

1 **Global patterns of genetic variation and association with clinical phenotypes at genes**
2 **involved in SARS-CoV-2 infection.**

3 Chao Zhang^{1¶}, Anurag Verma^{1¶}, Yuanqing Feng^{1¶}, Marcelo C. R. Melo², Michael McQuillan¹,
4 Matthew Hansen¹, Anastasia Lucas¹, Joseph Park¹, Alessia Ranciaro¹, Simon Thompson¹, Meghan
5 A. Rubel¹, Michael C. Campbell³, William Beggs¹, Jibril Hirbo⁴, Sununguko Wata Mpoloka⁵,
6 Gaonyadiwe George Mokone⁶, Regeneron Genetic Center⁷, Thomas Nyambo⁸, Dawit Wolde
7 Meskel⁹, Gurja Belay⁹, Charles Fokunang¹⁰, Alfred K. Njamnshi¹¹, Sabah A. Omar¹², Scott M.
8 Williams¹³, Daniel Rader¹, Marylyn D. Ritchie¹, Cesar de la Fuente Nunez², Giorgio Sirugo^{14*},
9 Sarah Tishkoff^{1,15*}.

10 ¹ *Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA*
11 *19104, USA.*

12 ² *Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical*
13 *Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, Penn*
14 *Institute for Computational Science, and Departments of Bioengineering and Chemical and Biomolecular*
15 *Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA*
16 *19104, USA.*

17 ³ *Department of Biology, Howard University, Washington DC 20059, USA*

18 ⁴ *Department of Medicine, Vanderbilt University*

19 ⁵ *University of Botswana, Biological Sciences, Gaborone, Botswana*

20 ⁶ *University of Botswana, Faculty of Medicine, Gaborone, Botswana*

21 ⁷ *Regeneron Genetics Center, Tarrytown, NY 10591, USA*

22 ⁸ *Department of Biochemistry, Kampala International University in Tanzania, Dar es Salaam, Tanzania*

23 ⁹ *Addis Ababa University Department of Microbial Cellular and Molecular Biology, Addis Ababa,*
24 *Ethiopia*

25 ¹⁰ *Department of Pharmacotoxicology and Pharmacokinetics, Faculty of Medicine and Biomedical*
26 *Sciences, The University of Yaoundé I, Yaoundé, Cameroon*

27 ¹¹ *Department of Neurology, Central Hospital Yaoundé; Brain Research Africa Initiative (BRAIN),*
28 *Neuroscience Lab, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, Yaoundé,*
29 *Cameroon*

30 ¹² *Center for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi,*
31 *Kenya*

32 ¹³ *Case Western Reserve University, Cleveland, OH*

33 ¹⁴ *Division of Translational Medicine and Human Genetics, University of Pennsylvania School of*
34 *Medicine, Philadelphia, PA 19104*

35 ¹⁵ *Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA*

36
37

38 **Joint Authorship**

39

40 ¶These authors contributed equally to this work

41

42 **Corresponding Authors**

43 *Correspondence: giorgio.sirugo@penntmedicine.upenn.edu, tishkoff@penntmedicine.upenn.edu

44

1 **Abstract**

2

3 The COVID-19 pandemic caused by SARS-COV-2 has had a devastating impact on population
4 health. We investigated global patterns of genetic variation and signatures of natural selection at
5 host genes relevant to SARS-CoV-2 infection (*ACE2*, *TMPRSS2*, *DPP4*, and *LY6E*). We analyzed
6 novel data from 2,012 ethnically diverse Africans, 15,997 individuals of European (7,061) and
7 African (8,916) ancestry recruited by the Penn Medicine BioBank (PMBB), and comparative data
8 from 2,504 individuals from the 1000 Genomes project. At *ACE2* we identified 41 non-
9 synonymous variants, found to be at low frequency in most populations. However, three non-
10 synonymous variants were frequent among Central African hunter-gatherers (CAHG) from
11 Cameroon, and signatures of positive selection could be detected on haplotypes encompassing
12 those variants. We also detected signatures of positive selection for variants at regulatory regions
13 upstream of *ACE2* in diverse African populations. At *TMPRSS2*, we identified 48 non-
14 synonymous variants, several of which are common in global populations, and 13 amino acid
15 changes that are fixed in the human lineage after divergence from Chimpanzee. At *DPP4* and
16 *LY6E* most variants were rare in global populations indicating that purifying selection is acting at
17 these loci. At all four loci, we identified common non-coding variants associated with gene
18 expression that vary in frequency across global populations. By analyzing electronic health records
19 from the PMBB we discovered genetic associations with clinical phenotypes, such as respiratory
20 failure with *ACE2* and upper respiratory tract infection with *DPP4*. Our study provides new
21 insights into global variation at genes potentially affecting susceptibility to SARS-CoV-2
22 infection.

23

24 **Keywords:** SARS-COV-2; COVID-19; genetic variation; genetic association; global populations;
25 Africans; natural selection; *ACE2*; *TMPRSS2*; *DPP4*; *LY6E*

26

27

28

29

30

31

32

33

34

1 Introduction

2
3 Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome
4 coronavirus 2 (SARS-CoV-2). Coronaviruses are enveloped, positive-sense, and single-stranded
5 RNA viruses, many of which are zoonotic pathogens that crossed over into humans. Seven
6 coronavirus species, including SARS-CoV-2, have been discovered that, depending on the virus
7 and host physiological condition, may cause mild or lethal respiratory disease. The novel SARS-
8 CoV-2 virus was initially identified in Wuhan, China, in December 2019¹, and due to high
9 transmission rates, including from asymptomatic subjects², quickly spread globally causing a
10 pandemic of historic proportions. In the US, the crude fatality rate of COVID-19 is ~ 1%, and
11 mortality increases significantly with age, with 70% of deaths being among individuals 70 years
12 old and above^{3, 4}. As is the case with other infectious diseases, COVID-19 progression appears to
13 exhibit sexual-dimorphism, with fatality rates 2-fold greater for men than women⁵. Patients with
14 COVID-19 can be clinically subdivided into three categories: asymptomatic/mild, severe (with
15 dyspnea, hypoxia), and critical (with respiratory failure, shock, or multiorgan dysfunction). The
16 rate of asymptomatic infection of SARS-CoV-2 may be as high as 40-45%⁶, and those who are
17 asymptomatic are unlikely to convert to acute symptoms even though they may transmit virus for
18 up to 2 weeks. Symptomatic patients may present dry cough, followed by sputum, hyposmia, nasal
19 congestion, nausea, diarrhea, fever and dyspnea, although initial presentation is known to be
20 variable (for example fever or dyspnea may be absent at admission in hospital)⁷. There is
21 considerable variation in disease prevalence and severity across populations and communities. For
22 example, in Chicago, more than 50% of COVID-19 cases and nearly 70% of COVID-19 deaths
23 are in African Americans (who make up 30% of the population of Chicago)⁸. More generally,
24 minority populations in the US appear to have been disproportionately affected by COVID-19;^{8, 9}.
25 In addition, adverse outcomes including death, have been associated with underlying
26 cardiometabolic comorbidities (e.g., hypertension, diabetes, cardiovascular disease, chronic
27 kidney disease)¹⁰⁻¹³. Liver impairment is common in patients with COVID-19, and elevated
28 alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels are relatively
29 frequent at presentation². The extent to which pre-existing chronic liver conditions affect COVID-
30 19 related complications remains to be elucidated. Smell and taste sensations as well as increased
31 incidence of ischemic stroke have been observed in individuals with COVID-19¹⁴⁻¹⁷.

1 Several host genes play a role in SARS-CoV-2 infection¹⁸. The *ACE2* gene, encoding the
2 angiotensin-converting enzyme-2 protein, was reported to be a main binding site for SARS-CoV
3 during an outbreak in 2003, and evidence showed stronger binding affinity to SARS-CoV-2, which
4 enters the target cells via ACE2 receptors^{18; 19}. The *ACE2* gene is located on the X chromosome,
5 its expression level varies among populations²⁰, and it is ubiquitously expressed in the lung, blood
6 vessels, gut, kidney, testis, and brain, all organs that appear to be affected as part of the COVID-
7 19 clinical spectrum. SARS-CoV-2 infects cells through a membrane fusion mechanism, which,
8 in the case of SARS-CoV, is known to induce down-regulation of *ACE2*²¹. Such down-regulation
9 has been shown to cause inefficient counteraction of angiotensin II effects, leading to enhanced
10 pulmonary inflammation and intravascular coagulation²¹. Additionally, altered expression of
11 *ACE2* has been associated with cardiovascular and cerebrovascular disease, which is highly
12 relevant to COVID-19 as several cardiovascular conditions are associated with severe disease.
13 Type II transmembrane serine protease (*TMPRSS2*), located on the outer membrane of host target
14 cells, binds to and cleaves ACE2, resulting in activation of spike proteins on the viral envelope,
15 and facilitating membrane fusion and endocytosis²². Two additional genes, dipeptidyl peptidase
16 (*DPP4*), and lymphocyte antigen 6 complex locus E (*LY6E*), have been shown to play an important
17 role in the entry of SARS-CoV2 virus into host cells. *DPP4* is a known functional receptor for the
18 Middle East Respiratory Syndrome coronavirus (MERS-CoV), causing a severe respiratory illness
19 with high mortality²³. Lastly, *LY6E* (lymphocyte antigen 6 complex, locus E) encodes a
20 glycosylphosphatidylinositol (GPI)-anchored cell surface protein which is a critical antiviral
21 immune effector that controls coronavirus infection and pathogenesis²⁴. Mice lacking *LY6E* in
22 hematopoietic cells were susceptible to murine coronavirus infection²⁴.

23 In this study, we characterized genetic variation at *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* in
24 ethnically diverse human populations by analyzing 2,012 novel genomes from ethnically diverse
25 Africans (referred to as the “African Diversity” dataset), 2,504 genomes from the 1000 Genomes
26 project, and whole exome sequencing of 15,997 individuals of European and African ancestry
27 from the Penn Medicine BioBank (PMBB) dataset. The African diversity dataset includes
28 populations with diverse subsistence patterns (hunter-gatherers, pastoralists, agriculturalists) and
29 speaking languages belonging to the four major language families in Africa (Khoesan, Niger-
30 Congo (of which Bantu is the largest subfamily), Afroasiatic, and Nilo-Saharan). We identify
31 functionally relevant variation, compare the patterns of variation across global populations, and

1 provide insight into the evolutionary forces underlying these patterns of genetic variation. In
2 addition, we perform an association study using the variants identified from whole-exome
3 sequencing at the four genes (*ACE2*, *TMPRSS2*, *DPP4*, and *LY6E*) and clinical traits derived from
4 electronic health record (EHR) data linked to the subjects enrolled in the Penn Medicine BioBank
5 (PMBB). The EHR data includes diseases related to organ dysfunctions associated with severe
6 COVID-19 such as respiratory, cardiovascular, liver and renal complications. Our study of genetic
7 variation in SARS-CoV-2 receptors and their partners provides novel data to investigate infection
8 susceptibility within and between populations and indicates that variants in these genes may play
9 a role in comorbidities relevant to COVID-19 severity.

10 **Material and Methods**

11

12 *Genomic data*

13 The genomic data used in this study were from three sources: the Africa 6K project
14 (referred to as the the “African Diversity” dataset) which is part of the TopMed consortium²⁵, the
15 1000 Genomes project (1KG)²⁶, and the Penn Medicine BioBank (PMBB). From the Africa 6K
16 project, a subset of 2012 high coverage (>30X) whole genome sequences of ethnically diverse
17 African populations (Figure S1) were included. The African samples were collected from
18 individuals from five countries (Cameroon, Ethiopia, Kenya, Botswana and Tanzania), speak
19 languages belonging to four different language families spoken in Africa (Afroasiatic, Nilo-
20 Saharan, Niger-Congo, and Khoesan) and have diverse subsistence practices (*e.g.*, hunter-
21 gatherers, agriculturalists, and pastoralists). IRB approval was obtained from the University of
22 Maryland and the University of Pennsylvania. Written informed consent was obtained from all
23 participants and research/ethics approval and permits were obtained from the following institutions
24 prior to sample collection: COSTECH, NIMR and Muhimbili University of Health and Allied
25 Sciences in Dar es Salaam, Tanzania; the University of Botswana and the Ministry of Health in
26 Gaborone, Botswana; the University of Addis Ababa and the Federal Democratic Republic of
27 Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee;
28 and the Cameroonian National Ethics Committee and the Cameroonian Ministry of Public Health.
29 Whole genome sequencing (WGS) was performed to a median depth of 30X using DNA isolated
30 from blood, PCR-free library construction and Illumina HiSeq X technology, as described

1 elsewhere²⁵. In the 1KG data set, 2504 genome sequences from phase 3²⁶ were included in our
2 analysis.

3 The PMBB participants were recruited through the University of Pennsylvania Health
4 System by enrolling at the time of clinic visit. Patients participate by donating either blood or a
5 tissue sample and allowing researchers access to their EHR information. This academic biobank
6 has DNA extracted from blood that has been genotyped using an Illumina Infinium Global
7 Screening Array-24 Kit *version 2* and whole exome sequencing (WES) using the IDT xgen exome
8 research panel v1.0. The study cohort consisted of 15,977 individuals total, with 7,061 of European
9 ancestry (EA) and 8,916 of African ancestry (AA) (Table S1). Genetic ancestry of these samples
10 was determined by performing quantitative discriminant analyses (QDA) on eigenvectors. The
11 1000 Genomes datasets with super population ancestry labels (EUR, AFR, EAS, SAS, Other) were
12 used as QDA training datasets to determine the genetic ancestry labels for the PMBB population.
13 We identified and removed 117 related individuals using a kinship coefficient of 0.25.

14

15 *Variant annotations*

16 We used Ensembl Variant Effect Predictor (VEP) for variant annotations²⁷. VEP classifies
17 variants into 36 types including non-synonymous, synonymous, and stop loss variants. For
18 pathogenicity predictions, we used CADD²⁸, SIFT²⁹, PolyPhen³⁰, Condel³¹, and REVEL scores
19 in Ensembl. For whole-genome sequencing datasets (African Diversity and 1KG), we annotated
20 genetic variants at *ACE2* (chrX:15,561,033-15,602,158), *TMPRSS2* (chr21:41,464,305-
21 41,531,116), *DPP4* (chr2:161,992,245-162,074,215) and *LY6E* (chr8:143,017,982-143,023,832),
22 and 10 Mb flanking these genes (Table S2). For whole-exome genomes from the PMBB dataset,
23 annotations were restricted to coding regions only. For gene-based association analysis using the
24 PMBB dataset, we collapsed all the predicted non-synonymous variants with REVEL score > 0.5
25 and putative loss of function variants (pLOFs) with MAF < 0.01. We assigned variants as pLoFs
26 if the variant was annotated as stop_lost, missense_variant, start_lost, splice_donor_variant,
27 inframe_deletion, frameshift_variant, splice_acceptor_variant, stop_gained, or inframe_insertion.
28 All genome coordinates followed the GRCh38 assembly.

29

30 *Characterization of putative regulatory variation*

1 We identified regulatory variants likely to impact the target genes. For all four genes (*ACE2*,
2 *TMPRSS2*, *DPP4* or *LY6E*), we extracted the variants located within ± 10 kb distance to their TSS
3 as well as enhancers supported by RNA Pol2 ChIA-PET data from ENCODE³². These variants
4 were further filtered by overlapping with DNase-seq and CHIP-seq peaks from Roadmap³³,
5 ENCODE³², Remap2³⁴; or overlapping with significant single-tissue expression quantitative trait
6 locus (eQTLs) (P-value<0.001) from the GTEx V8 database³⁵. We visualized the location of these
7 regulatory and eQTL variants using the UCSC genome browser and highlighted the variants using
8 Adobe Illustrator.

10 *Electronic Health Record Phenotypes*

11 In this analysis, we focused on the phenotypes characterized as primary organ dysfunctions
12 in the early studies on COVID-19. Broadly, we centered our analyses on these four broad clinical
13 conditions/phenotypes: respiratory injury/failure, acute liver injury/failure, acute cardiac
14 injury/failure, and acute kidney injury/failure. These disease classes are well characterized in
15 human disease ontologies such as Monarch Disease Ontology (MONDO). MONDO merges
16 multiple disease resources such as SNOMED, ICD-9, and ICD-10. We leveraged the existing
17 mappings between ICD-9/10 codes (which are how the data are coded in the EHR) and the
18 MONDO disease classes for the conditions described above. We identified 12 MONDO classes
19 that are closely related to four conditions of interest ([Table S1](#)). By using ICD-9 and ICD-10 data
20 from the EHR of the PMBB participants, we mapped the ICD codes to 12 MONDO disease classes.
21 Details on the ICD code mapping to MONDO disease classes are provided in [Table S3](#). Individuals
22 were defined as cases if they had at least one instance of any ICD code mapped to a MONDO
23 disease class or as controls if they had no instance of the code in that disease class. A clinical
24 expert on our team manually reviewed the MONDO and ICD-9/10 mappings.

25 We also used EHR phenotypes defined by groupings of ICD-9 and ICD-10 codes into
26 clinically relevant groups, called phecodes, used in prior PheWAS studies³⁶. Individuals with two
27 or more instances of a phecode were defined as cases, whereas those with no instance of a phecode
28 were defined as controls. Individuals with only one instance were excluded for that phecode. A
29 total of 1860 phecodes were included in the study.

30 Additionally, we extracted data on 34 clinical laboratory measures for PMBB participants
31 from the EHRs. We derived a median value for each laboratory measure based on all clinical tests

1 ever done within the Penn Medicine health system. Any measurement value that falls more than
2 three standard deviations from the normal were labeled as outliers and removed.

3

4 *Association Testing*

5 We used the R SKAT package for conducting a gene-based dispersion test and Biobin^{37; 38}
6 for gene burden analysis. Here, multiple genetic variations in a gene region were collapsed to
7 generate a gene burden/dispersion score and regression methods were used to test for association
8 between the genetic score and a phenotype or trait. We performed three separate burden analysis
9 for 12 MONDO disease classes (Table S3), 1860 phecode, and 34 clinical lab measures. Briefly,
10 the variants annotated as non-synonymous (REVEL score ≥ 0.5) and pLoFs within each of the
11 four candidate genes were collapsed into their respective gene regions (*ACE2*, *TMPRSS2*, *DPP4*
12 and *LY6E*). For both statistical dispersion and burden tests, models were adjusted by the first four
13 principal components of ancestry, sex, and decade of birth. For multiple hypothesis correction, a
14 conservative Bonferroni adjustment was used to derive a significant p-value threshold (p-value <
15 0.0001). We also performed a univariate statistical test for each of the rare variants from these four
16 candidate gene regions to study the effects of each single nucleotide variant (SNV) on the disease
17 phenotype.

18

19 *Structural analysis of nonsynonymous variations on ACE2-S protein binding interface*

20 The fast response from the structural biology community to the COVID-19 pandemic led
21 to the exceptionally fast determination and publication of over 900 as of Jan. 2021
22 (<https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true>)
23 protein structures related to SARS-Cov-2. Using experimentally determined structures of the
24 ACE2 protein complexed with the receptor binding domain (RBD) of SARS-CoV-2 spike
25 glycoprotein, we assessed possible impacts of nonsynonymous coding variants on the ACE2-
26 binding interface with SARS-CoV-2-RBD. Among the multiple entries available in the Protein
27 Data Bank (PDB), we chose to focus on the structure of the full-length human ACE2 bound to
28 RBD (PDB ID 6M17³⁹) determined with Cryo-Electron Microscopy (cryo-EM), as it presented
29 multiple advantages to our study. Unlike other PDB entries that only feature sections of ACE2,
30 usually focusing on the part of the enzymatic domain responsible for RBD binding, 6M17 presents
31 the full length ACE2 in its dimeric form. This allowed us to identify the 3D protein location of all

1 nonsynonymous coding variants identified in this study. Moreover, ACE2 was expressed in a
2 human cell line, maintaining important glycosylation sites and allowing the cryo-EM structure to
3 be used to identify their positions and compositions³⁹. All structural analysis and figures were
4 prepared using VMD⁴⁰.

5

6 *Detecting signatures of natural selection*

7 We used two methods (the McDonald–Kreitman test⁴¹ and the Dn/Ds test⁴²) to test for
8 signals of selection acting on the four candidate genes over long time scales, and two methods
9 (EHH and iHS) to detect recent (e.g. last ~10,000 years before present) signatures of positive
10 selection .

11 For the McDonald–Kreitman test (MK-test)⁴¹, we set up a two-way contingency table to
12 statistically compare the number of nonsynonymous (Dn) and synonymous (Ds) fixed differences
13 between humans and chimpanzees with the number of nonsynonymous (Pn) and synonymous (Ps)
14 polymorphisms among individuals within a population. Based on neutral theory, the ratio of
15 nonsynonymous to synonymous changes should be constant throughout evolutionary time, i.e. the
16 ratio observed among individuals within species (Pn/Ps) should be equal to the ratio observed
17 between species (Dn/Ds). Under a hypothesis of positive selection in the hominin lineage after
18 divergence from our closest ancestor, the chimpanzee, the ratio of nonsynonymous to synonymous
19 variation within species is expected be larger than the ratio of nonsynonymous to synonymous
20 variation between species (i.e. $Dn/Ds > Pn/Ps$). If there is positive diversifying selection among
21 human populations but conservation of fixed differences between species, the ratio of
22 nonsynonymous to synonymous variation between species should be lower than the ratio of
23 nonsynonymous to synonymous variation within species (i.e. $Dn/Ds < Pn/Ps$). The chimpanzee
24 sequence (Clint_PTRv2/panTro6) used in the analysis was obtained from the UCSC genome
25 browser. We used Fisher’s exact test to detect significance of the MK-test. We used transcripts
26 ENST00000252519.8, ENST00000398585.7, ENST00000360534.8, ENST00000521003.5 to
27 calculate Dn, Ds, Pn and Ps for *ACE2*, *TMPRSS2*, *DPP4* and *LY6E*, respectively.

28 We also used the ratio of substitution rates at non-synonymous and synonymous sites
29 (dN/dS) to infer selection pressures on the four candidate genes, as the dN/dS ratio has more power
30 to detect recurrent positive selection⁴³. This measure quantifies selection pressures by comparing
31 the rate of substitutions at synonymous sites (dS), which are neutral or close to neutral, to the rate

1 of substitutions at non-synonymous sites (dN), which are more likely to experience selection. The
2 dN/dS estimation used here follows Nei et al⁴². The number of synonymous sites, s , for codon i in
3 one protein is given by

$$s = \sum_{i=1}^{i=3} f_i$$

4 where f_i is defined as the proportion of synonymous changes at the i th position of a codon. For a
5 sequence of r codons, the total number of synonymous sites, S is given by

$$S = \sum_{j=1}^r s_j$$

7 where s_j is the value of s at the j th codon, and the total number of non-synonymous sites, $N = 3r -$
8 S . The total number of synonymous and non-synonymous differences between two sequences, S_d
9 and N_d respectively, are given by

$$S_d = \sum_{j=1}^r s_{dj}$$

11 and

$$N_d = \sum_{j=1}^r n_{dj}$$

12
13 where s_{dj} and n_{dj} are the numbers of synonymous and non-synonymous differences between two
14 sequences for the j th codon, and r is the number of codons compared. The proportions of
15 synonymous (pS) and non-synonymous (pN) differences are estimated by the equations $pS = S_d /$
16 S and $pN = N_d / N$. The numbers of synonymous (dS) and non-synonymous (dN) substitutions per
17 site are estimated using the Jukes-Cantor formula as below:
18

$$dS = \frac{-3 \ln(1 - \frac{4pS}{3})}{4}$$

19 and

$$dN = \frac{-3 \ln(1 - \frac{4pN}{3})}{4}$$

20
21 In our analysis, for each population, we estimated the total number of synonymous (S_d) and non-
22 synonymous (N_d) differences, and then calculated dN/dS . If dN/dS is larger than one, it suggests
23

1 positive diversifying selection influencing variation at the gene. If dN/dS is less than one it suggests
2 the gene is evolutionary conserved.

3 Genomic regions that have undergone recent positive selection are characterized by
4 extensive linkage disequilibrium (LD) on haplotypes containing the mutation under selection. We
5 used the extended haplotype homozygosity (EHH)⁴⁴ and the integrated Haplotype Score (iHS)
6 methods⁴⁵ to identify regions with extended haplotype homozygosity greater than expected under
7 a neutral model. iHS is based on the differential levels of LD surrounding a positively selected
8 allele compared to the ancestral allele at the same position. For the iHS analyses, we normalized
9 scores with respect to all values observed at sites with a similar derived allele frequency within
10 40Mb regions flanking the four target genes. SNPs with absolute values larger than 2 are within
11 the top 1% of observed values and are marked as extreme SNPs or candidate SNPs under positive
12 selection. An extreme positive iHS score ($iHS > 2$) means that haplotypes on the ancestral allele
13 background are longer compared to the derived allele background. An extreme negative iHS score
14 ($iHS < -2$) means that the haplotypes on the derived allele background are longer compared to the
15 haplotypes associated with the ancestral allele. All of the above processes were performed with
16 selscan⁴⁶. SNPs with predicted functional effects on protein structure that are identified as potential
17 targets of selection (stop_lost, missense_variant, start_lost, splice_donor_variant,
18 inframe_deletion, frameshift_variant, splice_acceptor_variant, stop_gained, or inframe_insertion)
19 are highlighted. Haplotypes were phased by Eagle V2.4.1⁴⁷. The ancestral state of alleles was
20 obtained from Ensembl.

21 To identify potential regulatory variants under selection, we overlapped SNPs showing
22 signatures of selection using iHS with DNase I hypersensitivity peak clusters from ENCODE³²
23 and eQTLs from GTEx v8.³⁵ The overlapped SNPs were uploaded to the UCSC browser for
24 visualization. The ChIP-seq density dataset was obtained from <http://remap.univ-amu.fr/>³³.
25 DNase-seq and ChIP-seq clusters, layered H3K4Me3 (often found near Promoters), H3K4Me1
26 and H3K27Ac (often found near Regulatory Elements) data are from ENCODE³². The DNase-seq
27 tracks of large intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle
28 were from ENCODE⁴⁸.

29 We used d_i statistics to identify SNPs that are highly differentiated in allele frequency
30 between populations based on unbiased estimates of pairwise F_{ST} ⁴⁹. The d_i statistics were
31 performed cross the 40Mb regions. If the candidate SNP was within the top 5% of the 40Mb

1 regions in a specific population, the SNP was considered as a variant showing significant
2 differentiation between the target population and other populations. These variants are candidate
3 SNPs that show signals of local adaptation.

4 Haplotype networks were constructed by PopART⁵⁰ using the built-in minimum spanning
5 algorithm.

6 7 **Results**

8 9 **Coding variation at *ACE2* among global populations**

10
11 SARS-CoV-2 employs *ACE2* as a receptor for cellular entry¹⁸. To systematically
12 characterize genetic variation in the coding region of *ACE2* across global populations, we analyzed
13 whole-genome sequence data from 2,012 individuals from diverse African ethnic groups (referred
14 to as “African diversity panel (ADP)”), 2,504 samples from the 1KG project²⁶, and whole exome
15 sequence data from 15,977 individuals of European and African ancestry from the Penn Medicine
16 Biobank (PMBB) (Figure S1). In total, we identified 41 amino acid changing variants (Figure 1A,
17 and Table S4). Twenty-eight (69%), twenty (49%), eighteen (44%), and sixteen (40%) of the
18 nonsynonymous variants were predicted to be deleterious or likely deleterious by the CADD²⁸,
19 SIFT²⁹, PolyPhen³⁰ and Condel prediction³¹ methods (Table S4).

20 Among the 41 coding variants identified at *ACE2*, the majority are rare (minor allele
21 frequency, MAF < 0.05) in the pooled global population dataset (Figure 1A and Table S4).
22 However, there are variants that are common (MAFs \geq 0.05) in the Central African Hunter
23 Gatherer (CAHG) population from Cameroon (often referred to as “pygmies”) (Figure 1B). One
24 of these variants, rs138390800 (Lys341Arg), is a deleterious non-synonymous variant, and present
25 at high frequency (MAF = 0.164) in the CAHG, while it is rare in other African populations and
26 absent in non-African populations (Figure 1C). Two other nonsynonymous variants, rs147311723
27 (Leu731Phe) (MAF = 0.083) and rs145437639 (Asp597Glu) (MAF = 0.083), are also common
28 only in the CAHG population (Figures 1B and Table S4). These three non-synonymous variants
29 are the only common coding variants found at *ACE2* in any of the populations examined.

30 We then investigated the potential role of these 41 coding variants in the conformation of
31 the *ACE2* protein. The 41 coding variants are distributed across the entire *ACE2* protein (Figure
32 1D and Table S4), including its receptor-binding domain (RBD) region which binds to the SARS-
33 CoV-2 spike protein, dimerization interface, and transmembrane helix. In particular, two novel

1 non-synonymous variants Gly354Asp (chrX:15581230) and Ser43Asn (chrX:15600784) are both
2 found directly in the RBD binding region of *ACE2* (Figure 1D and Table S4); the former is only
3 found in low frequency in one population, the Fulani from Cameroon (MAF = 0.008), and the
4 latter is also an African specific variant that is at low frequency in only three East African
5 populations, two of which are Afroasiatic speaking populations from Kenya (MAF = 0.031) and
6 Ethiopia (MAF = 0.012) (Table S4). The variant Arg708Trp (rs776995986) occurs in the region
7 identified as the *TMPRSS2* cleavage site in *ACE2*⁵¹ and is found only in the Afroasiatic speaking
8 populations from Ethiopia (MAF = 0.004). Importantly, the presence of Arginine residues has been
9 shown to be important in “multibasic” cleavage sites¹⁸. Therefore, due to the drastic change in
10 physicochemical properties of the residue, this variation could be expected to interfere in
11 *TMPRSS2* cleavage efficiency, though it warrants experimental validation. Finally, two variants
12 are located at glycosylation sites. Variant Asn546Ser (rs756905974, chrX:15572228), which
13 causes the loss of a conserved glycosylation site on the *ACE2* protein, is found only in the SAS
14 populations (MAF = 0.001). Variant Lys26Arg (rs4646116), found in individuals from the
15 European (EUR) (MAF = 0.005), African (AFR) (MAF = 0.001) and South Asian (SAS) (MAF =
16 0.002) populations from the 1KG dataset (Table S4), occurs near both the conserved *ACE2*
17 glycosylation site Asn90 and the RBD binding site. The modification to a similarly positively
18 charged positive residue could suggest a role for electrostatic interactions, though no direct
19 interference with RBD binding could be deduced without further studies.

20

21 **Regulatory variation at *ACE2* among global populations**

22 In contrast to coding variants which have direct effects on protein structure in all cells
23 expressing a gene, the effects of regulatory genetic variants are relatively difficult to determine⁵².
24 Expression quantitative trait locus (eQTL) analysis has been used to identify genetic variants
25 associated with gene expression. We first extracted 2,053 eQTLs significantly associated with
26 *ACE2* gene expression ($P < 0.001$) from the GTEx database³⁵ (Table S5). To narrow down
27 candidate functional variants, we focus on the eQTLs located in the promoter regions of target
28 genes or in enhancers supported by chromatin interaction data⁵³.

29 We identified six eQTLs (rs4830977, rs4830978, rs5936010, rs4830979, rs4830980 and
30 rs5934263) located in a strong DNase peak at 73.3 kb upstream of *ACE2* that have direct
31 interactions with *ACE2* based on RNA Pol2 ChIA-PET data (Figure 1E, S2 and Table S5). All six

1 SNPs are eQTLs of *ACE2* and all of them have positive normalized effect sizes ($NES > 0.2$) and
2 significant p-values ($P < 0.00008$) in brain, tibial nerve, tibial artery, pituitary and prostate cells
3 (Figure S3 and Table S5). In non-African populations, these six eQTLs are in high LD ($R^2 = 0.91$
4 $- 1.0$) (Figure S4) and, thus there are two common haplotypes: “CCGGAT” and “ATCATC”. The
5 frequency for the “ATCATC” haplotype ranges from 0.31 - 0.47 in all populations except the East
6 Asian population, which has a frequency of 0.068 at all 6 SNPs (Figure S5). In African
7 populations, LD is lower ($R^2 > 0.5$; Figure S4), and there are three common haplotypes:
8 “CCGGAT” (0.564), “ATCATC” (0.308), and “CCCGAC” (0.116). Of note, every allele in the
9 haplotype “CCGGAT” is correlated with higher expression of *ACE2* in the cortex of the brain
10 while alleles in haplotype “ATCATC” are correlated with lower expression of *ACE2*; other
11 haplotypes have alleles with both positive and negative effect sizes in different tissues (Table S5).
12 Haplotype “CCCGAC” is only present in populations with African ancestry and its frequency is
13 highest in the Botswana Khoesan (0.38) and Cameroon CAHG (0.38) populations. We also
14 identified one variant (rs186029035) located in strong TF and DNase clusters (ENCODE) in the
15 16th intron of *ACE2*. This variant is only common in the Cameroon CAHG population and,
16 therefore, there is no eQTL data for this SNP in the GTEx database ($MAF = 0.153$, Table S2).

17 18 **Signatures of natural selection at *ACE2***

19
20 As indicated above, most of the non-synonymous variants at *ACE2* are rare in global
21 populations and many of them are predicted to be deleterious, indicating that this gene is under
22 strong purifying selection. To formally test for signatures of natural selection at *ACE2*, we first
23 examined the ratio of non-synonymous and synonymous variants at each gene using the dN/dS
24 test⁴². The dN/dS for all pooled samples was 0.77, indicating that *ACE2* is under purifying selection
25 globally (Table S6 and Figure S6). However, in the East Asian population, we observed seven
26 non-synonymous variants (all of them are rare) and only one synonymous variant, and the dN/dS
27 value is 1.85, indicating an excess of non-synonymous variation. In other populations, the dN/dS
28 ratio ranges from 0 to 0.79 (Table S6 and Figure S6). Thus, *ACE2* appears to be under strong
29 purifying selection in most populations but may be under weak purifying selection in the East
30 Asian population. We next applied the MK-test⁴¹ which compares the ratio of fixed non-
31 synonymous sites between humans and chimpanzee ($D_n = 8$) and fixed synonymous sites ($D_s = 6$)
32 to the ratio of polymorphic nonsynonymous sites among populations ($P_n = 41$) relative to

1 polymorphic synonymous sites ($P_s = 14$) and found that it is not significant (odds ratio (OR) = 0.45,
2 $P = 0.94$, two-sided Fisher's exact test, **Table S7 and Figure S7, S8**).

3 Because the above-mentioned methods are more suitable for detecting signals of natural
4 selection acting over long time scales^{54,55}, we then tested for signatures of recent positive selection
5 at *ACE2* in global populations using the iHS test⁴⁵ to detect extended haplotype homozygosity
6 (EHH)⁴⁴, which identifies regions of extended linkage disequilibrium (LD) surrounding a
7 positively selected locus. We first focused on the three common non-synonymous variants in the
8 CAHG population from Cameroon (rs138390800, rs147311723 and rs145437639; MAF: 0.083
9 - .164), and a common putative regulatory variant (rs186029035, located in TF and DNase clusters
10 in the 16th intron of *ACE2*, MAF = 0.153). The derived alleles of these variants exist on three
11 different haplotype backgrounds: rs147311723 and rs145437639 are on the same haplotype
12 backgrounds, while rs138390800 and rs186029035 are on similar, but distinct, haplotype
13 background (**Figure 2A**). The derived alleles of the corresponding SNPs on each haplotype
14 background show EHH extending longer than 2 Mb, while the ancestral alleles of these SNPs
15 harbor haplotypes extending less than 0.3 Mb (**Figure 2B**). We then calculated the integrated
16 haplotype score (iHS) of each of these variants, to determine whether these extended haplotypes
17 are unusually long compared to other SNPs with a similar allele frequency; the iHS values were
18 not significant for any of these variants (**Figure S9 and Table S8**). However, if selection were
19 acting on multiple haplotypes simultaneously (as shown above), the EHH and iHS tests would not
20 be well powered to detect selection⁵⁶. We also used the d_i statistic⁴⁹ to measure if allele frequencies
21 at these candidate SNPs were strongly differentiated between Cameroon CAHG and other
22 populations. The d_i values of SNPs rs138390800 and rs186029035 were in the top 1.4% and 1.7%,
23 respectively, of d_i values for all SNPs examined, indicating that that allele frequencies at these
24 variants are amongst the most highly differentiated in the CAHG population. However, it should
25 be noted that in the CAHG these four variants (rs138390800, rs147311723, rs145437639 and
26 rs186029035) are in complete LD based on D' ($D' = 1$) with the 6 eQTLs described above (**Figure**
27 **S10, S11**), indicating that the alleles are on the same haplotype background. Thus, it is not possible
28 to distinguish if the non-synonymous variants are targets of selection or if they are “hitchhiking”
29 to high frequency due to selection on flanking regulatory variants. Given the high LD in the region,
30 it is possible that multiple functional variants on the same haplotype backgrounds have been under
31 selection.

1 We then investigated signatures of recent positive selection at candidate regulatory variants
2 near *ACE2* in the global datasets. In total, there are 234 variants that had high iHS scores ($|iHS| >$
3 2) in at least one population extending over an ~200 kb region (Table S9), and 48% (n=113) of
4 these variants are either eQTLs or located at DNase hypersensitive regions which are in high LD
5 based on D' (Figure S10, S11, and Table S9). Among the region near the TSS (<10kb from *ACE2*),
6 there are two variants in high LD ($D' = 1$) that had high iHS scores in the San population from
7 Botswana (Figure 3A); rs150147953 is located in a DNase peak in multiple tissues including lung,
8 intestine and heart and rs2097723 is an eQTL of *ACE2* in the brain (Figure 1E, S12; Table S9).
9 We also identified strong selection signals at the region 50 – 120 kb upstream of *ACE2* (chrX:
10 15650000-15720000) in the AFR, San from Botswana, and Niger-Congo-speaking populations
11 from Cameroon, as well as Afroasiatic- and Nilo-Saharan-speaking populations from Kenya
12 (Figure 3A and S9). Two SNPs in this region (rs5936010 and rs5934263) have elevated iHS scores
13 ($|iHS| > 2$) in the San population from Botswana and the Afroasiatic population from Kenya
14 (Figure 3A) and are part of the 6 eQTLs described above, located within a strong enhancer
15 interacting with the promoter of *ACE2* (Figure 1E). Two additional eQTLs, that are in complete
16 LD with the 6 eQTLs ($D' = 1$; Figure S11), rs4830984 and rs4830986, had high iHS scores in four
17 of the five African populations listed above (all but the Kenya Afroasiatic; Figure 3A).

18 We performed haplotype network analysis to examine phylogenetic relationships among
19 haplotypes at *ACE2* in global populations derived for SNPs showing signatures of natural selection
20 (Figure 3B and 3C). We identified two haplotype clades: one (clade 1) is nearly specific to
21 Africans and the other (clade 2) encompasses global populations (Figure 3B). In the CAHG,
22 haplotypes containing the rs138390800 (Lys341Arg) non-synonymous variant and the
23 rs186029035 regulatory variant are in clade 1, whereas haplotypes containing the rs147311723
24 (Leu731Phe) and rs145437639 (Asp597Glu) non-synonymous variants are located in clade 2
25 (Figure 3B). Haplotypes containing the two regulatory variants (rs5936010 and rs5934263) located
26 50 – 120 kb upstream of *ACE2* are shared in global populations, and the nearby regulatory variants
27 rs4830984 and rs4830986 are sub-lineages on those haplotype backgrounds (Figure 3B and 3C).

28

29 **Associations between genetic variations in *ACE2* and clinical disease phenotypes**

30

1 We examined associations of genetic variation at *ACE2* with clinical phenotypes using the
2 PMBB cohort that consists of exome-sequencing data from 15,977 participants between the ages
3 of 19 and 89 years (52% female) with extensive clinical data available through their electronic
4 health records (EHR). Of these, 7061 individuals were of European ancestry (42%) and 8916 were
5 of African ancestry (55%) (Table S1).

6 To test for association between rare coding variants and clinical phenotypes, we applied a
7 gene-based approach^{37; 57} and single variant analysis. First, we performed a gene-based analysis
8 by collapsing the coding region variants with MAF < 0.01 that are annotated as non-synonymous
9 or putative loss-of-function (pLOF) variants. We tested for association with 12 phenotypes,
10 encompassing COVID-relevant disease classes affecting different organ systems, defined by EHR
11 based diagnosis codes (Table S10). For the gene-based approach, we applied two statistical tests:
12 a) a burden test (i.e. the cumulative effect of rare variants in a gene) that uses logistic regression
13 and b) a sequence kernel association test (SKAT)⁵⁸. Thus, it can compute effect estimates but may
14 suffer from loss of power when gene variants have effects in opposite directions (i.e., protective
15 and higher risk variants). This limitation can be overcome by parallel analysis with SKAT, a
16 powerful approach to model mixed effect variants. However, this approach does not provide effect
17 estimates. Therefore, we reported outcomes using both methods. Ancestry specific analysis of
18 gene-based tests identified seven associations in African ancestry (AA) and three associations in
19 European ancestry (EA) populations that reached statistical significance levels after multiple
20 hypothesis correction ($p < 1 \times 10^{-04}$) for the SKAT model. None of the gene burden models reached
21 a significance level of $p < 1 \times 10^{-04}$. The effect size from the logistic regression model was used to
22 indicate a protective or increased risk effect on disease phenotype. In the AA population, the most
23 significant associations were with hepatic encephalopathy and respiratory failure (Figure 1F and
24 Table 1). The association with respiratory failure is interesting as it is one of the key severe clinical
25 features reported for COVID-19^{10; 59-62}. However, the same association was not significant in the
26 EA population, which could be explained by lack of power due to lower number of coding variants
27 at *ACE2* in EA. Within the EA population, the most significant associations included hepatic coma,
28 respiratory syncytial virus infectious disease, and cirrhosis of the liver (Table 1).

29 We also examined an extended list of ~1800 phecodes derived from the EHR and 33 EHR-
30 based quantitative lab measurements and performed a phenome-wide association study
31 (PheWAS)^{37;57;63}. After multiple testing correction, we identified one association in the AA and

1 five associations in the EA populations reaching study-wide significance ($p < 1 \times 10^{-5}$) (Table S10
2 and S5). Myocarditis, a rare cardiovascular disease caused by viral infection, was the top PheWAS
3 association in the AA population but not significant in the EA population. Although the population
4 difference for this specific association is unclear, recent studies have reported a link between
5 SARS-CoV-2 induced cardiac injury among COVID-19 patients⁶⁴, which it was suggested might
6 be mediated by *ACE2*. This observation would be consistent with *ACE2* expression in heart tissue,
7 and its upregulation in cardiomyocytes⁶⁴⁻⁶⁶. Among respiratory diseases, cough and allergic rhinitis
8 reached nominal significance ($p < 0.01$) in the AA population (Table S10). In the EA population,
9 we identified a nominal association with influenza, asthma, emphysema, cough, and painful
10 respiration (p -value <0.01). Our findings in the EA cohort are consistent with other studies of *ACE2*
11 in subjects derived from the UK biobank⁶⁷. Among the median measure of 33 EHR-based
12 quantitative lab measurements that we investigated (see Methods), only the internationalized
13 normalized ratio (INR) derived from the prothrombin time test showed a nominal association with
14 increase in INR above 1.11, potentially relevant to blood clotting abnormalities observed in
15 COVID patients (Table 2).

16 To further evaluate the individual effect of each rare coding variant in *ACE2*, we performed
17 a single variant association analysis on the rare variants in the genes identified from gene-based
18 tests. A Fishers exact test was used to account for the small sample size when testing the impact
19 of rare single variants on phenotypes. The *ACE2* variant rs147311723, which is only present in
20 African populations, was most significantly associated with respiratory infection ($p < 0.05$,
21 OR=1.95 [1.06 - 3.6]). Another African specific *ACE2* variant rs138390800 did not reach
22 statistical significance but showed modestly increased risk of respiratory failure ($p=0.1$; OR=2.29
23 [0.83 - 6.33]).

24 For the six eQTLs identified near *ACE2* (rs4830977, rs4830978, rs5936010, rs4830979,
25 rs4830980 and rs5934263), we performed a PheWAS with clinical data by ancestry. We found
26 that two of the six eQTLs (rs5936010 and rs5934263) (targets of positive selection in both
27 Afroasiatic populations from Kenya and Khoesan populations from Botswana) are significantly
28 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),
29 respectively, in the AA population (Figure 1G, and Table S11). Among the respiratory disorders,
30 all six eQTLs had nominal associations ($p < 0.01$) with acute sinusitis and dypnea (shortness of
31 breath) in AA and bronchiectasis in EA. Further, we noticed a difference in the effect of association

1 among the six eQTLs with respiratory disorders examined in AA. Variants rs5936010 (OR=1.11)
2 and rs5934263 (OR = 1.19) were associated with increased risk of respiratory disorder, whereas
3 the rest of the four eQTL variants were associated with decreased risk.

4 5 **Genetic variation at *TMPRSS2* and its potential role in SARS-COV-s2 infection** 6 **susceptibility**

7
8 The trans-membrane protease serine 2 (*TMPRSS2*) protein enhances the spike protein-
9 driven viral entry of SARS-CoV-2 into cells¹⁸. At this gene, we identified forty-eight
10 nonsynonymous variants. Among the non-synonymous variants, only two (rs12329760
11 [Val197Met] and rs75603675 [Gly8Val]) have high MAF (> 0.05) in the pooled global dataset
12 (Figure 4A, and Table S12). While rs75603675 is highly variable in non-East-Asian populations
13 (AFR = 0.3, AMR = 0.27, EUR = 0.4, and SAS = 0.2), it is not highly variable in East Asians
14 (MAF = 0.02) (Figure 4B and 4C, and Table S12). In addition, some non-synonymous variants
15 were common and specific to African populations. Notably, the non-synonymous variant
16 rs61735795 (Pro375Ser) had a high MAF in the Khoesan-speaking population from Botswana
17 (MAF = 0.18). This variant is present at low frequency in populations from Cameroon (MAF <
18 0.01) and Ethiopia (MAF < 0.03) and was absent in non-African populations. The non-
19 synonymous variant rs367866934 (Leu403Phe) is common in the Cameroonian CAHG population
20 (MAF = 0.15) and has low frequency (MAF = 0.02) in other populations from Cameroon, but it is
21 absent from non-Cameroonian populations (Figure 4B and Table S12). Another non-synonymous
22 variant rs61735790 (His18Arg) is common in the CAHG populations from Cameroon (MAF =
23 0.12) and the Nilo-Saharan populations from Ethiopia (MAF = 0.12) but is rare in other
24 populations (Figure 4B and Table S12).

25 We identified two regulatory SNPs (rs76833541 and rs4283504) in the promoter region of
26 the *TMPRSS2* gene that have been identified as eQTLs of *TMPRSS2* in testis (Figure 4D, S13, and
27 Table S5). The MAF of rs76833541 is higher in EUR (MAF = 0.16) than other populations (EAS
28 = 0.002, AFR = 0.006, AMR = 0.06 and SAS = 0.05) and the MAF of rs4283504 is more common
29 in EAS (MAF = 0.21) than other populations (EUR = 0.11, AFR = 0.04, AMR = 0.12 and SAS =
30 0.14) (Figure S14 and Table S2).

31 32 **Signatures of natural selection at *TMPRSS2*** 33

1 We applied the MK test at *TMPRSS2* and observed that Dn/Ds (13/2) is significantly larger
2 than Pn/Ps (48/45) among pooled human samples (OR = 6.1, P-val = 0.009, Fisher's exact test)
3 (Figure 5A, and Table S7) as well as in individual ethnic groups (OR ranged from 5.0 - 17),
4 indicating positive selection in the hominin lineage after divergence from chimpanzee. Notably,
5 there are 13 non-synonymous and 2 synonymous variants at *TMPRSS2* (ENST00000398585.7, see
6 Figure S15 for ENST00000332149.10) that were fixed in human populations. The non-
7 synonymous variants are located in different structural domains of *TMPRSS2*: amino acid A3P,
8 N10S, T46P, A70V, R103C, and M104T are located in the cytoplasmic region which may function
9 in intracellular signal transduction⁶⁸; L124I is located in the transmembrane region; N144K is
10 located in the extracellular region; S165N and S178G are located in the LDL-receptor class A
11 domain; E441Q and T515M are located in the Peptidase S1 domain which is involved in the
12 interaction with the SARS-CoV-2 spike protein¹⁸; S529G is located in the last amino acid position
13 of the protein (Figure 5B). In contrast to the MK test, the dN/dS ratio test was not significant in
14 any population, indicating no excess of non-synonymous to synonymous variation within
15 populations. (Table S6 and Figure S6).

16 We also tested for recent positive selection at *TMPRSS2* in all ethnic groups using iHS
17 (Figure S16 and Table S8). We found many SNPs (n = 153) with high iHS scores ($|iHS| > 2$) in
18 different ethnic groups in a 78 kb region encompassing the *TMPRSS2* gene which show high levels
19 of LD (ChrX:41454000-41541000; Figure S16, S17). We identified a non-synonymous variant
20 (rs150969307) that shows a signature of positive selection (iHS = 2.01) and is common only in the
21 Chabu hunter gatherer population from Ethiopia (MAF = 0.079) (Table S12). We found that more
22 than one third of SNPs with $|iHS|$ scores > 2.0 (62 of 153) are located in putative regulatory regions
23 (Figure S18 and Table S9).

24 25 **Associations between genetic variations in *TMPRSS2* and clinical disease phenotypes**

26

27 In the PMBB, gene-based analysis with 12 severe disease classes identified nominal
28 associations with respiratory failure, respiratory syncytial virus infectious disease, lower
29 respiratory tract infection and pneumonia in the AA population but no statistically significant
30 association in the EA population (Table 1, Figure 4E). As with *ACE2*, the clinical phenotype
31 associations with *TMPRSS2* in AA may be driven by an excess of rare variants in that population
32 and, hence, more carriers in comparison to EA. Among the diseases in the respiratory disorder

1 category, we identified a nominal association ($p < 0.01$) with allergic rhinitis in AA and with
2 obstructive bronchitis in EA populations (Table S10). Previously, a gene-based PheWAS with
3 EHR-derived disease codes in the UK biobank population, which consists of mostly individuals
4 of European descent, showed no statistically significant associations with *TMPRSS2*⁶⁹. Among
5 clinical lab measures, we identified nominal association with urine bilirubin levels ($p = 0.001$).
6 The PheWAS of the two regulatory eQTLs (rs76833541 and rs4283504) of *TMPRSS2* described
7 above identified association of rs76833541 with abnormal glucose ($p = 8.9 \times 10^{-4}$, OR=1.5) in EA
8 and rs4283504 with glucocorticoid deficiency ($p = 0.001$, OR=2.7) in AA (Figure 4F). We did not
9 identify any association between these two eQTLs and respiratory conditions (Figure 4F, and
10 Table S11).

11 **Patterns of variation at *DPP4* and *LY6E***

13

14 ***DPP4***

15 *DPP4* is a receptor for the Middle East Respiratory Coronavirus (MERS-Cov) and was
16 reported to interact with SARS-CoV-2⁷⁰. At this gene, we identified 47 non-synonymous variants
17 and one loss-of-function variant (Table S13). Among them, no variant was common in the pooled
18 global dataset (Figure 6A), suggesting this gene is extremely conserved during human evolutionary
19 history. Only one non-synonymous variant (rs1129599, Ser437Thr) was common in the Fulani
20 pastoralists from Cameroon (MAF = 0.081), was present at low frequency in other African
21 populations, and was absent in non-African populations (Figure 6B and 6C). In addition to the
22 nonsynonymous variants, one loss-of-function variant was identified at *DPP4*. The variant
23 rs149291595 (Q170*) has low MAF in some African populations (MAF < 0.05) but is absent in
24 non-African populations.

25 We identified four eQTLs (rs1861978, rs35128070, rs17574 and rs13015258) in the
26 promoter region of the *DPP4* gene (Figure 6D). Three of the variants (rs1861978, rs35128070 and
27 rs17574) are significant eQTLs in the transverse colon and rs13015258 is an eQTL in the lung (P
28 < 5.9×10^{-6} , Figure S20 and Table S5). The minor alleles of these three variants are rare in EAS (MAF
29 < 0.05) but common in all other populations (MAF > 0.15, Figure S21, and Table S2). The fourth
30 SNP, rs13015258, resides in the center of a cluster of DNase peaks identified in ENCODE (Figure
31 6D) with MAF ranging from 0.38 in the AMR population to 0.6 in other populations (Figure S21
32 and Table S2).

1

2 **Signatures of natural selection at *DPP4***

3 The MK-test result was not significant in either the pooled samples ($D_n = 3$, $D_s = 5$, $P_n =$
4 45 , $P_s = 33$ OR = 0.44, $P = 0.9$, two-sided Fisher's exact test) nor in each population separately
5 (Table S7 and Figure S8). For the the dN/dS test, we observed ratios ranging from 0 to 0.52 in
6 individual populations, indicating that *DPP4* is highly conserved (Table S6 and Figure S6) within
7 human populations. Using the iHS test, we identified 8 SNPs that had extreme high iHS scores
8 ($|iHS| > 2$) in the Khoesan populations from Botswana (Figure S22 and Table S8). Five of these
9 SNPs (rs10166124, rs2284872, rs2284870, rs7608798 and rs2160927) are in LD ($D' > 0.95$) with
10 each other (Figure S23). The SNP rs2284870 is located in a strong DNase peak in heart tissue
11 (Figure S24 and Table S9).

12

13 **Associations between genetic variations in *DPP4* and clinical disease phenotypes**

14 In the gene-based analysis among AA PMBB participants, we identified significant
15 associations (only in the SKAT model) with respiratory syncytial virus infectious disease and
16 upper respiratory tract disease (Figure 6E, Table 1 and S10). None of the gene-based models were
17 significant in the EA population. The PheWAS of four regulatory eQTLs identified the most
18 significant association with malignant neoplasm of the rectum (commonly referred as colon cancer)
19 for rs17574 ($p = 4.49 \times 10^{-04}$, OR = 1.8) and rs13015258 ($p = 0.002$, OR = 0.54) in AFR only.
20 Among respiratory disorders, rs35128070 had the most significant association with “abnormal
21 results of function study of pulmonary system” ($p=0.002$, OR=1.6) in the AFR population and we
22 observed a nominal association between rs17574 and “acute respiratory infections” ($p<0.01$,
23 OR=1.22) in the EA population (Figure 6F, and Table S11).

24

25 **LY6E**

26 Studies show that mice lacking *LY6E* were highly susceptible to a usually nonlethal mouse
27 coronavirus²⁴. At *LY6E* we observed twenty-eight non-synonymous variants and all of them,
28 except rs11547127 (MAF = 0.057), have MAF that are rare in the pooled global dataset (Figure
29 7A, and Table S14). However, some of the non-synonymous variants are common in specific
30 populations (Figure 7B). For instance, the non-synonymous variant rs111560737 (Asp104Asn)
31 was common in the southern African Khoesan population from Botswana (MAF = 0.36) and the
32 Chabu population from Ethiopia (MAF = 0.17) (Figure 7C). Three loss-of-function variants

1 (rs200177123 [stop gained, Ser59*], chr8:143020941, and chr8:143020946) were also identified
2 at *LY6E*, and all of them are rare. In the PMBB, only four pathogenic and likely pathogenic variants
3 were identified, and all were rare in both AA and European EA populations.

4 We identified three regulatory eQTLs (rs13252864, rs17061979 and rs114909654) located
5 within 2 kb of the transcription start site of *LY6E* (Figure 7D), all of which are significant in
6 esophageal mucosa ($P < 1e-5$, Figure S25 and Table S5), which has a high expression level of
7 *LY6E* (TPM=108, GTE_x). The minor alleles of rs13252864 and rs114909654 are common in
8 African populations (MAF > 0.15) while very rare in other populations (MAF < 0.02, Figure S26),
9 whereas the MAF of rs17061979 is relatively high in EAS (0.18) and SAS (0.13) and rare in other
10 populations (MAF < 0.05, Figure S26).

11 Signatures of natural selection at *LY6E*

12 The MK-test result was not significant in either the pooled samples ($D_n = 0$, $D_s = 4$, $P_n =$
13 9 , $P_s = 9$, $OR = 0$, $P = 0.9$, two-sided Fisher's exact test) nor in each population separately (OR
14 ranging from 0 to 0.52; Table S7 and Figure S8), indicating that *LY6E* is highly conserved. We
15 identified 19 variants that that had extreme high iHS scores ($|iHS| > 2$) (Table S8, Figure S27),
16 some of which are in LD in specific populations (Figure S28). One variant (rs867069115) shows
17 an extreme iHS score in the Hadza hunter-gatherer population from Tanzania ($iHS = -2.94$). This
18 variant is located in a regulatory region ~1.9kb downstream of *LY6E*, within DNase and TF peaks
19 in the lung, intestine, kidney, heart, stomach, pancreas and skeletal muscle from ENCODE (Figure
20 S29) and is common only in the Hadza population (MAF = 0.14), is rare in other African
21 population (MAF < 0.05) and is absent in all non-African populations (Table S2). SNP
22 rs10283236, which shows an extreme iHS value in the CEU population, is an eQTL of *LY6E*
23 located within DNase and TF clusters identified in ENCODE (~4.14kb downstream of
24 *LY6E*) active in many tissues including lung, kidney and small intestine.

25

26 Associations between genetic variations in *LY6E* and clinical disease phenotypes

27 We identified a nominal association between *LY6E* with pneumonia in the AA population
28 only ($p = 0.01$, Figure 7E, Table 2 and Table S10). *LY6E* also has nominal association ($p < 0.01$,
29 Table 2) with total cholesterol, prothrombin, and eosinophil levels among the AA population.
30 Association with prothrombin is not statistically significant in the EA population. The association
31 analysis of regulatory variants identified most significant association with "severe protein-calorie

1 malnutrition” ($p = 2.35 \times 10^{-05}$, OR = 1.9) and “acute post hemorrhagic anemia” ($p = 6.4 \times 10^{-04}$,
2 OR = 1.6) in the AA population. In the EA population, “chronic ulcer of skin” with rs13252864
3 ($p=0.001$, OR=2.2) was the most significant association (Figure 7F, and Table S11).

4 5 **Discussion**

6 Investigating global patterns of genetic variation at genes that play a role in SARS-CoV-2
7 infection could provide insights into potential differences in susceptibility to COVID-19 among
8 diverse human populations. However, African populations are under-represented in the majority
9 of current genetic studies of COVID-19 susceptibility and severity, despite the fact that they have
10 the highest genetic diversity among human populations^{71;72}. In this study, we present a
11 comprehensive analysis of human genes which play a key role in SARS-CoV-2 host receptor
12 binding and cellular invasion, i.e., *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E*. We characterized the
13 coding and non-coding variants in these candidate genes to examine population differences in
14 allele frequencies and signatures of natural selection in diverse ethnic populations. This included
15 novel sequence data from 2012 ethnically diverse African populations from five countries
16 (Cameroon, Ethiopia, Kenya, Botswana and Tanzania) in Africa practicing different lifestyles (e.g.
17 hunter-gatherers, agriculturalists, and pastoralists). Additionally, we analyzed the correlation of
18 common and rare genetic variants in these four genes with clinical traits derived from the dataset
19 of 15,997 individuals from the Penn Medicine BioBank (PMBB) with African and European
20 ancestry. We included 12 “organ dysfunction” categories defined by phenotype algorithms (see
21 Methods), ~1800 ICD diagnosis codes, and 33 laboratory test measures from the EHR. Our results
22 highlight the importance of including genomes from diverse ethnic groups in human genetic
23 studies.

24 At *ACE2* we identified 41 non-synonymous variants, most of which are rare, suggesting
25 that they are under purifying selection. Tests based on dN/dS indicate that East Asians have an
26 excess of non-synonymous variation at *ACE2*, indicating weak purifying selection has influenced
27 patterns of variation in that population. However, there are some variants that are common in
28 specific ancestry groups. Notably, we identified three common non-synonymous variants
29 (rs138390800, rs147311723, and rs145437639) at *ACE2* with MAF ranging from 0.083 to 0.164
30 in Central African hunter-gatherers (CAHG), which were the only common coding variants
31 (defined here as MAF > 0.05) found in global populations studied here and by others^{20;67;73;74}. We

1 observed that the derived alleles of the common non-synonymous SNPs (rs138390800,
2 rs147311723, rs145437639) and one putative regulatory variant (rs186029035) at *ACE2* in *CAHG*
3 show evidence of EHH, with the extended haplotypes extending longer than 2 Mb, though they
4 did not show deviation from neutrality based on the iHS test. However, we do not have much
5 power to detect a selection signal using this test because the SNPs are on three different haplotype
6 backgrounds in *CAHG*, possibly due to selection on existing variation (e.g. “soft selection”) which
7 decreases the power to detect significant iHS scores⁷⁵. Moreover, each haplotype is at a relatively
8 low frequency (0.083 to 0.164), which further reduces the power of the iHS test. The *CAHG* are
9 traditionally hunter-gatherers living in a rainforest ecosystem who consume wild animals. They
10 have high exposure to animal viruses and were reported to have relative resistance to viral
11 infection⁷⁶. Thus, it is possible that this locus is adaptive for protection from infectious diseases in
12 this population. Future *in vitro* or *in vivo* studies will be needed to determine the functional
13 significance of these variants.

14 At *TMPRSS2*, we identified forty-eight nonsynonymous variants, only two of which had a
15 high MAF (>.05) in the pooled global dataset (rs12329760 and rs75603675). However, some
16 variants have high MAF in two African hunter-gatherer populations. Notably, the non-
17 synonymous variant rs61735795 (Pro375Se) is only common in the Khoesan-speaking San
18 population from Botswana (MAF = 0.18) and the non-synonymous variant rs367866934
19 (Leu403Phe) is only common in the Cameroonian *CAHG* populations (MAF = 0.15). At *TMPRSS2*
20 we observed a strong signature of adaptive evolution in the human lineage after divergence from
21 Chimpanzee ~ 6 MYA⁷⁷. In total, 13 non-synonymous variants located on different structural
22 domains of *TMPRSS2* were fixed in human populations. Among them, E441Q and T515M are
23 located in the Peptidase S1 domain that plays an important role in acute respiratory syndrome
24 (SARS)-like coronavirus (SARS-CoV-2) infection⁷⁸ and six (A3P, N10S, T46P, A70V, R103C,
25 and M104T) are at the cytoplasmic amino terminal domains of *TMPRSS2* which plays an important
26 role in signal transduction. These variants at *TMPRSS2* could be potential candidates for future
27 studies to investigate their functional impact on susceptibility to pathogens in humans compared
28 to non-human primates.

29 SARS-CoV replication is significantly reduced in *ACE2* knockout mice⁷⁹ and cells with
30 low expression of *ACE2* were resistant to SARS-CoV2 infection⁸⁰. It has also been shown that
31 both SARS-CoV and SARS-CoV2 infection could down regulate *ACE2* expression^{22;79; 81}. The

1 expression of *ACE2* and *TMPRSS2* in nasal and bronchial epithelial cells is higher in adults than
2 children, and in healthy individuals compared with smokers or patients with chronic obstructive
3 pulmonary disease⁵¹. Therefore, differences in expression levels of *ACE2* and *TMPRSS2* could
4 influence the susceptibility and host reactions to SARS-CoV-2. Regulatory eQTLs that differ in
5 frequency across ethnically diverse populations may play a role in local adaptation and disease
6 susceptibility⁸². eQTL mapping has been used to identify population-specific regulatory variation
7 and revealed the association of regulatory alleles with complex traits such as multiple sclerosis⁸³,
8 malaria⁵⁴ and immune response to infection⁸⁴. We identified regulatory eQTLs associated with
9 *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* gene expression and highlighted the eQTLs showing highly
10 differentiated MAF among populations and/or signatures of natural selection. These eQTLs are
11 located in ChIP-seq and DNase peaks and have the potential to influence transcription factor
12 binding and, thus, change the promoter or enhancer activities in specific tissues^{85; 86}. Interestingly,
13 some of the eQTLs in the upstream regions of *ACE2* were under selection in African populations.
14 For example, rs5936010 and rs5934263, which are located within a strong enhancer interacting
15 with the promoter of *ACE2* as suggested by ChIA-PET, harbored significant iHS scores ($|iHS| >$
16 2) in both Afroasiatic populations from Kenya and the San population from Botswana. Further,
17 PheWAS of these eQTLs in the PMBB populations identified association of eQTLs at *ACE2* with
18 type 2 diabetes (rs5936010) and hypertension (rs5934263). These are known pre-existing
19 conditions that increases risk of severe illness due to COVID-19^{11; 87; 88}. Among respiratory
20 diseases, only one eQTL at *ACE2* had nominal association (rs4830977) with acute sinusitis. The
21 association was only identified in the AA population and had a protective effect (OR = 0.78 [0.66-
22 0.95]). The eQTLs we analyzed are from GTEx V8 database⁸⁹, and 84.6% of the donors are people
23 of European and Western Eurasian descent. Therefore, it is possible that we are missing some
24 regulatory variants that are only present in specific ancestry groups due to the lack of sample
25 diversity. Further experimental testing of predicted regulatory variants will provide insights into
26 differences in gene expression regulation at *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* among different
27 populations. In the future, eQTL mapping in diverse populations will be informative for identifying
28 novel trait associations that may differ in prevalence across ethnic groups⁹⁰.

29 The gene-based genetic association analyses of non-synonymous variants at *ACE2*,
30 *TMPRSS2*, *DPP4* and *LY6E* identified several associations with clinical phenotypes. We observed
31 that respiratory failure has significant association with *ACE2* and *TMPRSS2* among the PMBB AA

1 population. That is a particularly interesting finding as respiratory failure is one of the clinical
2 outcomes observed in some patients with COVID-19^{10; 59-62}. However, this association was not
3 significant in the EA population. This observation could be explained by the low number of coding
4 variants and carriers at *ACE2* and *TMPRSS2* among EA and, hence, low power to detect an
5 association. An association with myocarditis, a rare cardiovascular disease caused by viral
6 infection, was also observed in the AA population. Recent studies have reported a link between
7 SARS-CoV-2 induced cardiac injury such as myocarditis among COVID-19 patients⁹¹. Further,
8 *ACE2* has known expression in heart tissue, and it plays an important role in transcriptional
9 dysregulation in cardiomyocytes – cells that make up cardiac muscles⁶⁴⁻⁶⁶. We observed
10 association between *ACE2* and myocarditis only in the AA population but as noted above, we may
11 not have as much power to detect and association in EA. Blood clotting abnormalities in lungs and
12 other organs in COVID-19 patients have been reported by several studies⁹². In autopsies of
13 COVID-19 patients, thrombosis was found to be a prominent finding across multiple organs, even
14 in spite of extensive anticoagulation treatment and regardless of timing of clinical progression,
15 indicating that thrombosis might be at play in the early stages of disease⁹². One hypothesis to
16 explain this observation is that the dysfunction of endothelial cells may play an important role in
17 increased risk of thrombosis⁹³. We observed associations between the internationalized normalized
18 ratio (INR) derived from the prothrombin time test (PT) with *ACE2* and *LY6E* in a gene-based
19 association test. The INR test measures the time it takes blood to clot and is an important measure
20 for individuals with blood clotting disorders or on blood thinners.

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

29
30

31 **Description of Supplemental Data**

32 Supplemental file 1: Supplemental **figures S1-S29**.

- 1 Table S1. Penn Medicine Biobank (PMBB) participant characteristics
- 2 Table S2. Genetic variants identified around the four genes. “N” denotes variants were not
3 identified or called in the corresponding dataset. “0” denotes variants were identified in the
4 corresponding dataset, but the minor allele frequency is 0.
- 5 Table S3. ICD code mapping to MONDO disease classes.
- 6 Table S4. Coding variants identified at *ACE2*. “N” denotes variants were not identified or called
7 in the corresponding dataset. “0” denotes variants were identified in the corresponding dataset, but
8 the minor allele frequency is 0.
- 9 Table S5. Regulatory variants identified at the four candidate genes. eQTLs are extracted from
10 GTEx V8.
- 11 Table S6. Result of the dN/dS for four genes in both the pooled dataset and specific ethnic
12 groups.
- 13 Table S7. Results of the MK-test for four genes in both the pooled dataset and specific ethnic
14 groups.
- 15 Table S8. SNPs with significant selection signals in each ethnic group based on each method.
- 16 Table S9. Regulatory SNPs that overlap with significant selection signals at the four genes.
- 17 Table S10. Summary statistics from gene-based association results
- 18 Table S11. Summary statistics from PheWAS of eQTL variants
- 19 Table S12. Coding variants identified at *TMPRSS2*. “N” denotes variants were not identified or
20 called in the corresponding dataset. “0” denotes variants were identified in the corresponding
21 dataset, but the minor allele frequency is 0.
- 22 Table S13. Coding variants identified at *DPP4*. “N” denotes variants were not identified or called
23 in the corresponding dataset. “0” denotes variants were identified in the corresponding dataset, but
24 the minor allele frequency is 0.
- 25 Table S14. Coding variants identified at *LY6E*. “N” denotes variants were not identified or called
26 in the corresponding dataset. “0” denotes variants were identified in the corresponding dataset, but
27 the minor allele frequency is 0.
- 28
- 29
- 30 **Declaration of Interests**
- 31 No conflict of interest

1

2 **Acknowledgements**

3 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by
4 the National Heart, Lung and Blood Institute (NHLBI). Genome Sequencing for "NHLBI
5 TOPMed: Integrative Genomic Studies of Heart and Blood Related Traits in Africans (Africa6K) "
6 was performed at Broad Genomics (hhsn268201600034i). Core support including centralized
7 genomic read mapping and genotype calling, along with variant quality metrics and filtering were
8 provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract
9 HHSN268201800002I). Core support including phenotype harmonization, data management,
10 sample-identity QC, and general program coordination were provided by the TOPMed Data
11 Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We
12 gratefully acknowledge the studies and participants who provided biological samples and data for
13 TOPMed.

14 Funding for this study was provided by grant numbers: X01HL139409, 1R35GM134957,
15 R01GM113657, R01DK104339, ADA 1-19-VSN-02, R01AR076241 to SAT, and R01LM010098
16 to SW. Cesar de la Fuente-Nunez holds a Presidential Professorship at the University of
17 Pennsylvania, is a recipient of the Langer Prize by the AIChE Foundation and acknowledges
18 funding from the Institute for Diabetes, Obesity, and Metabolism, the Penn Mental Health AIDS
19 Research Center of the University of Pennsylvania, and the National Institute of General Medical
20 Sciences of the National Institutes of Health under award number R35GM138201. IRB approval
21 for this project was obtained from the University of Pennsylvania. Written informed consent was
22 obtained from all participants and research/ethics approval and permits were obtained from the
23 following institutions prior to sample collection: the University of Addis Ababa and the Federal
24 Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research
25 Ethics Review Committee; COSTECH, NIMR and Muhimbili University of Health and Allied
26 Sciences in Dar es Salaam, Tanzania; the University of Botswana and the Ministry of Health in
27 Gaborone, Botswana; the Cameroonian National Ethical Committee (NEC) and Cameroonian
28 Ministry of Health (MOH).

29 We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the
30 patient-participants of Penn Medicine who consented to participate in this research program. We
31 would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for
32 providing genetic variant data for analysis. The PMBB is approved under IRB
33 protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania.

34 We thank Alex Harris, Alexander Platt, and Srilakshmi Raj for discussing the evolutionary
35 analysis in the paper. We thank the following individuals and organizations for their essential work
36 in collecting samples for this project: Kenya: Lilian A. Nyndodo, Eva Aluvalla, Daniel Kariuki,
37 Fathya Abdo, and Hussein Musa; Ethiopia: Solomon Taye, Birhanu Mekauntie, and Alemayehu
38 Moges; Tanzania: Kweli Powell, Holly Mortensen, Mariki Euphrasia, Ruth Matiyas, John G.
39 Memra, Holliness Santa, Emanuel Kimario, Reginald Kavishe; Botswana: Michael Campbells, Ari
40 Ho-Foster, Maitseo M. M. Bolane, Maungo Moswang, Gaolape Mpoloka, Kingsley Motshegwe,
41 Mothusi Molatlhegi; Cameroon: Eric Mbunwe, Sali Django, Dickson Ndizi, Valentine Ngum
42 Ndze, Julius Fonsah, Eric Ngwang, Grace N. Tenjei, Meagan Rubel, Peter Kfu, BACUDA
43 (Association Culturelle pour le Développement Bagyeli/Bakola de l'Océan, CADDAP (Centre
44 d'Action pour le Développement Durable des Autochtones Pygmées), MBOSCUDA (Mbororo

1 Social and Cultural Development Association). We especially thank all African participants for
2 their important contributions to this study.

5 **Regeneron Genetics Center Banner Author List and Contribution Statements**

7 All authors/contributors are listed in alphabetical order.

9 **RGC Management and Leadership Team**

10 Goncalo Abecasis, Ph.D., Aris Baras, M.D., Michael Cantor, M.D., Giovanni Coppola, M.D.,
11 Aris Economides, Ph.D., Luca A. Lotta, M.D., Ph.D., John D. Overton, Ph.D., Jeffrey G. Reid,
12 Ph.D., Alan Shuldiner, M.D.

14 Contribution: All authors contributed to securing funding, study design and oversight. All
15 authors reviewed the final version of the manuscript.

17 **Sequencing and Lab Operations**

18 Christina Beechert, Caitlin Forsythe, M.S., Erin D. Fuller, Zhenhua Gu, M.S., Michael Lattari,
19 Alexander Lopez, M.S., John D. Overton, Ph.D., Thomas D. Schleicher, M.S., Maria
20 Sotiropoulos Padilla, M.S., Louis Widom, Sarah E. Wolf, M.S., Manasi Pradhan, M.S., Kia
21 Manoochehri, Ricardo H. Ulloa.

23 Contribution: C.B., C.F., A.L., and J.D.O. performed and are responsible for sample genotyping.
24 C.B, C.F., E.D.F., M.L., M.S.P., L.W., S.E.W., A.L., and J.D.O. performed and are responsible
25 for exome sequencing. T.D.S., Z.G., A.L., and J.D.O. conceived and are responsible for
26 laboratory automation. M.P., K.M., R.U., and J.D.O are responsible for sample tracking and the
27 library information management system.

29 **Clinical Informatics**

30 Nilanjana Banerjee, Ph.D., Michael Cantor, M.D. M.A., Dadong Li, Ph.D., Deepika Sharma,
31 MHI.

33 Contribution: All authors contributed to the development and validation of clinical phenotypes
34 used to identify study subjects and (when applicable) controls.

36 **Genome Informatics**

37 Xiaodong Bai, Ph.D., Suganthi Balasubramanian, Ph.D., Andrew Blumenfeld, Boris Boutkov,
38 Ph.D., Gisu Eom, Lukas Habegger, Ph.D., Alicia Hawes, B.S., Shareef Khalid, Olga
39 Krasheninina, M.S., Rouel Lanche, Adam J. Mansfield, B.A., Evan K. Maxwell, Ph.D., Mrunali
40 Nafde, Sean O’Keeffe, M.S., Max Orelus, Razvan Panea, Ph.D., Tommy Polanco, B.A., Ayesha
41 Rasool, M.S., Jeffrey G. Reid, Ph.D., William Salerno, Ph.D., Jeffrey C. Staples, Ph.D.

43 Contribution: X.B., A.H., O.K., A.M., S.O., R.P., T.P., A.R., W.S. and J.G.R. performed and are
44 responsible for the compute logistics, analysis and infrastructure needed to produce exome and
45 genotype data. G.E., M.O., M.N. and J.G.R. provided compute infrastructure development and
46 operational support. S.B., S.K., and J.G.R. provide variant and gene annotations and their
47 functional interpretation of variants. E.M., J.S., R.L., B.B., A.B., L.H., J.G.R. conceived and are

1 responsible for creating, developing, and deploying analysis platforms and computational
2 methods for analyzing genomic data.

4 **Research Program Management**

5 Marcus B. Jones, Ph.D., Michelle LeBlanc, Ph.D., Lyndon J. Mitnaul, Ph.D.

7 Contribution: All authors contributed to the management and coordination of all research
8 activities, planning and execution. All authors contributed to the review process for the final
9 version of the manuscript.

11 **Web Resources**

12 Variation type descriptions in Variant Effect Predictor (VEP):

13 https://uswest.ensembl.org/info/genome/variation/prediction/predicted_data.html

14 UCSC genome browser: <https://genome.ucsc.edu/>

16 **Data Availability**

17 Additional information for reproducing the results described in the article is available upon
18 reasonable request and subject to a data use agreement.

20 **References**

- 21 1. Tang, D., Comish, P., and Kang, R. (2020). The hallmarks of COVID-19 disease. *PLoS*
22 *Pathog* 16, e1008536.
- 23 2. Furukawa, N.W., Brooks, J.T., and Sobel, J. (2020). Evidence Supporting Transmission of
24 Severe Acute Respiratory Syndrome Coronavirus 2 While Presymptomatic or
25 Asymptomatic. *Emerg Infect Dis* 26.
- 26 3. Goldstein, J.R., and Lee, R.D. (2020). Demographic perspectives on the mortality of COVID-
27 19 and other epidemics. *Proceedings of the National Academy of Sciences of the United*
28 *States of America*.
- 29 4. Omori, R., Matsuyama, R., and Nakata, Y. (2020). The age distribution of mortality from novel
30 coronavirus disease (COVID-19) suggests no large difference of susceptibility by age.
31 *Sci Rep* 10, 16642.
- 32 5. Klein, S.L., Dhakal, S., Ursin, R.L., Deshpande, S., Sandberg, K., and Mauvais-Jarvis, F.
33 (2020). Biological sex impacts COVID-19 outcomes. *PLoS Pathog* 16, e1008570.
- 34 6. Oran, D.P., and Topol, E.J. (2021). Prevalence of Asymptomatic SARS-CoV-2 Infection. *Ann*
35 *Intern Med* 174, 286-287.
- 36 7. Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., Davidson, K.W.,
37 the Northwell, C.-R.C., Barnaby, D.P., Becker, L.B., Chelico, J.D., et al. (2020).
38 Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients
39 Hospitalized With COVID-19 in the New York City Area. *JAMA : the journal of the*
40 *American Medical Association* 323, 2052-2059.
- 41 8. Yancy, C.W. (2020). COVID-19 and African Americans. *JAMA : the journal of the American*
42 *Medical Association*.

- 1 9. Alcendor, D.J. (2020). Racial Disparities-Associated COVID-19 Mortality among Minority
2 Populations in the US. *J Clin Med* 9.
- 3 10. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong,
4 Y., et al. (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel
5 Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA : the journal of the American*
6 *Medical Association*.
- 7 11. Wu, Z., and McGoogan, J.M. (2020). Characteristics of and Important Lessons From the
8 Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of
9 72314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA : the*
10 *journal of the American Medical Association*.
- 11 12. Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L.,
12 Hui, D.S.C., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China.
13 *N Engl J Med* 382, 1708-1720.
- 14 13. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al.
15 (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in
16 Wuhan, China: a retrospective cohort study. *Lancet* 395, 1054-1062.
- 17 14. Nogueira, R., Abdalkader, M., Qureshi, M.M., Frankel, M.R., Mansour, O.Y., Yamagami, H.,
18 Qiu, Z., Farhoudi, M., Siegler, J.E., Yaghi, S., et al. (2021). EXPRESS: Global Impact of
19 the COVID-19 Pandemic on Stroke Hospitalizations and Mechanical Thrombectomy
20 Volumes. *Int J Stroke*, 1747493021991652.
- 21 15. Zhang, S., Zhang, J., Wang, C., Chen, X., Zhao, X., Jing, H., Liu, H., Li, Z., Wang, L., and
22 Shi, J. (2021). COVID19 and ischemic stroke: Mechanisms of hypercoagulability
23 (Review). *Int J Mol Med* 47.
- 24 16. Jha, N.K., Ojha, S., Jha, S.K., Dureja, H., Singh, S.K., Shukla, S.D., Chellappan, D.K.,
25 Gupta, G., Bhardwaj, S., Kumar, N., et al. (2021). Evidence of Coronavirus (CoV)
26 Pathogenesis and Emerging Pathogen SARS-CoV-2 in the Nervous System: A Review
27 on Neurological Impairments and Manifestations. *J Mol Neurosci*.
- 28 17. Oliveira, I.B., Pessoa, M.S., Lima, C.F., Holanda, J.L., and Coimbra, P.P.A. (2021).
29 Ischaemic stroke as an initial presentation in patients with COVID-19: evaluation of a
30 case series in an emergency in Brazil. *Neuroradiol J*, 1971400920987357.
- 31 18. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S.,
32 Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., et al. (2020). SARS-CoV-2 Cell
33 Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease
34 Inhibitor. *Cell* 181, 271-280 e278.
- 35 19. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020).
36 Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181,
37 281-292 e286.
- 38 20. Cao, Y., Li, L., Feng, Z., Wan, S., Huang, P., Sun, X., Wen, F., Huang, X., Ning, G., and
39 Wang, W. (2020). Comparative genetic analysis of the novel coronavirus (2019-
40 nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov* 6, 11.
- 41 21. Silhol, F., Sarlon, G., Deharo, J.C., and Vaisse, B. (2020). Downregulation of ACE2 induces
42 overstimulation of the renin-angiotensin system in COVID-19: should we block the renin-
43 angiotensin system? *Hypertens Res* 43, 854-856.
- 44 22. Glowacka, I., Bertram, S., Muller, M.A., Allen, P., Soilleux, E., Pfefferle, S., Steffen, I.,
45 Tsegaye, T.S., He, Y., Gnirss, K., et al. (2011). Evidence that TMPRSS2 activates the
46 severe acute respiratory syndrome coronavirus spike protein for membrane fusion and
47 reduces viral control by the humoral immune response. *Journal of virology* 85, 4122-
48 4134.
- 49 23. de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V.J. (2016). SARS and MERS:
50 recent insights into emerging coronaviruses. *Nat Rev Microbiol* 14, 523-534.

- 1 24. Pfaender, S., Mar, K.B., Michailidis, E., Kratzel, A., Hirt, D., V'Kovski, P., Fan, W., Ebert, N.,
2 Stalder, H., Kleine-Weber, H., et al. (2020). LY6E impairs coronavirus fusion and confers
3 immune control of viral disease. *bioRxiv*.
- 4 25. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G.,
5 Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse
6 genomes from the NHLBI TOPMed Program. *bioRxiv*.
- 7 26. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M.,
8 Korb, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global
9 reference for human genetic variation. *Nature* 526, 68-74.
- 10 27. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and
11 Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology* 17, 122.
- 12 28. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD:
13 predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids*
14 *Res* 47, D886-D894.
- 15 29. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein
16 function. *Nucleic Acids Res* 31, 3812-3814.
- 17 30. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human
18 missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7 20.
- 19 31. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome
20 of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American*
21 *journal of human genetics* 88, 440-449.
- 22 32. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human
23 genome. *Nature* 489, 57-74.
- 24 33. Chadwick, L.H. (2012). The NIH Roadmap Epigenomics Program data resource.
25 *Epigenomics* 4, 317-324.
- 26 34. Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon,
27 A., Lopez, F., and Ballester, B. (2020). ReMap 2020: a database of regulatory regions
28 from an integrative analysis of Human and Arabidopsis DNA-binding sequencing
29 experiments. *Nucleic Acids Res* 48, D180-D188.
- 30 35. Consortium, G.T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis:
31 multitissue gene regulation in humans. *Science* 348, 648-660.
- 32 36. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K.,
33 Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS:
34 demonstrating the feasibility of a phenome-wide scan to discover gene-disease
35 associations. *Bioinformatics (Oxford, England)* 26, 1205-1210.
- 36 37. Moore, C.B., Wallace, J.R., Frase, A.T., Pendergrass, S.A., and Ritchie, M.D. (2013).
37 BioBin: a bioinformatics tool for automating the binning of rare variants using publicly
38 available biological knowledge. *BMC Med Genomics* 6 Suppl 2, S6.
- 39 38. Basile, A.O., Wallace, J.R., Peissig, P., McCarty, C.A., Brilliant, M., and Ritchie, M.D.
40 (2016). Knowledge Driven Binning and Phewas Analysis in Marshfield Personalized
41 Medicine Research Project Using Biobin. *Pac Symp Biocomput* 21, 249-260.
- 42 39. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the
43 recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367, 1444-1448.
- 44 40. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J Mol*
45 *Graph* 14, 33-38, 27-38.
- 46 41. McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in
47 *Drosophila*. *Nature* 351, 652-654.
- 48 42. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of
49 synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3, 418-426.

- 1 43. Zhai, W., Nielsen, R., and Slatkin, M. (2009). An investigation of the statistical power of
2 neutrality tests based on comparative and population genetic data. *Mol Biol Evol* 26,
3 273-283.
- 4 44. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne,
5 E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and
6 characterization of positive selection in human populations. *Nature* 449, 913-918.
- 7 45. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive
8 selection in the human genome. *PLoS biology* 4, e72.
- 9 46. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to
10 perform EHH-based scans for positive selection. *Mol Biol Evol* 31, 2824-2827.
- 11 47. Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in
12 a UK Biobank cohort. *Nature genetics* 48, 811-816.
- 13 48. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, N., Adrian,
14 J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA
15 elements in the human and mouse genomes. *Nature* 583, 699-710.
- 16 49. Akey, J.M., Ruhe, A.L., Akey, D.T., Wong, A.K., Connelly, C.F., Madeoy, J., Nicholas, T.J.,
17 and Neff, M.W. (2010). Tracking footprints of artificial selection in the dog genome.
18 *Proceedings of the National Academy of Sciences of the United States of America* 107,
19 1160-1165.
- 20 50. Leigh, J.W., and Bryant, D. (2015). POPART: full-feature software for haplotype network
21 construction. *Methods Ecol Evol* 6, 1110-1116.
- 22 51. Heurich, A., Hofmann-Winkler, H., Gierer, S., Liepold, T., Jahn, O., and Pohlmann, S.
23 (2014). TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by
24 TMPRSS2 augments entry driven by the severe acute respiratory syndrome coronavirus
25 spike protein. *Journal of virology* 88, 1293-1307.
- 26 52. Gloss, B.S., and Dinger, M.E. (2018). Realizing the significance of noncoding functionality in
27 clinical genomics. *Exp Mol Med* 50, 97.
- 28 53. Duggal, G., Wang, H., and Kingsford, C. (2014). Higher-order chromatin domains link
29 eQTLs with the expression of far-away genes. *Nucleic Acids Res* 42, 87-96.
- 30 54. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A.,
31 Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the
32 human lineage. *Science* 312, 1614-1620.
- 33 55. Kryazhimskiy, S., and Plotkin, J.B. (2008). The population genetics of dN/dS. *PLoS genetics*
34 4, e1000304.
- 35 56. Scheinfeldt, L.B., and Tishkoff, S.A. (2013). Recent human adaptation: genomic
36 approaches, interpretation and insights. *Nature reviews Genetics* 14, 692-702.
- 37 57. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Team,
38 N.G.E.S.P.-E.L.P., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified
39 approach for rare-variant association testing with application to small-sample case-
40 control whole-exome sequencing studies. *American journal of human genetics* 91, 224-
41 237.
- 42 58. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association
43 testing for sequencing data with the sequence kernel association test. *American journal*
44 *of human genetics* 89, 82-93.
- 45 59. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y.,
46 et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel
47 coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395, 507-513.
- 48 60. Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., Castelli, A., Cereda, D.,
49 Coluccello, A., Foti, G., Fumagalli, R., et al. (2020). Baseline Characteristics and
50 Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the
51 Lombardy Region, Italy. *JAMA : the journal of the American Medical Association*.

- 1 61. Arentz, M., Yim, E., Klaff, L., Lokhandwala, S., Riedo, F.X., Chong, M., and Lee, M. (2020).
2 Characteristics and Outcomes of 21 Critically Ill Patients With COVID-19 in Washington
3 State. *JAMA : the journal of the American Medical Association*.
- 4 62. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et
5 al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan,
6 China. *Lancet* 395, 497-506.
- 7 63. Verma, S.S., Josyula, N., Verma, A., Zhang, X., Veturi, Y., Dewey, F.E., Hartzel, D.N.,
8 Lavage, D.R., Leader, J., Ritchie, M.D., et al. (2018). Rare variants in drug target genes
9 contributing to complex diseases, phenome-wide. *Sci Rep* 8, 4624.
- 10 64. Shi, S., Qin, M., Shen, B., Cai, Y., Liu, T., Yang, F., Gong, W., Liu, X., Liang, J., Zhao, Q., et
11 al. (2020). Association of Cardiac Injury With Mortality in Hospitalized Patients With
12 COVID-19 in Wuhan, China. *JAMA Cardiol*.
- 13 65. Siripanthong, B., Nazarian, S., Muser, D., Deo, R., Santangeli, P., Khanji, M.Y., Cooper,
14 L.T., Jr., and Chahal, C.A.A. (2020). Recognizing COVID-19-related myocarditis: The
15 possible pathophysiology and proposed guideline for diagnosis and management. *Heart*
16 *Rhythm* 17, 1463-1471.
- 17 66. Tucker, N.R., Chaffin, M., Bedi, K.C., Jr., Papangelis, I., Akkad, A.D., Arduini, A., Hayat, S.,
18 Eraslan, G., Muus, C., Bhattacharyya, R.P., et al. (2020). Myocyte-Specific Upregulation
19 of ACE2 in Cardiovascular Disease: Implications for SARS-CoV-2-Mediated Myocarditis.
20 *Circulation* 142, 708-710.
- 21 67. Cirulli, E.T., Riffle, S., Bolze, A., and Washington, N.L. (2020). Revealing variants in SARS-
22 CoV-2 interaction domain of ACE2 and loss of function intolerance through analysis of
23 >200,000 exomes. *bioRxiv*.
- 24 68. Hooper, J.D., Clements, J.A., Quigley, J.P., and Antalis, T.M. (2001). Type II
25 transmembrane serine proteases. Insights into an emerging class of cell surface
26 proteolytic enzymes. *The Journal of biological chemistry* 276, 857-860.
- 27 69. (!!! INVALID CITATION !!! 71).
- 28 70. Vankadari, N., and Wilce, J.A. (2020). Emerging WuHan (COVID-19) coronavirus: glycan
29 shield and structure prediction of spike glycoprotein and its interaction with human
30 CD26. *Emerg Microbes Infect* 9, 601-604.
- 31 71. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo,
32 J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and
33 history of Africans and African Americans. *Science* 324, 1035-1044.
- 34 72. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human
35 Genetic Studies. *Cell* 177, 26-31.
- 36 73. Hou, Y., Zhao, J., Martin, W., Kallianpur, A., Chung, M.K., Jehi, L., Sharifi, N., Erzurum, S.,
37 Eng, C., and Cheng, F. (2020). New insights into genetic susceptibility of COVID-19: an
38 ACE2 and TMPRSS2 polymorphism analysis. *BMC medicine* 18, 216.
- 39 74. Benetti, E., Tita, R., Spiga, O., Ciolfi, A., Birolo, G., Bruselles, A., Doddato, G., Giliberti, A.,
40 Marconi, C., Musacchia, F., et al. (2020). ACE2 gene variants may underlie
41 interindividual variability and susceptibility to COVID-19 in the Italian population.
42 *European journal of human genetics : EJHG*.
- 43 75. (!!! INVALID CITATION !!! 70;79).
- 44 76. Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nedelec, Y., Pacis, A., Barakatt, M., Gravel, S.,
45 Zhou, X., Nsobya, S.L., et al. (2014). Adaptive, convergent origins of the pygmy
46 phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy*
47 *of Sciences of the United States of America* 111, E3596-3603.
- 48 77. Glazko, G.V., and Nei, M. (2003). Estimation of divergence times for major lineages of
49 primate species. *Mol Biol Evol* 20, 424-434.

- 1 78. David, A., Khanna, T., Beykou, M., Hanna, G., and Sternberg, M.J.E. (2020). Structure,
2 function and variants analysis of the androgen-regulated *TMPRSS2*, a drug
3 target candidate for COVID-19 infection. *bioRxiv*.
- 4 79. Kuba, K., Imai, Y., Rao, S., Gao, H., Guo, F., Guan, B., Huan, Y., Yang, P., Zhang, Y.,
5 Deng, W., et al. (2005). A crucial role of angiotensin converting enzyme 2 (ACE2) in
6 SARS coronavirus-induced lung injury. *Nat Med* 11, 875-879.
- 7 80. Letko, M., Marzi, A., and Munster, V. (2020). Functional assessment of cell entry and
8 receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 5,
9 562-569.
- 10 81. Verdecchia, P., Cavallini, C., Spanevello, A., and Angeli, F. (2020). The pivotal link between
11 ACE2 deficiency and SARS-CoV-2 infection. *Eur J Intern Med* 76, 14-20.
- 12 82. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson,
13 W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene
14 expression traits across diverse populations. *PLoS Genet* 14, e1007586.
- 15 83. Pala, M., Zappala, Z., Marongiu, M., Li, X., Davis, J.R., Cusano, R., Crobu, F., Kukurba,
16 K.R., Gloudemans, M.J., Reinier, F., et al. (2017). Population- and individual-specific
17 regulatory variation in Sardinia. *Nature genetics* 49, 700-707.
- 18 84. Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S., and Reich, D. (2015). No evidence that
19 selection has been less effective at removing deleterious mutations in Europeans than in
20 Africans. *Nature genetics* 47, 126-131.
- 21 85. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De
22 Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity
23 QTLs are a major determinant of human expression variation. *Nature* 482, 390-394.
- 24 86. Brown, C.D., Mangravite, L.M., and Engelhardt, B.E. (2013). Integrative modeling of eQTLs
25 and cis-regulatory elements suggests mechanisms underlying cell type specificity of
26 eQTLs. *PLoS genetics* 9, e1003649.
- 27 87. Zhang, Y., Cui, Y., Shen, M., Zhang, J., Liu, B., Dai, M., Chen, L., Han, D., Fan, Y., Zeng,
28 Y., et al. (2020). Association of diabetes mellitus with disease severity and prognosis in
29 COVID-19: A retrospective cohort study. *Diabetes research and clinical practice* 165,
30 108227.
- 31 88. Apicella, M., Campopiano, M.C., Mantuano, M., Mazoni, L., Coppelli, A., and Del Prato, S.
32 (2020). COVID-19 in people with diabetes: understanding the reasons for worse
33 outcomes. *Lancet Diabetes Endocrinol* 8, 782-792.
- 34 89. Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical
35 Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci,
36 Nih/Nhgri, Nih/Nimh, Nih/Nida, et al. (2017). Genetic effects on gene expression across
37 human tissues. *Nature* 550, 204-213.
- 38 90. Zhong, Y., De, T., Alarcon, C., Park, C.S., Lec, B., and Perera, M.A. (2020). Discovery of
39 novel hepatocyte eQTLs in African Americans. *PLoS genetics* 16, e1008662.
- 40 91. Knowlton, K.U. (2020). Pathogenesis of SARS-CoV-2 induced cardiac injury from the
41 perspective of the virus. *J Mol Cell Cardiol* 147, 12-17.
- 42 92. Biswas, S., Thakur, V., Kaur, P., Khan, A., Kulshrestha, S., and Kumar, P. (2020). Blood
43 clots in COVID-19 patients: Simplifying the curious mystery. *Med Hypotheses*, 110371.
- 44 93. Rapkiewicz, A.V., Mai, X., Carsons, S.E., Pittaluga, S., Kleiner, D.E., Berger, J.S., Thomas,
45 S., Adler, N.M., Charytan, D.M., Gasmi, B., et al. (2020). Megakaryocytes and platelet-
46 fibrin thrombi characterize multi-organ thrombosis at autopsy in COVID-19: A case
47 series. *EClinicalMedicine* 24, 100434.
- 48

49
50

Tables

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 P	Burden P	Burden OR	Burden SE	95% CI	Dataset
Hepatic Encephalopathy	<i>ACE2</i>	97	8045	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	6392	85	1	6.8E-07	0.1221	6.06	1.17	-0.48 - 4.09	AA
Respiratory Failure	<i>TMPRSS2</i>	199	6392	11	2	2.3E-06	0.0124	7.31	0.80	0.43 - 3.55	AA
Respiratory Failure	<i>ACE2</i>	199	6392	351	12	9.0E-05	0.0509	3.10	0.58	0 - 2.26	AA
Upper Respiratory Tract Disease	<i>DPP4</i>	144	6392	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	6392	11	1	3.9E-04	0.0217	11.63	1.07	0.36 - 4.55	AA
Pneumonia	<i>LY6E</i>	1120	6392	7	5	1.0E-02	0.0108	6.09	0.71	0.42 - 3.19	AA
Respiratory Syncytial Virus Infectious Disease	<i>ACE2</i>	56	6392	351	7	1.3E-02	0.1857	3.85	1.02	-0.65 - 3.34	AA
Lower Respiratory Tract Disease	<i>TMPRSS2</i>	693	6392	7	1	3.4E-02	0.1541	2.59	0.67	-0.36 - 2.26	AA
Pneumonia	<i>TMPRSS2</i>	1120	6392	11	4	4.8E-02	0.2029	2.23	0.63	-0.43 - 2.04	AA
Hepatic Coma	<i>ACE2</i>	16	6817	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	5859	274	3	2.3E-07	0.1650	3.61	0.92	-0.53 - 3.1	EA
Cirrhosis Of Liver	<i>ACE2</i>	10	6817	43	1	1.8E-04	0.0837	9.40	1.30	-0.3 - 4.78	EA
Acute Myocardial Infarction	<i>LY6E</i>	396	6494	8	1	1.8E-02	0.2936	3.65	1.23	-1.12 - 3.71	EA

Table 2. Association of *ACE2*, *DPP4*, *TMPRSS2*, and *LY6E* with clinical laboratory measures derived from the EHR.

Lab Name	Gene	Sample Size	Carriers	Beta	SE	P	Dataset
Urine Bilirubin	<i>TMPRSS2</i>	1410	2	3.13	0.95	0.001	EA
Total Cholesterol	<i>LY6E</i>	5800	8	41.11	15.29	0.007	AA
Prothrombin	<i>LY6E</i>	6220	10	3.47	1.31	0.008	AA
Eosinophil (%)	<i>LY6E</i>	7697	12	-1.34	0.56	0.016	AA
Eosinophil (THO/uL)	<i>LY6E</i>	7678	12	-0.09	0.04	0.022	AA
Prothrombin	<i>LY6E</i>	5944	7	5.17	2.52	0.040	EA
Prothrombin	<i>ACE2</i>	6220	330	1.11	0.55	0.045	AA

Figures

Figure 1. Genetic variation at *ACE2* and its disease association.

(A) Location of coding variants and their minor allele frequency (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 variants within rs138390800 at *ACE2* in diverse global ethnic groups is highlighted. Each pie denotes frequencies of alleles in the corresponding population. (D) Locations of identified non-synonymous 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 large intestine, small intestine, lung, kidney and heart are from ENCODE³². (F) Gene-based association result between coding variants at *ACE2* and 12 disease classes. The disease severity is shown on the x-axis and the y-axis represents the p-values. EA, European Ancestry; AA, African American ancestry. (G) PheWAS plot of six eQTL associated with *ACE2* and ~1800 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 dot represents an eQTL and the direction of effect of the association. The red dashed line denotes the 0.0001 cutoff, and the blue dashed line represent the 0.001 cutoff.

Figure 2. Natural selection signatures at *ACE2* in the Cameroon CAHG populations.

(A) Haplotypes over 150kb flanking *ACE2* in CAHG populations. The X-axis denotes 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 a top-left vertical black line suggest 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, 4 at the top of the plot denotes positions for rs147311723, rs186029035, rs145437639 and rs138390800, respectively. (B) Extended haplotype homozygosity (EHH) of rs138390800, rs186029035 and rs147311723 (rs145437639 is in strong LD with rs147311723) at *ACE2* in CAHG populations.

Figure 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 line denotes the gene body region of *ACE2*. Putatively causal tag SNPs were annotated in the plots. (B) Haplotype network over 150kb 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. In the left panel, dashed line denotes the boundary of clade 1 and clade 2. Black oval denotes haplotypes containing the corresponding variants. (C) Haplotype containing variants (rs5936010, rs5934263, rs4830984 and rs4830986) are highlighted. Red pie denotes haplotypes containing the derived allele of the corresponding variants, while green pie denotes haplotypes containing the ancestral allele of the corresponding variants

Figure 4. Genetic variation at *TMPRSS2* and its disease association.

(A) Location of coding variants and their minor allele frequency (MAF) at *TMPRSS2* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The geographic distribution of MAF of variant within rs75603675 at *TMPRSS2* in diverse global ethnic groups. (D) Two regulatory eQTLs located in the promoter region of the *TMPRSS2* gene. RNA Pol2 ChIA-PET data and DNase-seq data of large intestine, small intestine, lung, kidney and heart are from ENCODE³². (E) Gene-based association result between coding variants at *TMPRSS2* and 12 disease classes. The disease classes are shown on the x-axis and the y-axis represents the p-values. EA, European Ancestry; AA, African American ancestry. (F) PheWAS plot of the two eQTLs associated with *TMPRSS2* and ~1800 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 dot represents an eQTL and the direction of effect of the association. The red dashed line denotes the 0.0001 cutoff, and the blue dashed line represents the 0.001 cutoff.

Figure 5. Natural selection signatures of *TMPRSS2*.

(A) The result of the MK-test for *TMPRSS2* in the pooled dataset. Non-syn indicates non-synonymous variants; Syn indicates synonymous variants. “Fixed” denotes variants that were fixed between the human and the Chimpanzee; “Poly” represents polymorphic variants within human populations. OR, odds ratio. The transcript *ENST00000398585.7* was used for calculation. (B) Illustration of locations of variants that are divergent between the human and Chimpanzee lineages on the *TMPRSS2* protein domains. Boxes denote the protein domains of *TMPRSS2*. Red lines represent non-synonymous variants that occurred in the corresponding domains of *TMPRSS2*, with the amino acids and positions of the Human and the Chimpanzee annotated at the bottom of the lines. Blue lines denote synonymous variants. TM, transmembrane domain; LDLRA, LDL-receptor class A; SRCR, scavenger receptor cysteine-rich domain 2; Peptidase S1, Serine peptidase.

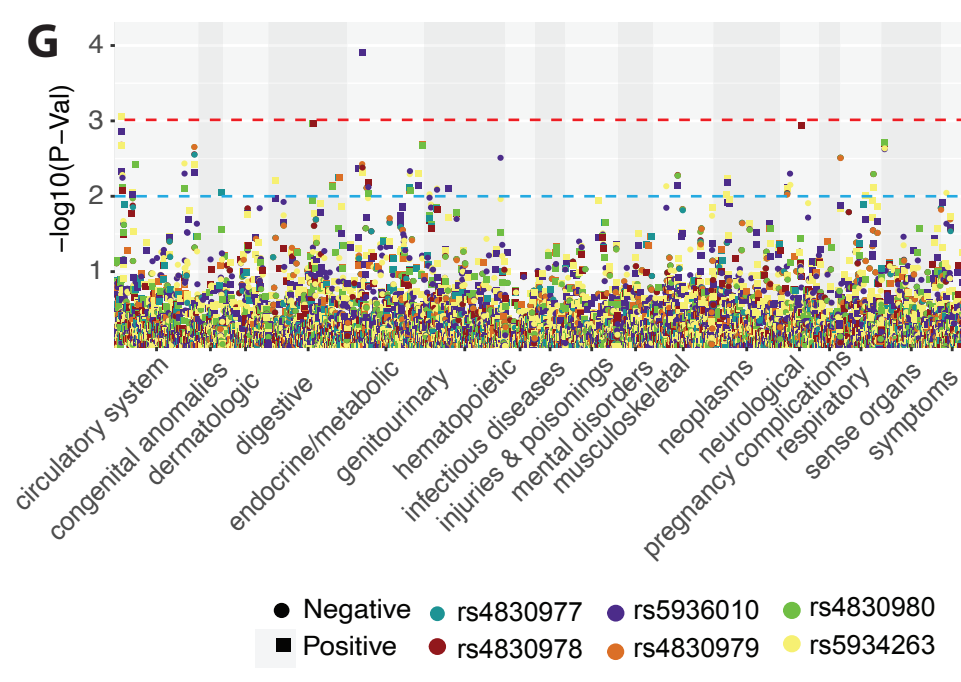
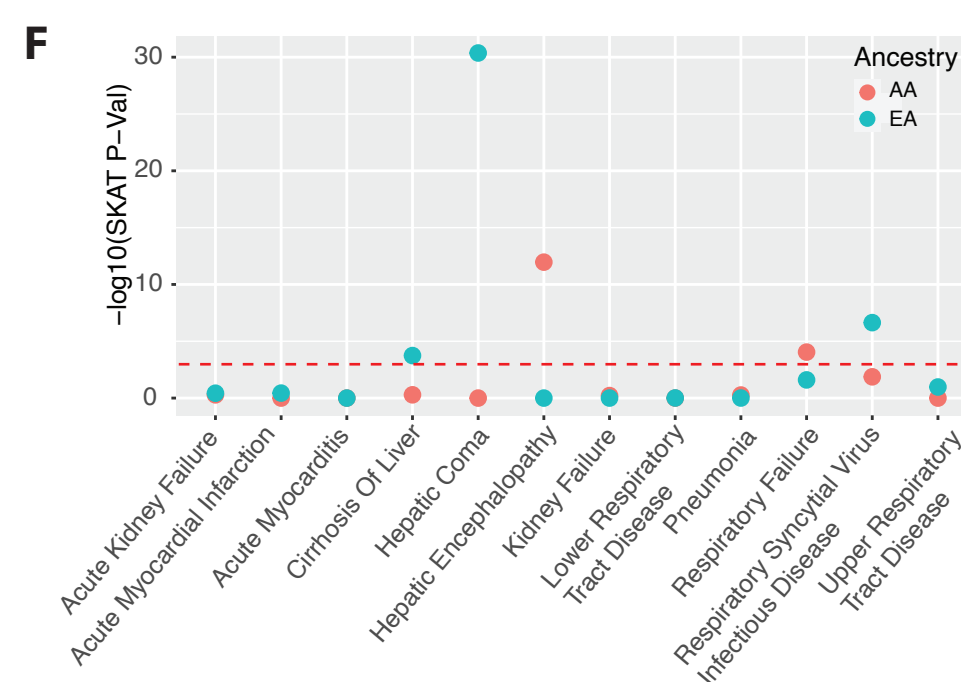
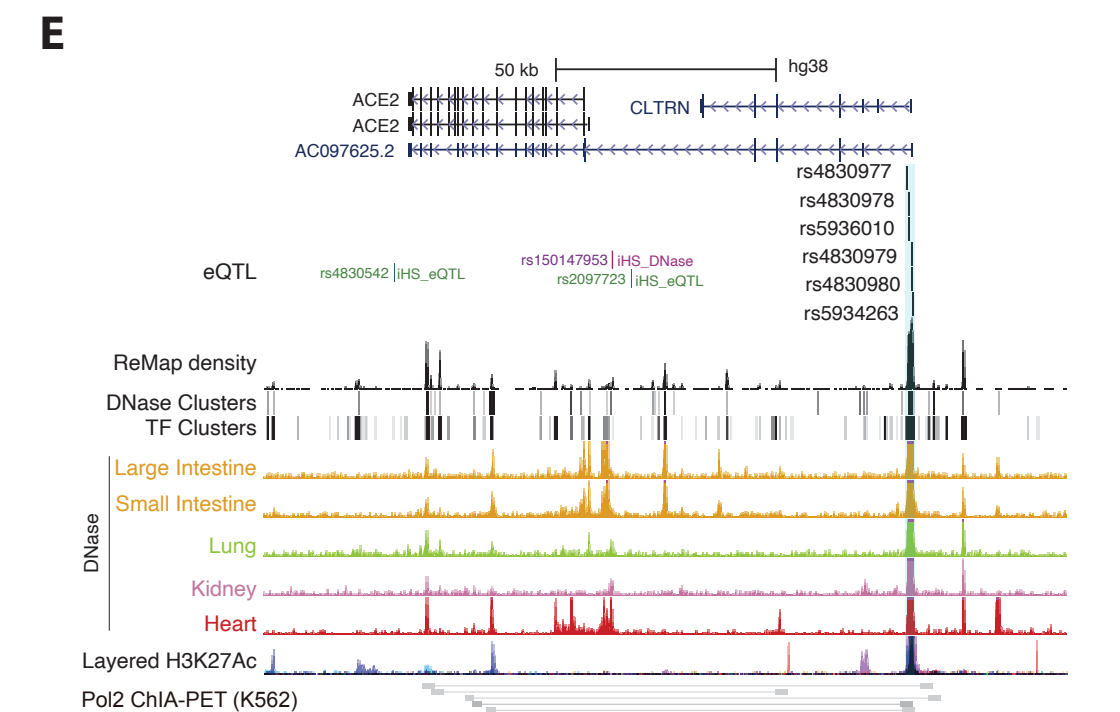
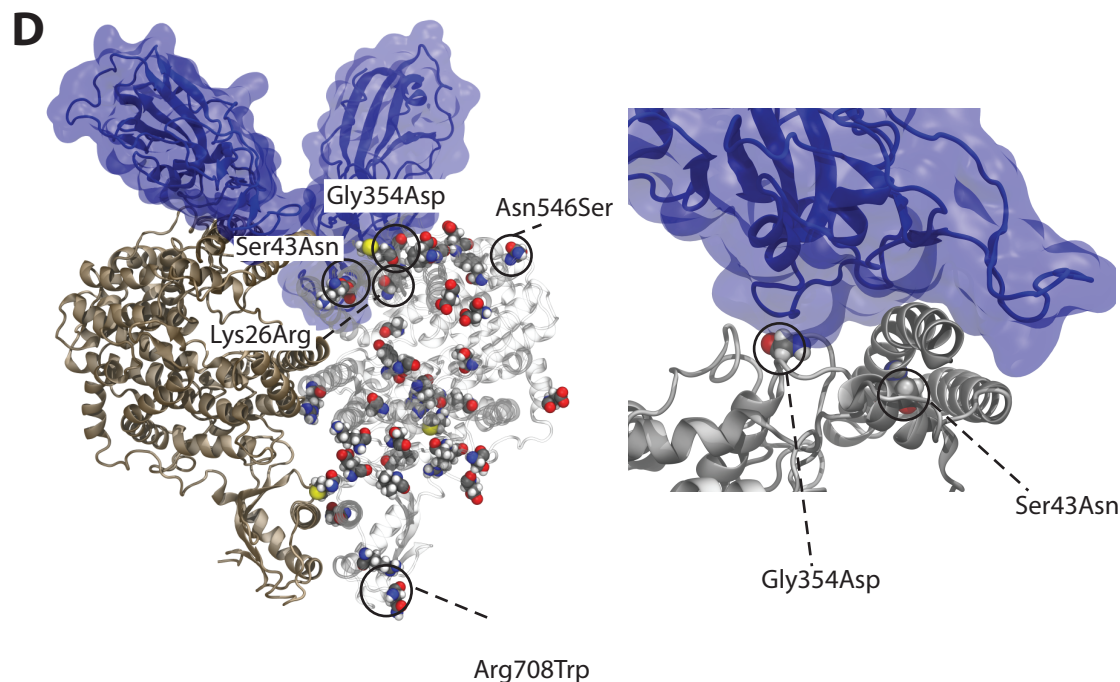
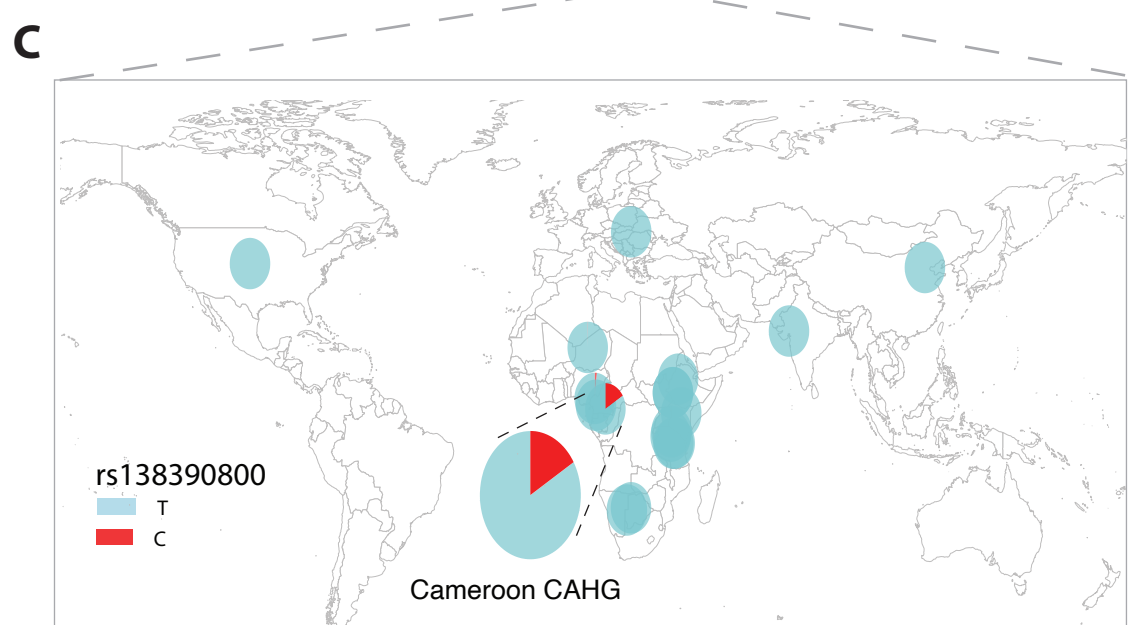
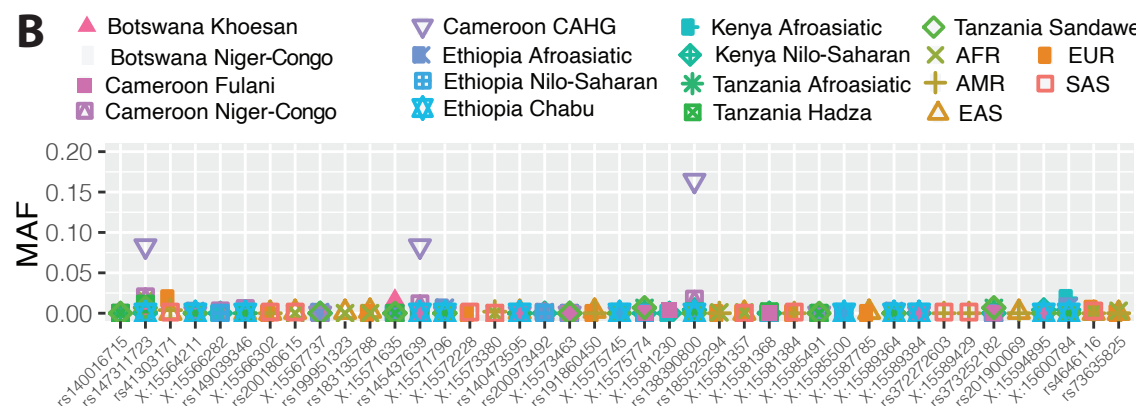
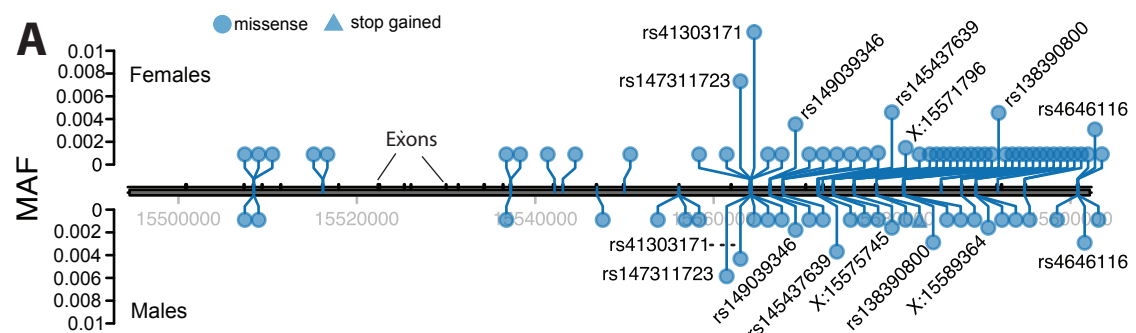
Figure 6. Genetic variation at *DPP4* and its disease association.

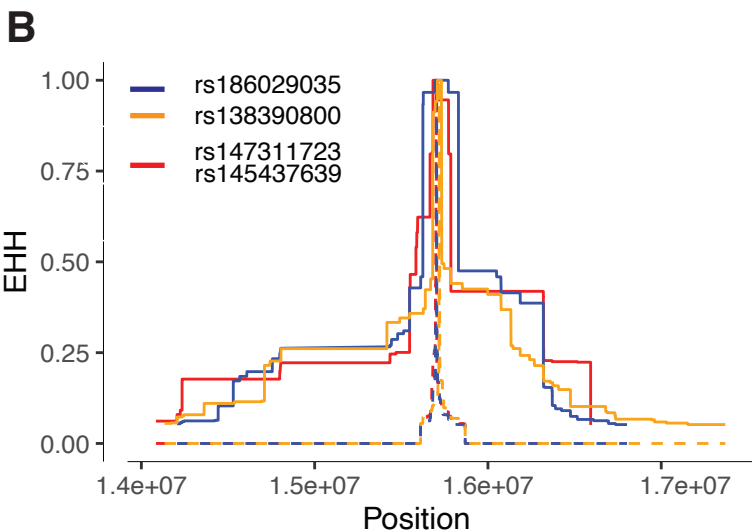
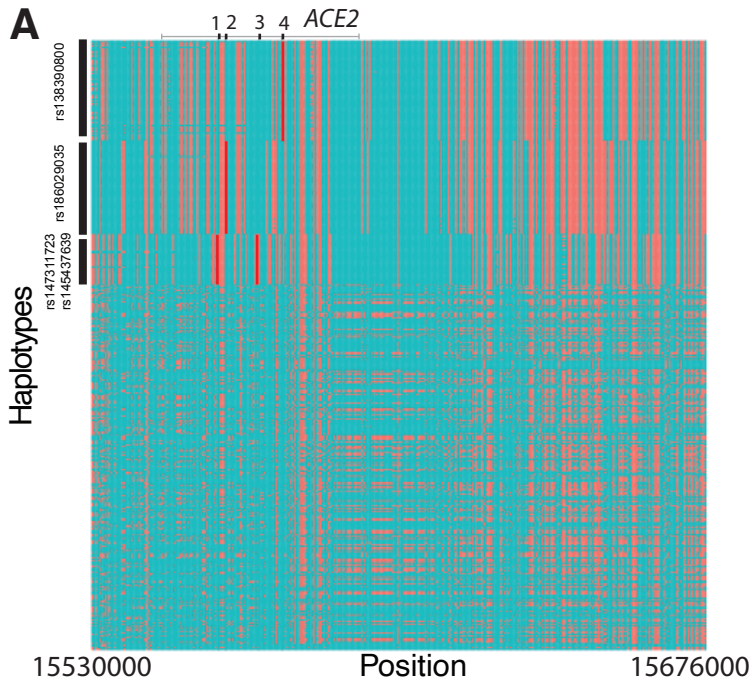
(A) Location of coding variants and their minor allele frequency (MAF) at *DPP4* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The MAF of variant within rs129559 at *DPP4* in diverse global ethnic groups. (D) Regulatory eQTLs located in *DPP4*. RNA Pol2 ChIA-PET data and DNase-seq data of large intestine, small intestine, lung, kidney and heart are from ENCODE³². (E) Gene-based association result between coding variants at *DPP4* and 12 disease classes. The disease classes are shown on the x-axis and the y-axis represents the p-values. EA, European Ancestry; AA, African American ancestry. (F) PheWAS plot of the four eQTLs associated with *DPP4* and ~1800 disease codes across 17 disease categories. The disease categories are shown on the x-axis and the y-axis

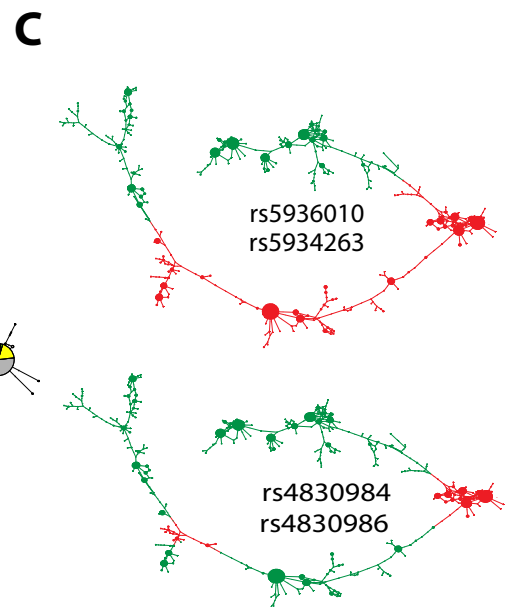
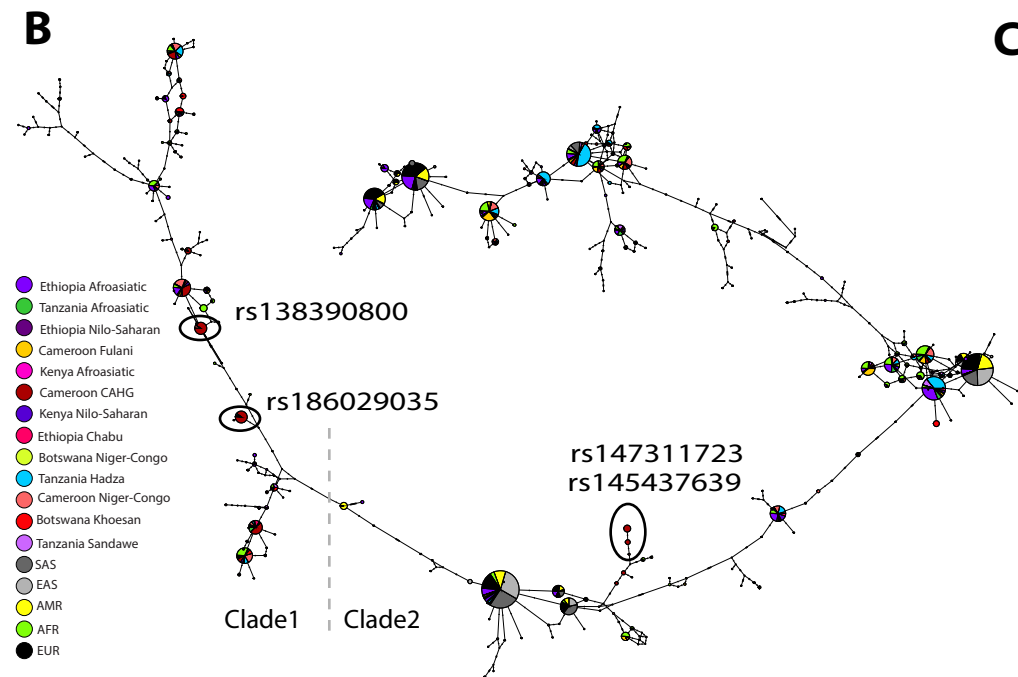
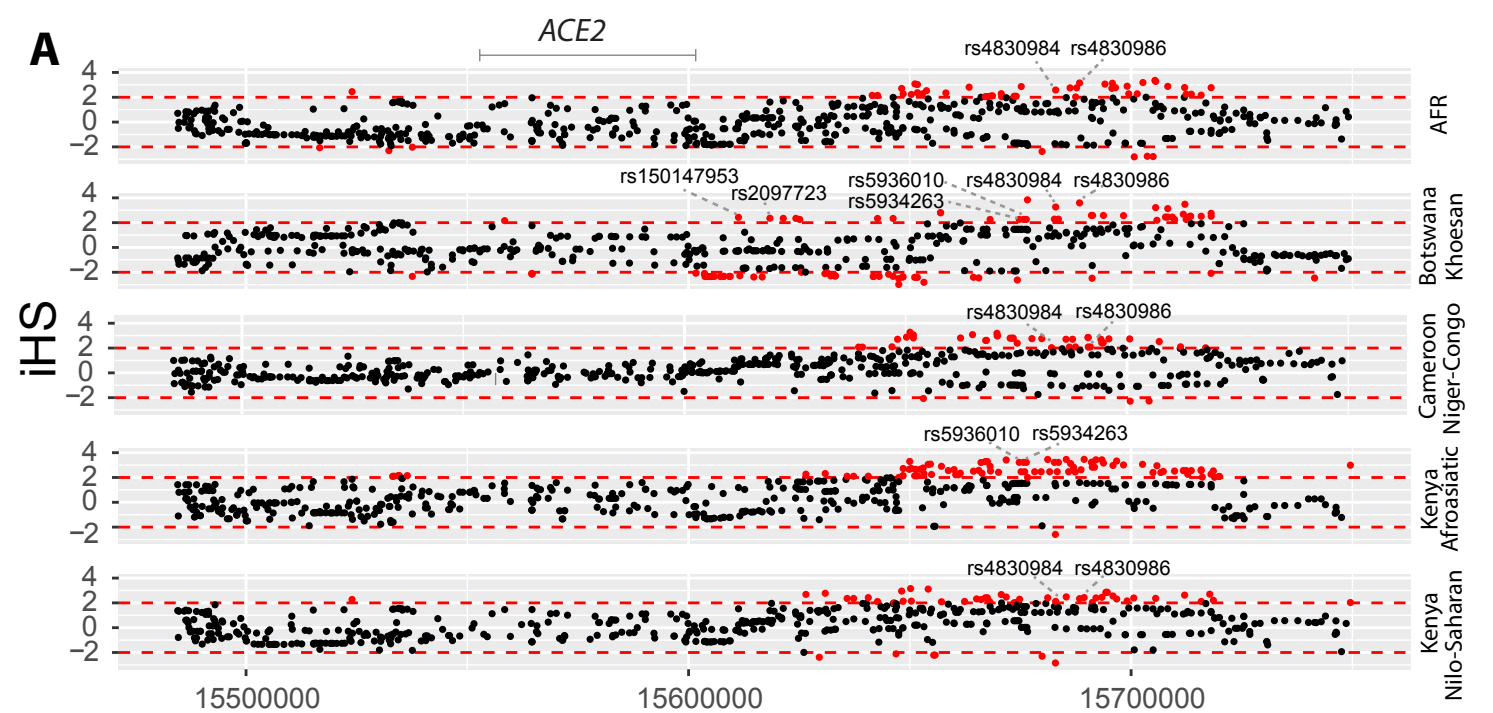
represents the $-\log_{10}$ of the p-values. The colored dot represents an eQTL and the direction of effect of the association. The red dashed line denotes the 0.0001 cutoff, and the blue dashed line represent the 0.001 cutoff.

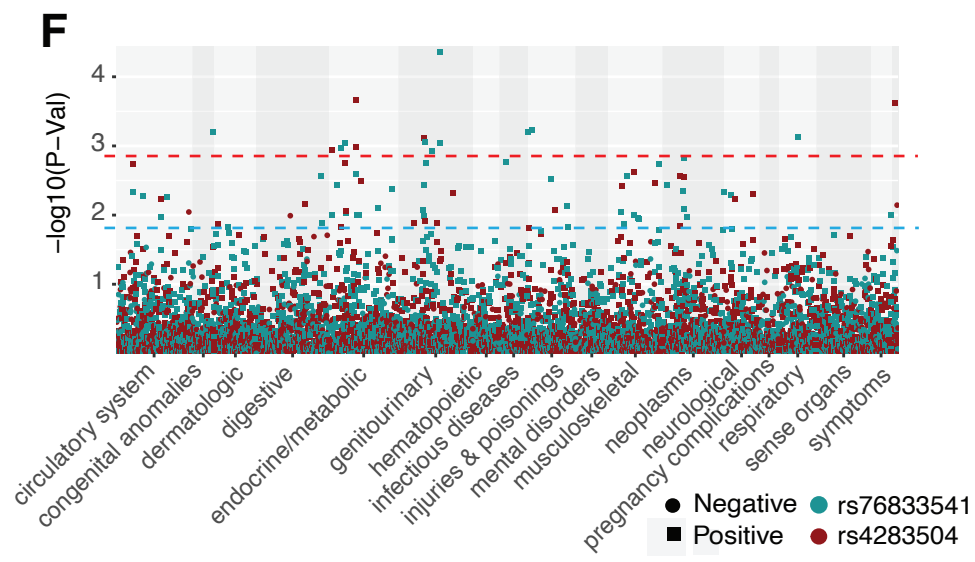
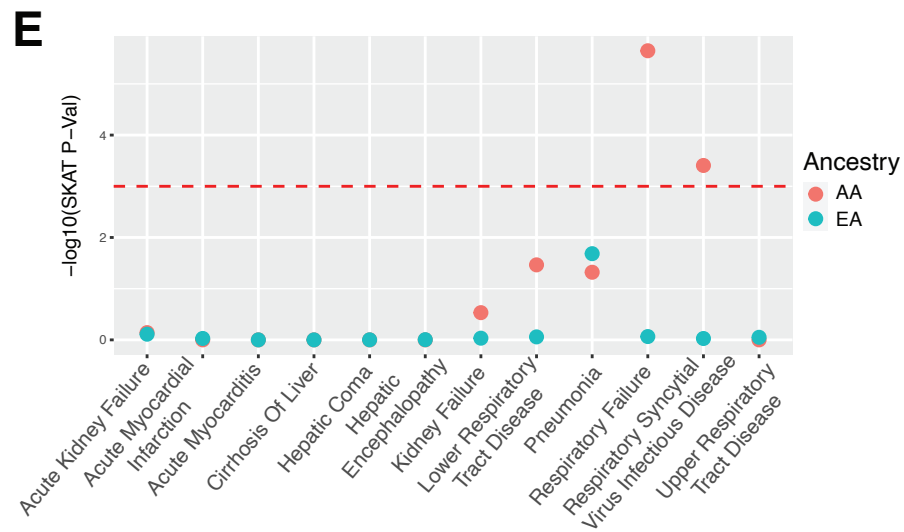
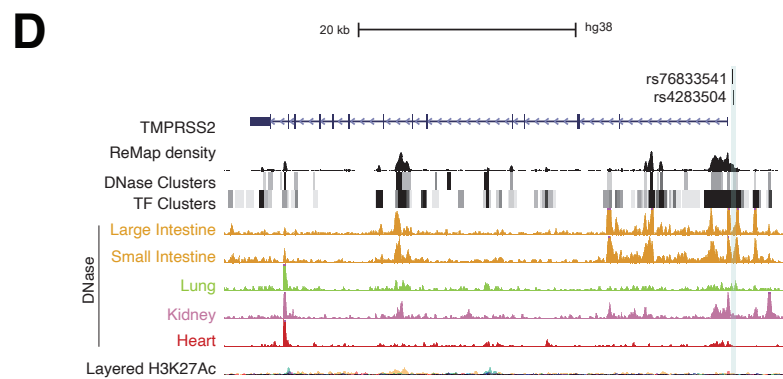
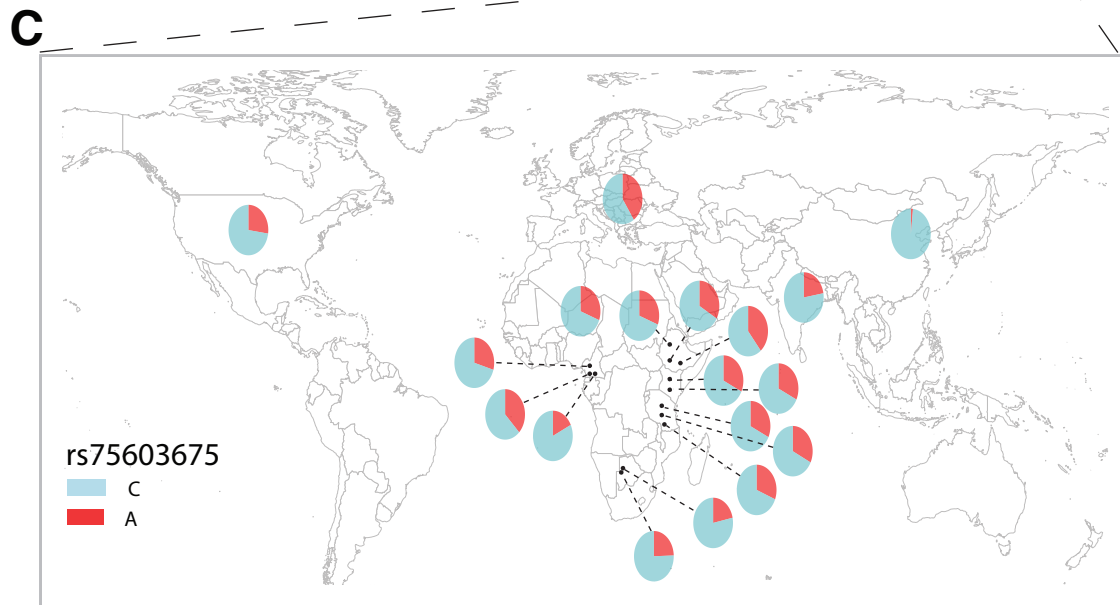
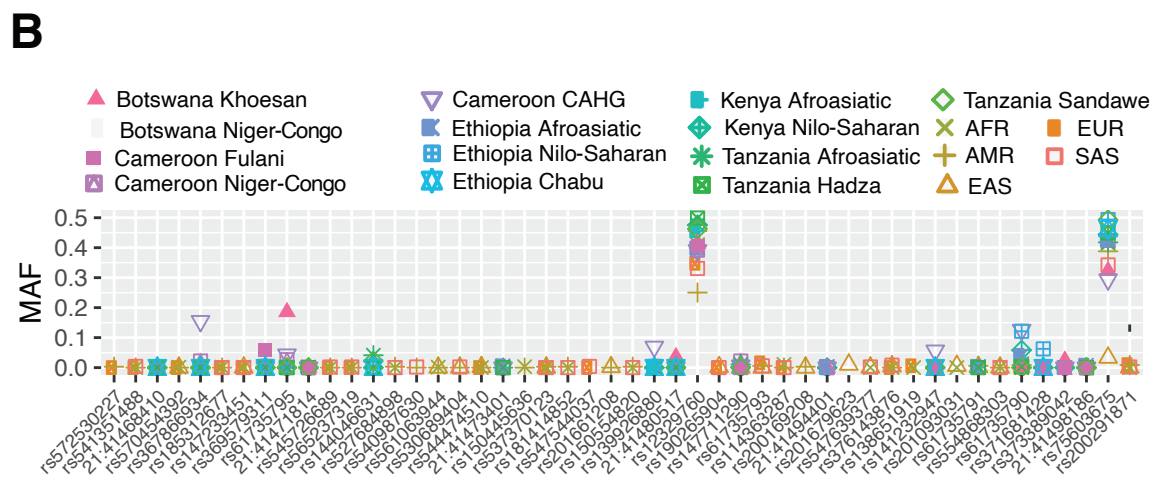
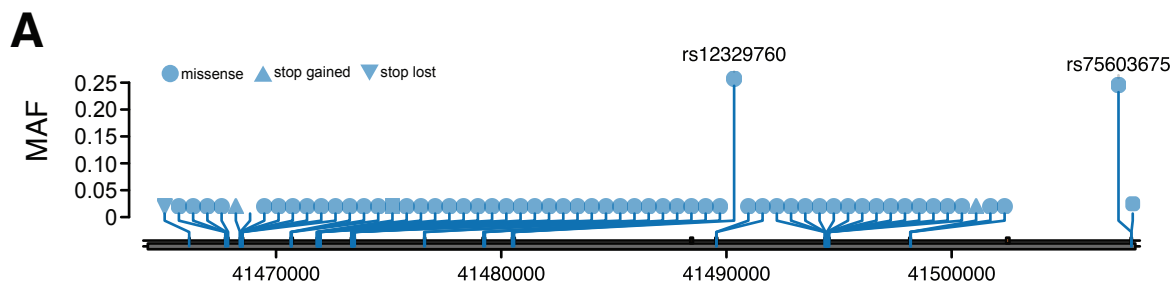
Figure 7. Genetic variation at *LY6E* and its disease association.

(A) Location of coding variants and their minor allele frequency (MAF) at *LY6E* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The MAF of variant rs111560737 at *LY6E* in diverse global ethnic groups. Each pie denotes frequencies of alleles in the corresponding population. (D) Three regulatory eQTLs identified at *LY6E*. RNA Pol2 ChIA-PET data and DNase-seq data of large intestine, small intestine, lung, kidney and heart are from ENCODE³². (E) Gene-based association result between coding variants at *LY6E* and 12 disease classes. The disease classes are shown on the x-axis and the y-axis represents the p-values. EA, European Ancestry; AA, African American ancestry. (F) PheWAS plot of the three eQTL associated with *LY6E* and ~1800 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 dot represents an eQTL and the direction of effect of the association. The red dashed line denotes the 0.0001 cutoff, and the blue dashed line represents the 0.001 cutoff.









A*TMPRSS2*

	Fixed	Poly.
Non-Syn	13	48
Syn	2	45

OR = 6.1

P-Val = 0.009

B