Research Article

The expression and polymorphism of entry machinery for COVID-19 in human: juxtaposing population groups, gender, and different tissues

Behrooz Darbani

The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Email: behdas@biosustain.dtu.dk; behroozdarbani@gmail.com

Tel.: +45-53578055

Keywords: **COVID-19, Gender, Host entry machinery, Polymorphism**

Abstract

Combating viral disease outbreaks has doubtlessly been one of the major public health challenges for the 21st century. Here, the host entry machinery required for COVID-19 (SARS-CoV-2) infection was examined for the gene expression profiles and polymorphism. The four human population groups of Europeans, Africans, Asians, and Americans had specific and also a shared pool of variants for the X-linked locus of ACE2 receptor. Several specific and common ACE2 variants were of the utmost importance to the viral entry and infection. In the absence of gender bias for the gene expression profiles, the hemizygous rare variants of ACE2 describe the observed higher mortality rate in males. Finally, a personalized medicine strategy is conceived for isolating high-risk individuals in epidemic circumstances.

Background

The early 21st century has been prominent in contagious disease history. Our public health has been repeatedly threatened by the pathogenic respiratory beta-coronaviruses (CoV) including the Severe Acute Respiratory Syndrome CoV (SARS-CoV)[1], Middle East Respiratory Syndrome CoV (MERS-CoV)[2], and very recently by the pandemic burst of SARS-CoV-2[3]. These are enveloped viruses with a capped and polyadenylated positive-strand RNA genomes of ~30,000 nucleotides[4]. They have been considered as extraordinary threats to public health by virtue of progressive respiratory failure and mortality rates of 4−36% in human[5–7]. The current pandemic SARS-CoV-2[3] has resulted in a global shut down in March 2020 just after a three-month period of its outset. The global mortality rate also reached to 4.9% measured among 823,626 confirmed cases as of April 1st, 2020[8]. Therefore, the respiratory syndrome by coronaviruses calls for sustainable preventive approaches and effective treatment protocols to combat the current and also future outbreaks. To identify susceptible individuals or groups and also for developing treatment strategies, we need a deep understanding of the mechanisms virus requires for its entry and within-host spread.

Results and discussion

Here, the genetic diversity and expression of the host factors required for SARS-CoV/CoV-2 entry were explored among the human population groups and tissues, respectively. The angiotensin-converting enzyme 2 (ACE2)[9,10], the endosomal cysteine protease cathepsin B (CTSB) and L (CTSL)[10,11], and the transmembrane serine protease 2 (TMPRSS2)[10,12,13] are exploited by SARS-CoV/CoV-2 as the cellular-entry machinery. The Genotype-Tissue Expression (GTEx) data[14] was used to inspect the tissue expression profiles. The *Ace*2, *Tmprss*2, *Cts*B, and *Cts*L genes are found expressed in different tissues (Fig. 1a). The minor salivary glands, lung, small intestine, liver, kidney, and heart were among the other tissues where the genes expressed (Fig. 1a). *Ace*2 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2020 doi:10.20944/preprints202004.0076.v1

Peer-reviewed version available at *Int. J. Environ. Res. Public Health* **2020**, *17*; doi:10.3390/ijerph17103433

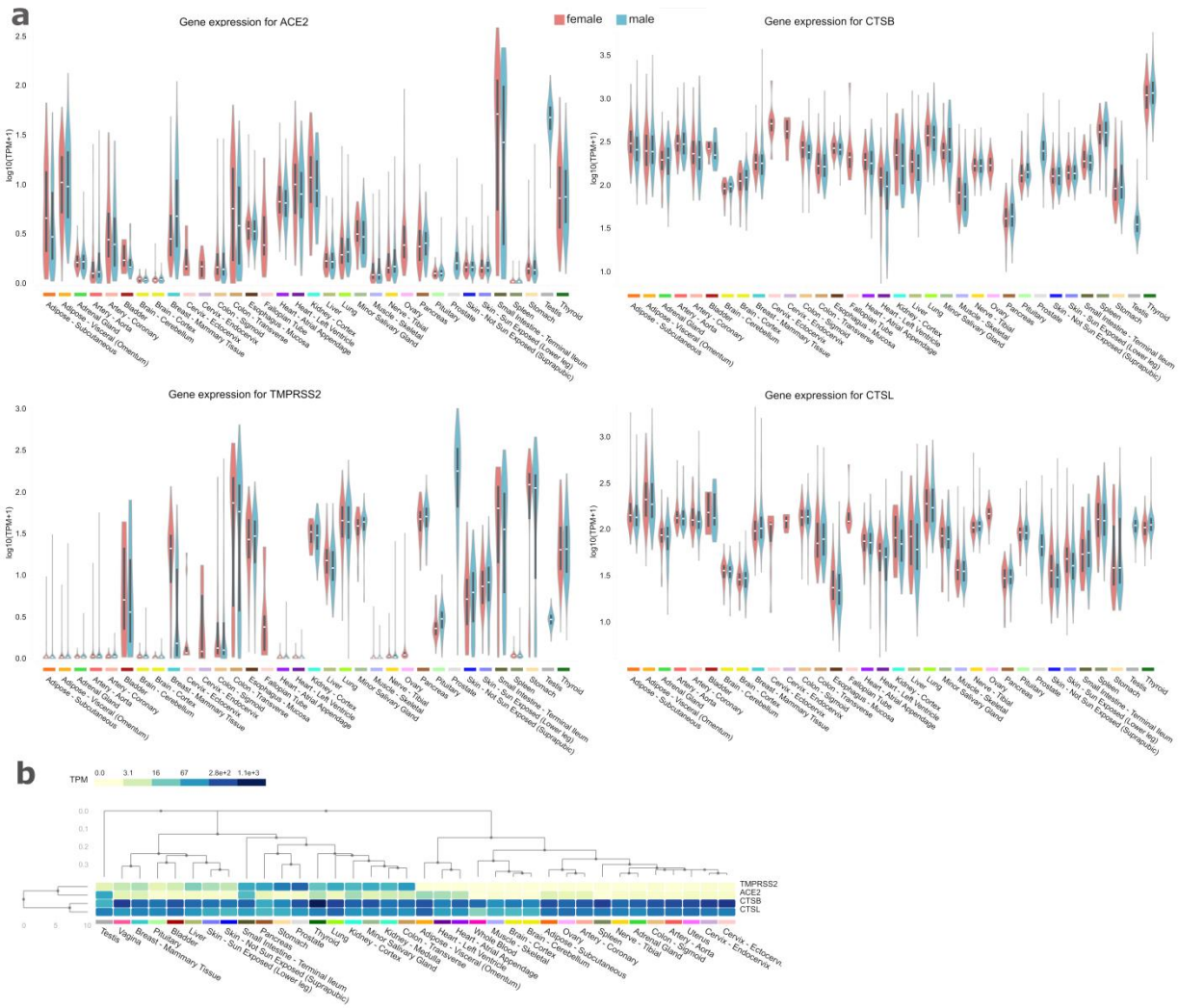

Fig. 1: Gene expression profiles for *Ace*2, *Tmprss*2, *Cts*B, and *Cts*L. a, Expression profiles across different tissues. Boxplots are shown as median and 25th–75th percentiles. **b,** Tissue-specific co-expression analysis on TPM measurements by using the BicMix biclustering. **a, b,** Data extracted from the Genotype-Tissue Expression (GTEx) project[14]. TPM: transcripts per million.

*Tmprss*2 showed no considerable expression in the brain and spleen as *Tmprss*2 in the heart (Fig. 1a). Among the three protease coding genes, *Tmprss*2 showed the highest correlated expression with *Ace*2 (Fig. 1b). Furthermore, lung, kidney, small intestine, and salivary glands were co-clustered together within a group of tissues with the highest expression levels considering all of the four genes (Fig. 1b). The additive effect of the co-expressed proteases on the viral entry suggests targeting of the TMPRSS2, CTSB, and CTSL proteases simultaneously, otherwise the ACE2 receptor for robust treatment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2020 doi:10.20944/preprints202004.0076.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2020, 17; doi:10.3390/ijerph17103433

By inspecting the human genetic variants pool available at https://www.ncbi.nlm.nih.gov/snp/, ~100 to 400 missense SNPs were extracted after filtering for the non-coding regions of *Ace*2, *Tmprss*2, *Cts*B, and *Cts*L genes (Table 1). In contrast to the two alleles of *Tmprss*2, *Cts*B, and *Cts*L present in both the males and females' genomes, there is one allele for the X-linked *Ace*2 locus in

Table 1: Chromosomal location and SNP characteristics of the human genes required for SARS-Cov/CoV-2 entry

Genes	*Ace*2	*Tmprss*2	*Cts*B	*Cts*L
Chromosome	X	21	8	9
Intronic-SNPs	15,391	10,293	94,70	1,101
Synonymous-SNPs	104	173	149	50
Missense SNPs, inframe insertion/deletion SNPs	265	394	311	102

Data extracted from the GenBank, Database for Single Nucleotide Polymorphisms (dbSNP) at https://www.ncbi.nlm.nih.gov/snp/.

males' genomes (Table 1). As the host receptor, ACE2 interacts with the receptor binding motif (RBM) of the viral S1 protein which is essential for SARS-CoV/CoV2 infection[9,10,15]. Further analyses on the *Ace*2 variants were carried out to compare the human population groups of Europeans, Africans, Asians, and Americans. The *Ace*2 locus had 265 missense SNPs, including inframe insertions and deletions, which were distributed in the exons I to XIII, XVII, and XVIII (Table 1, Fig. 2a). Of these, 194 SNPs were found with allele frequencies by considering the 1000 genomes project data, the exome aggregation consortium data, and the genome aggregation data (Fig. 2a). The global prevalence for rare variants of *Ace*2 ranged from 0.00001 to 0.016 and had an average and median of 2.14×10^{-04} and 3.59×10^{-05}, respectively. Several SNPs had significant frequency-changes at least in one of the population comparisons (Fig. 2a). Statistically significant population-specific SNPs were also detected which was more pronounced for Africans (Table 2). Interestingly, the amino acid substitutions in several variants can potentially influence the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2020 doi:10.20944/preprints202004.0076.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2020, 17; doi:10.3390/ijerph17103433

interaction between the ACE2 and the viral S1 protein, and thereby the viral infectivity. Different

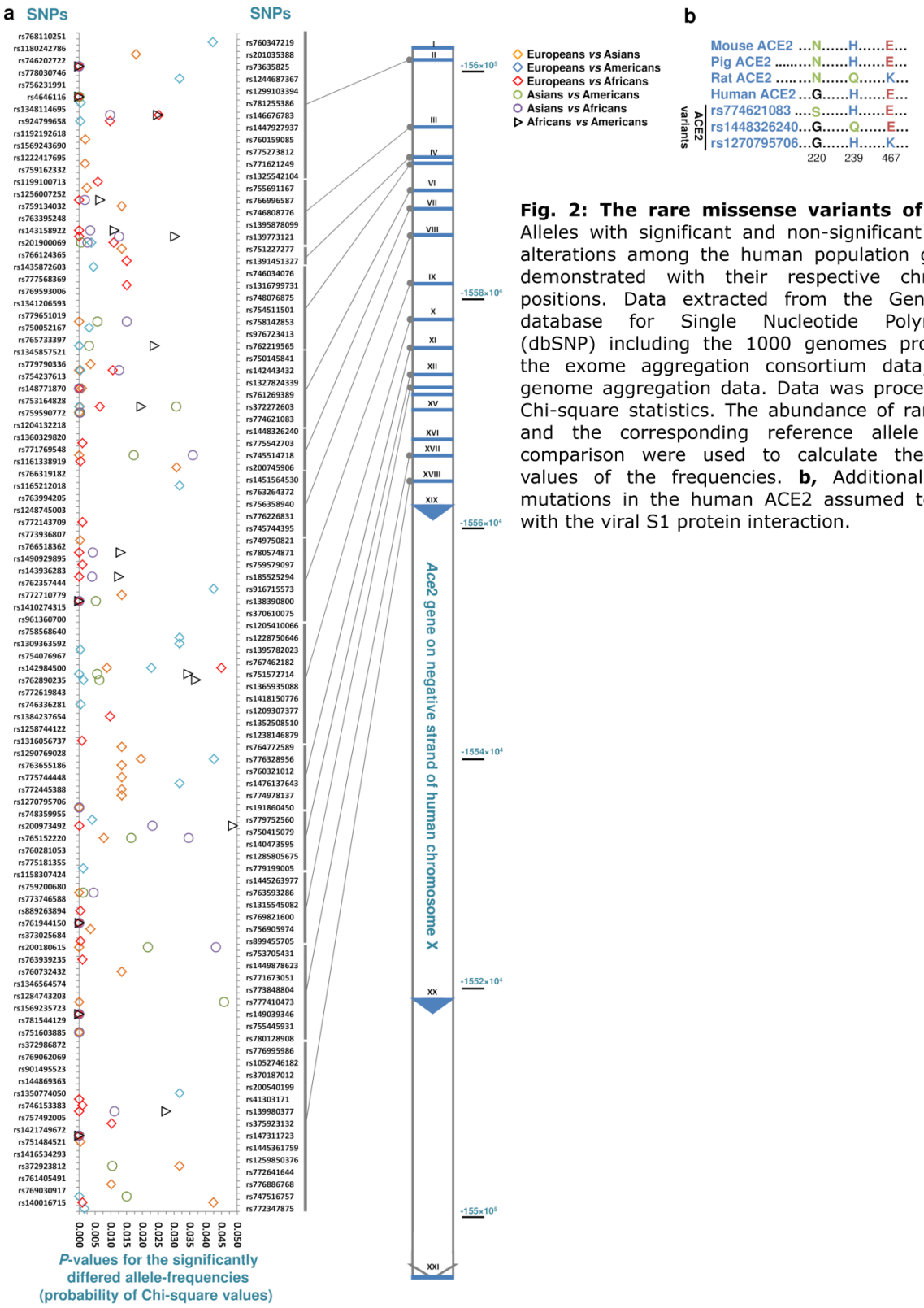

Fig. 2: The rare missense variants of ACE2. a, Alleles with significant and non-significant frequency alterations among the human population groups are demonstrated with their respective chromosomal positions. Data extracted from the GenBank, the database for Single Nucleotide Polymorphisms (dbSNP) including the 1000 genomes project data, the exome aggregation consortium data, and the genome aggregation data. Data was processed using Chi-square statistics. The abundance of rare variants and the corresponding reference allele in every comparison were used to calculate the expected values of the frequencies. **b,** Additional missense mutations in the human ACE2 assumed to interfere with the viral S1 protein interaction.

amino acid residues distributed all over the ACE2 receptor have been found very influential on the viral infectivity[15–17]. Accordingly, there are amino acid substitutions that can either facilitate or hinder the detectability of ACE2 by the viruses[15–17]. Using this knowledge, 13 variants (rs73635825 [S19P], rs778030746 [I21V], rs1244687367 [I21T], rs756231991 [E23K], rs1434130600 [A25T], rs4646116 [K26R], rs781255386 [T27A], rs778500138 [E35D], rs1199100713 [N64K], rs867318181 [E75G], rs763395248 [T92I], rs1395878099 [Q102P], and rs142984500 [H378R]) were found as the interaction-booster between ACE2 and S1. The amino acid exchanges at the positions of 25, 35, and 75 had no reported frequency. The abundance of S19P, I21T, K26R, T27A, N64K, and H378R variants significantly differed among the population groups (Fig. 2a). Interestingly, H378R and S19P were Europeans (Fig. 2a, Table 2, $p < 0.0449$, frequency: 0.00014) and Africans (Fig. 2a, Table 2, $p < 0.0000$, frequency: 0.0033) specific variants, respectively. Furthermore, another group of 13 SNPs including rs1348114695 [E35K], rs146676783 [E37K], rs760159085 [N51D], rs1569243690 [N51S], rs1325542104 [M62V], rs755691167 [K68E], rs1256007252 [F72V], rs766996587 [M82I], rs759579097 [G326E], rs370610075 [G352V], rs961360700 [D355N], rs751572714 [Q388L], and rs762890235 [P389H] were found as interaction-inhibitor variants. Of these, the variants E37K, N51D, K68E, F72V, M82I, G326E, Q388L, and P389H had significant changes in frequency among the population groups (Fig. 2a). The Q388L and M82I were also found as Americans (Fig. 2a, Table 2, $p < 0.0345$, frequency: 0.00016) and Africans (Fig. 2a, Table 2, $p < 0.0066$, frequency: 0.00021) specific variants, respectively. The impact on the viral infectivity was not clear for the Asians specific variants rs751603885 [R697G], rs763593286 [A532T], rs745514718 [S257N], and rs200180615 [E668K]. The ACE2 receptors in rat and to some extent in pig and mouse have weak interactions leading to lower levels of susceptibilities than in human against SARS-CoV[18,19]. Three additional SNPs

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2020 doi:10.20944/preprints202004.0076.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2020, 17; doi:10.3390/ijerph17103433

Table 2: Number of SNPs with significant variations in abundance

Comparisons	Total	Specific for the common population of the comparisons		
		SNPs in either of comparisons	SNPs in all of the comparisons (Frequency)	Population size (Avg./Median)
Europeans vs Asians, Americans or Africans	86	12	1 (0.000143)	122,845/147,472
Asians vs Europeans, Americans or Africans	55	43	4 (0.00013-0.00030)	40,150/45,859
Americans vs Africans, Asians or Europeans	52	29	2 (0.00016, 0.00018)	29,909/35,302
Africans vs Americans, Asians or Europeans	47	46	8 (0.00013-0.00336)	20,482/21,424

Data extracted from the GenBank, database for Single Nucleotide Polymorphisms (dbSNP) including the 1000 genomes project data, the exome aggregation consortium data, and the genome aggregation data available at https://www.ncbi.nlm.nih.gov/snp/. Data processed using Chi-square statistics (the frequencies of rare variants and their corresponding reference alleles were used to calculate their expected values in order to execute the test).

including rs774621083, rs1448326240, and rs1270795706, all perceived as interaction-inhibitor variants, were further selected by juxtaposing the human, rat, mouse, and pig ACE2s (Fig. 2b). The variant rs774621083 had a polar serine residue at position 220 instead of the non-polar glycine (Fig. 2b). The pig, mouse, and rat ACE2s have the polar amino acid asparagine at this position (Fig. 2b). This variant had a global abundance of 2.9×10^{-5} without significant difference among the population groups (Fig. 2a). As seen in rat, rs1270795706 had the positively charged residue of lysine at the position 467 and the Europeans specific variant rs1448326240 had the polar residue of glutamine at the position 239 in replacement of the negatively charged glutamic acid and the positively charged histidine, respectively (Fig. 2a, b).

Taken together, the human ACE2 has a rich pool of rare variants (Table 1, 2). These alleles, as 29 introduced in this study, can potentially be decisive in SARS-CoV/CoV2 recognition and infection (Fig. 2, Table 1). As an X-linked phenotype, the effectiveness of interaction-booster and interaction-inhibitor rare variants of ACE2 can be more definite in males than females. Accordingly, a gender bias has been observed towards a higher mortality rate in males accounting for up to ~70% of death caused by SARS-CoV2 or MERS[5,6,20]. The perceived effectiveness of the ACE2 rare variants might, in real, be even stronger due to the absence of gender bias for the expression of *Ace*2

and the protease coding genes *Tmprss*2, *Cts*B, and *Cts*L in asexual tissues (Fig. 1a). These results have the potential to advance the personalized medicine strategies, *e.g.*, by screening for the high-risk individuals that need isolation against the viral disease outbreaks.

Materials and methods

Gene expression data was extracted from the Genotype-Tissue Expression (GTEx) project[14] publicly available at https://www.gtexportal.org/home/. Tissue-specific co-expression analysis was performed on TPM measurements by using the BicMix biclustering. The human SNP data was extracted from the GenBank, the database for Single Nucleotide Polymorphisms (dbSNP) including the 1000 genomes project data, the exome aggregation consortium data, and the genome aggregation data. For every variant, data was obtained from the experiments and pooled. Data, *i.e.*, number of individuals carrying the reference or variant alleles, was processed using Chi-square statistics. The abundance of the rare variant (sum of individuals with rare variant) and the corresponding reference allele (sum of individuals with reference allele) in every comparison were used to calculate the expected values for the Chi-square test.

References:

1. Peiris, J. S. M., Guan, Y. & Yuen, K. Y. Severe acute respiratory syndrome. *Nature Medicine* **10**, S88–S97 (2004).
2. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
3. Wang, Z., Yang, B., Li, Q., Wen, L. & Zhang, R. Clinical Features of 69 Cases with Coronavirus Disease 2019 in Wuhan, China. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa272.

4. Rota, P. A. *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399 (2003).
5. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. *Zhonghua Liu Xing Bing Xue Za Zhi* **41**, 145–151 (2020).
6. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).
7. Channappanavar, R. *et al.* Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell Host Microbe* **19**, 181–193 (2016).
8. WHO. Coronavirus disease 2019 (COVID-19) Situation Report – 72: World Health Organization, April 1, 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.
9. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
10. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* (2020) doi:10.1016/j.cell.2020.02.052.
11. Simmons, G. *et al.* Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. *PNAS* **102**, 11876–11881 (2005).
12. Matsuyama, S. *et al.* Efficient Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2. *Journal of Virology* **84**, 12658–12664 (2010).
13. Shulla, A. *et al.* A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry. *J. Virol.* **85**, 873–882 (2011).
14. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).
15. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309**, 1864–1868 (2005).

16. Li, W. *et al.* Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *The EMBO Journal* **24**, 1634–1643 (2005).
17. Procko, E. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *bioRxiv* 2020.03.16.994236 (2020) doi:10.1101/2020.03.16.994236.
18. Li, W. *et al.* Efficient replication of severe acute respiratory syndrome coronavirus in mouse cells is limited by murine angiotensin-converting enzyme 2. *J. Virol.* **78**, 11429–11433 (2004).
19. Li, K. K. B. *et al.* Characterisation of animal angiotensin-converting enzyme 2 receptors and use of pseudotyped virus to correlate receptor binding with susceptibility of SARS-CoV infection. *Hong Kong Med J* **18 Suppl 3**, 35–38 (2012).
20. WHO. COVID-19 situation update for the WHO European Region: World Health Organization, March 23-29, 2020. http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19.

Figure Legends:

Fig. 1: Gene expression profiles for *Ace*2, *Tmprss*2, *Cts*B, and *Cts*L. a, Expression profiles across different tissues. Boxplots are shown as median and $25^{th}-75^{th}$ percentiles. **b**, Tissue-specific co-expression analysis on TPM measurements by using the BicMix biclustering. **a**, **b**, Data extracted from the Genotype-Tissue Expression (GTEx) project[14]. TPM: transcripts per million.

Fig. 2: The rare missense variants of ACE2. a, Alleles with significant and non-significant frequency alterations among the human population groups are demonstrated with their respective chromosomal positions. Data extracted from the GenBank, the database for Single Nucleotide Polymorphisms (dbSNP) including the 1000 genomes project data, the exome aggregation consortium data, and the genome aggregation data. Data was processed using Chi-square statistics. The abundance of rare variants and the corresponding reference allele in every comparison were

used to calculate the expected values of the frequencies. **b**, Additional missense mutations in the human ACE2 assumed to interfere with the viral S1 protein interaction.