scientific reports

OPEN

The major genetic risk factor for severe COVID-19 does not show any association among South Asian populations

Prajjval Pratap Singh[1], Anshika Srivastava[1], Gazi Nurun Nahar Sultana[2], Nargis Khanam[1], Abhishek Pathak[3], Prashanth Suravajhala[4], Royana Singh[5], Pankaj Shrivastava[6], George van Driem[7], Kumarasamy Thangaraj[8,9] & Gyaneshwer Chaubey[1✉]

With the growing evidence on the variable human susceptibility against COVID-19, it is evident that some genetic loci modulate the severity of the infection. Recent studies have identified several loci associated with greater severity. More recently, a study has identified a 50 kb genomic segment introgressed from Neanderthal adding a risk for COVID-19, and this genomic segment is present among 16% and 50% people of European and South Asian descent, respectively. Our studies on ACE2 identified a haplotype present among 20% and 60% of European and South Asian populations, respectively, which appears to be responsible for the low case fatality rate among South Asian populations. This result was also consistent with the real-time infection rate and case fatality rate among various states of India. We readdressed this issue using both of the contrasting datasets and compared them with the real-time infection rates and case fatality rate in India. We found that the polymorphism present in the 50 kb introgressed genomic segment (rs10490770) did not show any significant correlation with the infection and case fatality rate in India.

Since the beginning of COVID-19 pandemic, it has been observed that people with a different ethnic background and country or continent of origin have variable degrees of susceptibility[1,2]. Though there are a few well known factors for higher susceptibility, e.g. age and comorbidity[3,4], the hospitalisation of younger healthy people has also been reported[5]. Recent genome wide association study has identified a gene cluster at chromosome 3 as well as the ABO gene at chromosome 9 associated with the severe risk factor for COVID-19 among Europeans[6]. Subsequently, the COVID-19 Host Genetics Initiative has corroborated this result[7]. The worldwide meta-analysis of the COVID-19 Host Genetics Initiative has identified 13 genetic loci associated with higher susceptibility or higher severity[8].

Zeberg and Pääbo[9] have identified a risk haplotype of 50 kb size introgressed from Neanderthals, which they called the 'Neanderthal core haplotype'. This risk haplotype was found to be present with an allele frequency of 30% among South Asians, 8% in Europeans and 4% among admixed Americans. The peak carrier frequency was estimated among the Bangladeshi population, where 63% carried at least one copy of this haplotype. The study also cited twice the risk of mortality among people of Bangladeshi extraction living in the UK as opposed to the native population of Brittanic pedigree[10].

Conversely, three of our studies on ACE2, the gateway of SARS-CoV-2, identified a haplotype, shared among South Asians and East Eurasians, likely protecting them from severe risk[11–13]. Additionally, the spatial distribution of this haplotype showed strong association with the low infection as well as low case fatality rate (CFR)[13].

[1]Cytogenetics Laboratory, Department of Zoology, Banaras Hindu University, Varanasi, Uttar Pradesh 221005, India. [2]Centre for Advanced Research in Sciences (CARS), Genetic Engineering and Biotechnology Research Laboratory, University of Dhaka, Dhaka 1000, Bangladesh. [3]Department of Neurology, Institute of Medical Sciences, Banaras Hindu University, Varanasi, India. [4]Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research Statue Circle, Jaipur, Rajasthan, India. [5]Department of Anatomy, Institute of Science, Banaras Hindu University, Varanasi, Uttar Pradesh 221005, India. [6]Department of Home (Police), DNA Fingerprinting Unit, State Forensic Science Laboratory, Government of MP, Sagar, India. [7]Institut Für Sprachwissenschaft, Universität Bern, Länggassstrasse 49, 3012 Bern, Switzerland. [8]CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India. [9]Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India. ✉email: thangs@ccmb.res.in

		rs2285666 (T)	rs10490770 (C)	FreqMay2020	FreqAug2020	FreqDec2020	CFRAug2020	CFRAug2020
rs2285666	Pearson correlation Sig. (2-tailed)	1	− 0.428 0.098	**− 0.588*** 0.017	**− 0.558*** 0.025	**− 0.580*** 0.018	**− 0.649**** 0.007	**− 0.568** 0.022
rs10490770	Pearson correlation Sig. (2-tailed)	− 0.428 0.098	1	0.204 0.450	0.164 0.543	0.266 0.319	0.244 0.362	− 0.026 0.924

Table 1. Estimates of Pearson correlation coefficient for the rs2285666 and rs10490770 with the real-time COVID-19 cases as well as case fatality rate among Indian populations. The significant values are shown in bold letters. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

To resolve this discrepancy between the two sets of findings and the associated claims, we have extracted a SNP (rs10490770) reported to be associated with the high risk for COVID-19[9], from our published and unpublished genome wide datasets (Supplementary Table S1), and looked for existing association with the state-wise COVID-19 data of India.

Materials and methods

Zeberg and Pääbo[9] have mainly discussed about the SNP rs35044562. However, they reported 12 other SNPs present in the 'Neanderthal core haplotype' that are in high linkage disequilibrium ($r^2 > 0.98$) (Supplementary Table S2). SNP rs10490770 showed high LD ($r^2 = 0.99$) with the SNP rs35044562. The genome-wide genotype data by Illumina tagged rs2285666 and rs10490770 SNPs in their panel. Therefore, we searched the genotype datasets generated by this platform. The frequency data for both of the SNPs from various Indian populations were extracted using Plink 1.9[14], from 1000 genome project data phase 3[15], data published by the Estonian Biocentre[16–19] and our newly genotyped samples for various Indian states and Bangladesh (Supplementary Table S1). In addition to our previous study[13], more samples were added for the SNP rs2285666. The state-wise COVID-19 infection and CFR datasets were extracted from https://www.covid19india.org/. The regression estimations and plots were built by https://www.graphpad.com/quickcalcs/linear1/ and reverified by the Microsoft Excel regression calculations[20]. We have also used Pearson's correlation coefficient test[20] to evaluate the effect of both the SNPs. The spatial distribution of both SNPs were drawn using the web tool available at https://www.datawrapper.de/.

Results and discussion

In contrast to the conclusions drawn by Zeberg and Pääbo[9], our study on *ACE2* identified a haplotype that is frequent among South Asians and East Eurasians[11–13]. This haplotype is derived by a polymorphism rs2285666 responsible for elevated expression of *ACE2*. We have found high inverse correlation of this haplotype with the state-wise cases as well as the case-fatality rate (CFR) among Indian populations[13]. This correlation was significant at various timelines of the pandemic in India (Table 1). We verified the statistical tests with the updated data up to December 2020 and found these data to be consistent with previous observations (Fig. 1 and Supplementary Fig. S1). Thus, it is likely that the *ACE2* SNP rs2285666 has played a significant role in modulating the susceptibility to the disease among Indian populations.

In our search of the SNPs reported to be associated with high risk by Zeberg and Pääbo[9], we found rs10490770 from genome-wide datasets[17,18,21–23]. We applied the same tests done for the *ACE2* SNPs (Fig. 1). The state-wise frequency variation of this SNP did not show any association either with the number of cases or the CFR (Table 1 and Supplementary Fig. S1). We repeated these regression tests for the number of cases as well as the CFR data, obtained during all the three months. However, none of them showed any association with the rs10490770 ($p > 0.3$) (Table 1). It is interesting to note that this SNP (rs10490770) has been found to be associated with disease severity in the global data[8]. However, the lack of association for rs10490770 with COVID-19 cases or CFR in India is striking and suggests instead a complex susceptibility response among Indian populations. Along with the complex genetic structure[11,24], socio-economic status[25] and hygiene[26] may have contributed to such a complex scenario. Furthermore, a detailed clinical and genome-wide association study on Indian COVID-19 patients would be useful to resolve this complexity.

Zeberg and Pääbo[9] used the data of higher susceptibility to the disease among the Bangladeshi population living in UK[10] to support their findings. By considering the effect of sex, age, socio-economic deprivation and region, this report found that people of Bangladeshi origin had double the risk of mortality as compared to people of British origin. However, the higher mortality rate for Bangladeshi population in the UK needs more detailed investigation on comorbidity, genetic admixture as well as local environment and socio-economic circumstances in their particular British context. More importantly, a similar trend had also been observed among admixed Americans, where some of the same qualifications may apply *mutatis mutandis*[27–29]. Furthermore, it is notable that among the Bangladeshi samples analysed by us, the tribal populations of Bangladesh showed almost three times less frequency of rs10490770 (Supplementary Table S1). This is likely due to the different population histories of the caste and tribal populations of Bangladesh[30,31]. The Tibeto-Burman speakers of Bangladesh show a closer genetic affinity with the East and Southeast Asian populations, whereas the Indo-European speaking caste
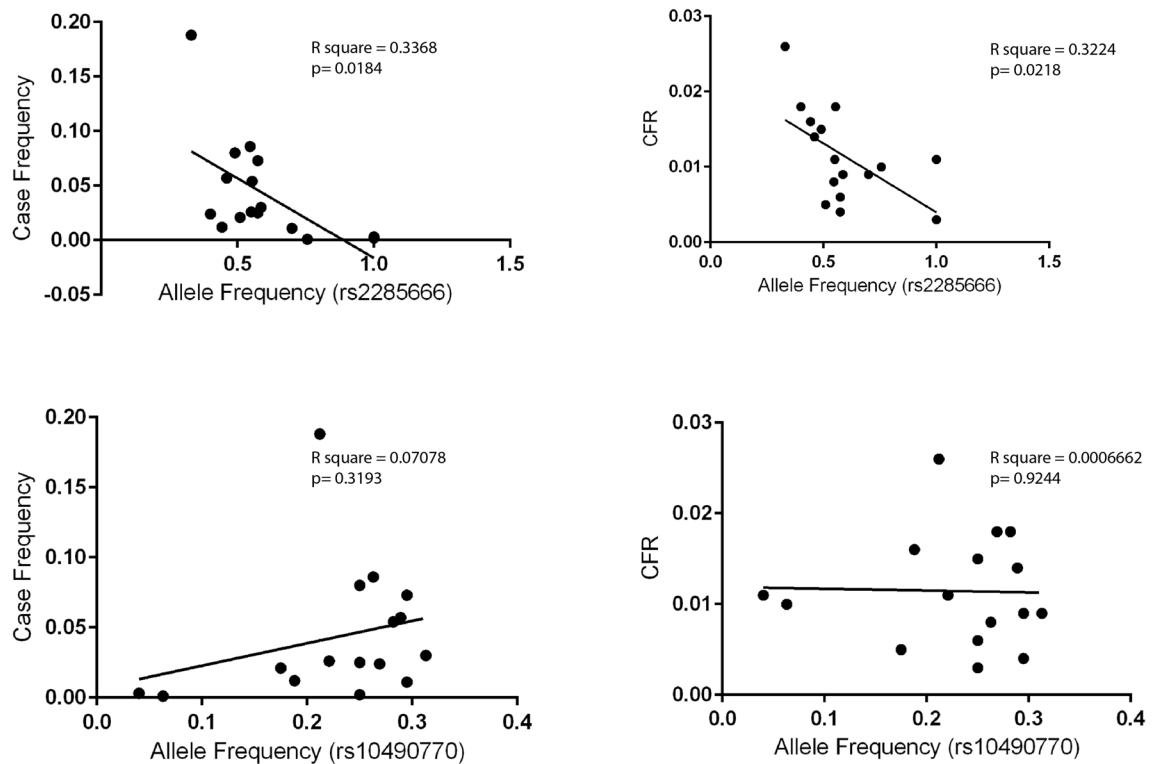
Figure 1. The regression analyses of rs2285666 (*ACE2*) and rs10490770 (*LZTFL1*) with the COVID-19 cases as well as case fatality rate in various states of India (Supplementary Table S1).

populations incline with the Indian populations. Therefore, it is advised explicitly to differentiate between the caste and tribal populations while making any statement about Bangladeshi populations. Significantly, our data also show that the incidence of the allele rs2285666 has been found to occur in the highest frequency of 100% in Indian populations such as the Nishi and Kokborok (Tripuri), who represent Trans-Himalayan language communities (Supplementary Fig. S1 and Supplementary Table S1). As a linguistic phylum, the Trans-Himalayan language family is widespread in parts of eastern Eurasia and includes languages such as Tibetan, Burmese, Mandarin, Cantonese and Hokkien.

Thus, our extensive analyses on real-time data did not show any association of rs10490770 with the state-wise infection rates as well as CFRs, suggesting that the risk allele for COVID-19 in Europe does not play a significant role in COVID-19 severity in South Asia.

Data availability
All datasets generated for this study are included in the article/Supplementary Material.

References

1. Mackey, K. *et al.* Racial and ethnic disparities in COVID-19–related infections, hospitalizations, and deaths: A systematic review. *Ann. Intern. Med.* **174**, 362–373 (2021).
2. Shelton, J. F. *et al.* Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. *medRxiv* https://doi.org/10.1101/2020.09.04.20188318 (2020).
3. Alberca, R. W., de Oliveira, L. M., Branco, A. C. C. C., Pereira, N. Z. & Sato, M. N. Obesity as a risk factor for COVID-19: An overview. *Crit. Rev. Food Sci. Nutr.* **1**, 15. https://doi.org/10.1080/10408398.2020.1775546 (2020).
4. Fang, L., Karakiulakis, G. & Roth, M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?. *Lancet Respir. Med.* **8**, 21 (2020).
5. Godri Pollitt, K. J. *et al.* COVID-19 vulnerability: The potential impact of genetic susceptibility and airborne transmission. *Hum. Genom.* **14**, 1–7 (2020).
6. Ellinghaus, D. *et al.* Genomewide association study of severe Covid-19 with respiratory failure. *N. Engl. J. Med.* https://doi.org/10.20944/preprints202007.0178.v2 (2020).
7. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715 (2020).
8. Ganna, A. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv* https://doi.org/10.1101/2021.03.10.21252820 (2021).
9. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).

10. Public Health England. *COVID-19: Review of disparities in risks and outcomes* (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908434/Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf

11. Srivastava, A. *et al.* Most frequent South Asian haplotypes of ACE2 share identity by descent with East Eurasian populations. *PLoS ONE* **15**, e0238255 (2020).

12. Singh, K. K., Chaubey, G., Chen, J. Y. & Suravajhala, P. Decoding SARS-CoV-2 hijacking of host mitochondria in pathogenesis of COVID-19. *Am. J. Physiol.* https://doi.org/10.1152/ajpcell.00224.2020 (2020).

13. Srivastava, A. *et al.* Genetic association of ACE2 rs2285666 polymorphism with COVID-19 spatial distribution in India. *Front. Genet.* **11**, 1163 (2020).

14. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer data sets. *Gigascience* **4**, S13742-015 (2015).

15. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

16. Estonian Biocentre Public_Data. https://evolbio.ut.ee/.

17. Chaubey, G. *et al.* "Like sugar in milk": Reconstructing the genetic history of the Parsi population. *Genome Biol.* **18**, 110 (2017).

18. Pathak, A. K. *et al.* The genetic ancestry of modern Indus valley populations from Northwest India. *Am. J. Hum. Genet.* **103**, 918–929 (2018).

19. Tätte, K. *et al.* The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci. Rep.* **9**, 3818 (2019).

20. Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). "Pearson correlation coefficient," in *Noise Reduction in Speech Processing, Springer Topics in Signal Processing*, vol. 2, Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-00296-0_5

21. Chaubey, G. *et al.* Population genetic structure in Indian austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Mol. Biol. Evol.* **28**, 1013–1024 (2011).

22. Metspalu, M. *et al.* Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89**, 731–744 (2011).

23. Tamang, R. *et al.* Reconstructing the demographic history of the Himalayan and adjoining populations. *Hum. Genet.* **137**, 129–139 (2018).

24. Metspalu, M., Mondal, M. & Chaubey, G. The genetic makings of South Asia. *Genet. Hum. Orig.* **53**, 128–133 (2018).

25. Rajkumar, R. P. The relationship between demographic, socioeconomic, and health-related parameters and the impact of COVID-19 on 24 regions in India: Exploratory cross-sectional study. *JMIR Public Health Surveill.* **6**, e23083 (2020).

26. Finlay, B. B. *et al.* The hygiene hypothesis, the COVID pandemic, and consequences for the human microbiome. *Proc. Natl. Acad. Sci. USA* **118**, e2010217118 (2021).

27. Kullar, R. *et al.* Racial disparity of coronavirus disease 2019 in African American communities. *J. Infect. Dis.* **222**, 890–893 (2020).

28. Yancy, C. W. COVID-19 and African Americans. *JAMA* **323**, 1891–1892 (2020).

29. Hooper, M. W., Nápoles, A. M. & Pérez-Stable, E. J. COVID-19 and racial/ethnic disparities. *JAMA* **323**, 2466–2467 (2020).

30. Sultana, G. N. N., Sharif, M. I., Asaduzzaman, M. & Chaubey, G. Evaluating the genetic impact of South and Southeast Asia on the peopling of Bangladesh. *Leg. Med. Tokyo Jpn.* **17**, 446–450 (2015).

31. Gazi, N. N. *et al.* Genetic structure of Tibeto-Burman populations of Bangladesh: Evaluating the gene flow along the Sides of Bay-of-Bengal. *PLoS ONE* **8**, e75064 (2013).

Acknowledgements

Author contributions

G.C., K.T., R.S. and P.S. conceived and designed this study. P.P.S., A.S., G.S., N.K., A.P., Pr.S., R.S., and P.S. collected the data for alleles and COVID-19. P.P.S., A.S., Pr.S., and G.C. analyzed the data. P.P.S., Pr.S., K.T., Gv.D. and G.C. wrote the manuscript from the inputs of other co-authors. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-91711-4.

Correspondence and requests for materials should be addressed to G.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.