Human-SARS-CoV-2 interactome and human genetic diversity: *TMPRSS2*-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans

Fernanda SG Kehdy[1], Murilo Pita-Oliveira[2], Mariana M Scudeler[2], Sabrina Torres-Loureiro[2], Camila Zolini[3,4], Rennan Moreira[3], Lucas A Michelin[3], Isabela Alvim[3], Carolina Silva-Carvalho[3], Vinicius C Furlan[3], Marla M Aquino[3], Meddly L. Santollala[5], Victor Borda[6], Giordano B Soares-Souza[3], Luis Jaramillo-Valverde[7], Andres Vasquez-Dominguez[7], Cesar Sanchez Neira[8], Renato S Aguiar[3], Ricardo A. Verdugo[9,10], Timothy D. O`Connor[11,12,13], Heinner Guio[8,14], Eduardo Tarazona-Santos[3], Thiago P Leal[3,15], Fernanda Rodrigues-Soares[2,15]

1 Laboratório de Hanseníase, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Rio de Janeiro, 21040-900, Brazil.
2 Departamento de Patologia, Genética e Evolução, Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, 38025-350, Uberaba, Minas Gerais, Brazil.
3 Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil.
4 Mosaico Translational Genomics Initiative. Belo Horizonte. Brazil.
5 Emerging Diseases and Climate Change Research Unit, School of Public Health and Administration, Universidad Peruana Cayetano Heredia, Lima, Peru.
6 Laboratório de Bioinformática, Laboratório Nacional de Computação Científica (LNCC). Petrópolis, Rio de Janeiro, Brazil
7 INBIOMEDIC Research and Technological Center, Lima, Peru
8 Instituto Nacional de Salud, Lima, Peru.
9 Programa de Genética Humana, Instituto de Ciencias Biomédicas, Facultad de Medicina, Universidad de Chile, Santiago, Chile.
10 Departamento de Oncología Básico Clínica, Facultad de Medicina, Universidad de Chile, Santiago, Chile.
11 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD.
12 Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD
13 Department of Medicine, University of Maryland School of Medicine, Baltimore, MD
14 Universidad de Huánuco, Huanuco, Peru
15 Both authors should be considered as Senior Authors

Correspondence to:
Fernanda Rodrigues-Soares, PhD

Universidade Federal do Triângulo Mineiro
Instituto de Ciências Biológicas e Naturais
Rua Vigário Carlos, 100, sala 314, Nossa Senhora da Abadia, CEP 38025-350
Uberaba – MG – Brazil
E-mail: fernanda.soares@uftm.edu.br
Phone: +55 34 3700-6847

ABSTRACT

The current search for host-susceptibility variants for COVID-19 contrasts with the fact that the study of the genetic architecture of Severe Acute Respiratory Syndrome (SARS) has been neglected. For human/SARS-CoV-2 interactome genes ACE2, TMPRSS2 and BSG, there is only one convincing evidence of association in Asians with influenza-induced SARS for TMPRSS2-rs2070788, tag-SNP of the eQTL rs383510. This case illustrates the importance of population genetics and of sequencing data in the design of genetic association studies in different human populations: the high linkage disequilibrium (LD) between rs2070788 and rs383510 is Asian-specific. Leveraging on a combination of genotyping and sequencing data for Native Americans (neglected in genetic studies), we show that while their frequencies of the Asian tag-SNP rs2070788 is, surprisingly, the highest worldwide, it is not in LD with the eQTL rs383510, that therefore, should be directly genotyped in genetic association studies of SARS in populations with Native American ancestry.

Keywords: TMPRSS2, ACE2, BSG, Native Americans, SARS-CoV-2, Genomics

Main Text

In the context of a global interest in host genetic determinants of COVID-19 susceptibility (Casanova & Su, 2020) we established a three-steps protocol to gain evidence about human genetic susceptibility to the new coronavirus 2019 disease (COVID-19):

(i) a systematic review of the literature about genes *ACE2* (angiotensin converting enzyme 2, Xp22.2), *TMPRSS2* (transmembrane serine protease 2, 21q22.3) and *BSG* (basigin, 19p13.3), which codify important proteins for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. SARS-CoV-2 spike S protein contains subunits S1 and S2, which bind the ACE2 cellular receptor, leading to an endosome formation around the virus. After this binding, TMPRSS2 host's transmembrane serine protease cleaves S1/S2 subunits and induces a conformational change in S2, facilitating the endosome formation and allowing the entrance of virus cellular into the cytoplasm. CD147 (also called basigin - BSG) is a transmembrane glycoprotein, encoded by *BSG* gene, discovered as a new SARS-CoV-2 cellular entry route (Wang et al., 2020). We

performed a systematic review under the terms “[gene name] genetics infection]", covering articles published until June 4th, 2020 in PubMed and in biorXiv during 2020 (Figure 1A). For the ACE2 and BSG viral receptors, there is no solid and direct evidence of association between genetic polymorphisms and any respiratory viral infections.

(ii) we annotated SNVs in *ACE2*, *TMPRSS2*, and *BSG* mining and integrating information from twenty-four biological and biomedical databases, using our bioinformatics tool (MASSA) [Multi-Agent System for SNP Annotation (Soares-Souza, 2014)], to identify functionally relevant variants (Table S1-A). MASSA integrates data with clinical findings from NCBI Databases like ClinVar and ClinGen. MASSA also includes approaches to distinguish between functional alleles, underlying clinical phenotypes and benign variants, cross-checking the data with multiple different databases. To ensure that collected variants are relevant for our analysis, MASSA performs some secondary filters, taking into account the frequency of alleles and SIFT and Polyphen predictions. The tool, in addition to performing the filters described above, searches for variants that have been cited in PubMed and also compares them to the OMIM database. From that, we've found 26 putatively functional variants for ACE2, 5 for TMPRSS2 and 17 for BSG gene, resulting in a total of 48 genetic variants.

(iii) we performed population genetics analysis of the 48 functionally relevant variants in the ACE2, TMPRSS2 and BSG genes in human populations to detect particular patterns of between-population genetic differentiation and independently of evidence of genetic association between ACE2, TMPRSS2 and BSG variants and infectious diseases, using published and unpublished data from different worldwide populations (Table S1-B), enriched for Latin Americans, who are mainly the product of admixture of Native Americans, Europeans and Africans. Unpublished data include the Peruvian Native Americans from the *Laboratório de Diversidade Genética Humana (UFMG)* and the whole genome sequenced Native Americans and admixed Peruvian populations from Peruvian Genome Project.

Main results

ACE2 and *BSG* allele frequencies and their regression analyses between population genomic ancestry (Native American, African, European and East Asian) and frequencies of functionally relevant SNPs are presented in Table S2 and S3, respectively. We did not

observe a particular pattern of inter-population genetic diversity for most of our 48 analyzed SNPs. Our most illustrative result regards *TMPRSS2* (Table S4). In our systematic review, the only genotype/infection association was reported by Cheng et al. (2015), between rs2070788-G (NC_000021.9:g.41470061G>A), a tag-SNP (i.e. in high linkage disequilibrium, r^2>0.80) of the regulatory e-QTL rs383510 (NC_000021.9:g.41486440T>A). Both SNPs were associated in Asiatic populations with severe pulmonary damage caused by influenza A(H7N9) in 2014 (OR 1.70 [1.13-2.55]) and rs2070788 was associated with severe pulmonary damage caused by the influenza A(H1N1) in 2009 (OR 1.54 [1.14–2.06]). The authors validated their finding by an in-vitro polymerase assay, showing that rs383510 maps on a region that regulates *TMPRSS2* expression, and therefore is a functionally relevant SNP tagged by rs2070788-G. This result and the role of TMPRSS2 in SARS-CoV-2 infection suggest that there are shared elements in the pathogenesis of SARS caused by different viral infections.

As in Cheng et al. (2015), the tag-SNP rs2070788 (https://www.ncbi.nlm.nih.gov/snp/rs2070788) is more commonly studied than the functional SNP rs383510 (https://www.ncbi.nlm.nih.gov/snp/rs383510), because the former is present in more SNP genome-wide arrays and has a TaqMan (Thermo Fisher, US) probe, while rs383510 does not. We examined our unpublished dataset of Native American and of admixed Latin Americans for the putative tag-SNP rs2070788 (genotyped with the Illumina Omni2.5 array) but not for rs383510 because there is no large dataset available for it. We realized that, interestingly, frequencies of the putative tag-SNP rs2070788-G are strongly correlated with population Native American ancestry (Figure 1B, Table S4), and its highest frequency worldwide are in Native Americans. Non-admixed Native American populations have frequencies between 76% and 94%, compared to around 50% in Europeans and 30-40% in Asians. Furthermore, the putative tag-SNP rs2070788-G is among the 5% most differentiated SNPs in Native Americans respect to Asians (the genetically closest continental group, Figure 1C). This result led us to hypothesize that Native Americans may have the highest frequencies of SARS susceptibility alleles in *TMPRSS2* and to test this hypothesis we designed a future association study between rs2070788 and COVID-19 in Peru (a country with predominant Native American ancestry).

Because Harris et al. (2018) published whole genome sequencing data for 150 Peruvian individuals with high Native American ancestry, we used those data to test the linkage disequilibrium between the putative tag-SNP rs2070788 and the functional SNP rs383510. Surprisingly for us, in these Native Americans, the continental group that, on average, shows the highest linkage disequilibrium in the human genome (Bosch et al., 2009), there is no linkage disequilibrium between rs2070788 and rs383510 (r^2=0.05, D'=0.61, Fig 1D). We verified that rs2070788 and rs383510 are in linkage disequilibrium only in Asian populations (Fig 1D) and therefore, the former is a tag-SNP of the latter functional SNP only in Asians. Thus, based on our current knowledge, there is no evidence that Native Americans have the highest frequency worldwide of *TMPRSS2* SARS susceptibility variants, as a superficial analysis would suggest. An association study in Native Americans should focus on the causative variant rs383510, to test its involvement in SARS induced by viral infection.

In summary, this case illustrates that, to properly design genetic association studies, it is compelling to: (i) consider the complexities of population genetics concepts such as differences not only in frequencies but also in linkage disequilibrium among different human populations, (ii) to have access to whole genome sequencing data for the broadest array of human populations, as we have in this case for Peruvians Native Americans. Moreover, if for any reason, including socioeconomic vulnerability, COVID-19 is more common in individuals with high Native American ancestries, the test of association between the rs383510 and COVID-19 phenotypes should be controlled for ancestry. Without considering differences in linkage disequilibrium (also for imputation in GWAS) and sequencing data, as well as ancestry, this is an example of how association studies may reach misleading conclusions in times when a search for susceptibility variants for SARS-CoV-2 is intense.

Methods

Systematic Review

The search was performed in Pubmed, using the terms "ace2 genetics infection", "tmprss2 genetics infection", "cd147 genetics infection" and "bsg genetics infection" covering articles published until June 4th, 2020. Additionally, we searched the biorxiv database for studies of 2020. Studies included should have covered any type of infection and information about genetic polymorphisms in *ACE2*, *TMPRSS2* or *BSG*.

*SNPs annotation - MASSA (*Multi-Agent System for SNP Annotation*)*

Variation lists of each gene were downloaded from human genome assembly Feb.2009 (GRCH37/hg19), from the UCSC Genome Browser (http://genome.ucsc.edu/). We selected the option "dbSNP Reference SNP (rs) identifier", that results in a list of all variations inside the genes. For *TMPRSS2*, *ACE2* and *BSG* genes we obtained 11,453, 534 and 4,258 SNPs respectively.

Functional annotation for all variants was performed with MASSA, that is a SNP annotation tool that queries and aggregates information from 24 public databases and two additional tools (SIFT and Polyphen) (Table S1-A). The population genetics analyses were performed on the functionally relevant SNPs.

Dataset Description

We studied 66 populations from five different datasets: i) 1000 Genomes Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); ii) The EPIGEN-Brazil project dataset (Kehdy et al., 2015): three Brazilian population-based cohorts (Salvador, Bambuí and Pelotas) that consist of 265 individuals genotyped with the Illumina HumanOmni5-4v1 array; iii) One hundred twenty nine Peruvian Native Americans genotyped on a Illumina HumanOmni2.5-8v1 array from the Laboratório de Diversidade Genética Humana (LDGH, unpublished) and iv) Native Americans and admixed Peruvian populations from Peruvian Genome Project (Harris et al. 2018) that include 150 whole genome sequenced individuals and 712 unpublished individuals genotyped on a Illumina HumanOmni2.5-8v1 array and v) Chilean population from ChileGenomico Project (http://www.chilegenomico.cl/) and Patagonia DNA Project. This dataset includes 9 whole genome sequenced individuals recruited from the north of Chile with Aymara ancestry, 9 individuals from the Metropolitan Region (the capital, in the center of the country) with Mapuche ancestry and 17 patagonians consisting, by ethnicity, of 3 Pehuenche, 3 Huilliche, 3 Chilote (putative Chono descendants), 4 Kaweskar, and 4 Yamana (Verdugo et al., 2020). The datasets descriptions are summarised at Table S1-B. All allele frequencies necessary to reproduce the analyses are available as supplementary tables.

Fst and Linkage Disequilibrium

Differentiation of the *TMPRSS2* gene between Native Americans and Asians was tested by calculating Fst. To perform this analysis, we integrated data from 71 SNPS of the *TMPRSS2* gene genotyped in 218 Peruvians from the Dr. Tarazona's group and from the Peruvian Genome Project with data from East Asia populations from 1000 Genomes Project (Table S1-B). The Fst for each *TMPRSS2* SNP was estimated in a pairwise fashion, between Peruvian Native-Americans and East Asians using the hierfstat R package (de Meeûs & Goudet, 2007) (Figure 1C).

The linkage disequilibrium [(Hill & Robertson, 1968) r2, D`, LD] between rs2070788 and rs383510 was calculated using genotyping data from East Asians populations from 1000 Genomes Project and sequencing data from Peruvian Native-Americans of Peruvian Genome Project (Table S1-B) using Haploview 4.2 (Barrett, Fry, Maller, & Daly, 2005) (Figure 1D).

Genomic ancestry and regression analysis

For genomic ancestry estimation, datasets from 1000 Genomes Project (CEU, IBS, FIN, GBR, TSI, LWK, YRI, JPT, ASW, CLM, MXL and PUR), EPIGEN-Brazil project (Salvador, Bambuí and Pelotas) and Native Americans and Admixed Peruvian populations (Aymaras, Ashaninkas, Awajun, Candoshi, Chachapoyas, Chopccas, Jacarus, Lamas, Matses, Moche, Qeros, Shipibo, Quechuas, Shimaa, Tallanes and Uros) were integrated into a single database containing autosomal genomewide SNPs shared by all these populations (Table S1-B). African, European, Asian and Native American genomic ancestries were estimated by ADMIXTURE (Alexander, Novembre, & Lange, 2009) on unsupervised mode using k=4.

Linear regression coefficient beta was estimated for each allele frequency in each studied population for the functionally relevant variants on each continental ancestry (Native American, European, African and East Asian) in R environment, using *logit* function (Cribari-Neto & Zeileis, 2010).

Acknowledgments

This work was funded by CNPq, CAPES, Department of Science and Technology of the Brazilian Ministry of Health (DECIT/MS), the Peruvian Genome Project from the Peruvian National Institute of Health and grants FONDEF D10I1007, D10E1007 and FONDEQUIP EQM140157 (CONICYT, Chile).

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–64. Retrieved from https://doi.org/10.1101/gr.094052.109

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview : analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), 263–265. Retrieved from https://doi.org/10.1093/bioinformatics/bth457

Bosch, E., Laayouni, H., Morcillo-Suarez, C., Casals, F., Moreno-Estrada, A., Ferrer-Admetlla, A., … Bertranpetit, J. (2009). Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: Most population isolates do not show increased LD. *BMC Genomics*, 10(1), 338. Retrieved 27 May 2020 from https://doi.org/10.1186/1471-2164-10-338

Casanova, J.-L., & Su, H. C. (2020). A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell*. Retrieved from https://doi.org/10.1016/j.cell.2020.05.016

Cheng, Z., Zhou, J., To, K. K.-W., Chu, H., Li, C., Wang, D., … Yuen, K.-Y. (2015). Identification of TMPRSS2 as a Susceptibility Gene for Severe 2009 Pandemic A(H1N1) Influenza and A(H7N9) Influenza. *The Journal of Infectious Diseases*, 212(8), 1214–21. Retrieved 29 March 2020 from https://doi.org/10.1093/infdis/jiv246

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2), 1–24. Retrieved from https://doi.org/10.18637/jss.v034.i02

de Meeûs, T., & Goudet, J. (2007). A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 7(6), 731–5. Retrieved 3 June 2015 from https://doi.org/10.1016/j.meegid.2007.07.005

Harris, D. N., Song, W., Shetty, A. C., Levano, K. S., Cáceres, O., Padilla, C., … Guio, H. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28), E6526–E6535. Retrieved 14 May 2020 from https://doi.org/10.1073/pnas.1720798115

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations.

Theoretical and Applied Genetics, 38(6), 226–231. Retrieved from https://doi.org/10.1007/BF01245622

Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., … Zolini, C. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, 112(28), 8696–8701. Retrieved from https://doi.org/10.1073/pnas.1504447112

Soares-Souza, G. B. (2014). *New approaches for database integration and development of bioinformatics tools for population genetics studies*. Federal University of Minas Gerais.

Verdugo, R. A., Di Genova, A., Herrera, L., Moraga, M., Acuña, M., Berríos, S., … Cifuentes, L. (2020). Development of a small panel of SNPs to infer ancestry in Chileans that distinguishes Aymara and Mapuche components. *Biological Research*, 53(1), 15. Retrieved 2 June 2020 from https://doi.org/10.1186/s40659-020-00284-5

Wang, K., Chen, W., Zhou, Y.-S., Lian, J.-Q., Zhang, Z., Du, P., … Chen, Z.-N. (2020). SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. *BioRxiv*, 2020.03.14.988345. Retrieved 14 April 2020 from https://doi.org/10.1101/2020.03.14.988345

Figure legends

Figure 1. (A) PRISMA flowchart of the systematic review; (B) Frequencies of the rs2070788 SNP and Native American ancestry in different populations (Populations form 1000 Genomes Project: ASW, Americans of African Ancestry in SW USA; CEU, Utah Residents(CEPH) with Northern and Western European Ancestry; CLM, Colombians from Medellin, Colombia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian Population in Spain; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; PUR, Puerto Ricans from Puerto Rico; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria); (C) Fst values distribution of Native Americans vs East Asian populations for 71 SNPs of *TMPRSS2* gene; (D) Linkage disequilibrium between rs2070788 and rs383510 in East Asian and Native American populations.