Supplementary Information for

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

Chao Zhang[1¶], Anurag Verma[1,14¶], Yuanqing Feng[1¶], Marcelo C. R. Melo[2], Michael McQuillan[1], Matthew Hansen[1], Anastasia Lucas[1], Joseph Park[1], Alessia Ranciaro[1], Simon Thompson[1], Meghan A. Rubel[1], Michael C. Campbell[3], William Beggs[1], Jibril Hirbo[4], Sununguko Wata Mpoloka[5], Gaonyadiwe George Mokone[6], Regeneron Genetic Center[7], Thomas Nyambo[8], Dawit Wolde Meskel[9], Gurja Belay[9], Charles Fokunang[10], Alfred K. Njamnshi[11], Sabah A. Omar[12], Scott M. Williams[13], Daniel Rader[1], Marylyn D. Ritchie[1], Cesar de la Fuente Nunez[2], Giorgio Sirugo[14]*, Sarah Tishkoff[1,15]*.

¶These authors contributed equally to this work

Corresponding Authors
*Correspondence: giorgio.sirugo@pennmedicine.upenn.edu, tishkoff@pennmedicine.upenn.edu

This PDF file includes:

Supplementary Methods

Genomic data

The genomic data used in this study were from three sources: the Africa 6K project (referred to as the the "African Diversity" dataset) which is part of the TopMed consortium[1], the 1000 Genomes project (1KG)[2], and the Penn Medicine BioBank (PMBB). From the Africa 6K project, a subset of 2012 high coverage (>30X) whole genome sequences of ethnically diverse African populations (Figure S1) were included. The African samples were collected from individuals from five countries (Cameroon, Ethiopia, Kenya, Botswana and Tanzania), speak languages belonging to four different language families spoken in Africa (Afroasiatic, Nilo-Saharan, Niger-Congo, and Khoesan) and have diverse subsistence practices (*e.g.,* hunter-gatherers, agriculturalists, and pastoralists). IRB approval was obtained from the University of Maryland and the University of Pennsylvania. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions prior to sample collection: COSTECH, NIMR and Muhimbili University of Health and Allied Sciences in Dar es Salaam, Tanzania; the University of Botswana and the Ministry of Health in Gaborone, Botswana; the University of Addis Ababa and the Federal Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee; and the Cameroonian National Ethics Committee and the Cameroonian Ministry of Public Health. Whole genome sequencing (WGS) was performed to a median depth of 30X using DNA isolated from blood, PCR-free library construction and Illumina HiSeq X technology, as described elsewhere[1]. In the 1KG data set, 2504 genome sequences from phase 3[2] were included in our analysis.

The PMBB participants were recruited through the University of Pennsylvania Health System by enrolling at the time of clinic visit. Patients participate by donating either blood or a tissue sample and allowing researchers access to their EHR information. This academic biobank has DNA extracted from blood that has been genotyped using an Illumina Infinium Global Screening Array-24 Kit *version 2* and whole exome sequencing (WES) using the IDT xgen exome research panel v1.0. The study cohort consisted of 15,977 individuals total, with 7,061 of European ancestry (EA) and 8,916 of African ancestry (AA) (Table S1). Genetic ancestry of these samples was determined by performing quantitative discriminant analyses (QDA) on

eigenvectors. The 1000 Genomes datasets with super population ancestry labels (EUR, AFR, EAS, SAS, Other) were used as QDA training datasets to determine the genetic ancestry labels for the PMBB population. We identified and removed 117 related individuals using a kinship coefficient of 0.25.

Variant annotations

We used Ensembl Variant Effect Predictor (VEP) for variant annotations[3]. VEP classifies variants into 36 types including non-synonymous, synonymous, and stop loss variants. For pathogenicity predictions, we used CADD[4], SIFT[5], PolyPhen[6] , Condel[7], and REVEL scores in Ensembl. For whole-genome sequencing datasets (African Diversity and 1KG), we annotated genetic variants at *ACE2* (chrX:15,561,033-15,602,158*)*, *TMPRSS2* (chr21:41,464,305-41,531,116), *DPP4* (chr2:161,992,245-162,074,215) and *LY6E* (chr8:143,017,982-143,023,832), and 10 Mb flanking these genes. For whole-exome genomes from the PMBB dataset, annotations were restricted to coding regions only. For gene-based association analysis using the PMBB dataset, we collapsed all the predicted non-synonymous variants with REVEL score > 0.5 and putative loss of function variants (pLOFs) with MAF < 0.01. We assigned variants as pLoFs if the variant was annotated by VEP as start_lost, splice_donor_variant, splice_acceptor_variant, frameshift_variant, stop_gained, stop_lost.. All genome coordinates followed the GRCh38 assembly.

Characterization of putative regulatory variation

We identified regulatory variants likely to impact the target genes. For all four genes (*ACE2*, *TMPRSS2*, *DPP4* or *LY6E*), we extracted the variants located within ±10kb distance to their TSS as well as enhancers supported by RNA Pol2 ChIA-PET data from ENCODE[8]. These variants were further filtered by overlapping with DNase-seq and ChIP-seq peaks from Roadmap[9], ENCODE[8], Remap2 [10]; or overlapping with significant single-tissue expression quantitative trait locus (eQTLs) (P-value<0.001) from the GTEx V8 database [11]. We visualized the location of these regulatory and eQTL variants using the UCSC genome browser and highlighted the variants using Adobe Illustrator.

Electronic Health Record Phenotypes

In this analysis, we focused on the phenotypes characterized as primary organ dysfunctions in the early studies on COVID-19. Broadly, we centered our analyses on these four broad clinical conditions/phenotypes: respiratory injury/failure, acute liver injury/failure, acute cardiac injury/failure, and acute kidney injury/failure. These disease classes are well characterized in human disease ontologies such as Monarch Disease Ontology (MONDO). MONDO merges multiple disease resources such as SNOMED, ICD-9, and ICD-10. We leveraged the existing mappings between ICD-9/10 codes (which are how the data are coded in the EHR) and the MONDO disease classes for the conditions described above. We identified 12 MONDO classes that are closely related to four conditions of interest. By using ICD-9 and ICD-10 data from the EHR of the PMBB participants, we mapped the ICD codes to 12 MONDO disease classes. Details on the ICD code mapping to MONDO disease classes are provided in Dataset S5. Individuals were defined as cases if they had at least one instance of any ICD code mapped to a MONDO disease class or as controls if they had no instance of the code in that disease class. A clinical expert on our team manually reviewed the MONDO and ICD-9/10 mappings.

We also used EHR phenotypes defined by groupings of ICD-9 and ICD-10 codes into clinically relevant groups, called phecodes, used in prior PheWAS studies[12]. Individuals with two or more instances of a phecode were defined as cases, whereas those with no instance of a phecode were defined as controls. Individuals with only one instance were excluded for that phecode. A total of 1860 phecodes were included in the study.

Additionally, we extracted data on 34 clinical laboratory measures for PMBB participants from the EHRs. We derived a median value for each laboratory measure based on all clinical tests ever done within the Penn Medicine health system. Any measurement value that falls more than three standard deviations from the normal were labeled as outliers and removed.

Association Testing

We used the R SKAT package for conducting a gene-based dispersion test and Biobin[13; 14] for gene burden analysis. Here, multiple genetic variations in a gene region were collapsed to generate a gene burden/dispersion score and regression methods were used to test for association between the genetic score and a phenotype or trait. We applied two statistical tests: a) a burden test (i.e. the cumulative effect of

rare variants in a gene) that uses logistic regression and b) a sequence kernel association test (SKAT)[15]. Thus, it can compute effect estimates but may suffer from loss of power when gene variants have effects in opposite directions (i.e., protective, and higher risk variants). This limitation can be overcome by parallel analysis with SKAT, a powerful approach to model mixed effect variants. However, this approach does not provide effect estimates. Therefore, we reported outcomes using both methods. Ancestry specific analysis of gene-based tests identified seven associations in African ancestry (AA) and three associations in European ancestry (EA) populations that reached statistical significance levels after multiple hypothesis correction ($p < 1 \times 10^{-04}$) for the SKAT model. None of the gene burden models reached a significance level of $p < 1 \times 10^{-04}$. The effect size from the logistic regression model was used to indicate a protective or increased risk effect on disease phenotype. We performed three separate SKAT and burden analysis for 12 MONDO disease classes (Dataset S5), 1860 phecode, and 34 clinical lab measures. Briefly, the variants annotated as non-synonymous (REVEL score >= 0.5) and pLoFs within each of the four candidate genes were collapsed into their respective gene regions *(ACE2, TMPRSS2, DPP4 and LY6E).* For both statistical dispersion and burden tests, models were adjusted by the first four principal components of ancestry, sex, and decade of birth. For multiple hypothesis correction, a conservative Bonferroni adjustment was used to derive a significant p-value threshold (p-value < 0.0001). We also performed a univariate statistical test for each of the rare variants from these four candidate gene regions to study the effects of each single nucleotide variant (SNV) on the disease phenotype.

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

The fast response from the structural biology community to the COVID-19 pandemic led to the exceptionally fast determination and publication of over 900 as of Jan. 2021 (https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true) protein structures related to SARS-Cov-2. Using experimentally determined structures of the ACE2 protein complexed with the receptor binding domain (RBD) of SARS-CoV-2 spike glycoprotein, we assessed possible impacts of nonsynonymous coding variants on the ACE2- binding interface with SARS-CoV-2-RBD. Among the multiple entries available in the Protein Data Bank (PDB), we chose to focus on the

structure of the full-length human ACE2 bound to RBD (PDB ID 6M17 [16]) determined with Cryo-Electron Microscopy (cryo-EM), as it presented multiple advantages to our study. Unlike other PDB entries that only feature sections of ACE2, usually focusing on the part of the enzymatic domain responsible for RBD binding, 6M17 presents the full length ACE2 in its dimeric form. This allowed us to identify the 3D protein location of all nonsynonymous coding variants identified in this study. Moreover, ACE2 was expressed in a human cell line, maintaining important glycosylation sites and allowing the cryo-EM structure to be used to identify their positions and compositions [16]. All structural analysis and figures were prepared using VMD [17].

Detecting signatures of natural selection

We used two methods (the McDonald–Kreitman test [18] and the dN/dS test [19]) to test for signals of selection acting on the four candidate genes over long time scales, and two methods (EHH and iHS) to detect recent (e.g. last ~10,000 years before present) signatures of positive selection .

For the McDonald–Kreitman test (MK-test) [18], we set up a two-way contingency table to statistically compare the number of nonsynonymous (Dn) and synonymous (Ds) fixed differences between humans and chimpanzees with the number of nonsynonymous (Pn) and synonymous (Ps) polymorphisms among individuals within a population. Based on neutral theory, the ratio of nonsynonymous to synonymous changes should be constant throughout evolutionary time, i.e. the ratio observed among individuals within species (Pn/Ps) should be equal to the ratio observed between species (Dn/Ds). Under a hypothesis of positive selection in the hominin lineage after divergence from our closest ancestor, the chimpanzee, the ratio of nonsynonymous to synonymous variation between species is expected be larger than the ratio of nonsynonymous to synonymous variation within species (i.e. Dn/Ds > Pn/Ps). If there is positive diversifying selection among human populations but conservation of fixed differences between species, the ratio of nonsynonymous to synonymous variation between species should be lower than the ratio of nonsynonymous to synonymous variation within species (i.e. Dn/Ds < Pn/Ps). The chimpanzee sequence (Clint_PTRv2/panTro6) used in the analysis was obtained from the UCSC genome browser. We used Fisher's exact test to detect significance of the MK-test. We used transcripts ENST00000252519.8, ENST00000398585.7,

ENST00000360534.8, ENST00000521003.5 to calculate Dn, Ds, Pn and Ps for *ACE2*, *TMPRSS2*, *DPP4* and *LY6E*, respectively.

We also used the ratio of substitution rates at non-synonymous and synonymous sites (dN/dS) to infer selection pressures on the four candidate genes, as the dN/dS ratio has more power to detect recurrent positive selection[20]. This measure quantifies selection pressures by comparing the rate of substitutions at synonymous sites (dS), which are neutral or close to neutral, to the rate of substitutions at non-synonymous sites (dN), which are more likely to experience selection. The dN/dS estimation used here follows Nei et al[19]. The number of synonymous sites, s, for codon i in one protein is given by

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

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

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

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

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

and

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

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

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

and

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

In our analysis, for each population, we estimated the total number of synonymous (S_d) and non-synonymous (N_d) differences, and then calculated *dN*/*dS*. If *dN*/*dS* is larger than one, it suggests positive diversifying selection influencing variation at the gene. If *dN*/*dS* is less than one it suggests the gene is evolutionary conserved. We used both the chimpanzee sequence (Clint_PTRv2/panTro6) and the human reference genome, combing our WGS data, to calculate *dN*/*dS,* separately (Dataset S2). We used the results based on the human reference in the main text.

Genomic regions that have undergone recent positive selection are characterized by extensive linkage disequilibrium (LD) on haplotypes containing the mutation under selection. We used the extended haplotype homozygosity (EHH)[21] and the integrated Haplotype Score (iHS) methods[22] to identify regions with extended haplotype homozygosity greater than expected under a neutral model. iHS is based on the differential levels of LD surrounding a positively selected allele compared to the ancestral allele at the same position. For the iHS analyses, we normalized scores with respect to all values observed at sites with a similar derived allele frequency within 40Mb regions flanking the four target genes. SNPs with absolute values larger than 2 are within the top 1% of observed values and are marked as extreme SNPs or candidate SNPs under positive selection. An extreme positive iHS score (iHS > 2) means that haplotypes on the ancestral allele background are longer compared to the derived allele background. An extreme negative iHS score (iHS < -2) means that the haplotypes on the derived allele background are longer compared to the haplotypes associated with the ancestral allele. All of the above processes were performed with selscan[23]. SNPs with predicted functional effects on protein structure that are identified as potential targets of selection (stop_lost, missense_variant, start_lost, splice_donor_variant, inframe_deletion, frameshift_variant, splice_acceptor_variant, stop_gained, or inframe_insertion) are highlighted. Haplotypes were phased by Eagle V2.4.1[24]. The ancestral state of alleles was obtained from Ensembl.

To identify potential regulatory variants under selection, we overlapped SNPs showing signatures of selection using iHS with DNase I hypersensitivity peak clusters from ENCODE [8] and eQTLs from GTEx v8. [11]. The overlapped SNPs were uploaded to the UCSC browser for visualization. The ChIP-seq density dataset was obtained from http://remap.univ-amu.fr/ [9]. DNase-seq and ChIP-seq clusters, layered H3K4Me3 (often found near Promoters), H3K4Me1 and H3K27Ac (often found near Regulatory Elements) data are from ENCODE[8]. The DNase-seq tracks of large intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle were from ENCODE [25].

We used d_i statistics to identify SNPs that are highly differentiated in allele frequency between populations based on unbiased estimates of pairwise F_{ST}[26]. The d_i statistics were performed cross the 40Mb regions. If the candidate SNP was within the top 5% of the 40Mb regions in a specific population, the SNP was considered as a variant showing significant differentiation between the target population and other populations. These variants are candidate SNPs that show signals of local adaptation.

Haplotype networks were constructed by PopART[27] using the built-in minimum spanning algorithm.

Supplementary Figures

Figure S1. Geographic information of ethnically diverse global populations included in our study. Cameroon CAHG (Cameroon Pygmy): the Central African hunter-gatherers (CAHG) from Cameroon; EUR, European populations; EAS, East Asian populations; SAS, South Asian populations; AFR, African Niger-Congo populations; AMR, Native American populations; AA, African American populations; EA, European American populations. AA and EA are from the PMBB dataset, and EUR, EAS, SAS, AMR and AFR are from the 1000 genomes project. All other populations are from the TOPMed Africa6K project, and they are from five countries (Cameroon, Ethiopia, Kenya, Botswana, and Tanzania), belonging to four major different language families (Afroasiatic, Nilo-Saharan, Niger-Congo (or Niger-Kordofanian), and Khoesan). The Chabu (or Sabue) population from Ethiopia, the Hadza (or Hadzabe) population from Tanzania, the Sandawe population from Tanzania, and the Fulani population from Cameroon are listed as separate ancestral groups in our studies since their different evolutionary histories with other ethnic groups. The numbers in brackets denote sample sizes.

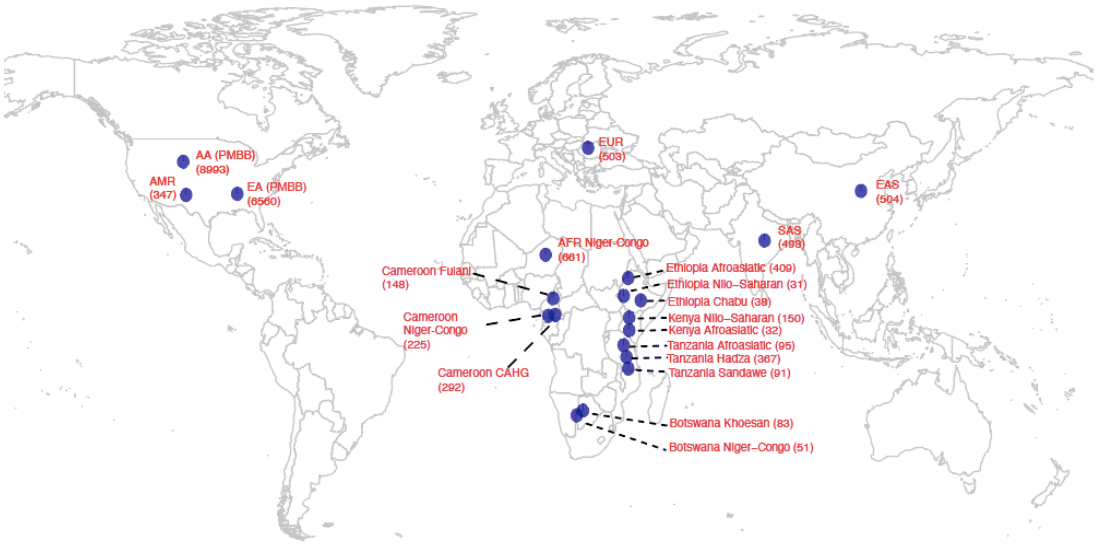


Figure S2. MAF of six eQTLs identified at *ACE2*.

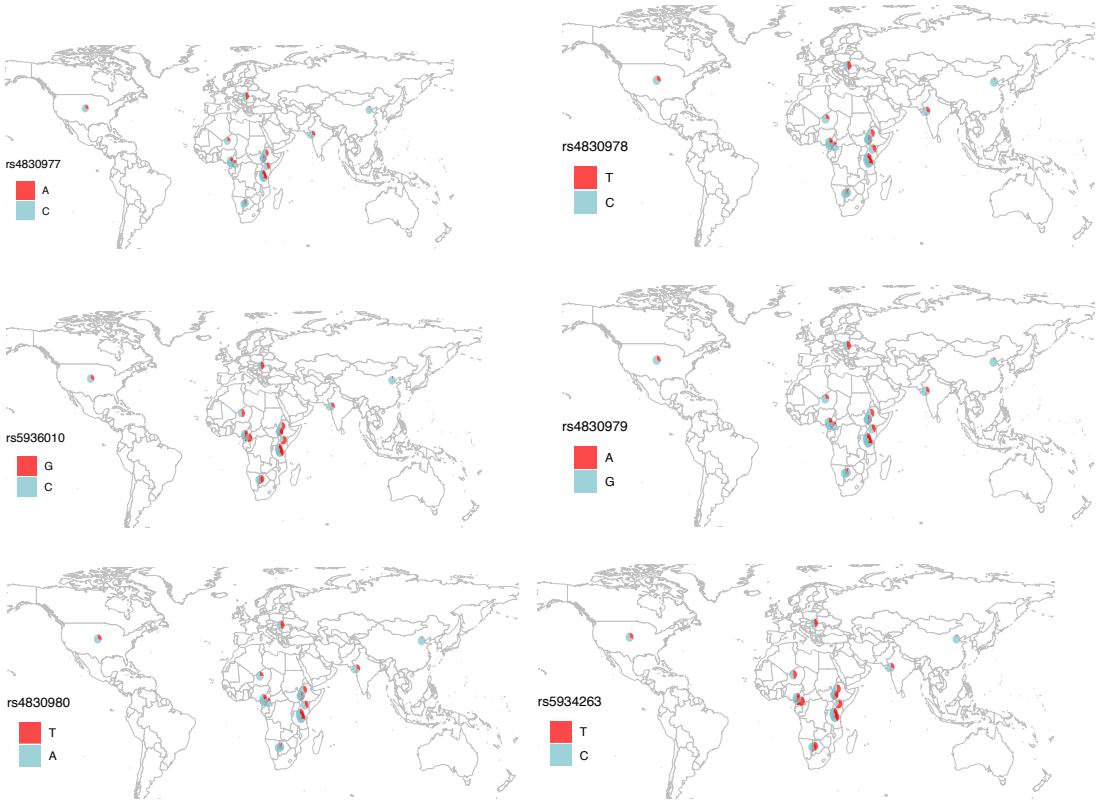


Figure S3. Normalized expression data of the six eQTLs at *ACE2* in frontal cortex from the GTEx database.

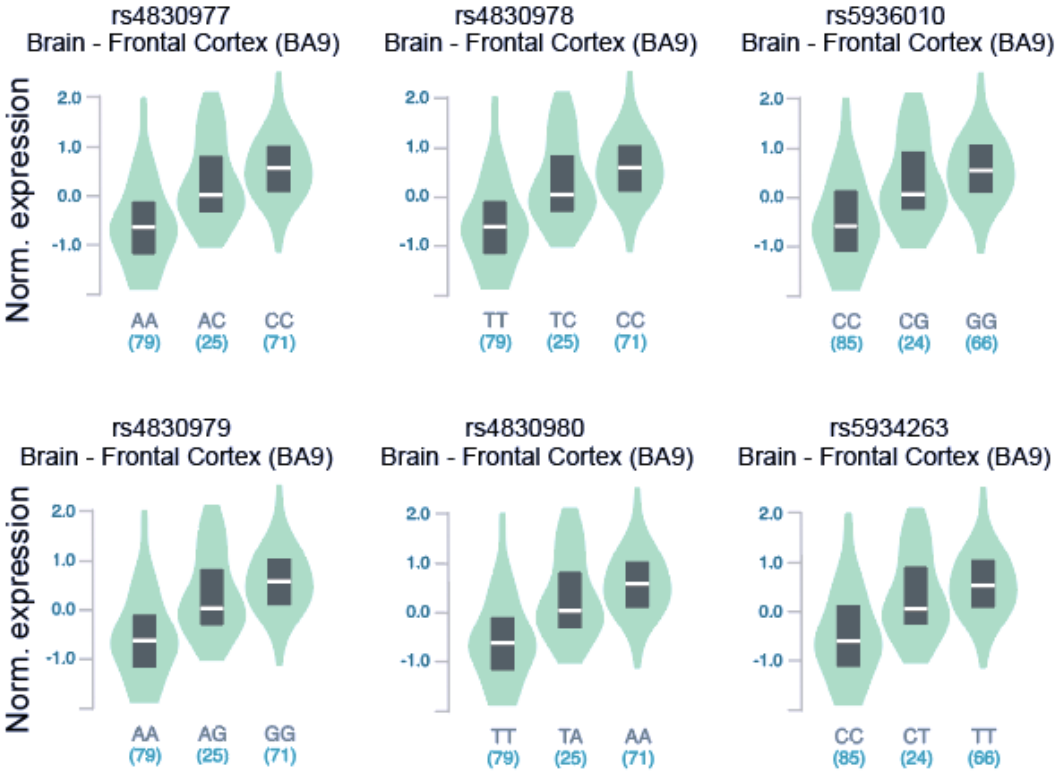


Figure S4. Linkage disequilibrium between six eQTLs at *ACE2*. R^2 were used to measure the LD.

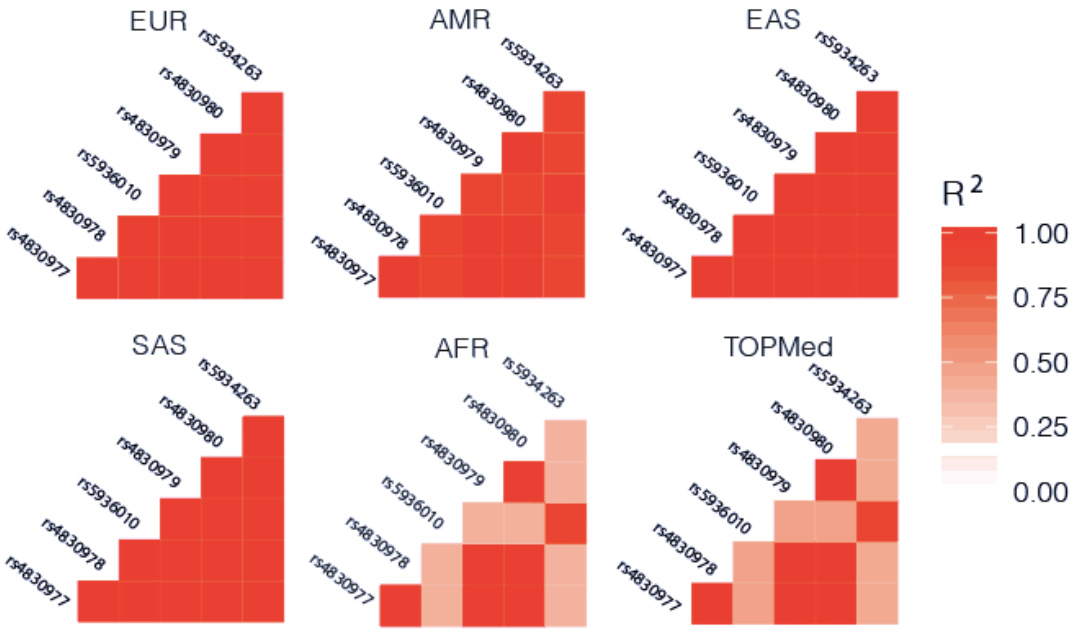


Figure S5. Results of dN/dS test for the four candidate genes. The dN/dS ratios of each ethnic group for *ACE2*, *TMPRSS2*, *DPP4*, and *LY6E* were plotted.

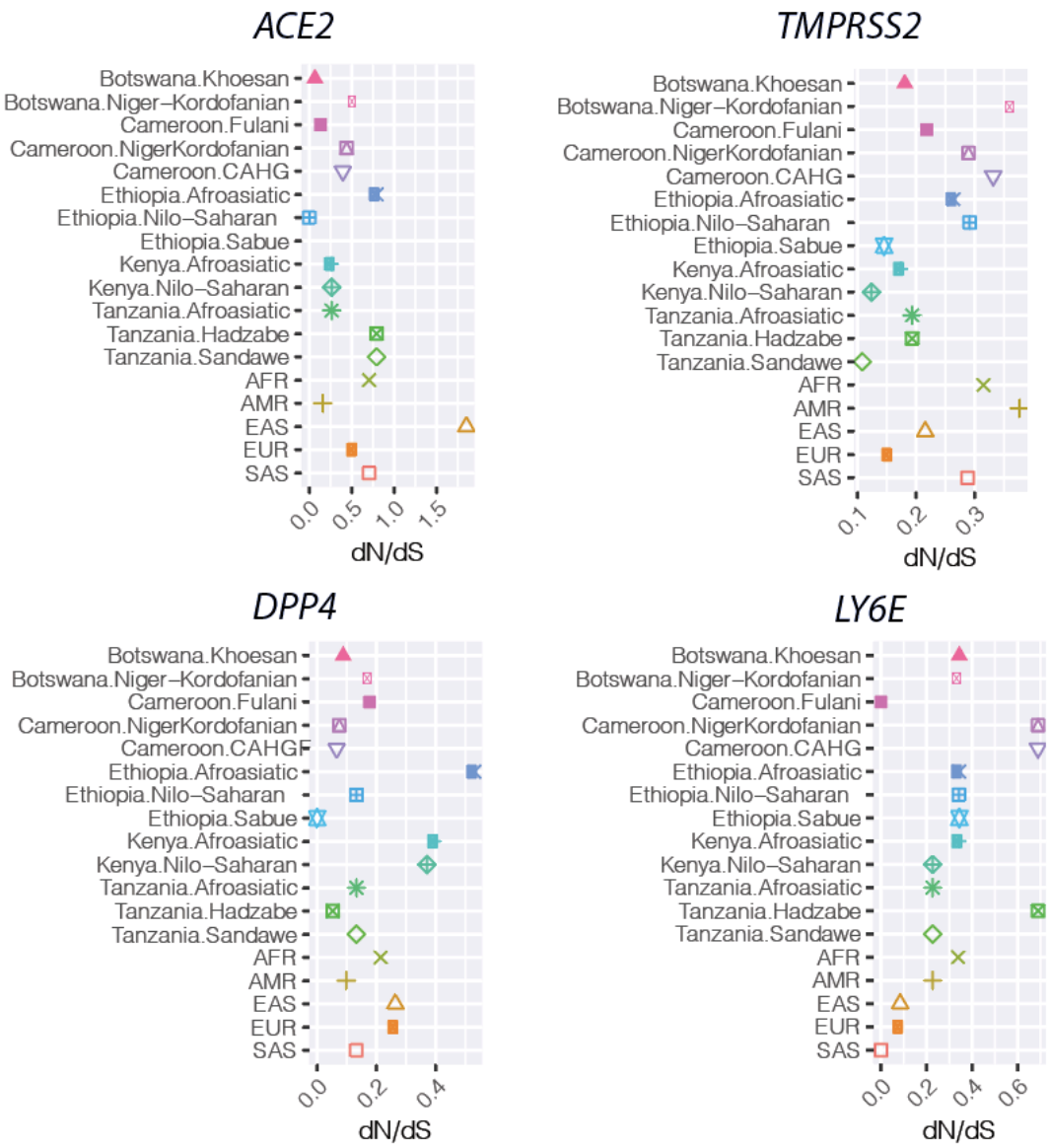


Figure S6. Schematic illustration of the McDonald–Kreitman test. Red dots denote non-synonymous variants and blue dots denote synonymous variants. Divergence of fixed variants mean that these variants were fixed in human lineage compared to the Chimpanzee. Polymorphism variants denotes that these variants were polymorphic within human populations. Dn, the number of divergence non-synonymous variants; Ds, the number of divergence synonymous variants; Pn, the number of polymorphism non-synonymous variants; Ps, the number of polymorphism synonymous variants; OR, odds ratio.

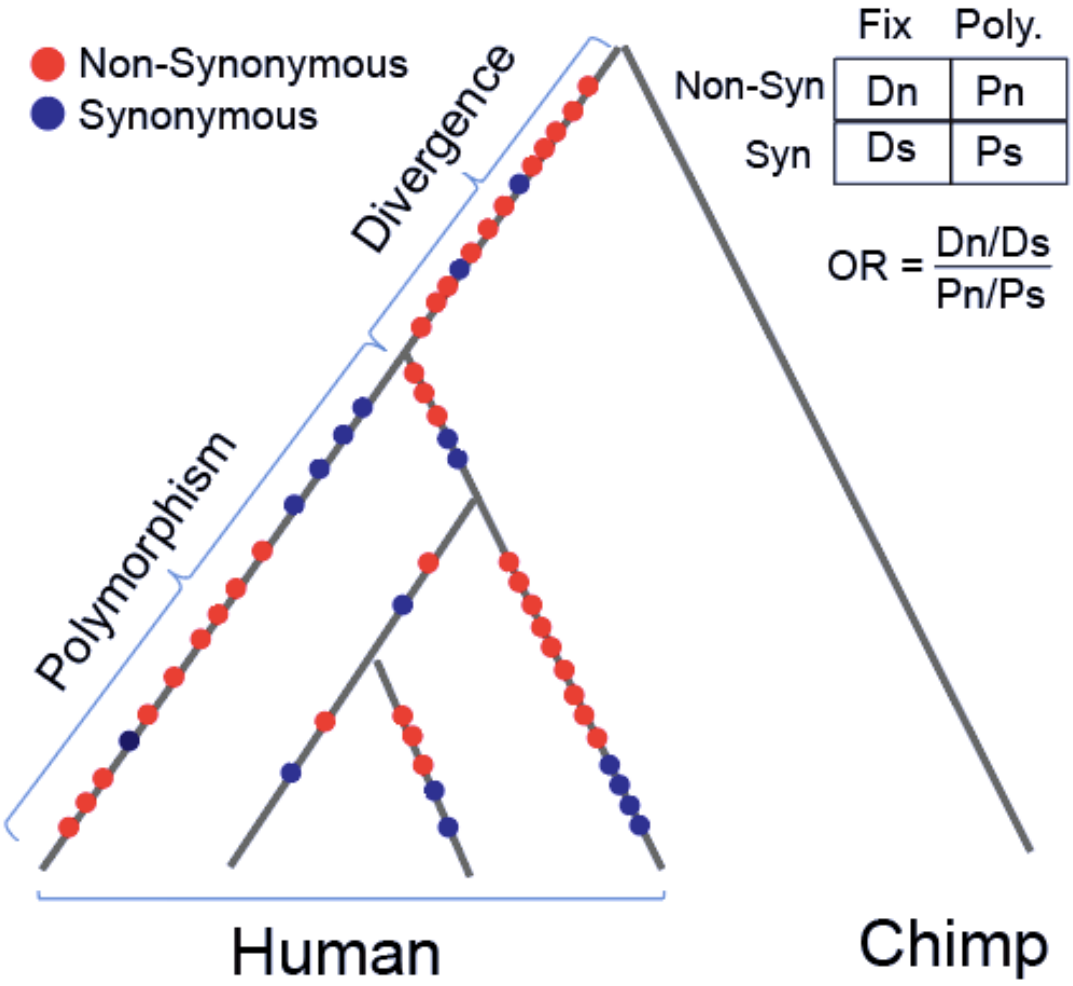


Figure S7. Results of MK-test for four genes. Odds ratios of (Dn/Ds) to (Pn/Ps) of each ethnic group for *ACE2* (A), *TMPRSS2* (B), *DPP4* (C), and *LY6E* (D) were plotted. Significance was tested by Fisher's exact test. No significant P-val was observed at three genes (*ACE2*, *DPP4* and *LY6E*). Odds ratios were not applied (NA) if no non-synonymous variants (Pn) were observed within individuals from a population.

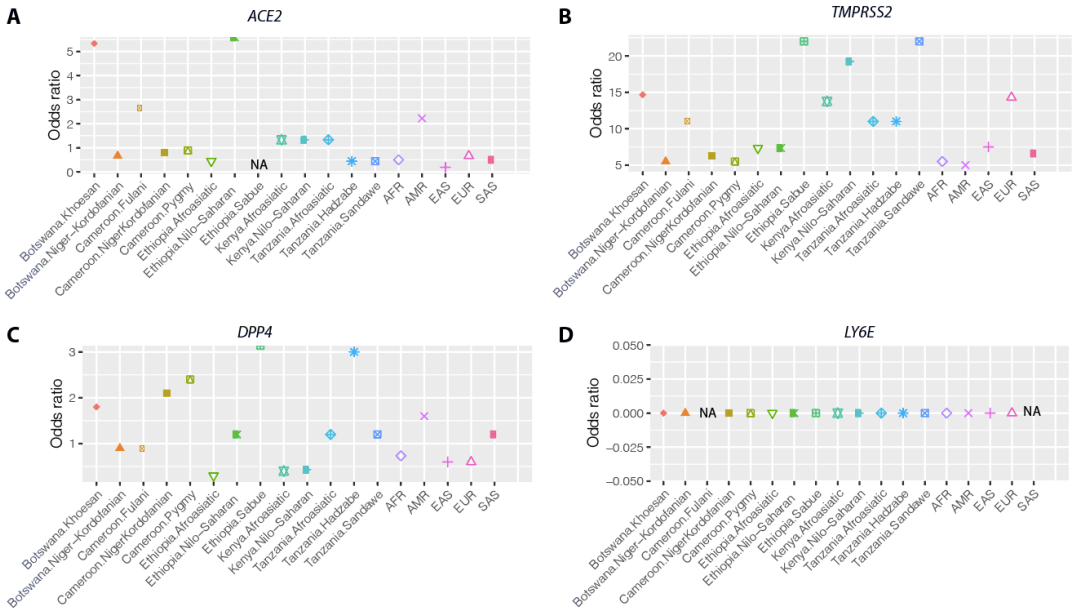


Figure S8. The iHS scores for SNPs within *ACE2* region in each ethnic group. Each dot represents a SNP. Dashed lines denote the empirical cutoff (|iHS|=2). Red dots mean that the corresponding SNPs harbor |iHS| scores > 2.

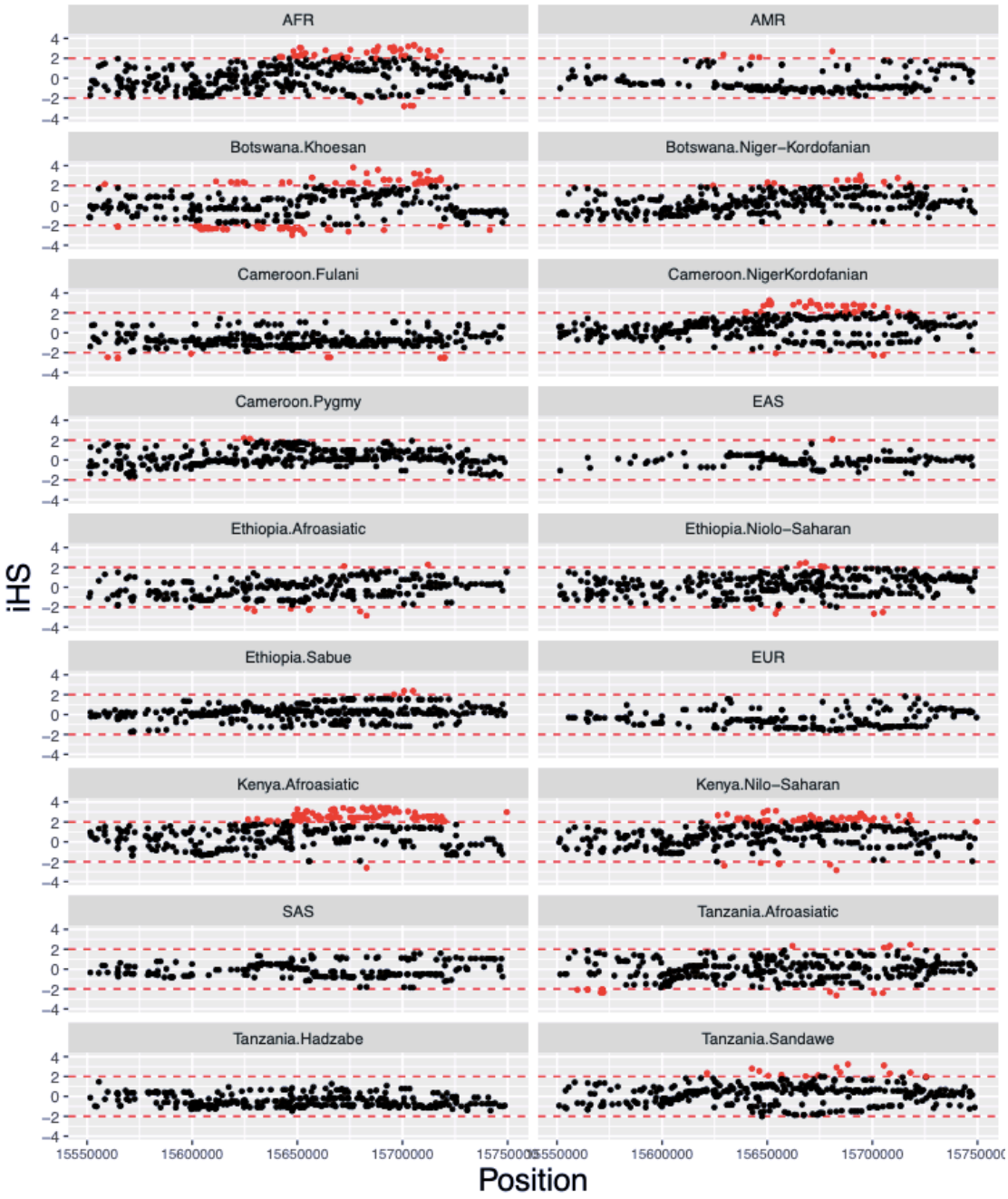


Figure S9. LD pattern for all variants near *ACE2*.

D prime was used to measure the LD. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding positions.

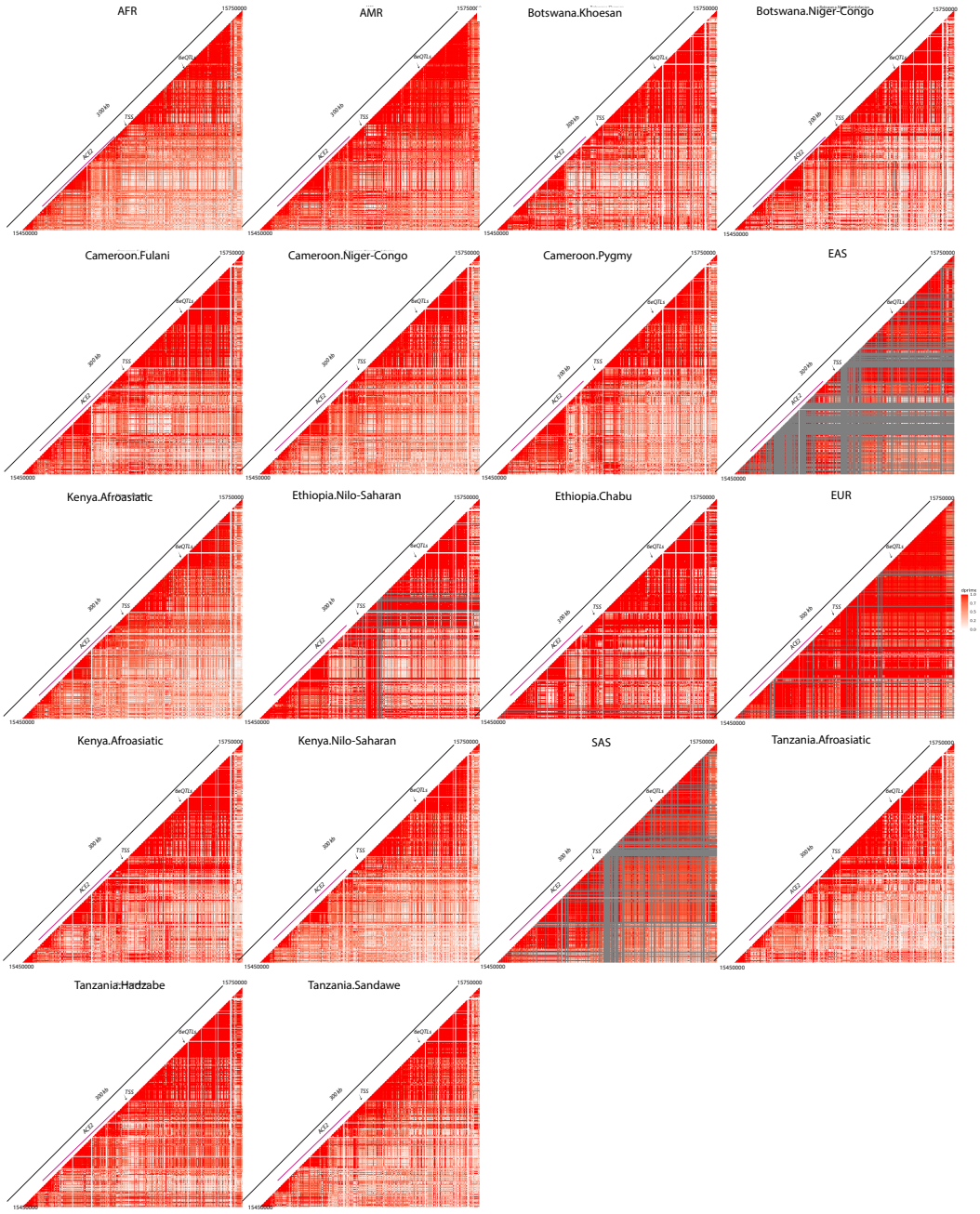


Figure S10. LD pattern between selected variants near *ACE2*. D prime was used to measure the LD. Variants included in the analysis are the four common variants (rs147311723, rs186029035, rs145437639, and rs138390800) identified in Cameroonian CAHG populations, 6 regulatory variants (rs4830977, rs4830978, rs5936010, rs4830979, rs4830980 and rs5934263), and four SNPs (rs150147953, rs2097723, rs4830984 and rs4830986) with significant selection signals at the upstream of *ACE2*. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding positions.

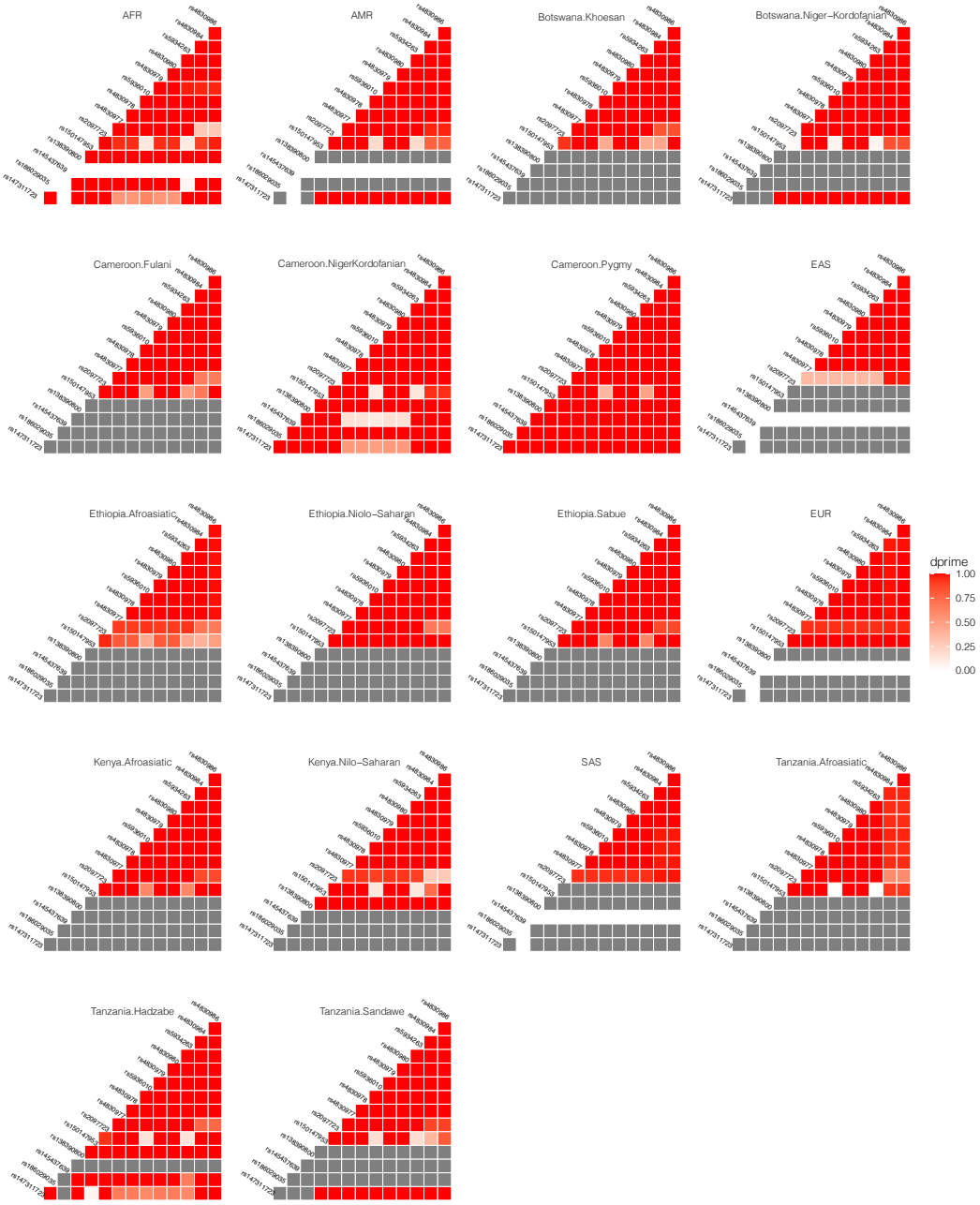


Figure S11. The intersection of SNPs with significant selection signals and regulatory regions at *ACE2*. SNPs with high iHS value (|iHS|>2) near *ACE2* locus overlapping DNase I hypersensitivity peaks from ENCODE (purple) or eQTLs from GTEx v8 (green) are shown in this figure. The SNPs discussed in the main text (rs150147953, rs2097723, rs5936010 and rs5934263) are highlighted with blue shadow. The DNase-seq tracks of large Intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle are also from ENCODE, and their signals are scaled to 1.5.

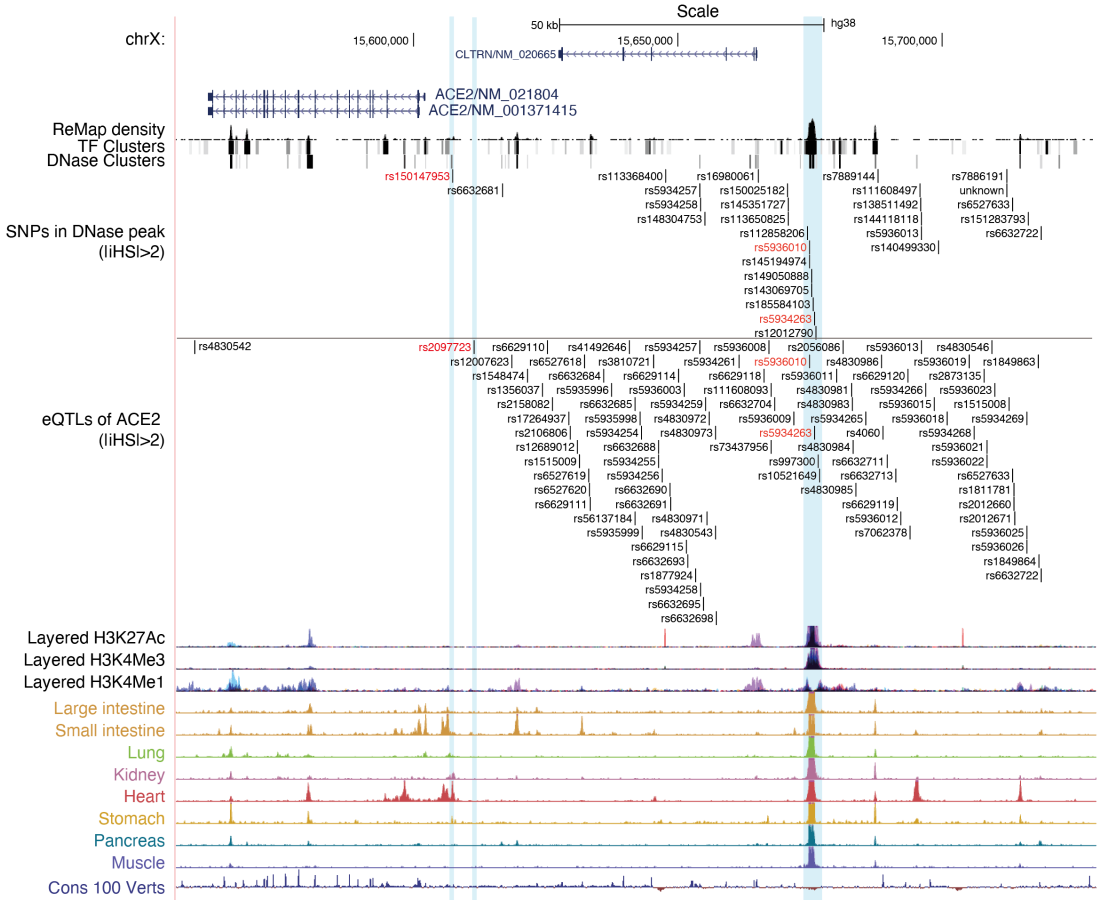


Figure S12. Normalized expression data of the two eQTLs (rs76833541 and rs4283504) at *TMPRSS2* from the GTEx database.

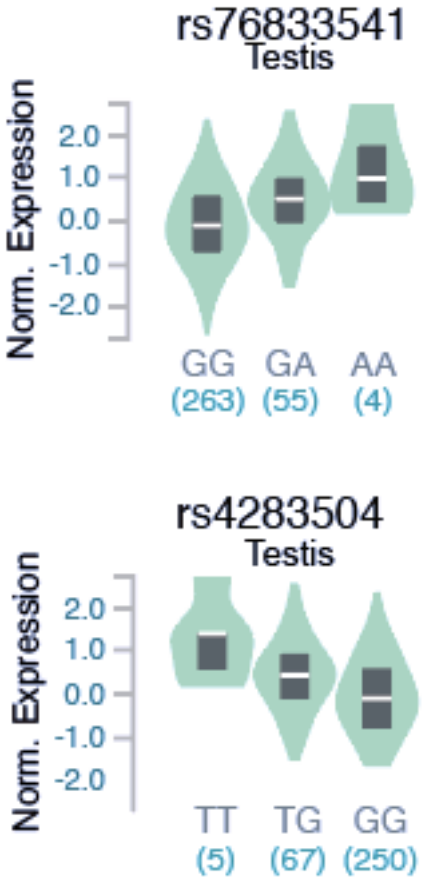


Figure S13. MAF of two regulatory variants (rs76833541 and rs4283504) at *TMPRSS2*.

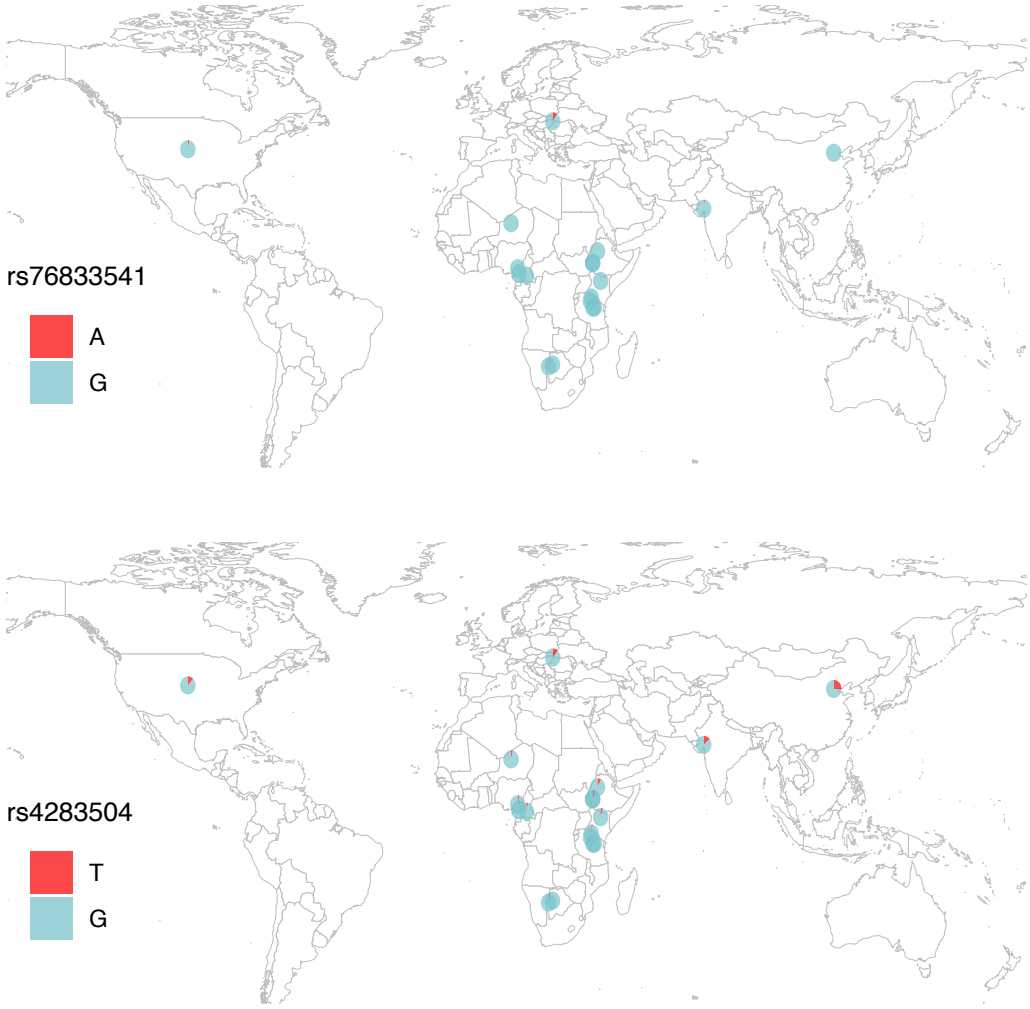


Figure S14. MK-test for transcript ENST00000332149.10 of *TMPRSS2*. There are 11 non-synonymous and 2 synonymous variants in ENST00000332149.10 that were fixed in human populations. These variants are located on different structure domain of *TMPRSS2*: amino acid T9P, A33V, R66C, and M67T in the cytoplasmic region; L87I in the transmembrane region; N107K in the extracellular region; S128N and S141G in the LDL-receptor class A domain; E374Q and T478M in the Peptidase S1 domain; S492G in the last residual); and A99A in the transmembrane region and G258G in the Peptidase S1 domain.

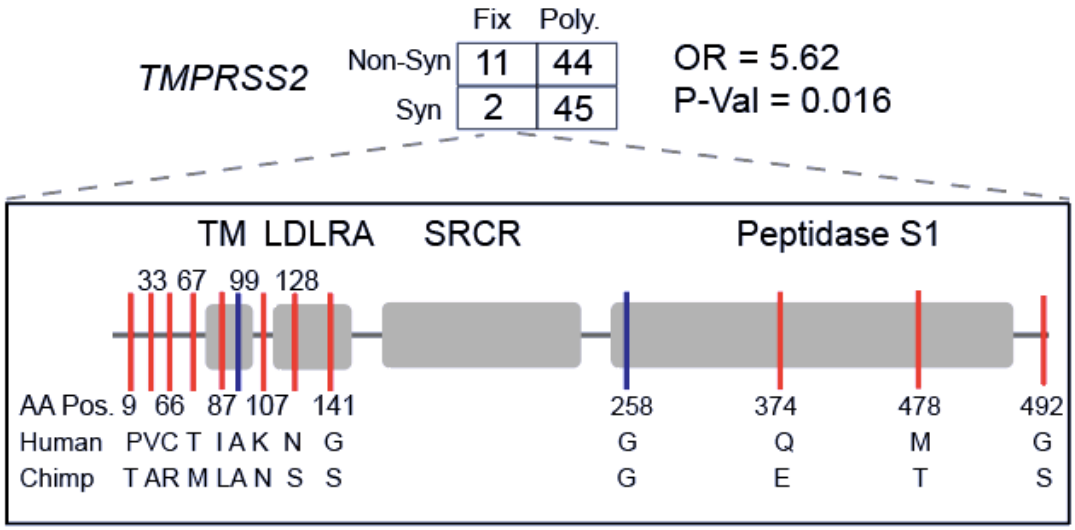


Figure S15. iHS score for SNPs within *TMPRSS2* in each ethnic group. Each dot represents a SNP. Dashed lines denote the empirical cutoff. Red dots mean that the corresponding SNPs harbor significant scores.

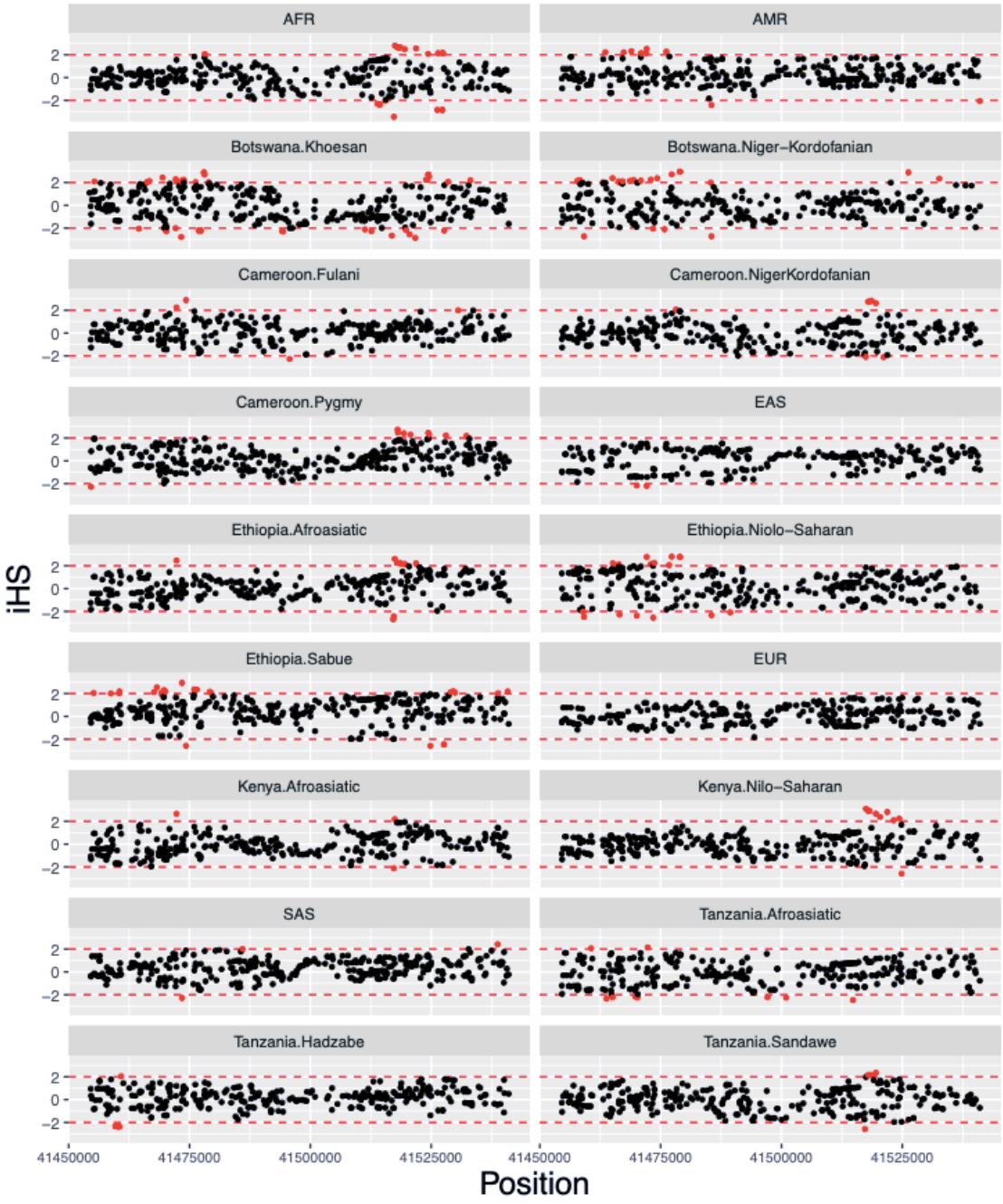


Figure S16. LD pattern between 153 SNPs at *TMPRSS2* showing iHS signals in diverse ethnic groups. D prime was used to measure the LD. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding position.

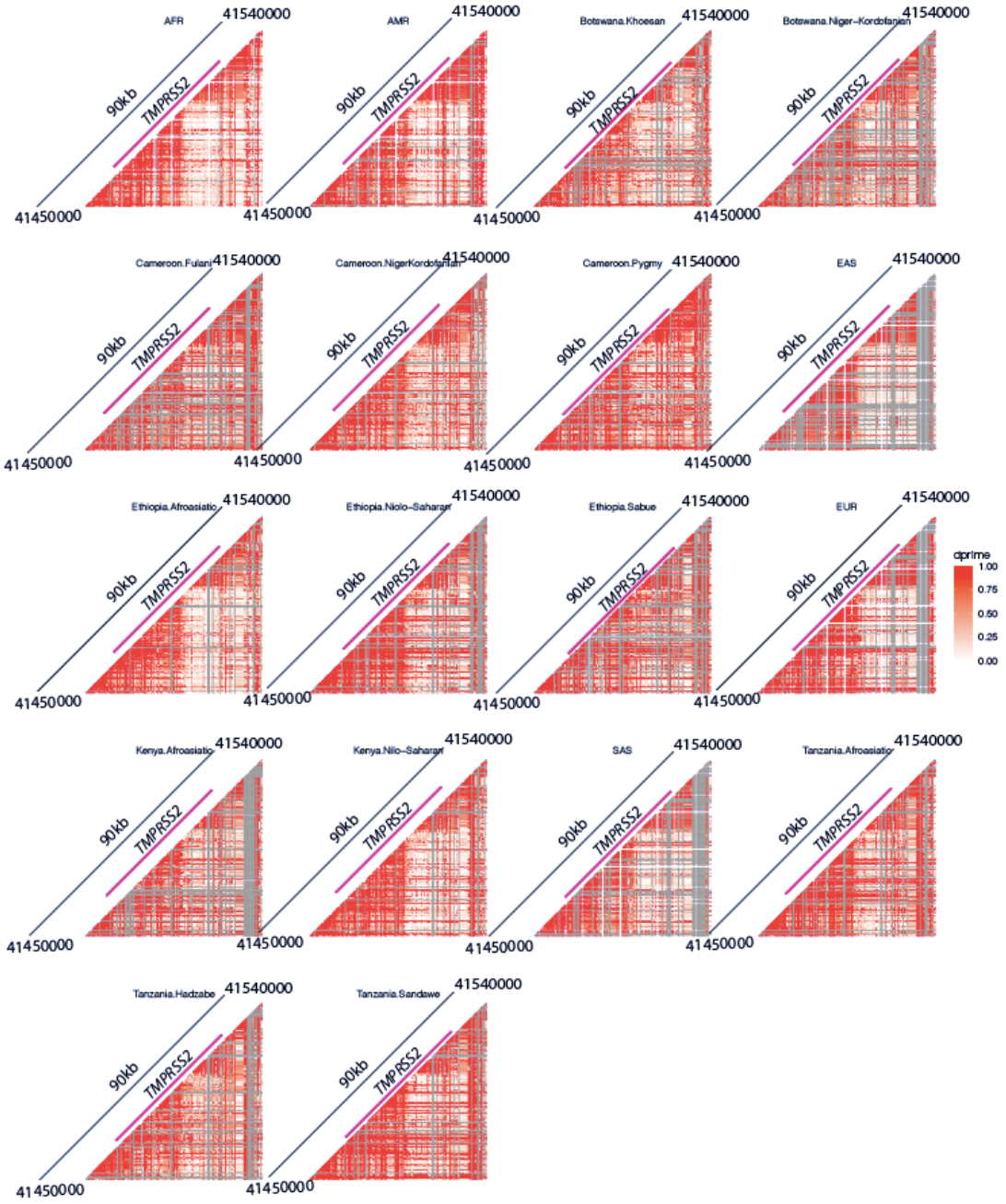


Figure S17. LD pattern between selected variants (rs111870470, rs112306677, rs116170128, rs9636988, rs150969307 and rs73372191) at *TMPRSS2*. D prime was used to measure the LD. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding position.

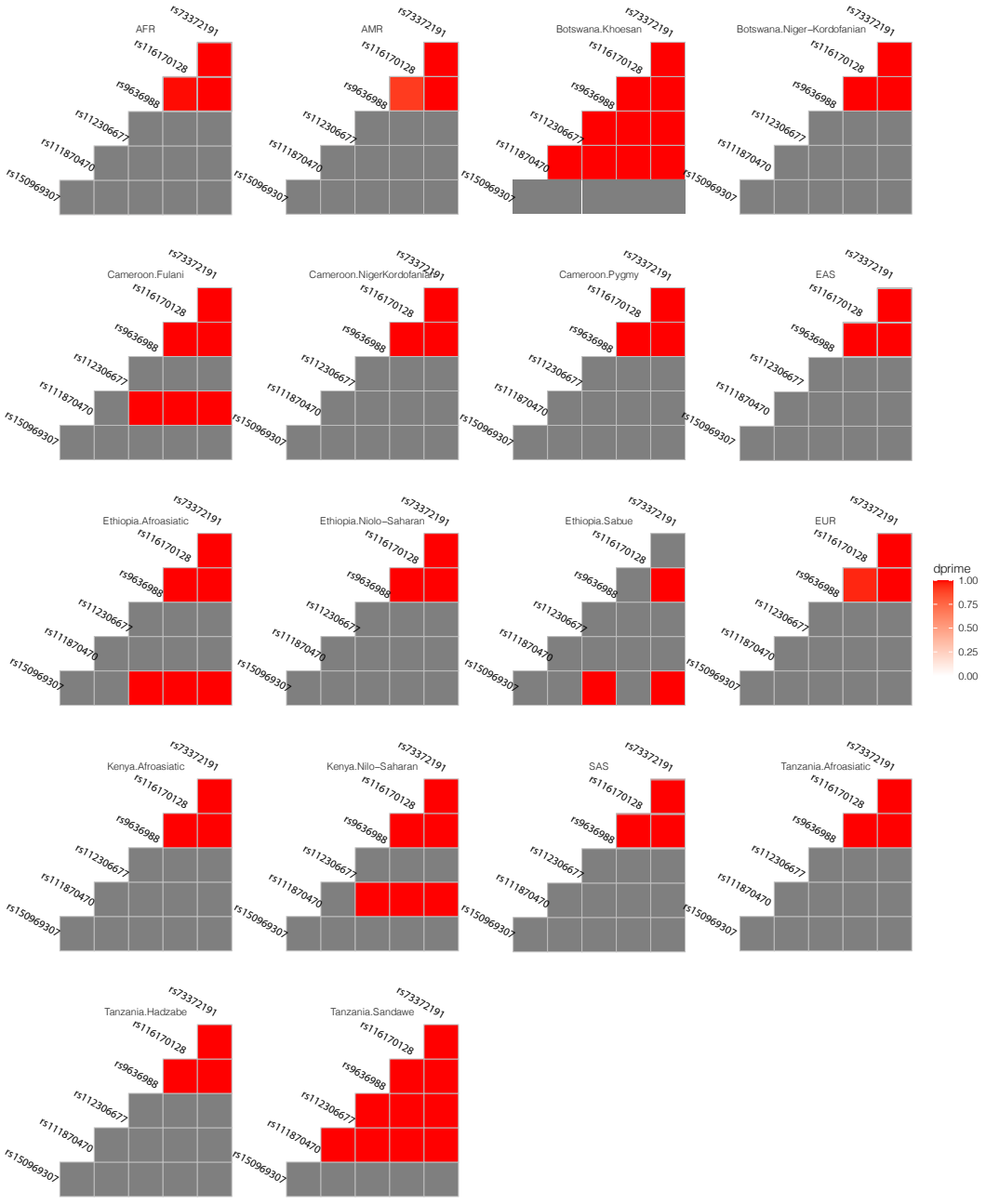


Figure S18. The intersection of SNPs with significant selection signals and regulatory regions at *TMPRSS2*. The SNPs rs435877, rs550390247, rs372713176, rs371744816, rs2838046 and rs777771526 are located in DNase peaks and transcription factor bind sites, and they are highlighted with light blue shadow. TF binding data are from ENCODE. Purple SNPs indicates the one with high iHS value (|iHS|>2) overlapping DNase I hypersensitivity peaks from ENCODE; green SNPs indicates they are significant eQTLs from GTEx v8 (green). The DNase-seq tracks of large Intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle are also from ENCOD, and their signals are scaled to 1.5.

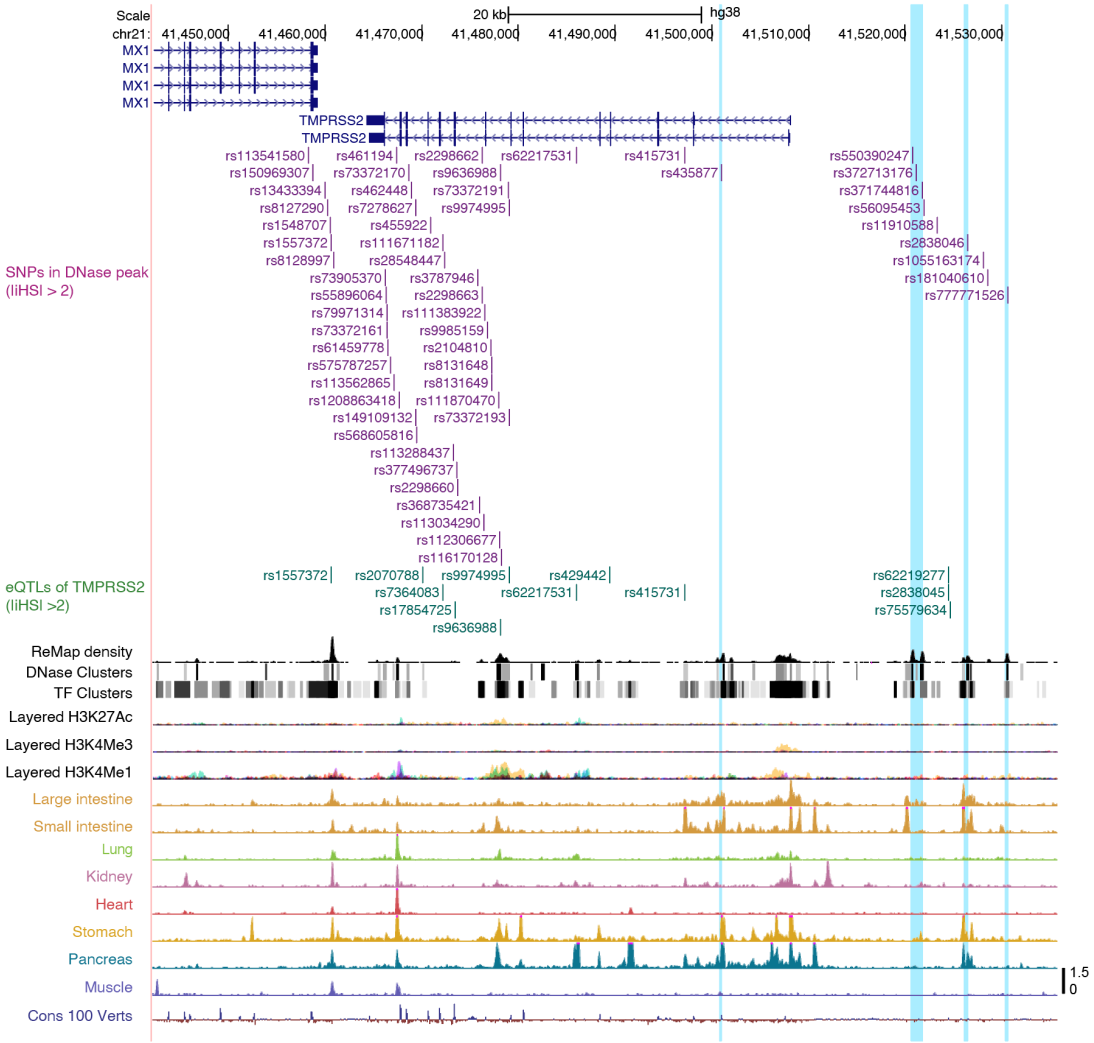


Figure 19. Genetic variation at *DPP4* and its disease association.
(A) Location of coding variants and their minor allele frequency (MAF) at *DPP4* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The MAF of variant within rs129559 at *DPP4* in diverse 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 ENCODE.

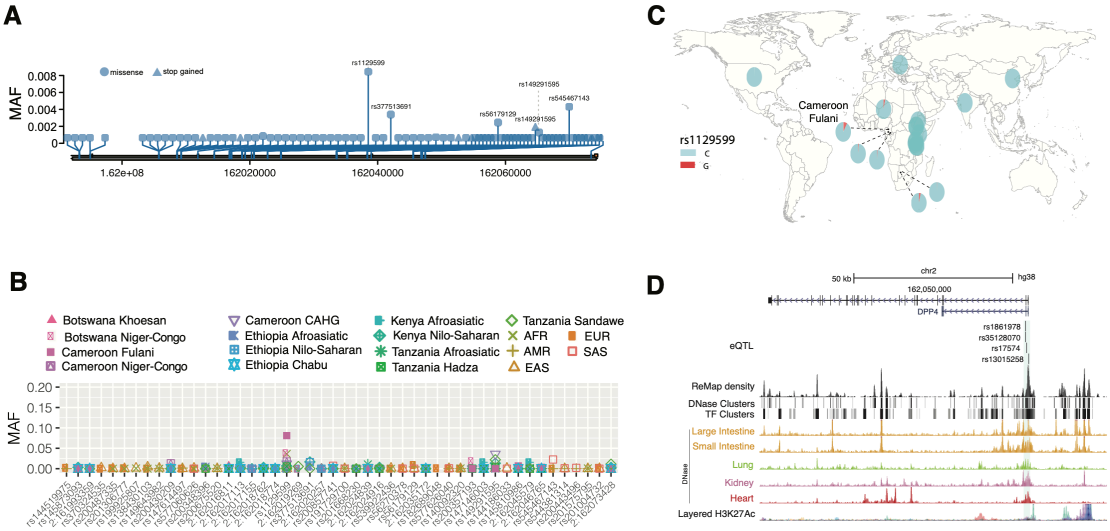


Figure S20. Normalzied expression data from GTEx show the significant association between eQTL allele frequency and *DPP4* gene expression.

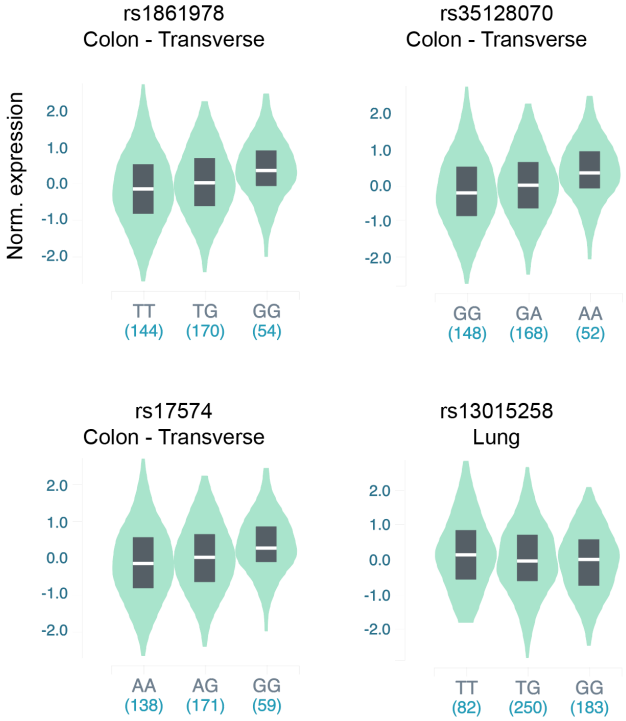


Figure S21. MAF of four regulatory variants at *DPP4*

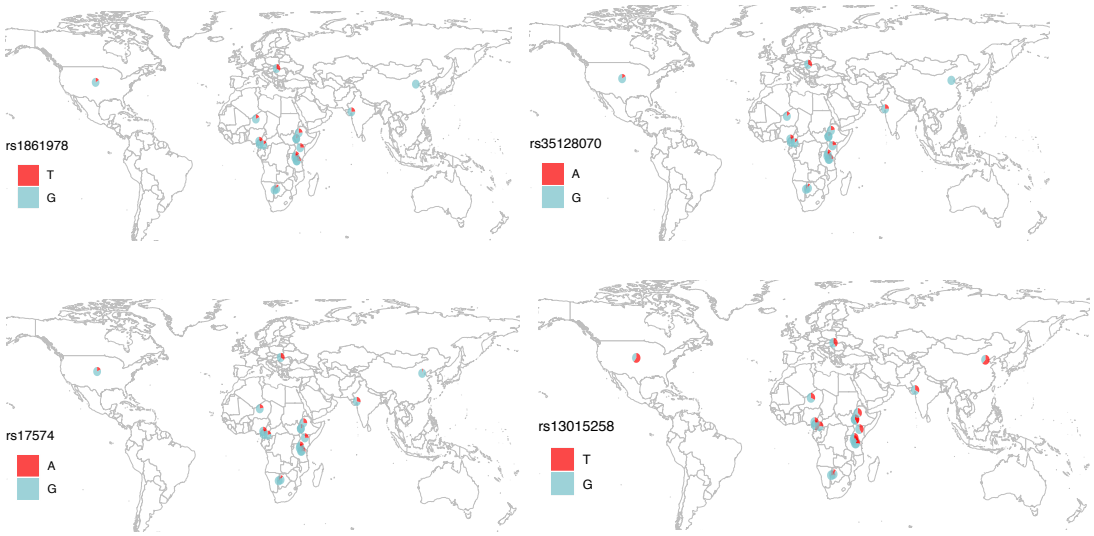


Figure 22. iHS scores for SNPs at *DPP4*. Each dot represents a SNP. Dashed lines denote the empirical cutoff. Red dots mean that the corresponding SNPs harbor significant scores.

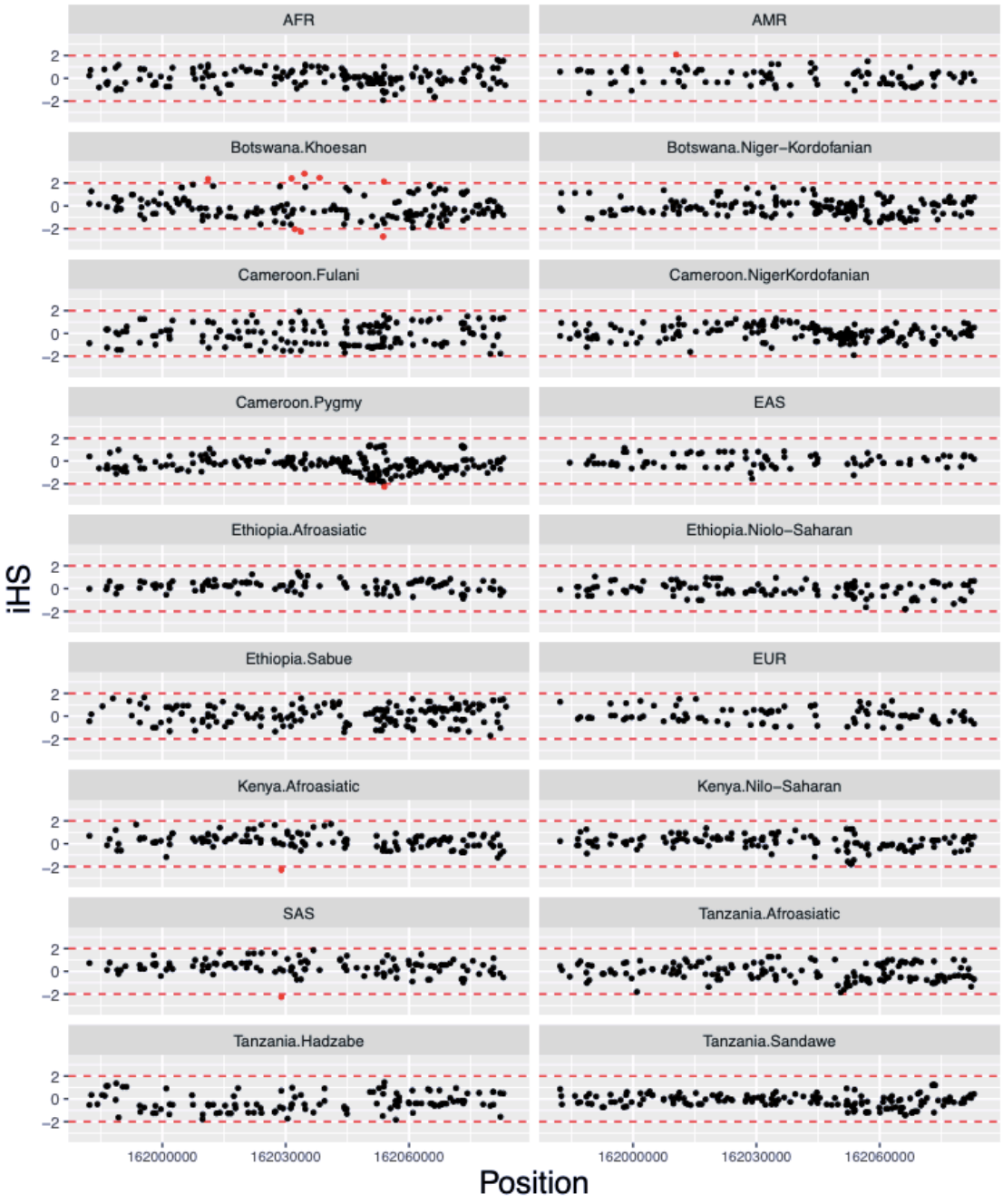


Figure S23. LD pattern between 11 SNPs at *DDP4* showing iHS signals in diverse ethnic groups. D prime was used to measure the LD. Eight of them were in the Khoesan populations from Botswana. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding positions.

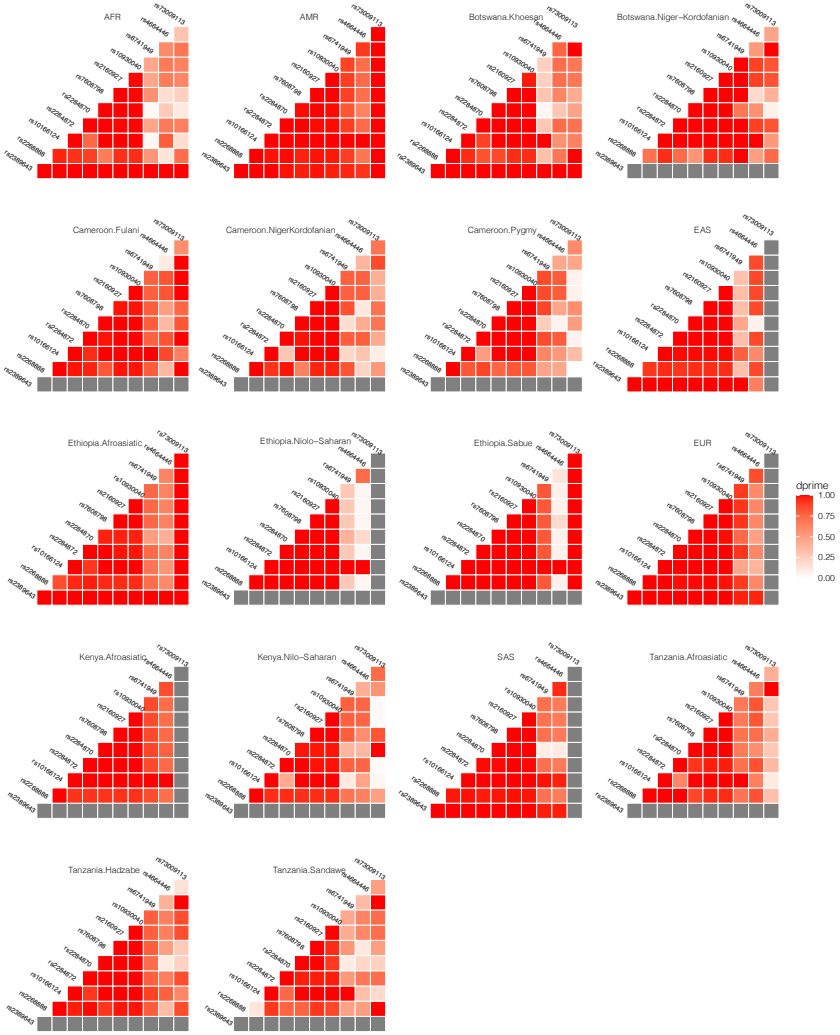


Figure S24. The intersection of SNPs with significant selection signals and regulatory regions at *DPP4*. The SNPs rs2098526 and rs2284870 highlighted with light blue shadow. Both SNPs have high iHS values and (|iHS|>2) overlap DNase I hypersensitivity peaks from ENCODE. The DNase-seq tracks of large Intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle are also from ENCODE, and their signals are scaled to 1.5.

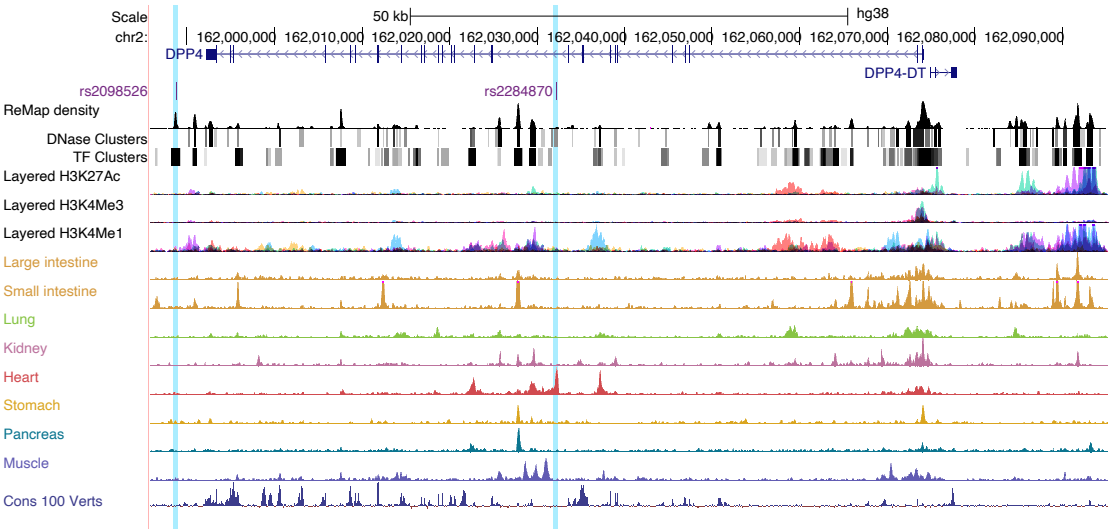


Figure S25. Genetic variation at *LY6E* and its disease association.
(A) Location of coding variants and their minor allele frequency (MAF) at *LY6E* identified from the pooled dataset. (B) MAF of coding variants in diverse global ethnic groups. (C) The MAF of variant rs111560737 at *LY6E* in diverse 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 ENCODE.

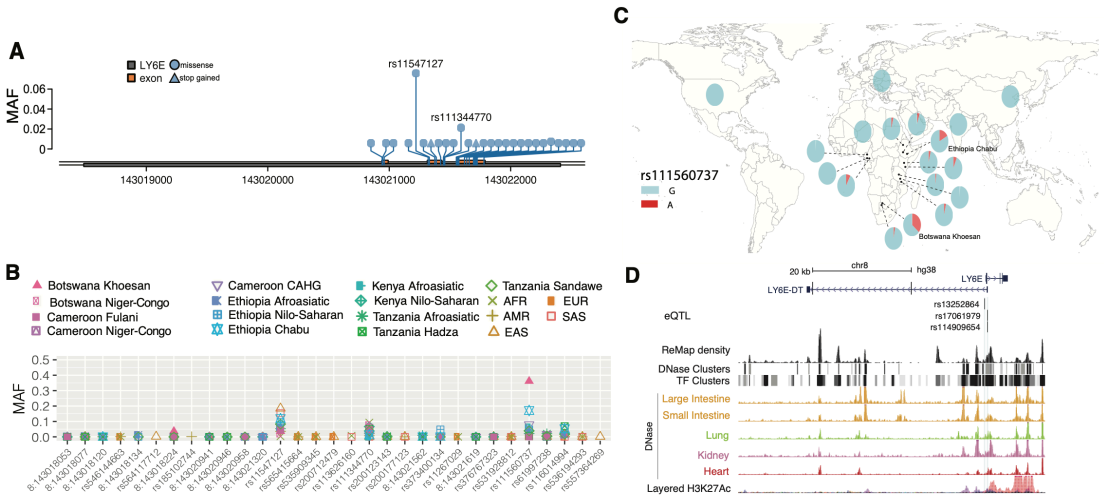


Figure S26. Normalized expression data of the three eQTLs (rs13252884, rs17061979 and rs114909654) of *LY6E* in frontal cortex from GTEx database.

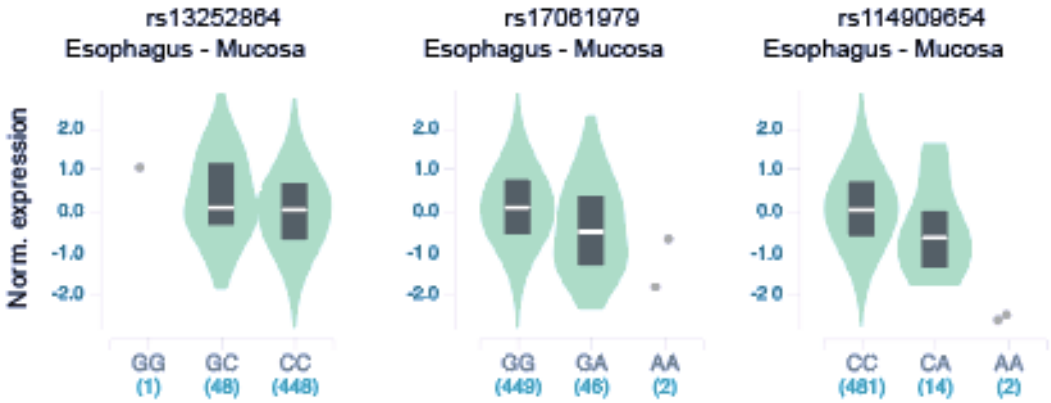


Figure S27. MAF of three regulatory variants at *LY6E*

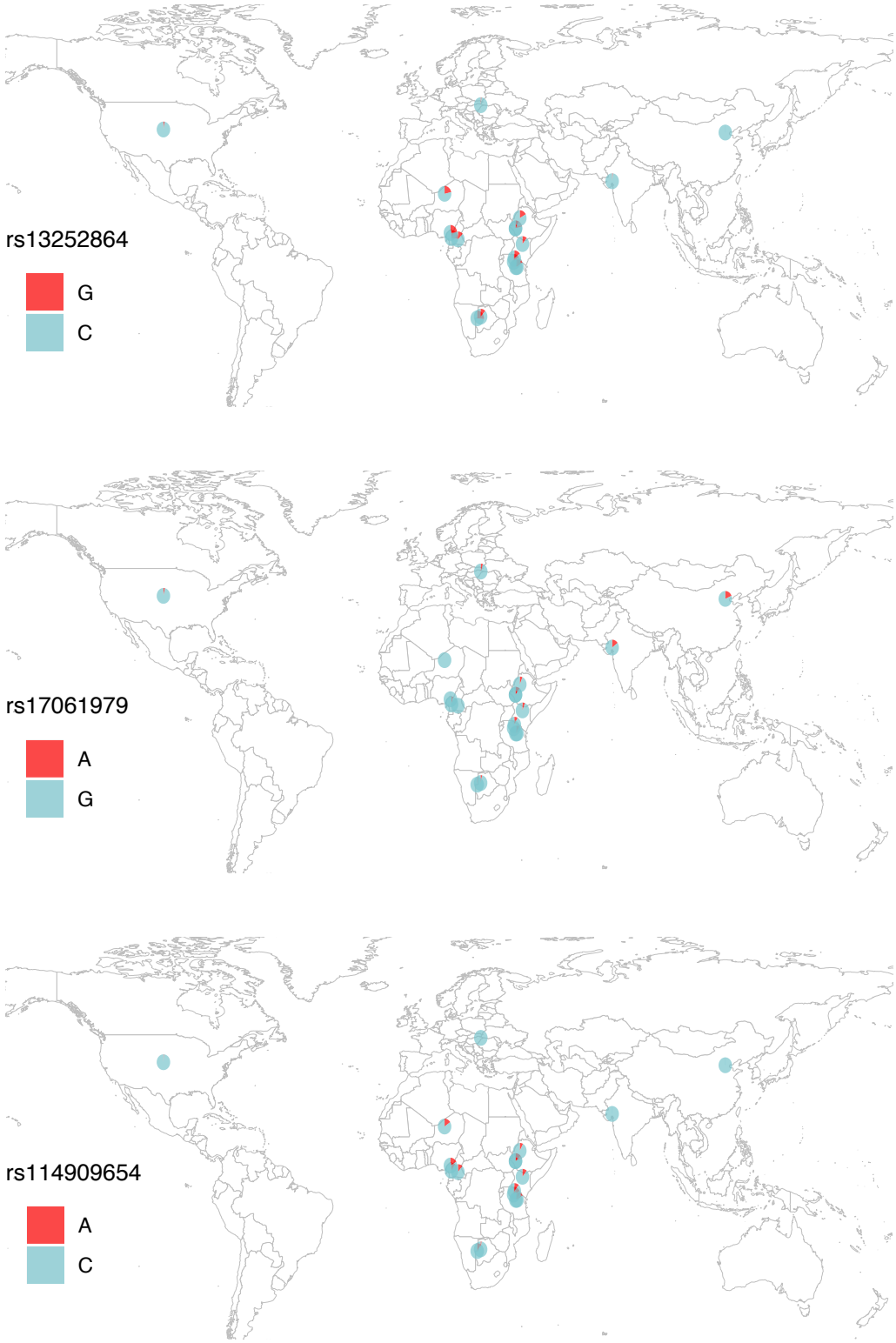


Figure S28. iHS scores for SNPs at *LY6E*. Each dot represents a SNP. Dashed lines denote the empirical cutoff. Red dots mean that the corresponding SNPs harbor significant scores.

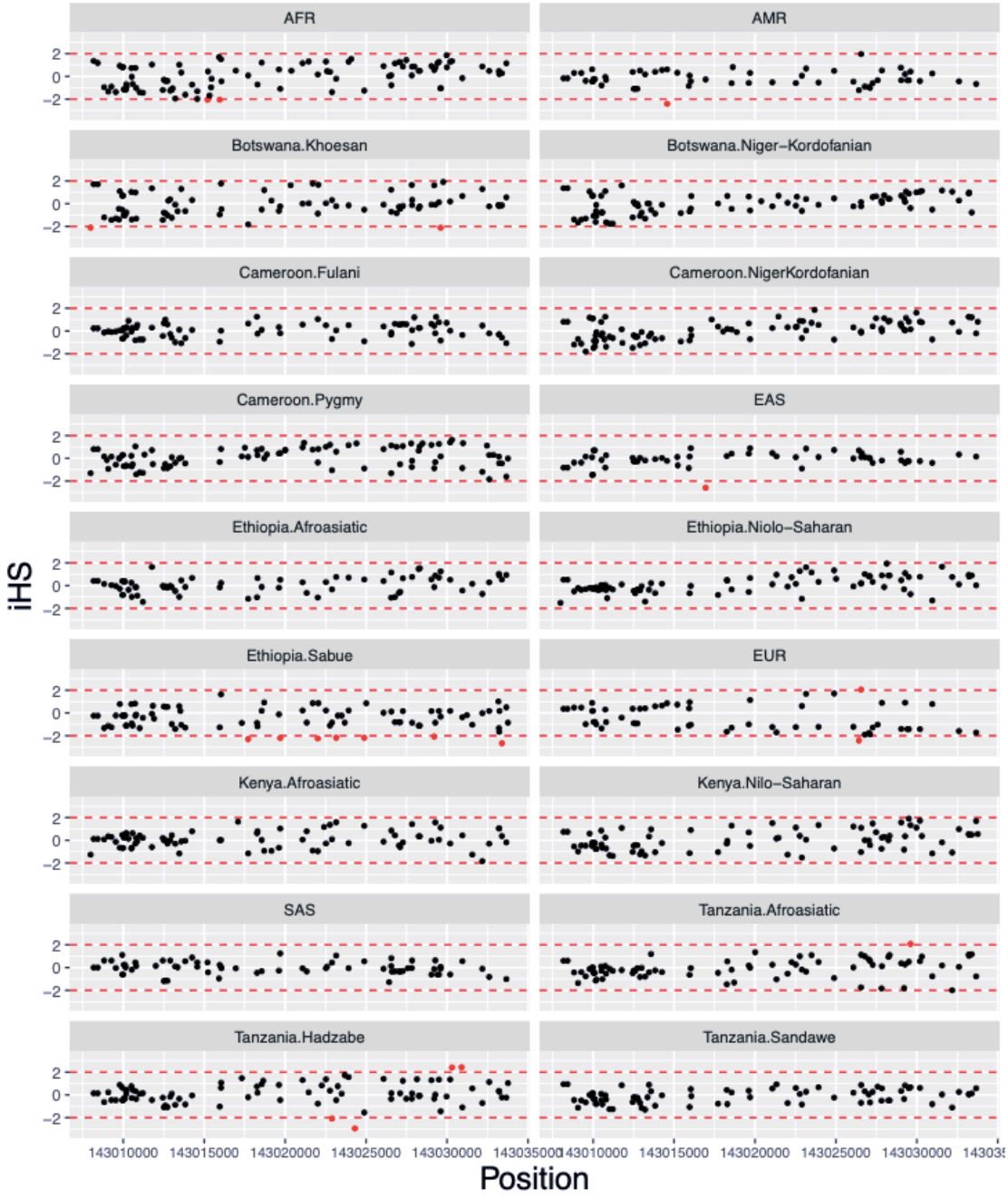


Figure S29. LD pattern between SNPs at *LY6E* showing iHS signals in diverse ethnic groups. D prime was used to measure the LD. Dark gray tiles in the LD heatmap plot denote no variant was observed at the corresponding positions.

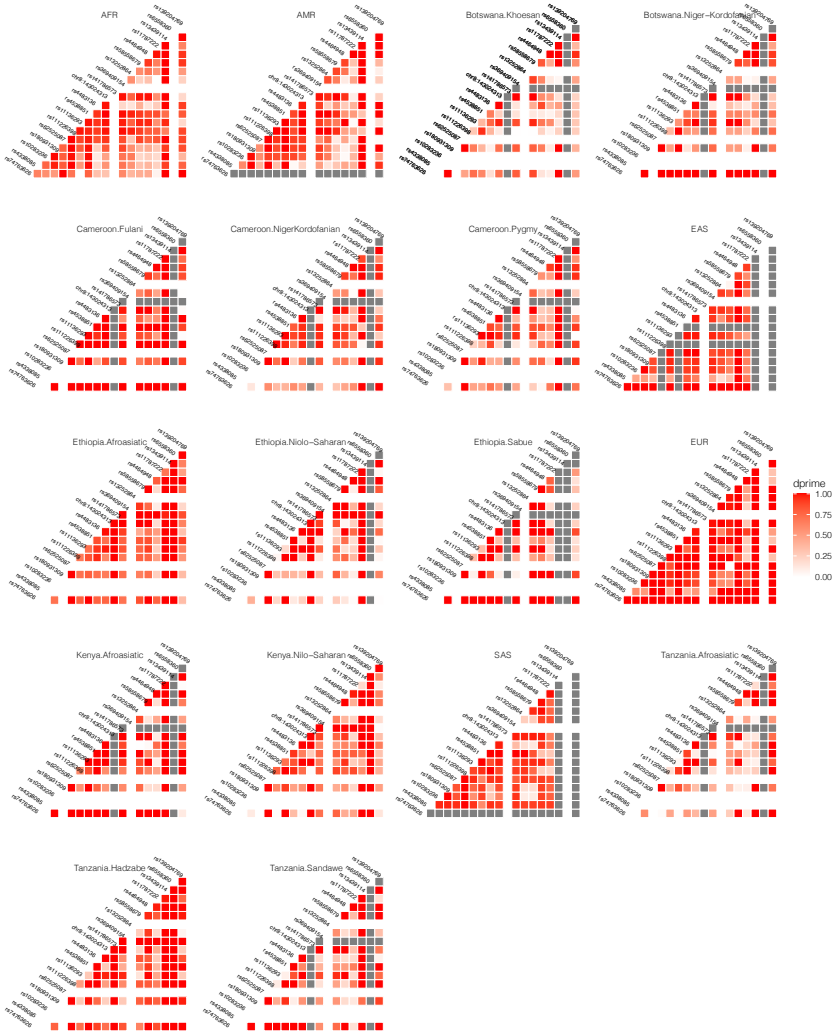
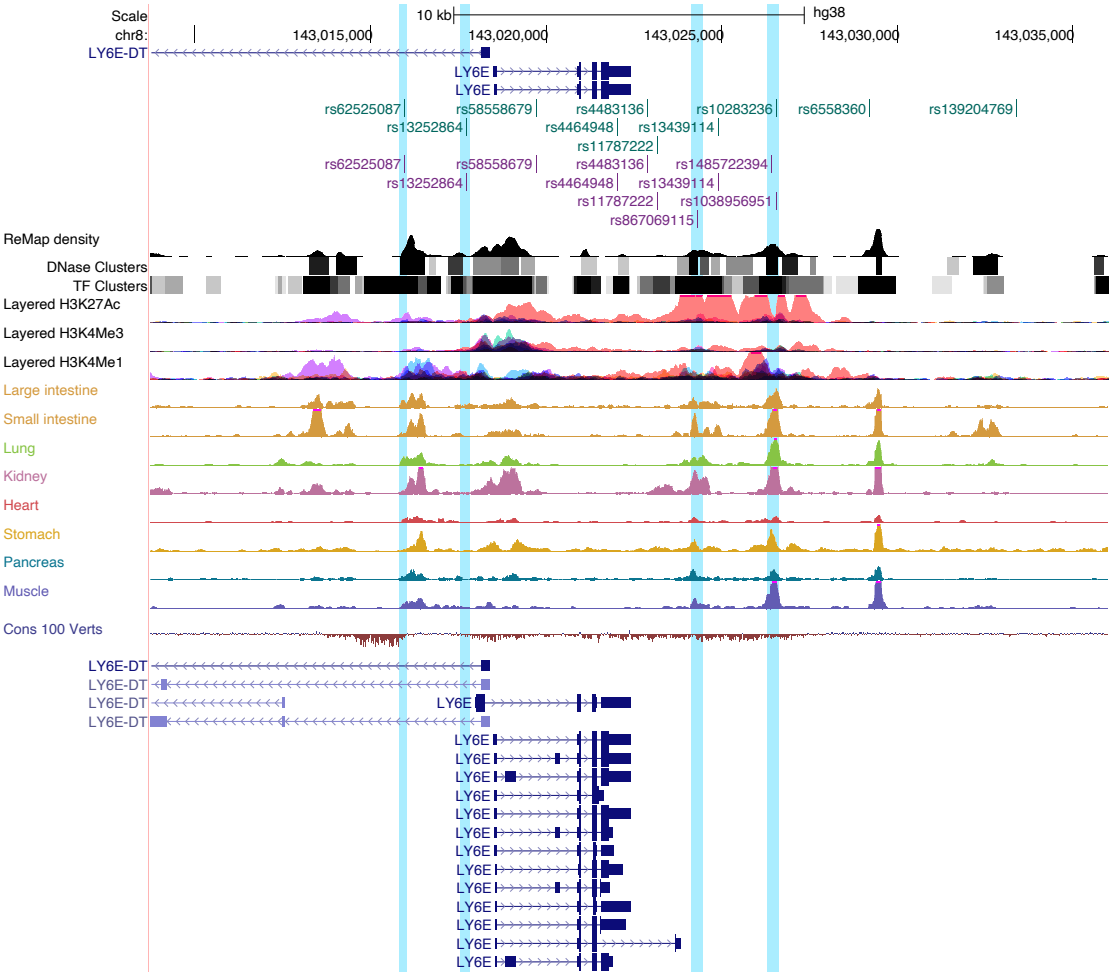


Figure S30. The intersection of SNPs with significant selection signals and regulatory regions at *LY6E*. SNPs with high iHS value (|iHS|>2) near DPP4 locus (< 10kb distance) overlapping DNase I hypersensitivity peaks from ENCODE (purple) or eQTLs from GTEx v8 (green) are shown in this figure. Potential regulatory elements are highlighted with blue shadow. The DNase-seq tracks of large Intestine, small intestine, lung, kidney, heart, stomach, pancreas and skeletal muscle are also from ENCODE.



Supplementary Table

Table S1. Penn Medicine Biobank (PMBB) participant characteristics

	PMBB (N=15,977)
Sex - number (%)	
Female	8379 (52%)
Male	7598 (48%)
Age - years	
Mean	62.7 (±17.2)
Range	19 - 89
Race	
Black (African Americans)	8916
White (European Americans)	7061

Legends of Dataset S1-S6

Dataset S1. Genetic variants identified around the *ACE2, TMPRSS2, DPP4 and LY6E*. For each gene, there are three sheets in the excel table: the first is for all genetic variants surrounding the corresponding gene, the second is only for coding variants, and the third is for GTEx eQTLs. The last sheet of the table summarizes the genetic variants that are eQTLs based on the GTEx dataset [6] and are located at promoter or enhancers regions based on the ENCODE dataset [68]. "N" denotes variants were not identified or called in the corresponding dataset. "0" denotes variants were identified in the corresponding dataset, but the minor allele frequency is 0.

Dataset S2. dN/dS tests for *ACE2, TMPRSS2, DPP4 and LY6E* in both the pooled dataset and specific ethnic groups.

Dataset S3. MK-tests for *ACE2, TMPRSS2, DPP4 and LY6E* in both the pooled dataset and specific ethnic groups.

Dataset S4. SNPs with significant selection signals for *ACE2, TMPRSS2, DPP4 and LY6E*. For each gene, there are three sheets in the excel table: the first summarizes iHS signatures in different population, the second shows variants with iHS signatures that overlap with DNase regions, and the third shows variants with iHS signatures that are GTEx eQTLs. The cell line or tissue codes used to identify DNase regions are listed on the last sheet of the table.

Dataset S5. Summary statistics from gene-based association results

Dataset S6. Summary statistics from PheWAS of eQTL variants

The Regeneron Genetic Center Authors and Contribution Statements

All authors/contributors are listed in alphabetical order.

RGC Management and Leadership Team

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

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

Sequencing and Lab Operations

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

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

Clinical Informatics

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

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

Genome Informatics

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

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 responsible for the compute logistics, analysis and infrastructure needed to produce exome and genotype data. G.E., M.O., M.N. and J.G.R. provided compute infrastructure development and operational support. S.B., S.K., and J.G.R. provide variant and gene annotations and their functional interpretation of variants. E.M., J.S., R.L., B.B., A.B., L.H., J.G.R. conceived and are responsible for creating, developing, and deploying analysis platforms and computational methods for analyzing genomic data.

Research Program Management

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

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

Reference

1. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv.
2. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
3. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome biology 17, 122.
4. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-D894.
5. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31, 3812-3814.
6. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7, Unit7 20.
7. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American journal of human genetics 88, 440-449.
8. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
9. Chadwick, L.H. (2012). The NIH Roadmap Epigenomics Program data resource. Epigenomics 4, 317-324.
10. Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon, A., Lopez, F., and Ballester, B. (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. Nucleic Acids Res 48, D180-D188.
11. Consortium, G.T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648-660.
12. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics (Oxford, England) 26, 1205-1210.
13. Moore, C.B., Wallace, J.R., Frase, A.T., Pendergrass, S.A., and Ritchie, M.D. (2013). BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genomics 6 Suppl 2, S6.
14. Basile, A.O., Wallace, J.R., Peissig, P., McCarty, C.A., Brilliant, M., and Ritchie, M.D. (2016). Knowledge Driven Binning and Phewas Analysis in Marshfield Personalized Medicine Research Project Using Biobin. Pac Symp Biocomput 21, 249-260.
15. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89, 82-93.

16. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444-1448.
17. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J Mol Graph 14, 33-38, 27-38.
18. McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-654.
19. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-426.
20. Zhai, W., Nielsen, R., and Slatkin, M. (2009). An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol 26, 273-283.
21. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-918.
22. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS biology 4, e72.
23. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol 31, 2824-2827.
24. Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nature genetics 48, 811-816.
25. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699-710.
26. Akey, J.M., Ruhe, A.L., Akey, D.T., Wong, A.K., Connelly, C.F., Madeoy, J., Nicholas, T.J., and Neff, M.W. (2010). Tracking footprints of artificial selection in the dog genome. Proceedings of the National Academy of Sciences of the United States of America 107, 1160-1165.
27. Leigh, J.W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. Methods Ecol Evol 6, 1110-1116.