candidate disease modulators, prompting for rapid experimental validations on large patient cohorts.

INTRODUCTION

As we write, Italy, Europe, and the entire world are facing one of the worst medical emergencies spanning centuries, the coronavirus disease 2019 (COVID-19) pandemia due to infection by SARS-CoV-2 virus. The early identification of risk factors for COVID-19 is an urgent medical need to provide the appropriate support to patients, including access to intensive care units.

Presently, Italy has one of the highest rate of SARS-CoV-2 infection in the world among large countries, with 371 cases per 100,000 people, one of the highest number of deaths and apparently also one of the highest mortality rates, 14.1% vs. an average value of 6.6% (as of May 16th, 2020, data from https://coronavirus.jhu.edu/map.html). These data may have different explanations, including: 1) the number of tests performed, 2) the structure of the population (Italy has the oldest population in Europe) [https://ec.europa.eu/eurostat/data/database], 3) the percentage of smokers, even though no significant association was found between smoking and severity of COVID-19 in a very recent study on the Chinese population [1], 4) the possible existence of a different virus strain [2], 5) a high population density in some hot spot areas of the infection, 6) the concentration of severe cases in a limited region of the country, potentially overwhelming the available intensive care units, 7) differences in environmental factors (e.g. air pollution), as well as 8) social factors, such as trust in the institutions and tendency to socialize [3]. However, there could also be some peculiar genetic characteristics of the Italian population that may have an impact on the susceptibility to viral infection, the

disease severity, and the number of patients shedding huge amounts of virus.

What is unquestionable is a more severe course of the disease associated with older age and high number of comorbidities and with the male sex (male:female ratio in case fatality rate among Italians 1.75, data from the Italian National Institute of Health: https://www.epicentro.iss.it/coronavirus/), a feature shared with the 2003 SARS epidemic and MERS [4–6]. Indeed, while males and females have similar susceptibility to both SARS-CoV-2 and SARS-CoV, males are more prone to have higher severity and mortality, independently of age [4]. Among the many possible factors impacting on sex-related differences in disease manifestations, including the fact that females are known to mount a stronger immune response to viral infections compared to males due to more robust humoral and cellular immune responses [7], we decided to center our attention on possible genetic components, with a particular focus on the Italian population.

It was recently demonstrated that both angiotensin I converting enzyme 2 (ACE2) and the transmembrane protease, serine 2 (TMPRSS2) are crucial for SARS-CoV-2 entry into host cells [8, 9]. As previously described for SARS-CoV, ACE2 is the main receptor also for the spike (S) protein of SARS-CoV-2, mediating viral attachment to target cells. Moreover, both coronaviruses use TMPRSS2 for protein S priming, i.e. the cleavage of protein S at the S1/S2 and the S2' sites, allowing fusion of viral and cellular membranes [9]. Both genes have been proposed to modulate susceptibility to SARS-CoV [10, 11], and are good candidates to mediate sex-related effects: *ACE2* is located on the X chromosome, while *TMPRSS2* expression is responsive to androgen/estrogen stimulation [12]. Controversial data have been reported on the level of expression of *ACE2* in the lung of males and females [13–15], however, it must also be taken into account the effect of estrogen drop in postmenopausal life and the possible compensating effect of hormone replacement therapy in some females.

With this background, we searched for possible genetic components of COVID-19 severity among Italians by looking at expression levels and genetic variants in *ACE2* and *TMPRSS2*, two crucial genes for viral infection.

RESULTS AND DISCUSSION

ACE2

For most X-chromosome genes, the double allelic dosage in females is balanced by the epigenetic silencing of one of the X chromosomes in early development [16]. However, the X-chromosome inactivation (XCI) is incomplete in humans and up to one third of genes are expressed from both alleles, with the degree of XCI escape varying between genes and individuals [17]. *ACE2* is one of the genes escaping X inactivation, but it belongs to a subgroup of X-chromosome genes escaping XCI showing an uncharacteristically heterogeneous pattern of male-female expression, with higher expression in males in several tissues [13]. Specifically concerning the lung, a recent analysis on published expression data, reported a substantial similar level of *ACE2* transcript in males and females [14], however, another study, using single-cell sequencing, found a higher expression of *ACE2* in Asian males [15]. Figure 1 reports data on *ACE2* mRNA expression levels in the lung as retrieved from the largest datasets available in the literature; no substantial differences were found between males and females, nor between younger and older females, thus confirming what already observed by Cai and colleagues [14].

Another possible sex-related effect might be due to the fact that males are hemizygous for the gene, therefore, in the presence of an *ACE2* allelic variant increasing disease susceptibility or severity, males will have all cells expressing the risk variant. Based on this hypothesis, we looked into the genetic variation in *ACE2*. A recent manuscript explored this same topic in different populations using data from public databases [18]. However, a specific analysis of the Italian population is lacking.

We have therefore exploited the available data on 3,984 exomes obtained from an Italian cohort representative of the whole country [19, 20] to extract the variants in exons and splice junctions of *ACE2*. Variants were filtered for quality and classified according to their predicted effect at protein level and on splicing. Concerning rare variants (i.e. those with a minor allele frequency, MAF, <1%; to be used in burden tests), we considered only null variants, abolishing or significantly impairing protein production (nonsense, out-of-frame ins/dels, and splicing variants), and missense variants predicted to be deleterious or possibly deleterious by all the 5 prediction algorithms used (see Supplementary Methods, paragraph "Definition of disrupting variants and statistical analysis"). Concerning common variants (i.e., MAF>5%), all were retained for comparing their frequency with those of the European (non Finnish) and East Asian populations, retrieved from the GnomAD repository.

No significant differences in the burden of rare deleterious variants were observed comparing the Italian population with Europeans and East Asians (Table 1A). Concerning common exonic variants, the

only striking difference, as also noticed by Cao and colleagues [18], was observed for the single nucleotide polymorphism (SNP) rs2285666 (also called G8790A), with the frequency of the rare A allele being 0.2 in Italians and Europeans, and 0.55 in East Asians (uncorrected P=2.2*10-16 for difference in Italians vs East Asians; corrected P=7.9*10-15; Table 1B). This variant was extensively studied as a potential risk factor

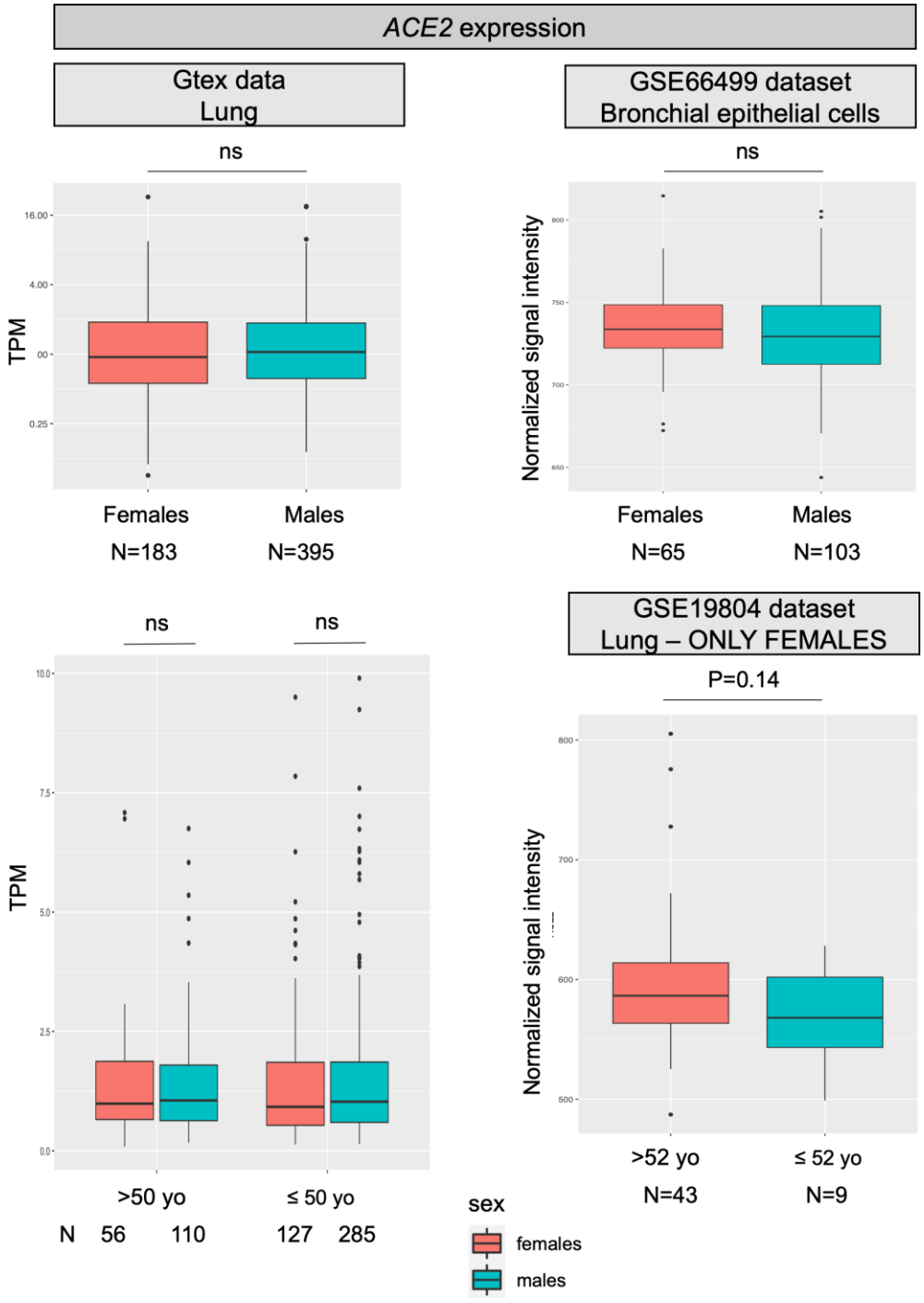

Figure 1. *ACE2* expression levels. All panels show *ACE2* mRNA expression levels in human normal lung samples stratified according to sex (or on sex and age). On left panels, data were retrieved for a total of 578 RNAseq experiments from the GTex repository. Expression levels are reported as transcripts per kilobase million (TPM). On the right, data were collected from two different datasets (GSE66499 and GSE19804) from the GEO database. Expression levels are reported as normalized signal intensities. P values were calculated by using either the Kruskal-Wallis or the student t test, using the R software (https://www.r-project.org/).

Table 1A. Burden of rare mutations in the *ACE2* gene in different populations.

Population	N alleles	T1	Freq T1	ITA	EUR	EAS
ITA	4422	7	0.0016	-	P=0.518	P=0.974
EUR	92545	200	0.0022	P=0.518	-	P=0.077
EAS	14840	21	0.0014	P=0.974	P=0.077	-

Total allele counts, carrier allele counts, and carrier frequencies are shown; only deleterious variants with MAF less than 1% were considered in the burden analysis. The 'deleterious' set is defined by missense variations predicted to be possibly damaging by all the 5 algorithms used (LRT score, MutationTaster, PolyPhen-2 HumDiv, PolyPhen-2 HumVar, and SIFT), and loss-of-function variants (nonsense, frameshift, and splicing variants affecting the donor/acceptor sites).
P values are presented as non-corrected; the number of statistical comparisons performed in Tables 1A, 1B, 2A, and 2B is collectively of 24, thus lowering the threshold for significance at P=0.0021 (Bonferroni threshold).
T1: alleles carrying damaging variants; Freq T1: frequency of T1 allele; ITA: Italian population; EUR: European population; EAS: East Asian population.

Table 1B. Common exon variants in the *ACE2* gene in different populations.

Variant ID	Consequence	A1/N alleles ITA	Freq ITA	A1/N alleles EUR	Freq EUR	A1/N alleles EAS	Freq EAS	ITA Vs EUR	ITA Vs EAS	EUR Vs EAS
rs2285666	c.439+4G>A	909/4408	0.206	17240/86164	0.200	7336/13387	0.548	0.331	**P<2.2e-16**	**P<2.2e-16**
rs35803318	p.Val749Val	235/4422	0.053	3935/88946	0.044	0/13918	0.0	P=0.0058	**P<2.2e-16**	**P<2.2e-16**

Total allele counts, carrier allele counts, and carrier frequencies are shown; only variants with MAF more than 5% were considered.
P values are presented as non-corrected; the number of statistical comparisons performed in Tables 1A, 1B, 2A, and 2B is collectively of 24, thus lowering the threshold for significance at P=0.0021 (Bonferroni threshold). Significant P values are indicated in bold.
A1: alleles carrying variants; Freq A1: frequency of A1 allele; ITA: Italian population; EUR: European population; EAS: East Asian population.

for hypertension, type 2 diabetes, and coronary artery disease [21, 22], hence possibly constituting a predisposing factor also for the comorbidities observed in COVID-19 patients. A single paper reports the association of the three rs2285666 genotypes with ACE2 protein level measured in serum by ELISA, with the A/A genotype having an expression level almost 50% higher than the G/G genotype, while heterozygous G/A individuals had intermediate levels [23]. Given the position of the variant, at nucleotide +4 in the donor splice site of intron 3 (c.439+4G>A), we calculated the predicted effect on splicing and indeed the substitution of G with an A is predicted to increase the strength of the splice site of about 9.2% (calculation made through the Human Splicing Finder v.3.1 webtool, http://www.umd.be/HSF/), consistently with the higher level of ACE2 protein in serum. It would be crucial to compare the frequency of this variant with *ACE2* expression in the lung and with susceptibility to viral infection and severity of COVID-19 manifestations. Of note, no eQTL for *ACE2* in the lung has been described so far in the GTEx database, and investigations on this topic are recommended.

TMPRSS2

TMPRSS2 is a gene well known to oncologists as genetic rearrangements producing a fusion between *TMPRSS2* and *ERG* (or, more rarely, other members of the ETS family) are the most frequent genetic lesions in prostate cancer patients [24]. As *TMPRSS2* is an androgen responsive gene, the fusion results in androgen dependent transcription of ERG in prostate tumor cells. Therefore, we can hypothesize that males might have higher *TMPRSS2* expression also in the lung, which might improve the ability of SARS-CoV-2 to enter cells by promoting membrane fusion. Looking into GTEx and GEO data, the overall expression of *TMPRSS2* in the lung is only slightly increased in males (P=0.029; Figure 2A). However, *TMPRSS2* expression is also promoted by estrogens [12], and therefore the situation might be different when considering individuals above 60 years, who are at higher risk of fatal events due to COVID-19, as in this group females will all be postmenopausal. According to this hypothesis, we checked the expression of the gene in lungs of males and females at different ages, but no

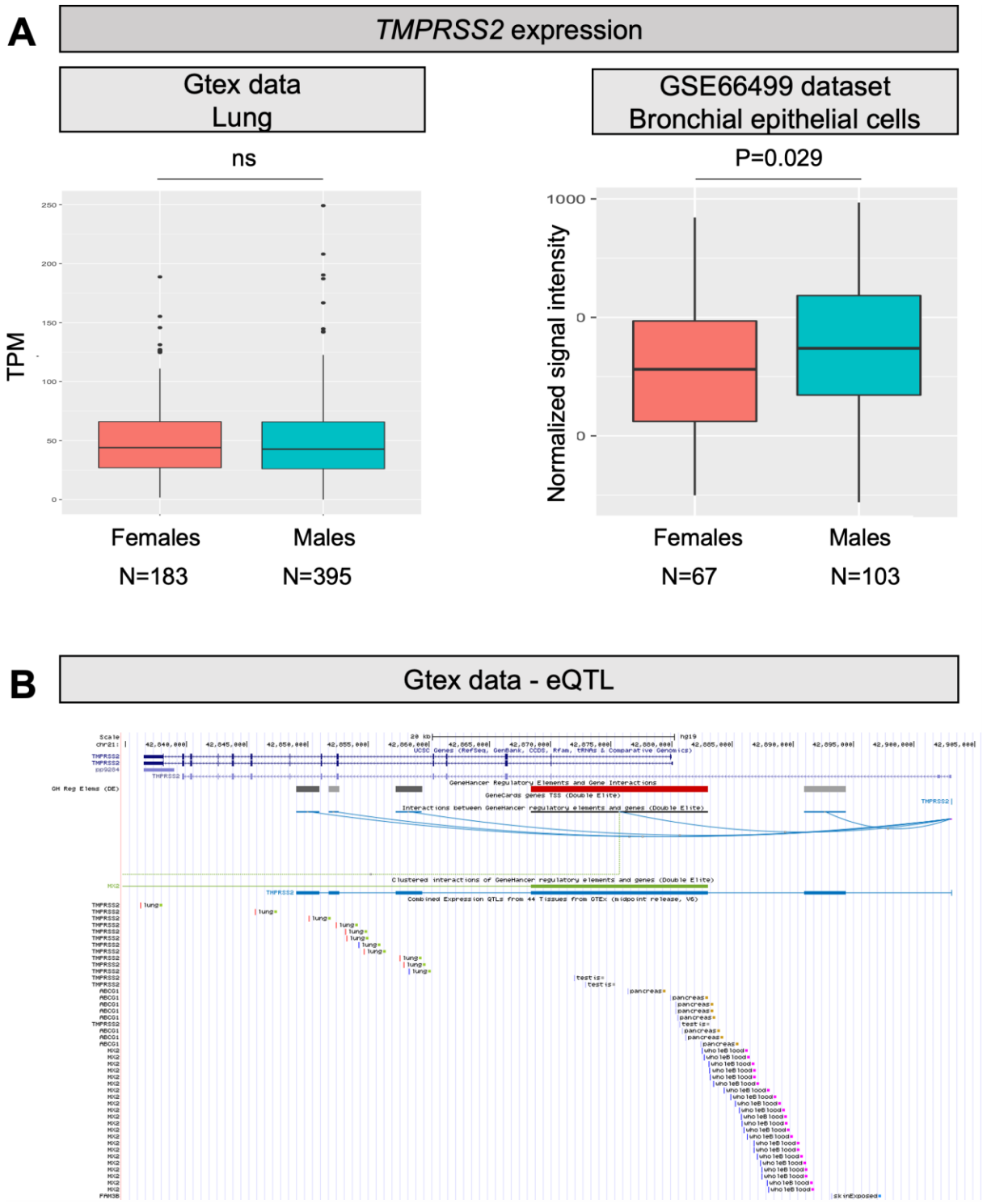

Figure 2. *TMPRSS2* expression levels and eQTLs. (**A**) Both panels show *TMPRSS2* mRNA expression levels in human normal lung samples stratified according to sex. On the left, data were retrieved for a total of 578 RNAseq experiments from the GTex repository. Expression levels are reported as transcripts per kilobase million (TPM). On the right, data were collected for a total of 170 microarray experiments from the GEO database. Expression levels are reported as normalized signal intensities. P values were calculated by using either the Kruskal-Wallis or the student t test. (**B**) Screenshot from the UCSC Genome browser (http://genome.ucsc.edu/; GRCh37/hg19) highlighting the *TMPRSS2* region (coordinates chr21: 42,835,000-42,905,000). The panel shows the following tracks: i) the ruler with the scale at the genomic level; ii) chromosome 21 nucleotide numbering; iii) the UCSC RefSeq track; iv) enhancers (grey and red bars) from GeneHancer database; v) interactions (curved lines) connecting GeneHancer regulatory elements and genes: all curved lines converge towards the androgen-responsive enhancer for the *TMPRSS2* gene described by Clinckemalie and colleagues [29].

substantial differences emerged between males and females (neither below, nor above 60 years of age; data not shown).

Finally, we explored genetic variation in *TMPRSS2* in search of variants, possibly already annotated as eQTL in the lung, which might have an impact on the serine protease expression as well as on its catalytic activity. Again, we used the available Italian exome data, as well as data deposited in GnomAD [25].

Firstly, we looked at the overall burden of deleterious rare variants, using the variant classification described above. Italians had a nominally significant decrease in the burden of deleterious variants compared to Europeans (uncorrected P=0.039, not significant after correction for multiple testing; Table 2A). This decrease was even more evident for the East Asian population (corrected P=8.6*10-4); however, in this case, we must consider that the number of individuals over 65 years of age in Italy is more than double the one in the Hubei province (22.7 vs. 10%, respectively) and this is a major determinant of disease lethality.

Focusing specifically on common exonic variants, 4 SNPs showed significantly (P<2.2*10-16) different frequencies when comparing the Italian population with East Asians (and with Europeans) (Table 2B); 3 of them are synonymous variants, whereas one is the missense substitution p.Val160Met, which impacts on a residue far from the serine protease catalytic triad. This variant was previously found significantly associated with genomic rearrangements involving *TMPRSS2*, with the risk of prostate cancer [26] and with shorter time to prostate cancer diagnosis for high-risk patients [27].

Concerning eQTLs, a number of variants significantly impacting on *TMPRSS2* expression in the lung (GTex data) are reported in the 3' region of the gene (Figure 2B). In Table 2C, a list of the most significant (P<1*10-8), together with their GnomAD frequencies in the East Asian and European populations, are reported. As for the Italian frequencies, we took advantage of the genome-wide association study (GWAS) performed on the above-described cohort (for a total of 3,284 individuals) [28]; in this case, we had to infer genotype frequencies by an imputation approach (for details, see Supplementary Methods). Interestingly, all these eQTLs appear to have extremely different frequencies among populations. In particular, 2 different haplotypes can be inferred from frequency data:

1) A frequent "European" haplotype (composed at least of SNPs rs463727, rs34624090, rs55964536, rs734056, rs4290734, rs34783969, rs11702475, rs35899679, and rs35041537), which is totally absent in the Asian population. Interestingly, this haplotype has been functionally linked to another eQTL (rs8134378), located at a known androgen-responsive enhancer for *TMPRSS2*, 13 kb upstream of the *TMPRSS2* transcription start site [29] (Figure 2B). Hence, this haplotype is expected to up regulate *TMPRSS2* gene expression in an androgen-specific way.

2) A second haplotype, predicted to be associated with higher *TMPRSS2* expression, is characterized by 3 SNPs (rs2070788, rs9974589, rs7364083), whose MAF is significantly increased in Europeans (9% increase in Italians respect to East Asians, corrected $P<6.8*10^{-9}$). Importantly, a small-scale GWAS, comparing the distribution of genetic variants in severe and mild cases of patients with A(H1N1)pdm09 influenza, identified rs2070788 as being associated with increased risk to both human A(H7N9) and severe A(H1N1)pdm09 influenza [11]. Of note, also in A(H7N9) influenza, the proportion of male patients was more than double that of female patients [30].

LIMITATIONS AND CONCLUSIONS

We are aware of the limitations of our study: first of all we focused our attention only to two candidate genes identified on the basis of their crucial role in viral infection and on the a priori probability that they might mediate sex-specific effects. A number of other X-linked genes (such as *IL13*, *IL4*, *IL10*, *XIST*, *TLR7*, *FOXP3)* and Y-linked genes (*SRY*, *SOX9*) may underlie sexually dimorphic immune responses [31]. Moreover, the number of non-genetic determinants of sex-biased severity and case fatality rates is huge and probably has to do not only with sex differences in both innate and adaptive immune responses [7], but also with gender and cultural habits in different countries. In particular, important gender-related factors might concern the social role of women (job, maternal and childcare role), the propensity to smoke, the hand hygiene compliance, as well as differences in the impact of the social role of women in the different countries.

In conclusion, we have explored possible genetic components impacting on COVID-19 severity, focusing on effects mediated by *ACE2* and *TMPRSS2* genes in the Italian population. From available data, it seems unlikely that sex-differences in *ACE2* levels can explain sex differences in disease severity. However, it remains to be evaluated if changes in *ACE2* levels in the lung correlate with susceptibility and severity of SARS-CoV-2 infection. Experimental data from patients with different disease manifestations are urgently needed. Among the analyzed hypotheses, the most interesting signals refer to sex-related differences in *TMPRSS2* expression and in genetic variation in *TMPRSS2*. In particular, we

Table 2A. Burden of rare mutations in the *TMPRSS2* gene in different populations.

Population	N alleles	T1	Freq T1	ITA	EUR	EAS
ITA	7968	30	0.0038	-	P=0.039	**P=3.6e-05**
EUR	129920	726	0.0056	P=0.039	-	**P=9.8e-16**
EAS	19979	25	0.0013	**P=3.6e-05**	**P=9.8e-16**	-

Total allele counts, carrier allele counts, and carrier frequencies are shown; only deleterious variants with MAF less than 1% were considered in the burden analysis. The 'deleterious' set is defined by missense variations predicted to be possibly damaging by all the 5 algorithms used (LRT score, MutationTaster, PolyPhen-2 HumDiv, PolyPhen-2 HumVar, and SIFT), and loss-of-function variants (nonsense, frameshift, and splicing variants affecting the donor/acceptor sites).
P values are presented as non-corrected; the number of statistical comparisons performed in Tables 1A, 1B, 2A, and 2B is collectively of 24, thus lowering the threshold for significance at P=0.0021 (Bonferroni threshold). Significant P values are indicated in bold.
T1: alleles carrying damaging variants; Freq T1: frequency of T1 allele; ITA: Italian population; EUR: European population; EAS: East Asian population.

Table 2B. Common exon variants in the *TMPRSS2* gene in different populations.

Variant ID	Consequence	A1/N alleles ITA	Freq ITA	A1/N alleles EUR	Freq EUR	A1/N alleles EAS	Freq EAS	ITA Vs EUR	ITA Vs EAS	EUR Vs EAS
rs2298659	p.Gly259Gly	1388/7968	0.174	28744/122880	0.234	5179/19478	0.266	**P<2.2e-16**	**P<2.2e-16**	**P<2.2e-16**
rs17854725	p.Ile256Ile	4131/7968	0.518	67712/122814	0.551	2544/19604	0.130	**P=1.16e-08**	**P<2.2e-16**	**P<2.2e-16**
rs12329760	p.Val160Met	1387/7968	0.174	29831/128604	0.232	7651/19934	0.384	**P<2.2e-16**	**P<2.2e-16**	**P<2.2e-16**
rs3787950	p.Thr75Thr	889/7968	0.112	9864/127666	0.077	2905/19600	0.148	**P<2.2e-16**	**P=1.39e-15**	**P<2.2e-16**

Total allele counts, carrier allele counts, and carrier frequencies are shown; only variants with MAF more than 5% were considered.
P values are presented as non-corrected; the number of statistical comparisons performed in Tables 1A, 1B, 2A, and 2B is collectively of 24, thus lowering the threshold for significance at P=0.0021 (Bonferroni threshold). Significant P values are indicated in bold.
A1: alleles carrying variants; Freq A1: frequency of A1 allele; ITA: Italian population; EUR: European population; EAS: East Asian population.

Table 2C. eQTL variants in the *TMPRSS2* gene in different populations.

Variant ID	P GTEx	NES GTEx	Freq ITA	Freq EUR	Freq EAS	ITA vs EUR	ITA vs EAS	EUR vs EAS
rs463727	5.0e-10	0.12	0.44	0.46	0.0051	P=0.038	**P<2.2e-16**	**P<2.2e-16**
rs2070788	8.9e-9	-0.11	0.55	0.53	0.66	P=0.003	**P=4.7e-15**	**P<2.2e-16**
rs9974589	7.4e-9	-0.12	0.55	0.53	0.66	P=0.002	**P=3.3e-15**	**P<2.2e-16**
rs34624090	9.2e-9	0.12	0.43	0.45	0.0051	P=0.005	**P<2.2e-16**	**P<2.2e-16**
rs7364083	3.3e-9	-0.12	0.56	0.53	0.65	**P=8.7e-05**	**P=1.9e-10**	**P<2.2e-16**
rs55964536	1.9e-9	0.12	0.46	0.49	0.0045	**P=4.6e-04**	**P<2.2e-16**	**P<2.2e-16**
rs734056	1.3e-9	0.12	0.47	0.49	0.0051	P=0.030	**P<2.2e-16**	**P<2.2e-16**
rs4290734	8.3e-10	0.12	0.47	0.49	0.0051	P=0.019	**P<2.2e-16**	**P<2.2e-16**
rs34783969	3.9e-10	0.12	0.47	0.49	0.0051	P=0.027	**P<2.2e-16**	**P<2.2e-16**
rs11702475	8.4e-10	0.12	0.47	0.49	0.0046	P=0.015	**P<2.2e-16**	**P<2.2e-16**
rs35899679	7.8e-9	0.11	0.44	0.46	0.0051	P=0.004	**P<2.2e-16**	**P<2.2e-16**
rs35041537	3.6e-9	0.12	0.44	0.47	0.0051	**P=8.6e-04**	**P<2.2e-16**	**P<2.2e-16**

P values are presented as non-corrected; the number of statistical comparisons performed in Table 2C is collectively of 36, thus lowering the threshold for significance at P=0.0013 (Bonferroni threshold). Significant P values are indicated in bold.
NES: normalized effect size; Freq: frequency of the minor allele; ITA: Italian population; EUR: European population; EAS: East Asian population.

identified an exonic variant (p.Val160Met) and 2 distinct haplotypes showing profound frequency differences between East Asians and Italians. The rare alleles of these haplotypes, all predicted to induce higher levels of *TMPRSS2*, are more frequent in the Italian than in the East Asian population; in one case, the haplotype could be regulated through androgens, thus possibly explaining the sex bias in COVID-19 severity, in the other case, a SNP belonging to the haplotype has been associated with increased susceptibility to influenza, possibly related to a higher susceptibility in Italians and Europeans.

Our data, beside suggesting possible explanations for the unusually high, relative to known data, lethality rates among Italians, provide reference frequencies in the general Italian population for candidate variants that can be compared to genetic data from patients infected by SARS-CoV-2 with different disease manifestations, as soon as they will be available on large numbers of patients. These studies will hopefully be of help in predicting the individual risk of infection and susceptibility to CoV-2 and in recognizing in advance infected individuals being at higher risk of poor prognosis.

MATERIALS AND METHODS

Gene expression data

Expression data for *ACE* and *TMPRSS2* genes were obtained through the: 1) genotype-tissue expression (GTEx) database (https://gtexportal.org/home/), which was also used to extract quantitative trait loci (eQTLs) for the two genes (all data based on RNAseq experiments); and 2) Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/). In particular, two GEO datasets were extracted and analyzed: 1) GSE66499, reporting microarray data on 152 normal lung samples from Caucasian individuals; 2) GSE19804, reporting microarray data on 60 normal lung samples from Taiwanese females (see also Supplementary Methods, paragraph "Datasets and statistical power estimations").

Genetic data

Genetic data for general European and East Asian populations were retrieved through the GnomAD repository, which contains data on a total of 125,748 exomes and 71,702 genomes (https://gnomad.broadinstitute.org/).

As for Italians, details on whole-exome sequencing (on 3,984 individuals) and genome-wide microarray genotyping (on 3,284 individuals) of the analyzed cohort are specified elsewhere [19, 20, 28], as well as in Supplementary Methods (paragraphs "Sequencing" and "Datasets and statistical power estimations"). Imputation procedures are detailed in Supplementary materials (paragraph "Dataset imputation").

Statistical analysis

Expression levels were compared by using either the Kruskal-Wallis test (RNAseq data) or the student t test (microarray data). Allele frequencies were compared using the chi square test. All calculations were performed using the R software (https://www.r-project.org/). P values are presented as non-corrected for multiple testing, but the Bonferroni-corrected threshold of significance is indicated below each set of comparisons presented in Tables. Power calculations have been described in Supplementary Methods (paragraph "Datasets and statistical power estimations").

AUTHOR CONTRIBUTIONS

All authors contributed to the study design. EMP did the genetic analysis, RA performed the statistical analysis, SD drafted the manuscript and supervised the entire study. All authors critically reviewed the manuscript and approved the final draft.

CONFLICTS OF INTEREST

No conflicts of interest to disclose

FUNDING

This work was supported by Ricerca Corrente (Italian Ministry of Health), intramural funding (Fondazione Humanitas per la Ricerca). Generous contributions of the Dolce and Gabbana Fashion Firm and of Banca Intesa San Paolo are gratefully acknowledged.

REFERENCES

1. Zhang JJ, Dong X, Cao YY, Yuan YD, Yang YB, Yan YQ, Akdis CA, Gao YD. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. Allergy. 2020. [Epub ahead of print]. https://doi.org/10.1111/all.14238 PMID:32077115

2. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020; 26:450–52. https://doi.org/10.1038/s41591-020-0820-9 PMID:32284615

3. Oksanen A, Kaakinen M, Latikka R, Savolainen I, Savela N, Koivula A. Regulation and trust: 3-month follow-up study on COVID-19 mortality in 25 european countries. JMIR Public Health Surveill. 2020; 6:e19218.

https://doi.org/10.2196/19218
PMID:32301734

4. Jin JM, Bai P, He W, Wu F, Liu XF, Han DM, Liu S, Yang JK. Gender differences in patients with COVID-19: focus on severity and mortality. Front Public Health. 2020; 8:152. https://doi.org/10.3389/fpubh.2020.00152 PMID:32411652

5. Channappanavar R, Fett C, Mack M, Ten Eyck PP, Meyerholz DK, Perlman S. Sex-based differences in susceptibility to severe acute respiratory syndrome coronavirus infection. J Immunol. 2017; 198:4046–53. https://doi.org/10.4049/jimmunol.1601896 PMID:28373583

6. Alghamdi IG, Hussain II, Almalki SS, Alghamdi MS, Alghamdi MM, El-Sheemy MA. The pattern of middle east respiratory syndrome coronavirus in Saudi Arabia: a descriptive epidemiological analysis of data from the saudi ministry of health. Int J Gen Med. 2014; 7:417–23. https://doi.org/10.2147/IJGM.S67061 PMID:25187734

7. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016; 16:626–38. https://doi.org/10.1038/nri.2016.90 PMID:27546235

8. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579:270–73. https://doi.org/10.1038/s41586-020-2012-7 PMID:32015507

9. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020; 181:271–80.e8. https://doi.org/10.1016/j.cell.2020.02.052 PMID:32142651

10. Kuba K, Imai Y, Rao S, Gao H, Guo F, Guan B, Huan Y, Yang P, Zhang Y, Deng W, Bao L, Zhang B, Liu G, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. Nat Med. 2005; 11:875–9. https://doi.org/10.1038/nm1267 PMID:16007097

11. Cheng Z, Zhou J, To KK, Chu H, Li C, Wang D, Yang D, Zheng S, Hao K, Bossé Y, Obeidat M, Brandsma CA, Song YQ, et al. Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic a(H1N1) influenza and a(H7N9) influenza. J Infect Dis. 2015; 212:1214–21. https://doi.org/10.1093/infdis/jiv246 PMID:25904605

12. Baena E, Shao Z, Linn DE, Glass K, Hamblen MJ, Fujiwara Y, Kim J, Nguyen M, Zhang X, Godinho FJ, Bronson RT, Mucci LA, Loda M, et al. ETV1 directs androgen metabolism and confers aggressive prostate cancer in targeted mice and patients. Genes Dev. 2013; 27:683–98. https://doi.org/10.1101/gad.211011.112 PMID:23512661

13. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, Cummings BB, Castel SE, Karczewski KJ, et al, and GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration & Visualization—EBI, and Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz. Landscape of X chromosome inactivation across human tissues. Nature. 2017; 550:244–48. https://doi.org/10.1038/nature24265 PMID:29022598

14. Cai G. Bulk and single-cell transcriptomics identify tobacco-use disparity in lung gene expression of ACE2, the receptor of 2019-nCov. Preprints. 2020; 2020020051. https://doi.org/10.20944/preprints202002.0051.v2

15. Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. bioRxiv. 2020. https://doi.org/10.1101/2020.01.26.919985

16. Lyon MF. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature. 1961; 190:372–3. https://doi.org/10.1038/190372a0 PMID:13764598

17. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005; 434:400–4. https://doi.org/10.1038/nature03479 PMID:15772666

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

Discov. 2020; 6:11.
https://doi.org/10.1038/s41421-020-0147-1
PMID:32133153

19. Do R, Stitziel NO, Won HH, Jørgensen AB, Duga S, Angelica Merlini P, Kiezun A, Farrall M, Goel A, Zuk O, Guella I, Asselta R, Lange LA, et al, and NHLBI Exome Sequencing Project. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015; 518:102–6.
https://doi.org/10.1038/nature13917
PMID:25487149

20. Paraboschi EM, Khera AV, Merlini PA, Gigante L, Peyvandi F, Chaffin M, Menegatti M, Busti F, Girelli D, Martinelli N, Olivieri O, Kathiresan S, Ardissino D, et al. Rare variants lowering the levels of coagulation factor X are protective against ischemic heart disease. Haematologica. 2019. [Epub ahead of print].
https://doi.org/10.3324/haematol.2019.237750
PMID:31699787

21. Chaoxin J, Daili S, Yanxin H, Ruwei G, Chenlong W, Yaobin T. The influence of angiotensin-converting enzyme 2 gene polymorphisms on type 2 diabetes mellitus and coronary heart disease. Eur Rev Med Pharmacol Sci. 2013; 17:2654–9.
PMID:24142614

22. Yang M, Zhao J, Xing L, Shi L. The association between angiotensin-converting enzyme 2 polymorphisms and essential hypertension risk: a meta-analysis involving 14,122 patients. J Renin Angiotensin Aldosterone Syst. 2015; 16:1240–4.
https://doi.org/10.1177/1470320314549221
PMID:25237167

23. Wu YH, Li JY, Wang C, Zhang LM, Qiao H. The ACE2 G8790A polymorphism: involvement in type 2 diabetes mellitus combined with cerebral stroke. J Clin Lab Anal. 2017; 31:e22033.
https://doi.org/10.1002/jcla.22033
PMID:27500554

24. Kron KJ, Murison A, Zhou S, Huang V, Yamaguchi TN, Shiah YJ, Fraser M, van der Kwast T, Boutros PC, Bristow RG, Lupien M. TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nat Genet. 2017; 49:1336–45.
https://doi.org/10.1038/ng.3930
PMID:28783165

25. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. 2019.

26. FitzGerald LM, Agalliu I, Johnson K, Miller MA, Kwon EM, Hurtado-Coll A, Fazli L, Rajput AB, Gleave ME, Cox ME, Ostrander EA, Stanford JL, Huntsman DG. Association of TMPRSS2-ERG gene fusion with clinical characteristics and outcomes: results from a population-based study of prostate cancer. BMC Cancer. 2008; 8:230.
https://doi.org/10.1186/1471-2407-8-230
PMID:18694509

27. Giri VN, Ruth K, Hughes L, Uzzo RG, Chen DY, Boorjian SA, Viterbo R, Rebbeck TR. Racial differences in prediction of time to prostate cancer diagnosis in a prospective screening cohort of high-risk men: effect of TMPRSS2 Met160Val. BJU Int. 2011; 107:466–70.
https://doi.org/10.1111/j.1464-410X.2010.09522.x
PMID:20735386

28. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, et al, and Myocardial Infarction Genetics Consortium, and Wellcome Trust Case Control Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41:334–41.
https://doi.org/10.1038/ng.327
PMID:19198609

29. Clinckemalie L, Spans L, Dubois V, Laurent M, Helsen C, Joniau S, Claessens F. Androgen regulation of the TMPRSS2 gene and the effect of a SNP in an androgen response element. Mol Endocrinol. 2013; 27:2028–40.
https://doi.org/10.1210/me.2013-1098
PMID:24109594

30. Jernigan DB, Cox NJ. H7N9: preparing for the unexpected in influenza. Annu Rev Med. 2015; 66:361–71.
https://doi.org/10.1146/annurev-med-010714-112311
PMID:25386931

31. Ghosh S, Klein RS. Sex drives dimorphic immune responses to viral infections. J Immunol. 2017; 198:1782–90.
https://doi.org/10.4049/jimmunol.1601166
PMID:28223406

SUPPLEMENTARY MATERIALS

Supplementary Methods

Sequencing

Whole-exome sequencing (WES) was performed at the Broad Institute (Boston, MA). Demographic characteristics, as well as exome capture methods, sequencing, variant annotation, and data processing of the samples were described previously [1].

Definition of disrupting variants and statistical analysis

Using WES data, we searched the *ACE2* and *TMPRSS2* genes for loss-of-function variants (nonsense, frameshift, splicing, or disrupting missense mutations). Missense variants were considered damaging if they were predicted to be deleterious or possibly deleterious by all the 5 prediction algorithms used: LRT (likelihood ratio test) [2], MutationTaster [3], PolyPhen-2 HumDiv, PolyPhen-2 HumVar [4], and SIFT [5].

The positions of mutations were based on the cDNA reference sequence for *ACE2* and *TMPRSS2* (NM_021804 and NM_005656) with the ATG initiation codon numbered as residue 1 (p.Met1).

Burden test analyses were performed considering only those variants having a minor allele frequency (MAF) <1%. Significance in the differences of MAFs between different populations were calculated using chi-square tests, with the R software (https://www.r-project.org/). A $P<0.05$ was considered to indicate statistical significance.

Dataset imputation

When missing from exome data, intronic variant frequencies in *TMPRSS2* were retrieved from SNP-array data obtained from the same Italian cohort. Genome-wide genotyping was performed at the Broad Institute. Genotyping details and data processing of the samples have been already described [6].

Imputation was performed remotely using the Michigan Imputation Server (https://imputationserver.sph.umich.edu) [7], using the 1000G Phase 3 v5 as reference panel, ShapeIT v2.r790 for the phasing step [8], and Minimac3 [7] as imputation software. The imputed dataset was then filtered to retain only those variants with $r^2>0.3$.

Datasets and statistical power estimations

For expression data analyses, we took advantage of microarray data reported in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/). We specifically searched for the wider datasets reporting expression data on normal lung tissues derived from individuals whose sex and geographical origin were specified (search done by keywords, filters based on the number of available samples in the dataset, and by a final manual inspection of the retrieved data). This search allowed the identification of two datasets: GSE66499 and GSE19804, for a total of 115 samples from male individuals, and 135 samples from female subjects. Indeed, it is difficult to provide an accurate power estimate for a microarray study. Among others, [9] suggested that a sample size of 20 is necessary, at a P value of 0.01 and 90% power, to detect a two-fold change in the 75% least variable genes in a microarray study. Based on this observation, the data available through the GSE66499 and GSE19804 datasets were considered reasonably powered to identify possible altered levels in the ACE2 and TMPRSS2 genes.

As for genotype data, from one side we took advantage of exome and SNP-array in-house data on ~3,500 individuals; [1, 6], from the other of exome and genome data on the largest dataset freely accessible online, i.e. the GnomAD repository (https://gnomad.broadinstitute.org/). For GnomAD data, we extracted allele/genotype frequencies available for East Asian and European individuals, for a total of at least 9,967 and 64,302 subjects, respectively. The use of such large cohorts ensured us to be sufficiently powered to detect significant differences in allele frequencies between the analyzed populations. As an example, a sample size of 2,000 pairs has an approximately 80% power of detecting a significant allele difference at $P<0.05$ if the frequency of the rare allele is 2%. For higher frequencies of 10% or more, the power of detection increases to more than 90%.

REFERENCES

1. Do R, Stitziel NO, Won HH, Jørgensen AB, Duga S, Angelica Merlini P, Kiezun A, Farrall M, Goel A, Zuk O, Guella I, Asselta R, Lange LA, et al, and NHLBI Exome Sequencing Project. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015; 518:102–6. https://doi.org/10.1038/nature13917 PMID:25487149

2. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19:1553–61. https://doi.org/10.1101/gr.092619.109 PMID:19602639

3. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7:575–6. https://doi.org/10.1038/nmeth0810-575 PMID:20676075

4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–9. https://doi.org/10.1038/nmeth0410-248 PMID:20354512

5. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–81. https://doi.org/10.1038/nprot.2009.86 PMID:19561590

6. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, et al, Myocardial Infarction Genetics Consortium, and Wellcome Trust Case Control Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41:334–41. https://doi.org/10.1038/ng.327 PMID:19198609

7. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016; 48:1284–87. https://doi.org/10.1038/ng.3656 PMID:27571263

8. Delaneau O, Coulonges C, Zagury JF. shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics. 2008; 9:540. https://doi.org/10.1186/1471-2105-9-540 PMID:19087329

9. Wei C, Li J, Bumgarner RE. Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics. 2004; 5:87. https://doi.org/10.1186/1471-2164-5-87 PMID:15533245