

SUPPLEMENTARY MATERIAL

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

Rosanna Asselta, PhD^{1,2*} Elvezia Maria Paraboschi, PhD^{1,2*}, Alberto Mantovani, MD^{1,2,3}
Stefano Duga, PhD^{1,2}

¹ Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy;1

² Humanitas Clinical and Research Center, IRCCS, Via Manzoni 56, 20089 Rozzano, Milan, Italy.

³ The William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK.

* These authors contributed equally to this work

Content:

- Supplementary Methods
- Supplementary References
- Supplementary Figure 1

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 [Do et al., 2015].

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) [Chun et al., 2009], MutationTaster [Schwarz et al., 2010], PolyPhen-2 HumDiv, PolyPhen-2 HumVar [Adzhubei et al., 2010], and SIFT [Kumar et al., 2009].

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 [Migen consortium, 2009].

Imputation was performed remotely using the Michigan Imputation Server (<https://imputationserver.sph.umich.edu>) [Das et al., 2016], using the 1000G Phase 3 v5 as reference panel, ShapeIT v2.r790 for the phasing step [Delaneau et al., 2008], and Minimac3 [Das et al., 2016] as imputation software. The imputed dataset was then filtered to retain only those variants with $r^2 > 0.3$.

SUPPLEMENTARY REFERENCES

Do R, Stitzel NO, Won HH, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2015; **518**: 102–6.

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009; **19**: 1553–61.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–6.

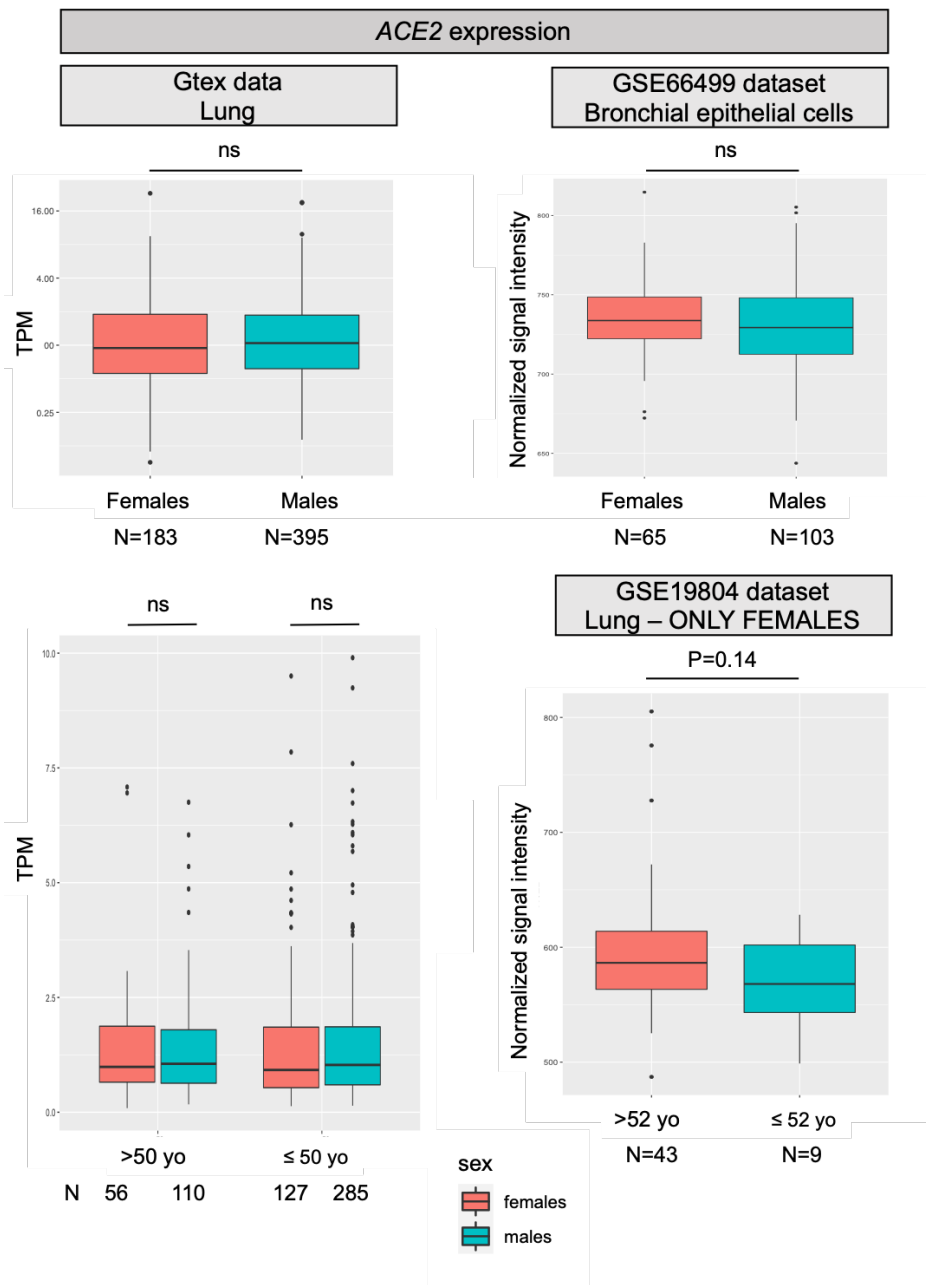
Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–9.

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.

Myocardial Infarction Genetics Consortium, Kathiresan S, Voight BF, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* 2009; **41**: 334–41.

Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016; **48**: 1284–7.

Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 2008; **9**: 540.



Supplementary 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/>).