




Impact of natural selection on global patterns of genetic variation and association with clinical phenotypes at genes involved in SARS-CoV-2 infection

Chao Zhang^{a,1}, Anurag Verma^{a,b,1} , Yuanqing Feng^{a,1}, Marcelo C. R. Melo^{c,d,e}, Michael McQuillan^a, Matthew Hansen^a , Anastasia Lucas^a, Joseph Park^a, Alessia Ranciaro^a, Simon Thompson^a, Meagan A. Rubel^{a,2}, Michael C. Campbell^f, William Beggs^a, Jibril Hirbo^g, Sununguko Wata Mpoloka^h, Gaonyadiwe George Mokoneⁱ, Regeneron Genetic Center^{j,3}, Thomas Nyambo^k, Dawit Wolde Meskel^l, Gurja Belay^l, Charles Fokunang^m, Alfred K. Njamshi^{n,o}, Sabah A. Omar^p, Scott M. Williams^q, Daniel J. Rader^a , Marylyn D. Ritchie^a, Cesar de la Fuente-Nunez^{c,d,e}, Giorgio Sirugo^{b,4}, and Sarah A. Tishkoff^{a,r,s,4}

Contributed by Sarah Tishkoff; received December 22, 2021; accepted March 29, 2022; reviewed by Rob Kulathinal and Xin Li

Human genomic diversity has been shaped by both ancient and ongoing challenges from viruses. The current coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has had a devastating impact on population health. However, genetic diversity and evolutionary forces impacting host genes related to SARS-CoV-2 infection are not well understood. We investigated global patterns of genetic variation and signatures of natural selection at host genes relevant to SARS-CoV-2 infection (angiotensin converting enzyme 2 [*ACE2*], transmembrane protease serine 2 [*TMPRSS2*], dipeptidyl peptidase 4 [*DPP4*], and lymphocyte antigen 6 complex locus E [*LY6E*]). We analyzed data from 2,012 ethnically diverse Africans and 15,977 individuals of European and African ancestry with electronic health records and integrated with global data from the 1000 Genomes Project. At *ACE2*, we identified 41 nonsynonymous variants that were rare in most populations, several of which impact protein function. However, three nonsynonymous variants (rs138390800, rs147311723, and rs145437639) were common among central African hunter-gatherers from Cameroon (minor allele frequency 0.083 to 0.164) and are on haplotypes that exhibit signatures of positive selection. We identify signatures of selection impacting variation at regulatory regions influencing *ACE2* expression in multiple African populations. At *TMPRSS2*, we identified 13 amino acid changes that are adaptive and specific to the human lineage compared with the chimpanzee genome. Genetic variants that are targets of natural selection are associated with clinical phenotypes common in patients with COVID-19. Our study provides insights into global variation at host genes related to SARS-CoV-2 infection, which have been shaped by natural selection in some populations, possibly due to prior viral infections.

SARS-CoV-2/COVID-19 | genetic variation | phenotype association | natural selection | African diversity

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Coronaviruses are enveloped, positive-sense, and single-stranded RNA viruses, many of which are zoonotic pathogens that crossed over into humans. Seven coronavirus species, including SARS-CoV-2, have been discovered that, depending on the virus and host physiological condition, may cause mild or lethal respiratory disease. There is considerable variation in disease prevalence and severity across populations and communities. Importantly, minority populations in the United States appear to have been disproportionately affected by COVID-19 (1, 2). For example, in Chicago, more than 50% of COVID-19 cases and nearly 70% of COVID-19 deaths are in African Americans (who make up 30% of the population of Chicago) (1). While social and economic factors are largely responsible for driving COVID-19 health disparities, investigating genetic diversity at host genes related to SARS-CoV-2 infection could help identify functionally important variation, which may play a role in individual risk for severe COVID-19 infection.

In this study, we focused on four key genes playing a role in SARS-CoV-2 infection (3). The *ACE2* gene, encoding the angiotensin-converting enzyme-2 protein, was reported to be a main binding site for severe acute respiratory syndrome coronavirus (SARS-CoV) during an outbreak in 2003, and evidence showed stronger binding affinity to SARS-CoV-2, which enters the target cells via ACE2 receptors (3, 4). The *ACE2* gene is located on the X chromosome (chrX); its expression level varies among populations (5); and it is ubiquitously expressed in the lung, blood vessels, gut, kidney, testis, and brain, all organs that appear to be affected as part of the COVID-19 clinical

Significance

Viruses are strong sources of natural selection pressure during human evolutionary history. Investigating genetic diversity and detecting signatures of natural selection at host genes related to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection help to identify functionally important variation. We conducted a large study of global genomic variation at host genes that play a role in SARS-CoV-2 infection with a focus on underrepresented African populations. We identified nonsynonymous and regulatory variants at *ACE2* that appear to be targets of recent natural selection in some African populations. We detected evidence of ancient adaptive evolution at *TMPRSS2* in the human lineage. Genetic variants that are targets of natural selection are associated with clinical phenotypes common in patients with coronavirus disease 2019.

The authors declare no competing interest.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

¹C.Z., A.V., and Y.F. contributed equally to this work.

²Present address: Janssen Research and Development Data Science, La Jolla, CA 92121.

³A complete list of the Regeneron Genetic Center can be found in *SI Appendix*.

⁴To whom correspondence may be addressed. Email: giorgio.sirugo@penmedicine.upenn.edu or tishkoff@penmedicine.upenn.edu.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2123000119/-DCSupplemental>.

Published May 17, 2022.

spectrum (6). SARS-CoV-2 infects cells through a membrane fusion mechanism, which in the case of SARS-CoV, is known to induce down-regulation of *ACE2* (7). Such down-regulation has been shown to cause inefficient counteraction of angiotensin II effects, leading to enhanced pulmonary inflammation and intravascular coagulation (7). Additionally, altered expression of *ACE2* has been associated with cardiovascular and cerebrovascular disease, which is highly relevant to COVID-19 as several cardiovascular conditions are associated with severe disease. *TMPRSS2*, located on the outer membrane of host target cells, binds to and cleaves *ACE2*, resulting in activation of spike proteins on the viral envelope and facilitating membrane fusion and endocytosis (8). Two additional genes, *DPP4* and *LY6E*, have been shown to play an important role in the entry of SARS-CoV-2 virus into host cells. *DPP4* is a known functional receptor for the Middle East respiratory syndrome coronavirus (MERS-CoV), causing a severe respiratory illness with high mortality (9, 10). *LY6E* encodes a glycosylphosphatidylinositol-anchored cell surface protein, which is a critical antiviral immune effector that controls coronavirus infection and pathogenesis (11). Mice lacking *LY6E* in hematopoietic cells were susceptible to murine coronavirus infection (11).

Previous studies of genetic diversity at *ACE2* and *TMPRSS2* in global human populations did not include an extensive set of African populations (5, 12–14). No common coding variants (defined here as minor allele frequency [MAF] > 0.05) at *ACE2* were identified in any prior population studies. However, few studies included diverse indigenous African populations whose genomes harbor the greatest diversity among humans. This leads to a substantial disparity in the representation of African ancestries in human genetic studies of COVID-19, impeding health equity as the transferability of findings based on non-African ancestries to African populations can be low (15). Including more African populations in studying the genetic diversity of genes involved in SARS-CoV-2 infection is extremely necessary. Additionally, the evolutionary forces underlying global patterns of genetic diversity at host genes related to SARS-CoV-2 infection are not well understood. Using methods to detect natural selection signatures at host genes related to viral infections helps identify putatively functional variants that could play a role in disease risk.

We characterized genetic variation and studied natural selection signatures at *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* in ethnically diverse human populations by analyzing 2,012 genomes from ethnically diverse Africans (referred to as the “African diversity” dataset), 2,504 genomes from the 1000 Genomes Project (1KG), and whole-exome sequencing of 15,977 individuals of European ancestry (EA) and African ancestry from the Penn Medicine BioBank (PMBB) dataset (*SI Appendix*, Fig. S1). The African diversity dataset includes populations with diverse subsistence patterns (hunter-gatherers, pastoralists, agriculturalists) and speaking languages belonging to the four major language families in Africa (Khoesan; Niger–Congo, of which Bantu is the largest subfamily; Afroasiatic; and Nilo-Saharan). We identify functionally relevant variation, compare the patterns of variation across global populations, and provide insight into the evolutionary forces underlying these patterns of genetic variation. In addition, we perform an association study using the variants identified from whole-exome sequencing at the four genes and clinical traits derived from electronic health record (EHR) data linked to the subjects enrolled in the PMBB. The EHR data include diseases related to organ dysfunctions associated with severe COVID-19, such as respiratory, cardiovascular, liver, and renal complications. Our study

of genetic variation in genes involved in SARS-CoV-2 infection provides data to investigate infection susceptibility within and between populations and indicates that variants in these genes may play a role in comorbidities relevant to COVID-19 severity.

Results

Coding Variation at *ACE2* among Global Populations. In total, we identified 41 amino acid changing variants (Fig. 1A and Dataset S1). Twenty-eight (69%), 20 (49%), 18 (44%), and 16 (40%) of the nonsynonymous variants were predicted to be deleterious or likely deleterious based on scores generated from Combined Annotation Dependent Depletion (CADD) (16), Sorting Intolerant From Tolerant (SIFT) (17), Polymorphism Phenotyping v2 (PolyPhen2) (18), and Consensus Deleteriousness Score of Missense Mutations (Condel) (19), respectively (Dataset S1).

Among the 41 coding variants identified at *ACE2*, all are rare (MAF < 0.05) in the pooled global population dataset (Fig. 1A and Dataset S1). However, there are variants that are common (MAFs \geq 0.05) in the central African hunter-gatherer (CAHG) population from Cameroon (often referred to as “pygmies”) (Fig. 1B). One of these variants, rs138390800 (Lys341Arg), is a deleterious nonsynonymous variant based on SIFT and CADD and present at high frequency (MAF = 0.164) in the CAHG, while it is rare in other African populations and absent in non-African populations (Fig. 1C). Two other nonsynonymous variants, rs147311723 (Leu731Phe; MAF = 0.083) and rs145437639 (Asp597Glu; MAF = 0.083), are also common only in the CAHG population (Fig. 1B and Dataset S1). These three nonsynonymous variants are the only common coding variants found at *ACE2* in any of the populations examined.

We then investigated the potential role of these 41 coding variants in the conformation of the *ACE2* protein. The 41 coding variants are distributed across the entire *ACE2* protein (Fig. 1D and Dataset S1), including its receptor binding domain (RBD) region, which binds to the SARS-CoV-2 spike protein, dimerization interface, and transmembrane helix. In particular, two nonsynonymous variants Gly354Asp (chrX:15581230_C_T) and Ser43Asn (chrX:15600784_C_T) are both found directly in the RBD binding region of *ACE2* (Fig. 1D and Dataset S1); the former is only found in low frequency in one population, the Fulani from Cameroon (MAF = 0.004), and the latter is also an African-specific variant that is at low frequency in only three East African populations, two of which are Afroasiatic-speaking populations from Kenya (MAF = 0.018) and Ethiopia (MAF = 0.008) (Dataset S1). The variant Arg708Trp (rs776995986) occurs in the region identified as the *TMPRSS2* cleavage site in *ACE2* (20) and is found only in the Afroasiatic-speaking populations from Ethiopia (MAF = 0.002). The presence of arginine residues has been shown to be important in “multibasic” cleavage sites (3). Therefore, due to the drastic change in physicochemical properties of the residue, this variation could be expected to interfere in *TMPRSS2* cleavage efficiency, although it warrants experimental validation. Finally, two variants are located at glycosylation sites. Variant Asn546Ser (rs756905974), which causes the loss of a conserved glycosylation site on the *ACE2* protein, is found only in the South Asian (SAS) populations (MAF = 0.001). Variant Lys26Arg (rs4646116), found in individuals from the European (EUR; MAF = 0.005), African (AFR; MAF = 0.001), and SAS (MAF = 0.002) populations from the 1KG dataset (Dataset S1), occurs near both the conserved *ACE2*

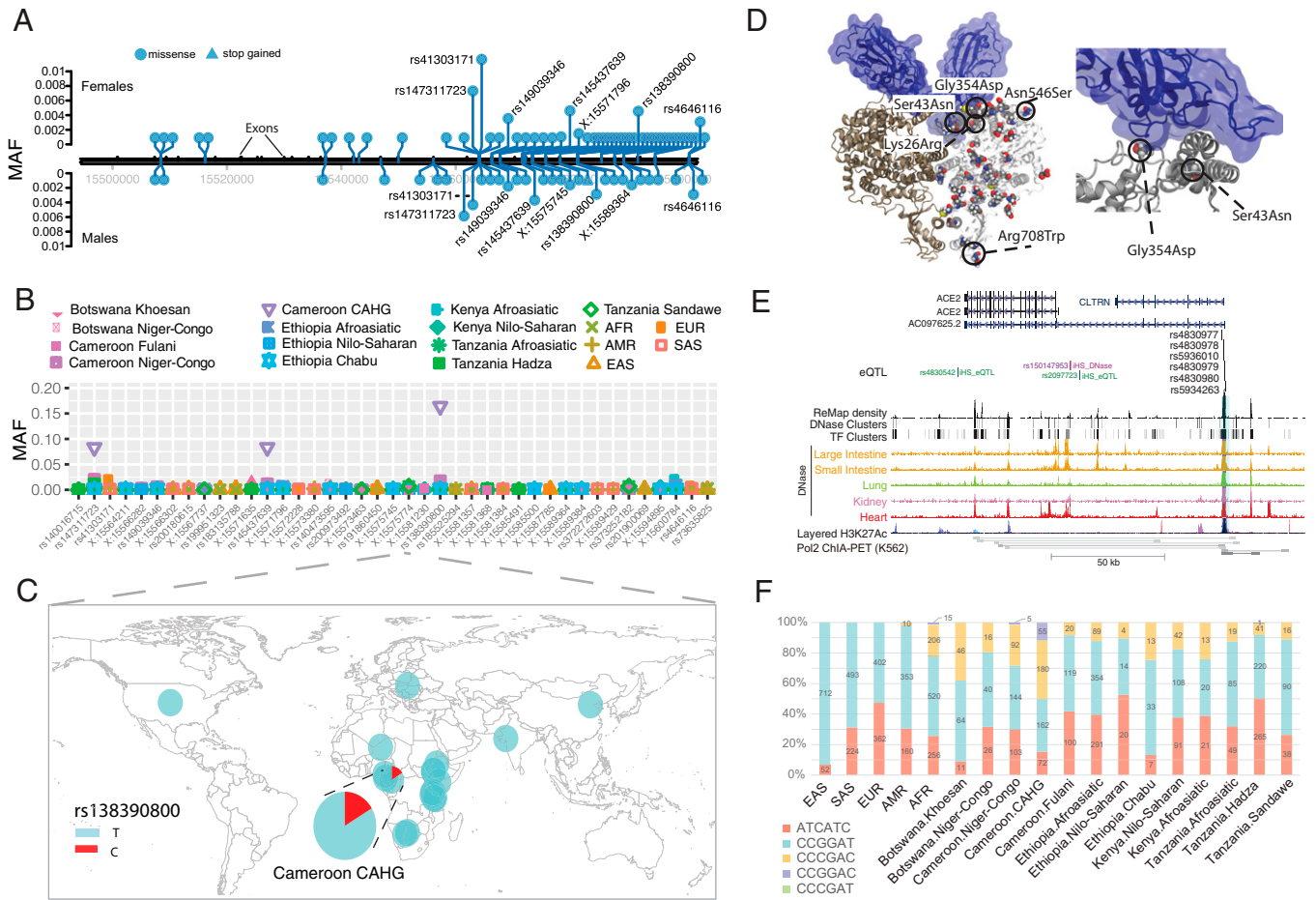


Fig. 1. Genetic variation at *ACE2*. (A) Location of coding variants and their MAF at *ACE2* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The geographic distribution of the MAF for rs138390800 at *ACE2* in diverse global ethnic groups is highlighted. Each pie denotes frequencies of alleles in the corresponding population. (D) Locations of identified nonsynonymous variants within the secondary structure of the *ACE2* protein. (E) Six regulatory eQTLs located in an upstream enhancer of *ACE2*. RNA Pol2 ChIA-PET data and DNase-seq data of the large intestine, small intestine, lung, kidney, and heart are from ENCODE (68). (F) Haplotype frequencies of the six eQTLs in global populations. The six eQTLs were ordered by their genomic position on the chromosome (i.e., with the following order: rs4830977, rs4830978, rs5936010, rs4830979, rs4830980, and rs5934263). Haplotypes with frequency <0.01 are not shown.

glycosylation site Asn90 and the RBD binding site. The modification to a similarly positively charged residue could suggest a role for electrostatic interactions, although no direct interference with RBD binding could be deduced without further studies.

Regulatory Variation at *ACE2* among Global Populations. In contrast to coding variants, which have direct effects on protein structure in all cells expressing a gene, the effects of regulatory genetic variants are relatively difficult to determine (21). We first extracted 2,053 expression quantitative trait loci (eQTLs) significantly associated with *ACE2* gene expression ($P < 0.001$) from the Genotype-Tissue Expression (GTEx) project database (6) (Dataset S1). To narrow down candidate functional variants, we focus on the eQTLs located in the promoter regions of target genes or in enhancers supported by chromatin interaction data (22) (Dataset S1).

We identified six eQTLs (rs4830977, rs4830978, rs5936010, rs4830979, rs4830980, and rs5934263) located in a strong deoxyribonuclease (DNase) peak 73.3 kb upstream of *ACE2* that have direct interactions with *ACE2* based on RNA Pol2 chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data (Fig. 1E, Dataset S1, and SI Appendix, Fig. S2). All six single nucleotide polymorphisms (SNPs) are eQTLs of *ACE2*, and all of them have positive normalized effect sizes (NESs; NES > 0.2) and

significant P values ($P < 0.00008$) in brain, tibial nerve, tibial artery, pituitary, and prostate cells (Dataset S1 and SI Appendix, Fig. S3). In non-African populations, these six eQTLs are in high linkage disequilibrium (LD; $R^2 = 0.91$ to 1.0) (SI Appendix, Fig. S4), and thus, there are two common haplotypes: “CCGGAT” and “ATCATC.” The frequency for the ATCATC haplotype ranges from 0.31 to 0.47 in all populations except the East Asian population, which has a frequency of 0.068 at all six SNPs (Fig. 1F). In African populations, LD is lower ($R^2 > 0.5$) (SI Appendix, Fig. S4), and there are three common haplotypes: CCGGAT (0.564), ATCATC (0.308), and “CCCGAC” (0.116). Of note, every allele in the haplotype CCGGAT is correlated with higher expression of *ACE2* in the cortex of the brain, while alleles in haplotype ATCATC are correlated with lower expression of *ACE2*; other haplotypes have alleles with both positive and negative effect sizes in different tissues (Dataset S1). Haplotype CCCGAC is only present in populations with African ancestry, and its frequency is highest in the Botswana Khoesan (0.38) and Cameroon CAHG (0.38) hunter-gatherer populations. We also identified one variant (rs186029035) located in a strong transcription factor (TF) binding and DNase region (based on the Encyclopedia of DNA Elements, ENCODE) in the 16th intron of *ACE2*. This variant is only common in the Cameroon CAHG population, and therefore, there are no eQTL data for this SNP in the GTEx database (MAF = 0.153) (Dataset S1).

Signatures of Natural Selection at *ACE2*. As indicated above, most of the nonsynonymous variants at *ACE2* are rare in global populations, and many of them are predicted to be deleterious, indicating that this gene is under strong purifying selection. To formally test for signatures of natural selection at *ACE2*, we first examined the ratio of nonsynonymous and synonymous variants at each gene using the ratio of nonsynonymous to synonymous substitutions (dN/dS) test (23) (*Materials and Methods*). The dN/dS for all pooled samples was 0.77, indicating that *ACE2* is under moderate purifying selection globally (*Dataset S2* and *SI Appendix, Fig. S5*). However, in the East Asian population, we observed seven nonsynonymous variants (all of them are rare) and only one synonymous variant, and the dN/dS value is 1.85, indicating an excess of nonsynonymous variation. In other populations, the dN/dS ratio ranges from 0 to 0.79 (*Dataset S2* and *SI Appendix, Fig. S5*). Thus, *ACE2* appears to be under strong purifying selection in most populations but may be under weak purifying selection in the East Asian population. We next applied the McDonald–Kreitman (MK) test (24), which compares the ratio of fixed nonsynonymous sites between humans and chimpanzee ($D_n = 8$) and fixed synonymous sites ($D_s = 6$) with the ratio of polymorphic nonsynonymous sites among populations ($P_n = 41$) relative to polymorphic synonymous sites ($P_s = 14$), and found that it is not significant (odds ratio [OR] = 0.45, $P = 0.94$, two-sided Fisher’s exact test) (*Dataset S3* and *SI Appendix, Figs. S6 and S7*).

Because the above-mentioned methods are more suitable for detecting signals of natural selection acting over long timescales (25, 26), we then tested for signatures of recent positive selection at *ACE2* in global populations using the integrated haplotype score (iHS) test (27) to detect extended haplotype homozygosity (EHH) (28), which identifies regions of extended LD surrounding a positively selected locus. We first focused on the three common nonsynonymous variants in the CAHG population from Cameroon (rs138390800, rs147311723, and rs145437639; MAF = 0.083 to 0.164) and a common putative regulatory variant (rs186029035, located in TF and DNase regions in the 16th intron of *ACE2*; MAF = 0.153). The derived alleles of these variants exist on three different haplotype backgrounds; rs147311723 and rs145437639 are on the same haplotype backgrounds, while rs138390800 and rs186029035 are on similar but distinct haplotype backgrounds (Fig. 2A). The derived alleles of the corresponding SNPs on each haplotype background show EHH extending longer than 2 Mb, while the ancestral alleles of these SNPs harbor haplotypes extending less than 0.3 Mb (Fig. 2B). We then calculated the iHS of each of these variants to determine whether these extended haplotypes are unusually long compared with other SNPs with a similar allele frequency; the iHS values were not significant for any of these variants (iHS values for rs138390800, rs147311723, and rs145437639 in CAHG were 0.5, -0.08 , and -0.55 , respectively) (*Dataset S4* and *SI Appendix, Fig. S8*). However, if selection was acting on multiple haplotypes simultaneously, the EHH and iHS tests would not be well powered to detect selection (29). We also used the d_i statistic (30) to measure if allele frequencies at these candidate SNPs were significantly differentiated between Cameroon CAHG and other populations. The d_i values of SNPs rs138390800 and rs186029035 were in the top 1.4% and 1.7%, respectively, of d_i values for all SNPs examined, indicating that that allele frequencies at these variants are among the most highly differentiated in the CAHG population, consistent with local adaptation. However, it should be noted that in the CAHG, these four variants (rs138390800, rs147311723, rs145437639, and rs186029035) are in complete LD based on D' ($D' = 1$) with the six eQTLs described above

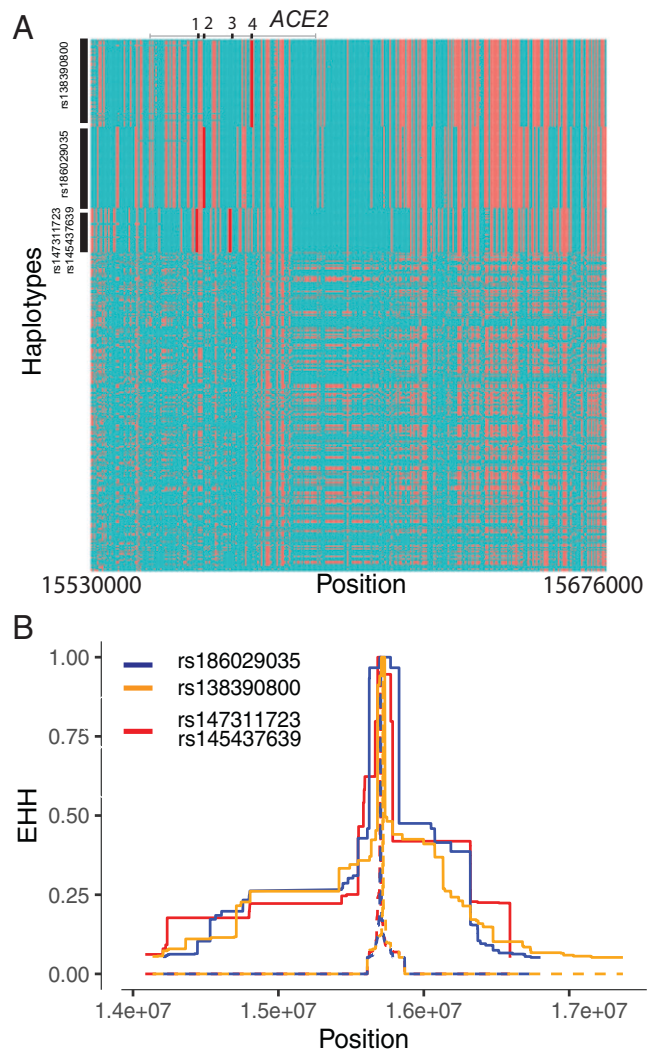


Fig. 2. Natural selection signatures at *ACE2* in the Cameroon CAHG populations. (A) Haplotypes over 150 kb flanking *ACE2* in CAHG populations. The x axis denotes the genetic variant position, and the y axis represents haplotypes. Each haplotype (one horizontal line) is composed of the genetic variants (columns). Red dots indicate the derived allele, while green dots indicate the ancestral allele. Haplotypes surrounded by the upper left vertical black lines suggest that these haplotypes carry derived allele(s) of the labeled variant near the corresponding black line. For example, the first black line denotes all the haplotypes that have the derived allele at rs138390800 (dark red line). Haplotypes carrying rs138390800, rs147311723, rs145437639, and rs186029035 show more homozygosity than other haplotypes; 1, 2, 3, and 4 along the top of the plot denote positions for rs147311723, rs186029035, rs145437639, and rs138390800, respectively. (B) EHH of rs138390800, rs186029035, and rs147311723 (rs145437639 is in strong LD with rs147311723) at *ACE2* in CAHG populations.

(*SI Appendix, Figs. S9 and S10*), indicating that the alleles are on the same haplotype background. Thus, it is not possible to distinguish if the nonsynonymous variants are targets of selection or if they are “hitchhiking” to high frequency due to selection on flanking regulatory variants. Given the high LD in the region, it is possible that multiple functional variants on the same haplotype backgrounds have been under selection.

We then investigated signatures of recent positive selection at candidate regulatory variants near *ACE2* in the global datasets. In total, there are 234 variants that had high iHS scores ($|iHS| > 2$) in at least one population extending over an ~ 200 -kb region (*Dataset S4*), and 48% ($n = 113$) of these variants are either eQTLs or located at DNase hypersensitive regions, which are in high LD based on D' (*Dataset S4* and *SI*

Appendix, Figs. S9 and S10). Among the region near the transcription start site (TSS) (<10 kb from *ACE2*), there are two variants in high LD ($D' = 1$) that had high *iHS* scores in the San population from Botswana (Fig. 3A); rs150147953 is located in a DNase peak in multiple tissues, including lung, intestine, and heart, and rs2097723 is an eQTL of *ACE2* in the brain (Fig. 1E, Dataset S4, and SI Appendix, Fig. S11). We also identified high *iHS* signals at the region 50 to 120 kb upstream of *ACE2* (chrX:15650000-15720000) in the AFR IKG population, the San from Botswana, and Niger–Congo-speaking populations from Cameroon as well as Afroasiatic- and Nilo-Saharan-speaking populations from Kenya (Fig. 3A

and SI Appendix, Fig. S7). Two SNPs in this region (rs5936010 and rs5934263) have elevated *iHS* scores ($|iHS| > 2$) in the San population from Botswana and the Afroasiatic population from Kenya (Fig. 3A) and are part of the six eQTLs described above, located within a strong enhancer interacting with the promoter of *ACE2* (Fig. 1E). Two additional eQTLs that are in complete LD with the six eQTLs ($D' = 1$) (SI Appendix, Fig. S10), rs4830984 and rs4830986, had high *iHS* scores in four of the five African populations listed above (all but Kenya Afroasiatic) (Fig. 3A).

We performed haplotype network analysis to examine phylogenetic relationships among haplotypes at *ACE2* in global

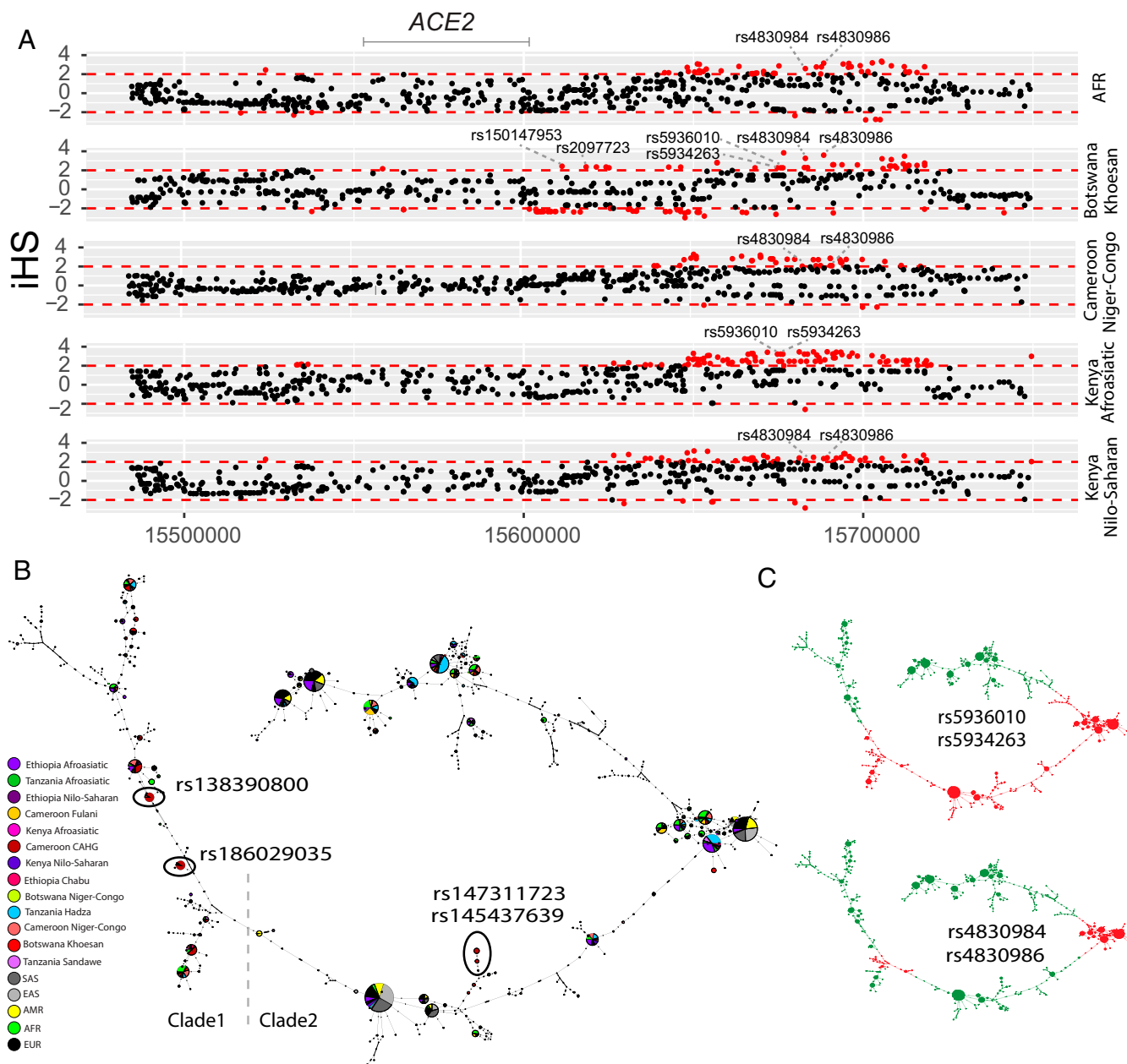


Fig. 3. Natural selection signatures at the upstream region of *ACE2* in African populations. (A) *iHS* signals at the upstream region of *ACE2* (chrX:15650000-15720000) in African populations. Each dot represents a SNP. Red dots denote SNPs that are significant ($|iHS| > 2$). The gray solid lines denote the gene body region of *ACE2*. Putatively causal tag SNPs are annotated in the plots. (B) Haplotype network over 150 kb flanking *ACE2* in diverse ethnic populations. The network was constructed with SNPs that showed *iHS* signals in all populations and overlapped with DNase regions or eQTLs. The four functional candidates identified in Cameroon CAHG were also included in the networks. Each pie represents a haplotype, each color represents a geographical population, and the size of the pie is proportional to that haplotype frequency. The dashed line denotes the boundary of clade 1 and clade 2. Black ovals denote haplotypes containing the corresponding variants. (C) Haplotypes containing variants rs5936010, rs5934263, rs4830984, and rs4830986 are highlighted. Red pies denote haplotypes containing the derived allele of the corresponding variants, while green pies denote haplotypes containing the ancestral allele of the corresponding variants.

populations for SNPs showing signatures of natural selection (Fig. 3 B and C). We identified two haplotype clades; one (clade 1) is nearly specific to Africans, and the other (clade 2) encompasses global populations (Fig. 3B). In the CAHG, haplotypes containing the rs138390800 (Lys341Arg) nonsynonymous variant and the rs186029035 regulatory variant are in clade 1, whereas haplotypes containing the rs147311723 (Leu731Phe) and rs145437639 (Asp597Glu) nonsynonymous variants are located in clade 2 (Fig. 3B). Haplotypes containing the two regulatory variants (rs5936010 and rs5934263) located 50 to 120 kb upstream of *ACE2* are shared in global populations, and the nearby regulatory variants rs4830984 and rs4830986 are sublineages on those haplotype backgrounds (Fig. 3 B and C).

Genetic Variation at *TMPRSS2* among Global Populations. The *TMPRSS2* protein enhances the spike protein-driven viral entry of SARS-CoV-2 into cells (3). At this gene, we identified 48 nonsynonymous variants. Among the nonsynonymous variants, only two (rs12329760 [Val197Met] and rs75603675 [Gly8Val]) have high MAF (>0.05) in the pooled global dataset (Fig. 4A and Dataset S1). While rs75603675 is highly variable in non-East Asian populations (AFR = 0.3, AMR = 0.27, EUR = 0.4, and SAS = 0.2), it is not highly variable in East Asians (MAF = 0.02) (Fig. 4 B and C and Dataset S1). In addition, some nonsynonymous variants were common and specific to African populations. Notably, the nonsynonymous variant rs61735795 (Pro375Ser) had a high MAF in the Khoesan-speaking population from Botswana (MAF = 0.18). This variant is present at low frequency in populations from Cameroon (MAF < 0.01) and Ethiopia (MAF < 0.03) and was absent in non-African populations. The nonsynonymous variant rs367866934 (Leu403Phe) is common in the Cameroonian CAHG population (MAF = 0.15) and has low frequency (MAF = 0.02) in other populations from Cameroon, but it is absent from non-Cameroonian populations (Fig. 4B and Dataset S1). Another nonsynonymous variant rs61735790 (His18Arg) is common in the CAHG populations from Cameroon (MAF = 0.12) and the Nilo-Saharan populations from Ethiopia (MAF = 0.12) but is rare in other populations (Fig. 4B and Dataset S1).

We identified two regulatory SNPs (rs76833541 and rs4283504) in the promoter region of the *TMPRSS2* gene that have been identified as eQTLs of *TMPRSS2* in testis (Fig. 4D, Dataset S1, and SI Appendix, Fig. S12). The MAF of rs76833541 is higher in EUR (MAF = 0.16) than other populations (EAS = 0.002, AFR = 0.006, AMR = 0.06, and SAS = 0.05), and the MAF of rs4283504 is more common in EAS (MAF = 0.21) than other populations (EUR = 0.11, AFR = 0.04, AMR = 0.12, and SAS = 0.14) (Dataset S1 and SI Appendix, Fig. S13).

Signatures of Natural Selection at *TMPRSS2*. We applied the MK test at *TMPRSS2* and observed that Dn/Ds (13/2) is significantly larger than Pn/Ps (48/45) among pooled human samples (OR = 6.1, *P* value = 0.009, Fisher's exact test) (Fig. 5A and Dataset S3) as well as in individual ethnic groups (OR ranged from 5.0 to 17), indicating positive selection in the hominin lineage after divergence from chimpanzee. Notably, there are 13 nonsynonymous and 2 synonymous variants at *TMPRSS2* (ENST00000398585.7) (SI Appendix, Fig. S14 shows ENST00000332149.10) that were fixed in human populations. The nonsynonymous variants are located in different structural domains of *TMPRSS2*. Amino acids A3P, N10S, T46P, A70V, R103C, and M104T are located in the cytoplasmic region, which may function in intracellular signal transduction (31). L124I is located in the transmembrane region, N144K is located in the extracellular region, S165N and S178G are located in the low-density lipoprotein (LDL) receptor class A domain, E441Q and T515M are located in the serine peptidase (Peptidase S1) domain that is involved in the interaction with the SARS-CoV-2 spike protein (3), and S529G is located in the last amino acid position of the protein (Fig. 5B). In contrast to the MK test, the dN/dS ratio test was not significant in any population, indicating no excess of nonsynonymous to synonymous variation within populations (Dataset S2 and SI Appendix, Fig. S5).

We also tested for recent positive selection at *TMPRSS2* in all ethnic groups using iHS (Dataset S4 and SI Appendix, Fig. S15). We found many SNPs (*n* = 153) with high iHS scores (*|iHS|* > 2) in different ethnic groups in a 78-kb region encompassing the *TMPRSS2* gene that show high levels of LD (chrX:

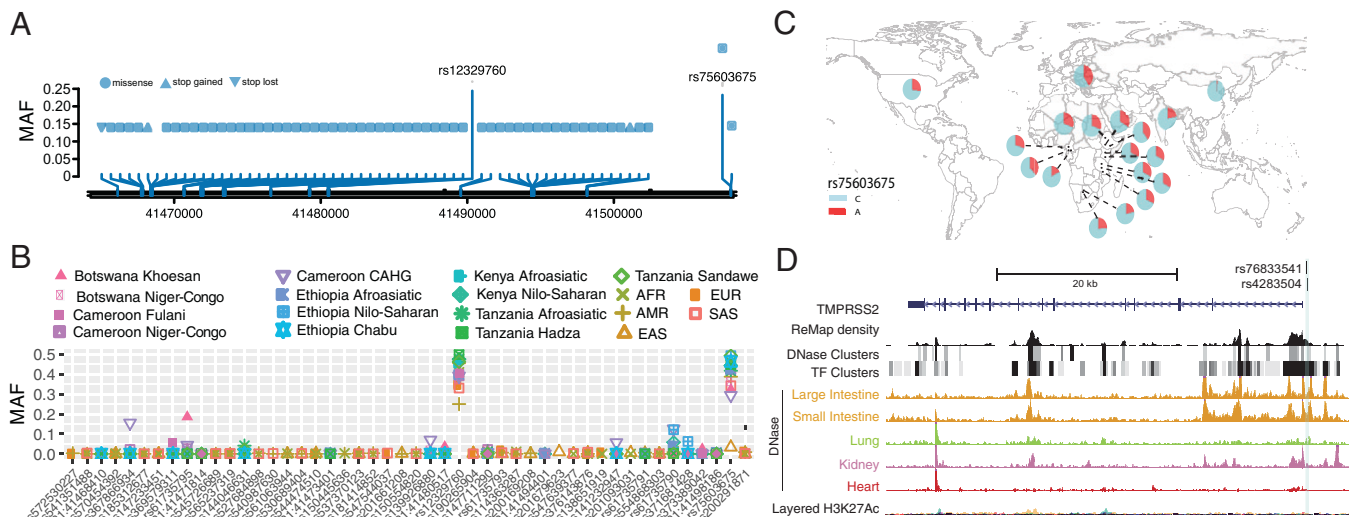


Fig. 4. Genetic variation at *TMPRSS2*. (A) Location of coding variants and their MAF at *TMPRSS2* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The geographic distribution of the MAF for rs75603675 at *TMPRSS2* in diverse global ethnic groups is highlighted. Each pie denotes frequencies of alleles in the corresponding population. (D) Two regulatory eQTLs located in the promoter region of the *TMPRSS2* gene. DNase-seq data of the large intestine, small intestine, lung, kidney, and heart are from ENCODE (68).

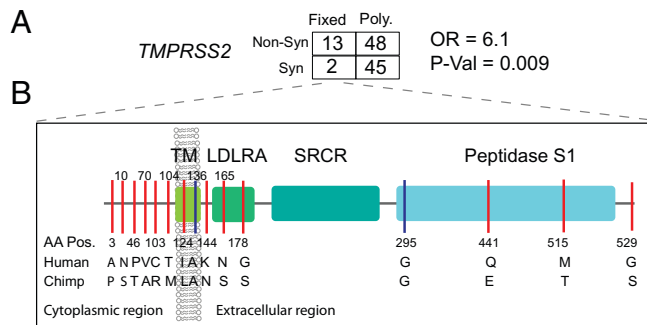


Fig. 5. Natural selection signatures at *TMPPRSS2*. (A) The result of the MK test for *TMPPRSS2* in the pooled dataset. Nonsyn indicates nonsynonymous variants; Syn indicates synonymous variants. “Fixed” denotes variants that were fixed between the human and the chimpanzee; “Poly” represents polymorphic variants within human populations. The transcript ENST00000398585.7 was used for the calculation. (B) Illustration of locations of variants that are divergent between the human and chimpanzee lineages on the *TMPPRSS2* protein domains. Boxes denote the protein domains of *TMPPRSS2*. Red lines represent nonsynonymous variants that occurred in the corresponding domains of *TMPPRSS2*, with the amino acids and positions of the human and the chimpanzee annotated at the bottom of the lines. Blue lines denote synonymous variants. LDLRA, LDL receptor class A; SRCR, scavenger receptor cysteine-rich domain 2; TM, transmembrane domain.

41454000–41541000) (*SI Appendix, Figs. S16 and S17*). We identified a nonsynonymous variant (rs150969307) that shows a signature of positive selection ($iHS = 2.01$) and is common only in the Chabu hunter-gatherer population from Ethiopia (MAF = 0.079) (*Dataset S4*). We found that more than one-third of SNPs with $|iHS|$ scores >2.0 (62 of 153) are located in putative regulatory regions (*Dataset S4 and SI Appendix, Fig. S18*).

Genetic Variation and Signatures of Natural Selection at *DPP4* and *LY6E*. *DPP4* is a receptor for MERS-CoV and was reported to interact with SARS-CoV-2 (10). At this gene, we identified 47 nonsynonymous variants and 1 loss-of-function variant (*Dataset S1*). There were no common nonsynonymous variants in the pooled global dataset (*SI Appendix, Fig. S19A*), suggesting that this gene is extremely conserved during human evolutionary history. Only one nonsynonymous variant (rs1129599, Ser437Thr) was common in the Fulani pastoralists from Cameroon (MAF = 0.081), was present at low frequency in other African populations, and was absent in non-African populations (*SI Appendix, Fig. S19 B and C*). In addition to the nonsynonymous variants, one loss-of-function variant was identified at *DPP4*. The variant rs149291595 (Q170*) has low MAF in some African populations (MAF < 0.05) but is absent in non-African populations.

We identified four eQTLs (rs1861978, rs35128070, rs17574, and rs13015258) in the promoter region of the *DPP4* gene (*SI Appendix, Figs. S19D and S20*). Three of the variants (rs1861978, rs35128070, and rs17574) are significant eQTLs in the transverse colon, and rs13015258 is an eQTL in the lung ($P < 5.9 \times 10^{-6}$) (*Dataset S1 and SI Appendix, Fig. S19*). The minor alleles of these three variants are rare in EAS (MAF < 0.05) but common in all other populations (MAF > 0.15) (*Dataset S1 and SI Appendix, Fig. S21*). The fourth SNP, rs13015258, resides in the center of a cluster of DNase peaks identified in ENCODE (*SI Appendix, Fig. S19D*), with MAF ranging from 0.38 in the AMR population to 0.6 in other populations (*Dataset S1 and SI Appendix, Fig. S21*).

The MK test result of *DPP4* was not significant in either the pooled samples ($D_n = 3$, $D_s = 5$, $P_n = 45$, $P_s = 33$; OR = 0.44, $P = 0.9$, two-sided Fisher’s exact test) or each population separately (*Dataset S3 and SI Appendix, Fig. S7*). For the

dN/dS test, we observed ratios ranging from 0 to 0.52 in individual populations, indicating that *DPP4* is highly conserved (*Dataset S2 and SI Appendix, Fig. S5*) within human populations. Using the iHS test, we identified eight SNPs that had extremely high iHS scores ($|iHS| > 2$) in the Khoesan populations from Botswana (*Dataset S4 and SI Appendix, Fig. S22*). Five of these SNPs (rs10166124, rs2284872, rs2284870, rs7608798, and rs2160927) are in LD ($D' > 0.95$) with each other (*SI Appendix, Fig. S23*). The SNP rs2284870 is located in a strong DNase peak in heart tissue (*Dataset S4 and SI Appendix, Fig. S24*).

Studies show that mice lacking *LY6E* were highly susceptible to a usually nonlethal mouse coronavirus (11). At *LY6E*, we observed 28 nonsynonymous variants, and all of them, except rs11547127 (MAF = 0.057), have MAFs that are rare in the pooled global dataset (*Dataset S1 and SI Appendix, Fig. S25A*). However, some nonsynonymous variants are common in specific populations (*SI Appendix, Fig. S25B*). For instance, the nonsynonymous variant rs111560737 (Asp104Asn) was common in the southern African Khoesan population from Botswana (MAF = 0.36) and the Chabu population from Ethiopia (MAF = 0.17) (*SI Appendix, Fig. S25C*). Three loss-of-function variants (rs200177123 [stop gained, Ser59*], chr8:143020941, and chr8:143020946) were also identified at *LY6E*, and all of them are rare.

We identified three regulatory eQTLs (rs13252864, rs17061979, and rs114909654) located within 2 kb of the transcription start site of *LY6E* (*SI Appendix, Fig. S25D*), all of which are significant in esophageal mucosa ($P < 1 \times 10^{-5}$) (*Dataset S1 and SI Appendix, Fig. S26*), which has a high expression level of *LY6E* (transcript per million, TPM = 108, GTEX). The minor alleles of rs13252864 and rs114909654 are common in African populations (MAF > 0.15) while very rare in other populations (MAF < 0.02) (*SI Appendix, Fig. S27*), whereas the MAF of rs17061979 is relatively high in EAS (0.18) and SAS (0.13) and rare in other populations (MAF < 0.05) (*SI Appendix, Fig. S27*).

The MK test for *LY6E* was not significant in either the pooled samples ($D_n = 0$, $D_s = 4$, $P_n = 9$, $P_s = 9$; OR = 0, $P = 0.9$, two-sided Fisher’s exact test) or each population separately (OR ranging from 0 to 0.52) (*Dataset S3 and SI Appendix, Fig. S7*). For the dN/dS test, we observed ratios ranging from 0 to 0.68 in individual populations, indicating that *LY6E* is highly conserved (*Dataset S2 and SI Appendix, Fig. S5*). We identified 19 variants that had extreme high iHS scores ($|iHS| > 2$) (*Dataset S4 and SI Appendix, Fig. S28*), some of which are in LD in specific populations (*SI Appendix, Fig. S29*). One variant (rs867069115) shows an extreme iHS score in the Hadza hunter-gatherer population from Tanzania ($iHS = -2.94$). This variant is located in a regulatory region ~ 1.9 kb downstream of *LY6E* within DNase and TF peaks in the lung, intestine, kidney, heart, stomach, pancreas, and skeletal muscle from ENCODE (*SI Appendix, Fig. S30*); is common only in the Hadza population (MAF = 0.14); is rare in other African populations (MAF < 0.05); and is absent in all non-African populations (*Dataset S1*). SNP rs10283236, which shows an extreme iHS value in the CEU population, is an eQTL of *LY6E* located within DNase and TF clusters identified in ENCODE (~ 4.14 kb downstream of *LY6E*) active in many tissues, including lung, kidney, and small intestine.

Associations between Genetic Variation in Host Genes and Clinical Disease Phenotypes. We examined associations of genetic variation at four host genes relevant to SARS-CoV-2 infection with clinical phenotypes using the PMBB cohort that

consists of exome-sequencing data from 15,977 participants between the ages of 19 and 89 y (52% female) with extensive clinical data available through their EHRs. Of these, 7,061 individuals were of European ancestry (EA) (42%), and 8,916 were of African ancestry (AA) (55%) (*SI Appendix, Table S1*).

Gene burden phenome-wide association study with clinical phenotypes. To test for the association between rare coding variants and clinical phenotypes, we applied a gene-based approach (32, 33) as well as single-variant analysis. First, we performed a gene-based analysis by collapsing the coding region variants with MAF < 0.01 that are annotated as nonsynonymous or putative loss-of-function (pLOF) variants. We examined ~1,800 phecodes derived from the EHR and performed a phenome-wide association study (PheWAS) with individual SNPs (32–34). After multiple testing correction, we identified one association in the AA population and five associations in the EA population reaching study-wide significance ($P = 6.6 \times 10^{-6}$ [0.05/(1,866 codes \times 4 genes)]) (*Dataset S5*). Myocarditis, a rare cardiovascular disease caused by viral infection, was the top PheWAS association with *ACE2* in the AA population but was not significant in the EA population. Although the population difference for this specific association is unclear, recent studies have reported that COVID-19 patients have a 16 times higher risk of myocarditis (35). Furthermore, *ACE2* is expressed in heart tissue and its upregulation in cardiomyocytes has an important role in both dilated cardiomyopathy and hypertrophic cardiomyopathy (36–38).

Gene burden association analyses with 12 COVID-19-relevant organ dysfunctions. We tested for the association of rare coding variants with 12 phenotypes, encompassing COVID-19-relevant disease classes affecting different organ systems, defined by EHR-based diagnosis codes (*Dataset S5*). In the AA population, the phenotypes with the most significant associations with *ACE2* were hepatic encephalopathy and respiratory failure (Fig. 6A and Table 1). The association with respiratory failure is interesting as it is one of the key severe clinical features reported for COVID-19 (39–43). However, the same association was not significant in the EA population, which could be explained by lack of power due to a lower number of coding variants at *ACE2* in EA. Within the EA population, the most significant associations with *ACE2* included hepatic coma, respiratory syncytial virus infectious disease, and cirrhosis of the liver (Table 1). In the gene-based analysis of variants in *DPP4*, we identified significant associations (only in the sequence kernel association test [SKAT] model) with respiratory syncytial virus infectious disease and upper respiratory tract disease in the AA population (Fig. 6A, Table 1, and *Dataset S5*); this observation was not observed in the EA population. None of the associations with *DPP4*, *TMPRSS2*, and *LY6E* reach statistical significance in gene burden analysis.

PheWAS of eQTLs near COVID-19 host immunity-related genes. Lastly, we conducted PheWAS of eQTLs identified near the host genes. For the six eQTLs identified near *ACE2*, rs5936010 and rs5934263 (targets of positive selection in both Afroasiatic populations from Kenya and Khoesan populations from Botswana) were significantly associated with type 2 diabetes ($P = 1.23 \times 10^{-4}$, OR = 1.1) and hypertension ($P = 8.8 \times 10^{-4}$, OR = 1.13), respectively. The association was only observed in the AA population (Fig. 6B and *Dataset S6*). The PheWAS of the two regulatory eQTLs (rs76833541 and rs4283504) near *TMPRSS2* described above identified association of rs76833541 with abnormal glucose ($P = 8.9 \times 10^{-4}$, OR = 1.5) in EA and rs4283504 with glucocorticoid deficiency ($P = 0.001$, OR = 2.7) in AA (Fig. 6B). The PheWAS of four

regulatory eQTLs near *DPP4* identified the association with malignant neoplasm of the rectum (commonly referred as colon cancer). The SNP rs17574 was associated with increased risk ($P = 4.49 \times 10^{-04}$, OR = 1.8) of colon cancer; however, the observation was only observed among AA individuals. The association analysis of regulatory variants near *LY6E* identified significant associations with “severe protein-calorie malnutrition” (rs114909654, $P = 2.35 \times 10^{-05}$, OR = 1.9) and “acute post hemorrhagic anemia” (rs114909654, $P = 6.4 \times 10^{-04}$, OR = 1.6) in the AA population. In the EA population, “chronic ulcer of skin” with rs13252864 ($P = 0.001$, OR = 2.2) was the most significant association (Fig. 6B and *Dataset S6*). Among EHR-derived phenotypes for respiratory disorders, the rs35128070 eQTL near *DPP4* was associated with “abnormal results of function study of pulmonary system” ($P = 0.002$, OR = 1.6) in the AA population (Fig. 6B and *Dataset S6*).

Discussion

Investigating global patterns of genetic variation at genes that play a role in SARS-CoV-2 infection could provide insights into potential differences in susceptibility to COVID-19 among diverse human populations. However, African populations are underrepresented in the majority of current genetic studies of COVID-19 susceptibility and severity, despite the fact that they have the highest genetic diversity among human populations (44, 45) and have high burdens of infectious disease (44, 45).

Three Nonsynonymous Variants That Are Common and Specific to CAHG at ACE2 Are on Haplotypes with Signatures of Positive Selection. Several studies have investigated patterns of genetic variation at *ACE2*, a receptor of SARS-CoV-2 entry (5, 12–14). None of these studies identified any common coding variation at *ACE2*, suggesting that *ACE2* is evolutionarily conserved. However, these studies did not include an extensive set of African populations. At *ACE2*, we identified 41 nonsynonymous variants, most of which are rare, suggesting that they are under purifying selection. Tests based on dN/dS indicate that East Asians have an excess of nonsynonymous variation at *ACE2*, indicating that weak purifying selection has influenced patterns of variation in that population. However, we identified three common nonsynonymous variants (rs138390800, rs147311723, and rs145437639) at *ACE2* with MAF ranging from 0.083 to 0.164 in CAHGs, which were the only common coding variants (defined here as MAF > 0.05) found in global populations studied here and by others (5, 12–14). We observed that the derived alleles of the common nonsynonymous SNPs (rs138390800, rs147311723, rs145437639) and one putative regulatory variant (rs186029035) at *ACE2* in CAHG show evidence of EHH, with the extended haplotypes extending longer than 2 Mb, although they did not show deviation from neutrality based on the iHS test. However, we do not have much power to detect a selection signal using this test because the SNPs are on three different haplotype backgrounds in CAHG, possibly due to selection on existing variation (e.g., “soft selection”), which decreases the power to detect significant iHS scores. Moreover, each haplotype is at a relatively low frequency (0.083 to 0.164), which further reduces the power of the iHS test. Allele frequencies at two of the putative functional variants are among the most highly differentiated between the CAHG population and other populations, consistent with local adaptation, as indicated by the d_i values of SNPs rs138390800 and rs186029035, which were in the top 1.4 and 1.7%, respectively, of d_i values for all SNPs examined. The CAHGs are traditionally hunter-gatherers living in a rainforest ecosystem who consume wild

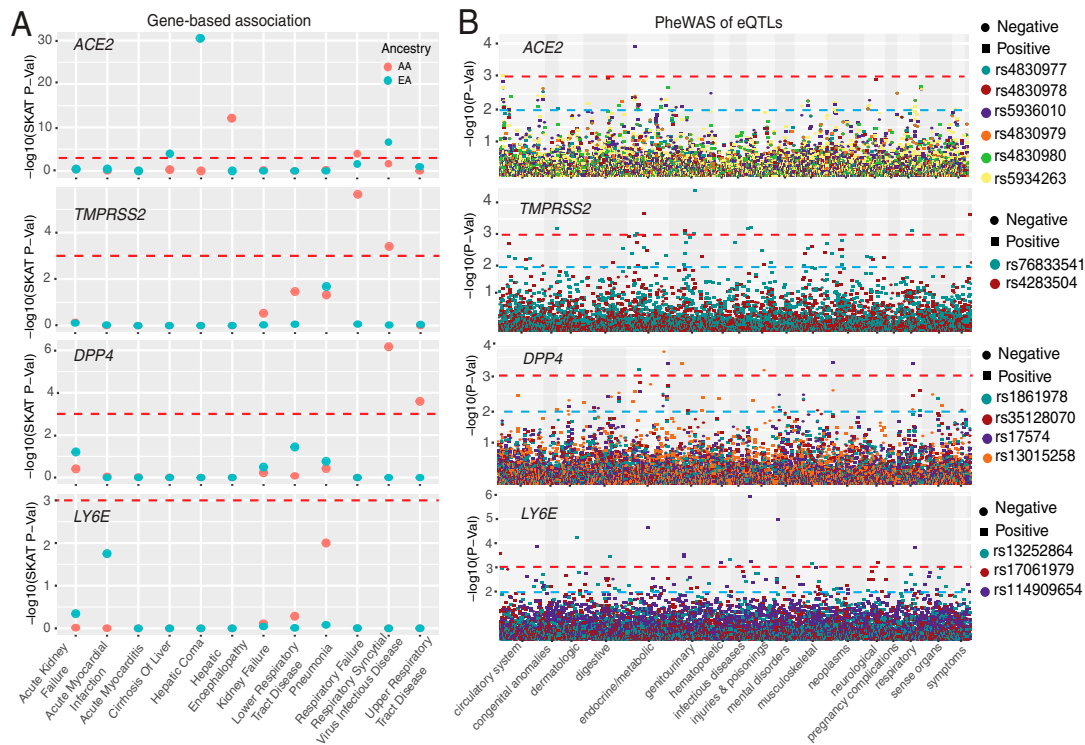


Fig. 6. Associations between genetic variations at four genes and clinical disease phenotypes. (A) Gene-based association result between coding variants at four genes and 12 disease classes. The disease classes are shown on the x axis, and the y axis represents the P values. (B) PheWAS plot of the eQTLs associated with four genes and $\sim 1,800$ disease codes across 17 disease categories. The disease categories are shown on the x axis, and the y axis represents the $-\log_{10}$ of the P values. The colored dots represent eQTLs and the direction of effect of the association. The red dashed lines denote the 0.0001 cutoff, and the blue dashed lines represent the 0.001 cutoff.

animals. They have high exposure to animal viruses and were reported to have relative resistance to viral infection (46). Thus, it is possible that this locus is adaptive for protection from infectious diseases in this population. Future *in vitro* or *in vivo* studies will be needed to determine the functional significance of these variants.

TMPRSS2 Shows Adaptive Evolution in the Human Lineage after Divergence from Chimpanzee. At *TMPRSS2*, we identified 48 nonsynonymous variants, only 2 of which had a high MAF (>0.05) in the pooled global dataset (rs12329760 and rs75603675). However, some variants have high MAF in two African hunter-gatherer populations. Notably, the nonsynonymous variant rs61735795 (Pro375Se) is only common in the Khoesan-speaking San population from Botswana (MAF = 0.18), and the nonsynonymous variant rs367866934 (Leu403-Phe) is only common in the Cameroonian CAHG populations (MAF = 0.15). At *TMPRSS2*, we observed a signature of adaptive evolution in the human lineage after divergence from chimpanzee ~ 6 Mya (47). In total, 13 nonsynonymous variants located on different structural domains of *TMPRSS2* were fixed in human populations. Among them, E441Q and T515M are located in the Peptidase S1 domain that plays an important role in acute respiratory syndrome-like (SARS) coronavirus (SARS-CoV-2) infection (48), and six (A3P, N10S, T46P, A70V, R103C, and M104T) are at the cytoplasmic amino terminal domains of *TMPRSS2*, which plays an important role in signal transduction. By contrast, the coding regions of *DPP4* and *LY6E* are evolutionarily conserved.

eQTLs for Four Genes Vary in Frequency among Populations and Show Signatures of Natural Selection and Associations with Clinical Phenotypes. SARS-CoV replication is significantly reduced in *ACE2* knockout mice (49), and cells with

low expression of *ACE2* were resistant to SARS-CoV-2 infection (50). It has also been shown that both SARS-CoV and SARS-CoV-2 infection could down-regulate *ACE2* expression (8, 49, 51). The expression of *ACE2* and *TMPRSS2* in nasal and bronchial epithelial cells is higher in adults than children and in healthy individuals compared with smokers or patients with chronic obstructive pulmonary disease (51, 52). Therefore, differences in expression levels of *ACE2* and *TMPRSS2* could influence the susceptibility and host reactions to SARS-CoV-2. Previous studies identified eQTLs influencing *ACE2* gene expression, showing differences in allele frequency among different populations (5, 53). For instance, the C allele at the eQTL rs1978124 (53) and the G allele at the eQTL rs4646127 (5), both associated with high *ACE2* expression, are close to 100% frequency in East Asians but are $<80\%$ frequency in other populations. Recently, a rare eQTL (rs190509934, MAF = 0.3%) has been identified that is associated with decreased *ACE2* expression and reduced risk of severe COVID-19 disease (54).

We systematically identified regulatory eQTLs associated with *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* gene expression and highlighted the eQTLs showing highly differentiated MAF among populations and/or signatures of natural selection. Regulatory eQTLs that differ in frequency across ethnically diverse populations may play a role in local adaptation and disease susceptibility (55). These eQTLs are located in chromatin immunoprecipitation sequencing (ChIP-seq) and DNase peaks, and they have the potential to influence transcription factor binding and thus, change the promoter or enhancer activities in specific tissues (56, 57). Notably, some of the eQTLs in the upstream regions of *ACE2* were under selection in African populations. For example, rs5936010 and rs5934263, located within a

Table 1. Associations of *ACE2*, *DPP4*, *TMPRSS2*, and *LY6E* with 12 disease classes derived from EHR data

Disease phenotype	Gene	Cases	Controls	Carrier controls	Carrier cases	SKAT <i>P</i>	Burden <i>P</i>	Burden OR	Burden SE	95% CI	Dataset
Hepatic encephalopathy	<i>ACE2</i>	97	8,045	441	5	1.1E-12	0.0043	5.73	0.61	0.55–2.94	AA
Respiratory syncytial virus infectious disease	<i>DPP4</i>	56	6,392	85	1	6.8E-07	0.1221	6.06	1.17	–0.48–4.09	AA
Respiratory failure	<i>TMPRSS2</i>	199	6,392	11	2	2.3E-06	0.0124	7.31	0.80	0.43–3.55	AA
Respiratory failure	<i>ACE2</i>	199	6,392	351	12	9.0E-05	0.0509	3.10	0.58	0–2.26	AA
Upper respiratory tract disease	<i>DPP4</i>	144	6,392	85	3	2.5E-04	0.0978	4.16	0.86	–0.26–3.11	AA
Respiratory syncytial virus infectious disease	<i>TMPRSS2</i>	56	6,392	11	1	3.9E-04	0.0217	11.63	1.07	0.36–4.55	AA
Hepatic coma	<i>ACE2</i>	16	6,817	318	1	4.3E-31	0.0019	10.45	0.76	0.87–3.83	EA
Respiratory syncytial virus infectious disease	<i>ACE2</i>	40	5,859	274	3	2.3E-07	0.1650	3.61	0.92	–0.53–3.1	EA
Cirrhosis of liver	<i>ACE2</i>	10	6,817	43	1	1.8E-04	0.0837	9.40	1.30	–0.3–4.78	EA

strong enhancer interacting with the promoter of *ACE2* as suggested by ChIA-PET, harbored significant iHS scores ($|iHS| > 2$) in both Afroasiatic populations from Kenya and the San population from Botswana. Further, PheWAS of these eQTLs in the PMBB populations identified association of eQTLs at *ACE2* with type 2 diabetes (rs5936010) and hypertension (rs5934263). These are known preexisting conditions that increase risk of severe illness due to COVID-19 (58–60). Among respiratory diseases, only one eQTL at *ACE2* had nominal association (rs4830977) with acute sinusitis. The association was only identified in the AA population and had a protective effect (OR = 0.78 [0.66 to 0.95]). The eQTLs we analyzed are from the GTEx V8 database (61), and 84.6% of the donors are people of European and western Eurasian descent. Therefore, it is possible that we are missing some regulatory variants that are only present in specific ancestry groups due to the lack of sample diversity. Further experimental testing of predicted regulatory variants will provide insights into differences in gene expression regulation at *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* among different populations.

***ACE2* and *TMPRSS2* Are Significantly Associated with Respiratory, Cardiac, and Blood Phenotypes in the PMBB Dataset.** The gene-based genetic association analyses of nonsynonymous variants at *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* identified several associations with clinical phenotypes. We observed that respiratory failure has significant association with *ACE2* and *TMPRSS2* among the PMBB AA population. That is a particularly interesting finding as respiratory failure is one of the clinical outcomes observed in some patients with COVID-19 (39–43). However, this association was not significant in the EA population. This observation could be explained by the low number of coding variants and carriers at *ACE2* and *TMPRSS2* among EA and hence, low power to detect an association. An association with myocarditis, a rare cardiovascular disease caused by viral infection, was also observed in the AA population. Recent studies have reported a link between SARS-CoV-2–induced cardiac injury, such as myocarditis, among COVID-19 patients (62). Further, *ACE2* has known expression in heart tissue, and it plays an important role in transcriptional dysregulation in cardiomyocytes—cells that make up cardiac muscles (36–38). We observed an association between *ACE2*

and myocarditis only in the AA population, but as noted above, we may not have as much power to detect an association in EA. Blood clotting abnormalities in lungs and other organs in COVID-19 patients have been reported by several studies (63). In autopsies of COVID-19 patients, thrombosis was found to be a prominent finding across multiple organs, even despite extensive anticoagulation treatment and regardless of the timing of clinical progression, indicating that thrombosis might be at play in the early stages of disease (63). One hypothesis to explain this observation is that the dysfunction of endothelial cells may play an important role in increased risk of thrombosis (64). We observed associations between the internationalized normalized ratio (INR) derived from the prothrombin time test with *ACE2* and *LY6E* in a gene-based association test. The INR test measures the time it takes blood to clot and is an important measure for individuals with blood clotting disorders or on blood thinners.

Characterizing the genetic variation and clinical phenotype associations at these four genes that play a key role in SARS-CoV-2 infection could be relevant for understanding individual differences in infection susceptibility. We performed evolutionary analyses to dissect the forces underlying global patterns of genetic variation and identified variants that may be targets of selection. It will be important to determine the functional effects of these candidate adaptive variants using *in vitro* and *in vivo* approaches in future studies. Additional studies will be needed to investigate the impact of genetic variation in modulating susceptibility/resistance to SARS-CoV-2 infection and other coronaviruses across ethnically diverse populations.

Materials and Methods

Genomic Data and Populations. The genomic data used in this study were from three sources: the Africa 6K project (referred to as the African diversity dataset), which is part of the Trans-Omics in Precision Medicine consortium (65); the 1KG (66); and the PMBB. From the Africa 6K project, a subset of 2,012 high-coverage (>30×) whole-genome sequences of ethnically diverse African populations (SI Appendix, Fig. S1) was included (SI Appendix). Institutional review board (IRB) approval was obtained from the University of Maryland and the University of Pennsylvania. Written informed consent was obtained from all participants, and research/ethics approval and permits were obtained from the following institutions prior to sample collection: Tanzania Commission for Science and Technology,

National Institute for Medical Research, and Muhimbili University of Health and Allied Sciences in Dar es Salaam, Tanzania; the University of Botswana and the Ministry of Health in Gaborone, Botswana; the University of Addis Ababa and the Federal Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee; and the Cameroonian National Ethics Committee and the Cameroonian Ministry of Public Health. The PMBB participants were recruited through the University of Pennsylvania Health System by enrolling at the time of clinic visit. Patients participated by donating either blood or a tissue sample and allowing researchers access to their EHR information, and all participants provided written informed consent. The PMBB is approved under IRB protocol 813913.

Variant Annotation. We used Ensembl Variant Effect Predictor (VEP) for variant annotations (67) (*SI Appendix*). For gene-based association analysis using the PMBB dataset, we collapsed all the predicted nonsynonymous variants with rare exome variant ensemble learner (REVEL) score > 0.5 and pLOFs with MAF < 0.01 . We assigned variants as pLOFs if the variant was annotated by VEP as start_lost, splice_donor_variant, splice_acceptor_variant, frameshift_variant, stop_gained, and stop_lost. All genome coordinates followed the GRCh38 assembly.

Characterization of Putative Regulatory Variation. We extracted variants located within a ± 10 -kb distance to their TSS as well as enhancers supported by RNA Pol2 ChIA-PET data from ENCODE (68). These variants were further filtered by overlapping with DNase I hypersensitive sites sequencing (DNase-seq) and ChIP-seq peaks from Roadmap (69), ENCODE (68), and Remap2 (70) or overlapping with significant single-tissue eQTLs (P value < 0.001) from the GTEx V8 database (6).

EHR Phenotypes and Association Testing. We focused on the phenotypes characterized as primary organ dysfunctions in the early studies on COVID-19 (Dataset S5 and *SI Appendix*). Broadly, we centered our analyses on these four broad clinical conditions/phenotypes: respiratory injury/failure, acute liver injury/failure, acute cardiac injury/failure, and acute kidney injury/failure. These disease classes are well characterized in human disease ontologies, such as Monarch Disease Ontology (*SI Appendix*).

We used the R SKAT package for conducting a gene-based dispersion test and Biobin for gene burden analysis (32, 71, 72). Here, multiple genetic variations in a gene region were collapsed to generate a gene burden/dispersion score, and regression methods were used to test for association between the genetic score and a phenotype or trait (*SI Appendix*).

Structural Analysis of Nonsynonymous Variations on the ACE2-S Protein Binding Interface. We determined the three-dimensional protein location of all nonsynonymous coding variants identified in this study using experimentally determined structures of the ACE2 protein complexed with the RBD of SARS-CoV-2 spike glycoprotein based on cryoelectron microscopy available in the Protein Data Bank [ID code 6M17 (73)]. All structural analysis and figures were prepared using VMD (74) (*SI Appendix*).

Detecting Signatures of Natural Selection. We used two methods [the McDonald-Kreitman test (24) and the dN/dS test (23)] to test for signals of selection acting on the four candidate genes over long timescales and two methods [EHH (28) and iHS (27)] to detect recent (e.g., last $\sim 10,000$ y before present) signatures of positive selection (*SI Appendix*). We used d_i statistics to identify SNPs that are highly differentiated in allele frequency between populations based on unbiased estimates of pairwise F_{ST} (30) (*SI Appendix*). Haplotype networks were constructed by PopART (75) using the built-in minimum spanning algorithm.

Description of Supplemental Data. The supplemental data include *SI Appendix, SI Methods, Figs. S1–S30, and Table S1*, the legends of Datasets S1–S6 and the Regeneron Genetic Center authors and contribution statements.

Data Availability. All data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program were supported by the National Heart, Lung and Blood Institute (NHLBI). Core support, including centralized genomic read mapping

and genotype calling, along with variant quality metrics and filtering were provided by TOPMed Informatics Research Center Grant 3R01HL117626-02S1 (Contract HHSN2682018000021). Core support, including phenotype harmonization, data management, sample-identity quality control, and general program coordination, were provided by NIH funded TOPMed Data Coordinating Center Grants R01HL120393 and U01HL120393 (Contract HHSN2682018000011). Funding for this study was provided by ADA 1-19-VSN-02 (to S.A.T.) and by NIH grants R01LM010098 (to S.M.W.), X01HL139409 (to S.A.T.), 1R35GM134957 (to S.A.T.), R01GM113657 (to S.A.T.), R01DK104339 (to S.A.T.), and R01AR076241 (to S.A.T.). A.V. is supported by NIH grant R01GM138597. C.F.-N. holds a Presidential Professorship at the University of Pennsylvania, is a recipient of the Langer Prize by the American Institute of Chemical Engineers Foundation and acknowledges funding from NIH (R35GM138201), the Defense Threat Reduction Agency (DTRA; HDTRA11810041 and HDTRA1-21-1-0014) and from a Health-Tech Accelerator Award, the Nemirovsky Prize, the Institute for Diabetes, Obesity, and Metabolism, the Mental Health AIDS Research Center and the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the NIH under CTSA award number UL1TR001878. Genome Sequencing for "NHLBI TOPMed: Integrative Genomic Studies of Heart and Blood Related Traits in Africans (Africa6K)" was performed at Broad Genomics (hhsn268201600034i). We acknowledge the studies and participants who provided biological samples and data for TOPMed and PMBB. We thank Alex Harris, Alexander Platt, and Srilakshmi Raj for discussing the evolutionary analysis in the paper. We thank the following individuals and organizations for their assistance in collecting samples for this project: Kenya: Lilian A. Nyndodo, Eva Aluvalla, Daniel Kariuki, Fathya Abdo, and Hussein Musa; Ethiopia: Solomon Taye, Birhanu Mekauntie, and Alemayehu Moges; Tanzania: Kweli Powell, Holly Mortensen, Mariki Euphrasia, Ruth Maviyas, John G. Memra, Holliness Santa, Emanuel Kimario, and Reginald Kavishe; Botswana: Michael Campbell, Ari Ho-Foster, Maitseo M. M. Bolane, Maungo Moswang, Gaolape Mpoloka, Kingsley Motshegwe, and Mthusi Molathegi; and Cameroon: Eric Mbuwne, Sali Django, Dickson Ndizi, Valentine Ngum Ndze, Julius Fonsah, Eric Ngwang, Grace N. Tenjei, Meagan Rubel, Peter Kfu, Association Culturelle pour le Développement Bagyeli/Bakola de l'Océan, Centre d'Action pour le Développement Durable des Autochtones Pygmées, and Mbororo Social and Cultural Development Association. We especially thank all African participants for their important contributions to this study.

Author affiliations: ^aDepartment of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; ^bDivision of Translational Medicine and Human Genetics, Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA 19104; ^cMachine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; ^dDepartments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104; ^ePenn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA 19104; ^fDepartment of Biological Sciences, University of Southern California, Los Angeles, CA 90089; ^gDepartment of Medicine, Vanderbilt University, Nashville, TN 37232; ^hBiological Sciences, University of Botswana, Gaborone, Botswana; ⁱFaculty of Medicine, University of Botswana, Gaborone, Botswana; ^jRegeneron Genetics Center, Tarrytown, NY 10591; ^kDepartment of Biochemistry, Kampala International University in Tanzania, Dar es Salaam, Tanzania; ^lDepartment of Microbial Cellular and Molecular Biology, Addis Ababa University, Addis Ababa, Ethiopia; ^mDepartment of Pharmacotoxicology and Pharmacokinetics, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, Yaoundé, Cameroon; ⁿDepartment of Neurology, Central Hospital Yaoundé, Yaoundé, Cameroon; ^oBrain Research Africa Initiative, Neuroscience Laboratory, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, Yaoundé, Cameroon; ^pCenter for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya; ^qDepartment of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106; ^rDepartment of Biology, University of Pennsylvania, Philadelphia, PA 19104; and ^sCenter for Global Genomics and Health Equity, University of Pennsylvania, Philadelphia, PA 19104

Author contributions: C.Z., A.V., Y.F., G.S., and S.A.T. designed research; C.Z., A.V., and Y.F. performed research; C.Z., A.V., Y.F., M.C.R.M., M.M., and C.F.-N. analyzed data; M.H., A.L., J.P., S.T., M.A.R., M.C.C., W.B., J.H., S.W.M., G.G.M., R.G.C., T.N., D.W.M., G.B., C.F., A.K.N., S.A.O., S.M.W., D.J.R., M.D.R., and S.A.T. contributed samples and data to the study; and C.Z., A.V., Y.F., M.M., G.S., and S.A.T. wrote the paper.

Reviewers: R.K., Temple University; and X.L., Chinese Academy of Sciences Shanghai Institute of Nutrition and Health.

1. C. W. Yancy, COVID-19 and African Americans. *JAMA* **323**, 1891–1892 (2020).
2. D. J. Alencor, Racial disparities-associated COVID-19 mortality among minority populations in the US. *J. Clin. Med.* **9**, 2442 (2020).
3. M. Hoffmann *et al.*, SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
4. A. C. Walls *et al.*, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
5. Y. Cao *et al.*, Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, 11 (2020).
6. GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
7. F. Silhol, G. Sarlon, J. C. Deharo, B. Vaisse, Downregulation of ACE2 induces overstimulation of the renin-angiotensin system in COVID-19: Should we block the renin-angiotensin system? *Hypertens. Res.* **43**, 854–856 (2020).
8. I. Glowacka *et al.*, Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response. *J. Virol.* **85**, 4122–4134 (2011).
9. E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: Recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
10. N. Vankadari, J. A. Wilce, Emerging Wuhan (COVID-19) coronavirus: Glycan shield and structure prediction of spike glycoprotein and its interaction with human CD26. *Emerg. Microbes Infect.* **9**, 601–604 (2020).
11. S. Pfander *et al.*, LY6E impairs coronavirus fusion and confers immune control of viral disease. *Nat. Microbiol.* **5**, 1330–1339 (2020).
12. Y. Hou *et al.*, New insights into genetic susceptibility of COVID-19: An ACE2 and TMPRSS2 polymorphism analysis. *BMC Med.* **18**, 216 (2020).
13. E. Benetti *et al.*, ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *Eur. J. Hum. Genet.* **28**, 1602–1614 (2020).
14. E. T. Cirulli, S. Riffle, A. Bolze, N. L. Washington, Revealing variants in SARS-CoV-2 interaction domain of ACE2 and loss of function intolerance through analysis of >200,000 exomes. *bioRxiv* [Preprint] (2020). <https://doi.org/10.1101/2020.04.07.030544> (Accessed 4 May 2020).
15. C. Zhang, M. E. B. Hansen, S. A. Tishkoff, Advances in integrative African genomics. *Trends Genet.* **38**, 152–168 (2022).
16. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
17. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
18. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **7**, 7.20 (2013).
19. A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011).
20. A. Heurich *et al.*, TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by TMPRSS2 augments entry driven by the severe acute respiratory syndrome coronavirus spike protein. *J. Virol.* **88**, 1293–1307 (2014).
21. B. S. Gloss, M. E. Dinger, Realizing the significance of noncoding functionality in clinical genomics. *Exp. Mol. Med.* **50**, 1–8 (2018).
22. G. Duggal, H. Wang, C. Kingsford, Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.* **42**, 87–96 (2014).
23. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
24. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
25. P. C. Sabeti *et al.*, Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
26. S. Kryazhinskiy, J. B. Plotkin, The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
27. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
28. P. C. Sabeti *et al.*, International HapMap Consortium, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
29. L. B. Scheinfeldt, S. A. Tishkoff, Recent human adaptation: Genomic approaches, interpretation and insights. *Nat. Rev. Genet.* **14**, 692–702 (2013).
30. J. M. Akey *et al.*, Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1160–1165 (2010).
31. J. D. Hooper, J. A. Clements, J. P. Quigley, T. M. Antalis, Type II transmembrane serine proteases. Insights into an emerging class of cell surface proteolytic enzymes. *J. Biol. Chem.* **276**, 857–860 (2001).
32. C. B. Moore, J. R. Wallace, A. T. Frase, S. A. Pendergrass, M. D. Ritchie, BioBin: A bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Med. Genomics* **6** (suppl. 2), S6 (2013).
33. S. Lee *et al.*; NHLBI GO Exome Sequencing Project–ESP Lung Project Team, Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
34. S. S. Verma *et al.*, Rare variants in drug target genes contributing to complex diseases, phenome-wide. *Sci. Rep.* **8**, 4624 (2018).
35. T. K. Boehmer *et al.*, Association between COVID-19 and myocarditis using hospital-based administrative data – United States, March “369”/2020–January 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 1228–1232 (2021).
36. B. Siripanthong *et al.*, Recognizing COVID-19-related myocarditis: The possible pathophysiology and proposed guideline for diagnosis and management. *Heart Rhythm* **17**, 1463–1471 (2020).
37. N. R. Tucker *et al.*; Human Cell Atlas Lung Biological Network; Human Cell Atlas Lung Biological Network Consortium Members, Myocyte-specific upregulation of ACE2 in cardiovascular disease: Implications for SARS-CoV-2-mediated myocarditis. *Circulation* **142**, 708–710 (2020).
38. S. Shi *et al.*, Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA Cardiol.* **5**, 802–810 (2020).
39. D. Wang *et al.*, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
40. N. Chen *et al.*, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020).
41. G. Grasselli *et al.*, Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. *JAMA* **323**, 1574–1581 (2020).
42. M. Arentz *et al.*, Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. *JAMA* **323**, 1612–1614 (2020).
43. C. Huang *et al.*, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
44. S. A. Tishkoff *et al.*, The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
45. G. Sirugo, S. M. Williams, S. A. Tishkoff, The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
46. G. H. Perry *et al.*, Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3596–E3603 (2014).
47. G. V. Glazko, M. Nei, Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424–434 (2003).
48. A. David, T. Khanna, M. Beykou, G. Hanna, M. J. E. Sternberg, Structure, function and variants analysis of the androgen-regulated *TMPRSS2*, a drug target candidate for COVID-19 infection. *bioRxiv* [Preprint] (2020) <https://doi.org/10.1101/2020.05.26.116608> (Accessed 26 May 2020).
49. K. Kuba *et al.*, A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. *Nat. Med.* **11**, 875–879 (2005).
50. M. Letko, A. Marzi, V. Munster, Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
51. P. Verdecchia, C. Cavallini, A. Spanevello, F. Angeli, The pivotal link between ACE2 deficiency and SARS-CoV-2 infection. *Eur. J. Intern. Med.* **76**, 14–20 (2020).
52. S. Bunyavanich, A. Do, A. Vicencio, Nasal gene expression of angiotensin-converting enzyme 2 in children and adults. *JAMA* **323**, 2427–2429 (2020).
53. J. Chen *et al.*, Individual variation of the SARS-CoV-2 receptor ACE2 gene expression and regulation. *Aging Cell* **19**, e13168 (2020).
54. J. E. Horowitz *et al.*; Regeneron Genetics Center, Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nat. Genet.*, 10.1038/s41588-021-01006-7 (2022).
55. L. S. Mogil *et al.*, Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* **14**, e1007586 (2018).
56. J. F. Degner *et al.*, DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
57. C. D. Brown, L. M. Mangravite, B. E. Engelhardt, Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
58. Z. Wu, J. M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* **323**, 1239–1242 (2020).
59. Y. Zhang *et al.*; medical team from Xiangya Hospital to support Hubei, China, Association of diabetes mellitus with disease severity and prognosis in COVID-19: A retrospective cohort study. *Diabetes Res. Clin. Pract.* **165**, 108227 (2020).
60. M. Apicella *et al.*, COVID-19 in people with diabetes: Understanding the reasons for worse outcomes. *Lancet Diabetes Endocrinol.* **8**, 782–792 (2020).
61. 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–RPIC; 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; Genome Browser Data Integration & Visualization–UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group; A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
62. K. U. Knowlton, Pathogenesis of SARS-CoV-2 induced cardiac injury from the perspective of the virus. *J. Mol. Cell. Cardiol.* **147**, 12–17 (2020).
63. S. Biswas *et al.*, Blood clots in COVID-19 patients: Simplifying the curious mystery. *Med. Hypotheses* **146**, 110371 (2021).
64. A. V. Rapkiewicz *et al.*, Megakaryocytes and platelet-fibrin thrombi characterize multi-organ thrombosis at autopsy in COVID-19: A case series. *EClinicalMedicine* **24**, 100434 (2020).
65. D. Taliun *et al.*, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
66. A. Auton *et al.*; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
67. W. McLaren *et al.*, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
68. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
69. L. H. Chadwick, The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317–324 (2012).
70. J. Chènèby *et al.*, ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* **48**, D180–D188 (2020).
71. A. O. Basile *et al.*, Knowledge driven binning and phewas analysis in marshfield personalized medicine research project using Biobin. *Pac. Symp. Biocomput.* **21**, 249–260 (2016).
72. M. C. Wu *et al.*, Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
73. R. Yan *et al.*, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
74. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
75. J. W. Leigh, D. Bryant, POPART: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).