Research

The evolutionary landscape of host immunity genes involved in respiratory and other immune-related diseases, and their association with severe COVID-19 outcomes

Christopher N. Cross[1,2], Alessandro Lisi[3], Faith C. Simmonds[4], Kareem Washington[5], Thomas Heinbockel[1], and Michael C. Campbell[3,*]

[1] Department of Anatomy, College of Medicine, Howard University, Washington, DC 20059

[2] Department of Biostatistics, Yale University School of Public Health, New Haven, CT 06510

[3]Department of Biological Sciences (Human and Evolutionary Biology Section), University of Southern California, Los Angeles, California 90089

[4] Columbia Vagelos College of Physicians and Surgeons, Columbia University, New York, NY 10032.

[5] Department of Genetics and Human Genetics, College of Medicine, Howard University, Washington, DC 20059

*To whom correspondence should be addressed

***Corresponding author**

E-mail address: Michael C. Campbell (mc44680@usc.edu)

Abstract

Background

Given its high mortality and broad societal impacts, the COVID-19 pandemic is a particularly notable global outbreak of a respiratory illness in the 21st century. Although previous studies have identified several genes associated with COVID-19 susceptibility, relatively little is known about the genes contributing to severe COVID-19, including their evolutionary histories. In the current study, we analyzed *IL-4*, *TLR2*, *CCL2*, and *SLC11A1*—four immunity genes that have been implicated in severe COVID-19 and other immune-related diseases—in globally diverse populations from the 1000 Genomes Project. We also tested for associations between genetic variation in these genes and clinical COVID-19 phenotypes in more than 4,000 laboratory-confirmed COVID-19–positive individuals from Italy.

Results

Based on our analyses, we identified 72 single nucleotide polymorphisms (SNPs) across these genes as targets of positive selection, including several derived alleles shared with archaic Neanderthal and/or Denisovan genomes—a finding not previously reported in the literature. Furthermore, we found that common SNPs—implicated in respiratory diseases such as tuberculosis and chronic obstructive pulmonary disorder—were also under selection. Functional predictions based on *in silico* analyses revealed that a subset of selected alleles map to transcription factor binding sites and are predicted to affect binding affinity. In addition, our genetic association analyses uncovered significant correlations between derived alleles in the coding region of *TLR2* and COVID-19 severity. Interestingly, these candidate alleles occurred at relatively low frequency in western European and East Asian populations but were absent in populations of African and South Asian descent.

Conclusions

Overall, our study provides new insights into the evolution of biologically relevant immunity genes in the modern human lineage and highlights genetic variants that may underlie differential risk for severe COVID-19.

keywords: severe COVID-19, low-frequency alleles, natural selection, selection coefficient, genetic association analysis, antagonistic pleiotropy, and population genetics

Introduction

Novel coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was a public health crisis that resulted in more than 700 million positive cases and over 7 million deaths worldwide (https://www.worldometers.info/coronavirus/). While SARS-CoV-2 infection can lead to severe respiratory problems, other symptoms can include cardiovascular and neurological impairment [1–5]. Furthermore, data have shown that host genetics play a role in SARS-CoV-2 infection [6–11]. In particular, studies have identified host genes that encode receptors for coronavirus binding (*e.g.*, *ACE2*, *ANPEP*, *DPP4*) and proteases that facilitate entry of SARS-CoV-2 into cells (*e.g.*, *TMPRSS2*, *FURIN*, *TMPRSS11D*, *CTSL*, and *CTSB*), increasing COVID-19 susceptibility [7, 12, 13]. Genetic analyses have also demonstrated that alleles within these genes vary in frequency among populations [7, 13], potentially contributing to differential risk of infection in the absence of vaccination.

In addition, accumulating evidence has indicated that host genes contribute to severe COVID-19 after initial SARS-CoV-2 infection [12, 14–38]. For example, mRNA expression levels of Toll-like receptor genes, including *TLR2*, have been correlated with disease severity (commonly characterized by dysregulated inflammation and a cytokine storm) in SARS-CoV-2-positive patients [19, 39]; *in vivo* mice experiments further confirmed that the TLR2 protein was the innate sensor that triggered inflammatory cytokine expression in response to SARS-CoV-2 infection [19]. Prior analyses have also shown that cytokines, such as IL-4, IL-6, TNF-a, CXCL8, CXCL9, CXCL10, and CCL2, are present at elevated levels in the serum of COVID-19 patients with severe respiratory symptoms [40–44]. These findings suggest that COVID-19 severity may have a different genetic architecture involving cytokine and interleukin genes than COVID-19 susceptibility [45, 46]. In addition, multiple lines of evidence, including a meta-analysis of ~400 host genetic susceptibility studies, have implicated a similar set of genes (specifically, *IL-4*, *TLR2*, *CCL2* and *SLC11A1*) in the onset of other respiratory and immune-related diseases, such as psoriasis, osteoarthritis, pulmonary tuberculosis (TB), inflammatory bowel syndrome (IBD), influenza, pneumonia, and severe acute respiratory syndrome (SARS) [47–63]. Finally, a recent analysis identified 13 genome-wide loci associated with lung, autoimmune, or inflammatory diseases that were significantly correlated with severe COVID-19 [6]. Thus, we hypothesize that genes known to play a role in respiratory and/or other immune-related illnesses, including their severity, may contribute to acute manifestations of COVID-19.

Despite the far-reaching consequences of COVID-19, the role of host genetics in severe disease has not been extensively examined. Furthermore, little is known about the evolutionary histories of immunity genes that have been implicated in severe COVID-19 and other immune-mediated diseases [56–58, 64, 65]. To overcome these gaps in knowledge, we examined *IL-4*, *TLR2*, *CCL2,* and *SLC11A1*—genes associated with an array of immune-mediated illnesses that also have been of long-standing interest to the

authors—in 2,039 individuals from globally diverse populations in the 1000 Genomes Project to: 1) characterize patterns of nucleotide and haplotype variation within and between populations, and 2) infer the microevolutionary forces that have shaped these genetic patterns over time. In addition, we performed association analyses of clinical phenotypes and known mutations in our four genes implicated in respiratory and/or other immune-related illnesses in >4,000 laboratory-confirmed SARS-CoV-2-positive individuals enrolled in the GEN-COVID Multicenter Study in Italy. Overall, our study aims to characterize the evolutionary histories of key host immunity genes and provide new information regarding severe COVID-19 susceptibility.

Results

Patterns of Nucleotide Variation

We identified a total of 260 bi-allelic SNPs across the 8,690-base pair (bp) region of *IL-4* (**Figure 1**; **Table S1**). Of these 260 SNPs, 20 were found in exons, while the remaining 240 polymorphisms were present in non-coding regions (**Figure 1; Table S1**). Among the 20 exonic SNPs, nine were synonymous, 10 were nonsynonymous, and one was a stop gain mutation (*rs*550558636; **Table S1**). Interestingly, we found that nonsynonymous SNPs mainly occurred at very low frequency (< 1%) in all populations (**Table S1**). Similarly, most of the synonymous variants were either at low frequency (< 1%) or absent in populations, except for one SNP (*rs*2243251) which had a minor allele frequency (MAF) that ranged from 14% to 27%; the stop gain mutation was present only in the Mende of Sierra Leone with an MAF < 1% (**Table S1)**.

In the non-coding regions of *IL-4*, we identified five common SNPs (MAF ≥ 5%) in pooled populations (**Table S1)**; one of these SNPs (*rs*2070874) was located in the 5′-UTR, while the other four polymorphisms (*rs*2243281, *rs*734244, *rs*2243253, and *rs*2243261) were found in intronic regions. Furthermore, the MAF at these SNPs showed considerable variation across different populations. For example, the minor T-allele at *rs*2070874 exhibited frequencies ranging from 37.5% to 58% and from 73% to 86% in populations of African and East Asian descent, respectively (**Table S1)**. Prior studies have also linked the *rs*2070874 T-allele to severe respiratory illness and heightened immune responses [57, 66–68].

Across the 21,981-bp length of the intronless *TLR2* gene, we identified a total of 627 bi-allelic SNPs (**Figure 1**; **Table S2**). Within the single exon of this gene, we found 23 synonymous variants, 47 nonsynonymous variants, and two stop-gain variants. We also uncovered 541 and 14 polymorphisms in the 5′-UTR and 3′-UTR, respectively (**Table S2**). Furthermore, we identified 36 common SNPs in the 5′-UTR, 20 of which were found in all populations (none were present in the 3′-UTR). Interestingly, the MAF at these common sites differed widely among diverse populations (**Table S2**). For example, the minor T-allele at *rs*6811715 ranged from 33.9% to 49.2% frequency in populations of African descent but occurred at

even higher frequency in non-Africans (between 47% and 53% in Europeans; between 50% and 60% in East Asians; **Table S2**); this allele was absent in South Asian populations. The minor T-alleles at *rs*13115230, *rs*11099894, *rs*7694512 also varied between 9% and 24% in populations of African descent but were observed at even higher frequency (≥ 29%) in non-African populations (**Table S2**). Similarly, the derived G-allele at *rs*1898830—which has been correlated with respiratory disease [69–71]—was present at relatively high frequency (from 29% to ~60%) in non-African populations, particularly in Asia (**Table S2**). In contrast, the *rs*1898830 G-allele occurred at much lower frequency (from 7% to 23.3%) in African and African-descended populations (**Table S2**).

At *CCL2*, a total of 34 bi-allelic SNPs were identified across the 1,919-bp length of this gene (Figure 1). Of these bi-allelic SNPs, four were in exons, 11 in introns, two in the 5′-UTR, and 17 in the 3′-UTR (**Table S3**). Among the four exonic variants, two were synonymous (*rs*4586 and *rs*368861512) and two were nonsynonymous (*rs*148285031 and *rs*190996557). Furthermore, of the 34 SNPs that we identified, only four—*rs*13900 (3′-UTR), *rs*2857657 (intron), rs4586 (exon) and *rs*28730833 (intron)—had an MAF ≥ 5% (**Table S3**). In addition, the MAF of these common SNPs varied dramatically among populations. For example, the minor T-allele at *rs*13900 ranged from 20% to 28% frequency in populations of African descent while the same allele varied from 27% to 41% frequency in South Asians, from 29% to 36% in Europeans, and from 41% to 65% in East Asians (**Table S3**). Similarly, the rs4586 T-allele ranged from 13% to 37% frequency in African and African-descended populations but varied from 35% to 63% frequency in non-Africans (**Table S3**). Prior studies have also implicated *rs*13900 and rs4586 in gene expression regulation and/or respiratory disease susceptibility [65, 72, 73].

At *SLC11A1*, we identified 464 bi-allelic SNPs across the 14,622-bp length of this gene (**Figure 1**; **Table S4**). Of these SNPs, 75 were located in exons, 299 in introns, six in intronic splice regions, two in the 5′-UTR, and 82 in the 3′-UTR (**Figure 1; Table S4**). We also observed eight common SNPs in exons, 41 in introns, and eight in the 3′-UTR (**Table S4)**. Furthermore, the MAF at common exonic SNPs differed considerably between African and non-African populations. For example, the minor *rs*17221959 T-allele occurred at a frequency no greater than 15.5% in populations of African descent, whereas this allele varied between 28.3% and 40.3% frequency in non-Africans (**Table S4)**. Conversely, the minor *rs*7576974 T-allele spanned from 0.08% to 26% frequency in populations of African descent but occurred at <0.02% frequency in non-Africans (**Table S4)**.

We also noted frequency differences at common intronic variants at *SLC11A1* across populations. For example, the minor *rs*2290708 T-allele varied between 19% to 35% frequency in Europeans but was present at <17.9% frequency in other global populations (**Table S4**). Likewise, the T-alleles at *rs*13011285 and *rs*17228995 ranged from 20.7% to 34.1% in European populations but occurred at lower frequency in populations of African, South Asian, and East Asian descent (**Table S4**). Furthermore, the derived

*rs*2279014 T-allele in the 3'-UTR—associated with decreased risk for *Mycobacterium avium* complex infection [74]—was found at ~40% frequency or higher in most populations, except in East Asians where the frequency of this allele varied between 18.8% and 30% (**Table S4**).

Tests of Neutrality

To investigate if patterns of nucleotide variation evolved neutrally, we calculated Fay and Wu's *H* (*H*) and Tajima's D (D_T) statistics for *IL-4*, *TLR2*, *CCL2* and *SLC11A1* in each population with DnaSP [75] (**Tables S5; Tables S6; Tables S7; Tables S8**). We also simulated the expected *H* and D_T values under different models of population growth to assess the statistical significance of observed values (**Tables S9-S16**). Using this approach, we found a mix of positive and negative *H* values at *IL-4* in non-African populations with significant negative *H* statistics observed in East Asians. In contrast, populations of African descent generally exhibited positive *H* values (**Table S9**). At *TLR2*, *H* statistics were consistently negative in non-Africans with significant values occurring in European, South Asian, and East Asian populations. We also observed a pattern of significantly negative *H* statistics at *TLR2* in populations of African descent (**Table S10**). Our analyses of *CCL2* revealed a pattern of negative *H* statistics in non-African populations (**Table S11**). We observed a similar trend in Africans with significant values occurring in the East African Luhya and West African Yoruba. At *SLC11A1*, significantly negative *H* values were found in both African and non-African populations (**Table S12**).

Finally, although the D_T statistics were a mix of negative and positive values at *IL-4*, *TLR2*, and *CCL2* in both Africans and non-Africans, these values were generally more positive than expected under different population growth models (**Tables S13-S15**). By contrast, at *SLC11A1*, we observed significantly negative D_T statistics in populations of African descent and consistently negative D_T statistics in non-Africans (though these values were not significant; **Table S16**).

Long-range haplotype homozygosity

We calculated *i*HS and nS_L statistics for each population to examine patterns of haplotype homozygosity surrounding the *IL-4*, *CCL2*, *TLR2*, and *SLC11A1* genes using the selscan and Polaris packages [76, 77]. Based on these analyses, we found outlier | nS_L | statistics at several *IL-4* variants—specifically, *rs*199693275, *rs*734244, *rs*58349986, *rs*56408050 and *rs*56174361—primarily in non-African populations (**Table 1; Table S17; Figures S1-S2**). Furthermore, we observed short extended haplotype homozygosity (EHH) around *rs*58349986 and *rs*56174361 in non-Africans, particularly in Europe and Asia**.**

At *TLR2*, we identified SNPs with extreme | *i*HS | and/or | nS_L | statistics in African and non-African populations (**Table 1**; **Tables S18-S19; Figures S4-S5**). In several cases, we also detected unusually long haplotypes around either the derived or ancestral alleles at these loci. (**Figure S6**). Most notably, we

observed long-range EHH surrounding the derived *rs*7694512 T-allele and derived *rs*11099894 T-allele, spanning up to 700,000 kbs, in African and South Asian populations (**Figure S6**). Additionally, the derived G-allele at *rs*1898830—which has been previously correlated with a decrease risk for TB development [71]—was found on an extended haplotype background in these populations (**Figure S6**).

At *CCL2*, we identified two SNPs (*rs*2857657 and *rs*4586) that exhibited extreme | *i*HS | or | nS_L |statistics in global populations (**Tables S20-S21; Figures S7-S8**). Specifically, *rs*2857657 had an extreme positive *i*HS statistic, indicating selection on the derived G-allele, in the Afro-Caribbean and Iberian populations (**Tables S20-S21; Figures S7-S8**), while *rs*4586 exhibited negative outlier *i*HS statistics favoring the ancestral C-allele in populations of African, European, and East Asian descent (**Tables S20-S21**). As with *TLR2*, we plotted the decay of haplotype homozygosity from these core SNPs using the EHH statistic [78]. This analysis revealed unusually long haplotype structure—spanning ~200 kbs around the derived *rs*2857657 G-allele—in the Afro-Caribbean and Iberian populations (**Figure S9**). Furthermore, we detected extensive EHH on chromosomes carrying the ancestral rs4586 C-allele (relative to chromosomes with the derived allele at the same site) in East Asian populations (**Figure S9**).

Finally, we observed distinct patterns of haplotype homozygosity around the *SLC11A1* gene when comparing African and non-African populations. Notably, several SNPs within *SLC11A1* exhibited outlier | iHS | and/or | nS_L | statistics in South Asian, European, and East Asian populations (**Tables S22 and S23; Figures S10 and S11**). We further examined the decay of haplotype homozygosity around these outlier SNPs using the EHH statistic (**Figure S12**). Based on this analysis, we found extended haplotype homozygosity (EHH) around the ancestral C-allele of rs2279014 in South Asian, European, and East Asian groups (**Figure 2; Figure S12**). Similarly, we detected unusually long haplotypes around the derived rs2279015 G-allele in South and East Asians (**Figure 3; Figure S12**). These variants are of particular interest as both are predicted to affect transcription factor binding. Furthermore, the rs2279014 T-allele is linked to decreased susceptibility to *Mycobacterium avium* complex infection [74]. Comparatively, African and African-descended populations did not show outlier | *i*HS | and | nS_L | values at these loci.

Selection coefficient estimates

As a complementary analysis, we applied the CLUES2 algorithm to *IL-4*, *TLR2*, *CCL2*, and *SLC11A1* in each population at different time depths (i.e., 1,000, 500, and 200 generations ago) to assess temporal shifts in selection intensity [79–81]. Our results indicated that selection pressure on SNPs in *IL-4* likely began at a deep point in time. For example, we observed significant *s* estimates for derived alleles at rs56174361 in African populations at 1000 generations ago (corresponding to 28,000 ya; **Figure 4; Table S24**). Furthermore, we detected significant *s* for ancestral alleles at *rs*2227282 and *rs*2227284 in African populations (**Figure 4; Table S24**). In contrast, selection appeared to target the ancestral allele at

*rs*56174361, as well as the derived alleles at *rs*2227282 and *rs*2227284 in different populations (**Figure 4; Table S24**). Interestingly, this pattern of selection on derived and ancestral alleles at the same polymorphic site occurred across *IL-4*; in many cases, the inferred timing of these selection events dated to at least 1000 generations ago, suggesting deep and complex selection histories at this gene.

Similarly, at *TLR2*, we found evidence for strong selection favoring derived alleles at 1,000 generations ago (28,000 ya) in Africa. For example, in the Yoruba population from Nigeria, *s* estimates for *rs*2083287 and *rs*6811185 were 0.2942 (CI: 0.00664–0.0522) and 0.01293 (CI: -0.00166–0.02753), respectively (**Table S25**), indicating selection on derived alleles at these sites; however, we did not observe extensive EHH around these polymorphic sites. In contrast, we did observe non-significant *s* estimates for derived alleles at other SNPs (such as *rs*12510167, *rs*7694512, *rs*11099874, *rs*7694512, and *rs*1311523), which exhibited extensive EHH in populations of Africa descent. This finding suggests that strong selection may have occurred more recently. One notable exception to this pattern was *rs*13123230 which showed long-range EHH and a significant *s* estimate at 1,000 generations ago (28,000 ya) in the Esan from Nigeria (**Table S25; Figure S6**). However, upon further examination, we found that *s* was 0.01405 (CI: 0.00612–0.02199) at 1,000 generations ago, decreased to 0.00386 (CI: -0.17166–0.17938) at 500 generations ago (14,000 ya), and then dramatically increased to 0.01683 (CI: -0.26538–0.29903) at 200 generations ago (5,600 ya; **Table S25**), indicating a more recent increase in selection pressure. Similarly, outside of Africa, we observed progressively stronger *s* estimates from 28,000 ya to 5,600 ya at multiple SNPs in populations, including sites such as rs62323856 and rs74341183 which are polymorphic in the Neandertal and/or Denisovan genomes (**Table S25**).

At *CCL2*, we found that the strength of selection at loci in this gene differed among populations (**Table S26**). For instance, we observed a significant *s* estimate for the ancestral C-allele at rs4586 in Africans (Mende and Yoruba) and East Asians (Japanese) but not in other populations (**Table S26**). We also found significant *s* estimates at rs13900 favoring the derived allele in Africans (Luhya) and East Asians (Han Chinese and Japanese), as well as significant *s* for *rs*2857657 in the Tuscan Italian population (**Table S26**).

At *SLC11A1*, we detected significant *s* statistics at several polymorphic sites in African and European populations at least 28,000 ya (**Table S27**). This deeper timeframe of selection is further corroborated by the absent of extensive EHH in these populations [78]. In contrast, we observed instances of strong selection at sites at a more recent time depth in South Asian and East Asian populations (**Table S27**). For example, rs2695343 exhibited significant *s* estimates favoring the derived allele at least 200 generations ago (5,600 ya). Intriguingly, we also detected significant *s* estimates for derived allele at rs2279015 and rs1059823 that progressively increased over time (from 28,000 ya to 5,600 ya) in South and East Asian populations (**Table S27**), which likely contributed to the observed long-range EHH in these

groups (**Table S27**). Finally, unlike the previous three genes, we found very few instances in which both ancestral and derived alleles at the same site were targeted by selection in different populations (**Table S27**).

Population differentiation

We assessed among-population differentiation by calculating F_{ST} statistics at polymorphic sites within our genes. To identity F_{ST} outliers, we compared observed values to an empirical distribution consisting of F_{ST} estimates calculated at ~8.1 million SNP randomly chosen across the genome. Based on this approach, we identified nine SNPs at *IL-4* (*rs*2243288, *rs*2243289, *rs*2243290, *rs*2243284, *rs*2243282, *rs*2243270, *rs*561556119, rs2227282, *rs*2227284) that had outlier F_{ST} statistics (**Table S1**). Notably, the derived alleles at these sites (specifically, *rs*2243288, rs2243290, rs2243284, rs2243282) were also present in the Neandertal and/or Denisovan genomes (**Table S1**). At *TLR2*, we detected outlier F_{ST} statistics at *rs*1439164, *rs*6811185, *rs*2083287, *rs*1898831, and *rs*4315760 (**Table S2**). We also identified a single outlier F_{ST} estimate at *rs*2279015 in the *SLC11A1* gene (**Table S4**); however, we did not detect extreme F_{ST} values at loci in *CCL2* (**Table S3**). Interestingly, the majority of the highly differentiated SNPs listed above, including those sites of archaic origin, were inferred to be under selection in contemporary populations based on our *i*HS, nS_L, EHH, and/or CLUES2 analyses (**Tables S17-S23; Figures S1-S12**).

Distribution of haplotype variation

At *IL-4*, we inferred 296 distinct haplotypes, nine of which were common (H44, H74, H85, H103, H115, H197, H234, H241, and H287) in populations of African descent (**Tables S28-S30**). Together, these common haplotypes comprised 51% to 58% of all haplotypes in these populations (**Tables S28-S30**). In contrast, eight common haplotypes (H44, H69, H82, H176, H197, H198, H234, and H241) comprised ~80% to 93% of all haplotypes in non-African populations (**Tables S28-S30**). Further analyses also revealed that a subset of these common haplotypes (H44, H69, H82, H176, and H234) accounted for more than 78% of all haplotypes in Europeans and South Asians; these same haplotypes only constituted between 47.4% and 74.6% of the total haplotypes in East Asian populations (**Tables S28-S30**).

At *TLR2*, we extracted a total of 599 distinct haplotypes in pooled global populations (**Tables S31-S33)**; eight common haplotypes (H11, H30, H231, H265, H361, H402, H438, and H580) comprised ~38% to 44% of the variation present in African and African-descended populations (**Tables S31-S33**). In comparison, 12 common haplotypes (H2, H40, H65, H96, H98, H108, H227, H231, H482, H507, H532, and H540) constituted ~49% to 80% of the haplotypes in non-Africans (**Tables S31-S33)**. Our analyses also showed that a few of these haplotypes largely contributed to the variation in these populations. For example, H231 accounted for ~40% to 46% of all the haplotypes in South Asians, while H231 comprised

~24% to 25% and ~24% to 32.9% of the haplotypes in Europeans and East Asians, respectively (**Tables S31-S33**). In contrast, the maximum frequency of haplotypes was 15% in populations of African descent (**Tables S31-S33)**.

At *CCL2*, we identified 35 distinct haplotypes in pooled global populations (**Tables S34-S36)**, five of which were common (H2, H9, H11, H19, and H29). Our results also showed that H9, H11, H19, and H29 constituted 92%-97% and 97% to 100% of the total variation in African and non-African populations, respectively (**Tables S34-S36**), indicating a strong skew toward a small set of dominant haplotypes at this gene.

At *SLC11A1*, we found a total of 619 distinct haplotypes in pooled global populations (**Tables S37-S39**); 11 common haplotypes (H102, H159, H179, H204, H215, H254, H269, H334, H445, H484, and H491) comprised between ~28% and 54% of all haplotypes in populations of African descent (**Tables S37-S39**). In contrast, eight common haplotypes (H150, H223, H284, H321, H334, H340, H484, and H491) comprised between ~50% and 75% of all haplotypes in non-Africans (**Tables S37-S39**). Notably, only two common haplotypes (specifically. H334 and H484) were shared between African and non-African populations (**Tables S37-S39**), revealing a high degree of divergence between these broad groups at this gene.

Inferred haplotype relationships

We reconstructed the evolutionary relationships among haplotypes at each gene using the median joining network method [82](**Figure 5**). This analysis revealed a divergent geographic distribution of several common *IL-4* haplotypes (**Figure 5; Table S40**). Specifically, H44, H82, and H69 were found mainly in European and South Asian populations, while H234 and H197 were more prevalent in East Asians (**Figure 5**). We also observed African-specific haplotypes (i.e., H74, H85, H103, H115, and H287) at low to moderate frequency; the presence of reticulations among these African haplotypes suggests historical recombination (**Figure 5**)[83]. Our previous analyses also showed that both ancestral and derived alleles at the same SNP (e.g., *rs*2243288, *rs*2243290, *rs*2243284, and *rs*2243282) were subject to positive selection in different populations. Many of the ancestral alleles at these sites reside on haplotypes H44, H82, and H69 which are common in European and South Asian populations (**Figure 5**). Conversely, many of the derived alleles at the same SNPs were found on a different set of haplotypes (H176, H197, H234, and H241) that were prevalent in East Asians (**Figure 5**).

At *TLR2*, H98, H231 and H540 were the most frequent haplotypes in the phylogeny and were surrounded by low-frequency haplotypes in a "star-like" pattern, suggestive of a recent expansion of variation (**Figure 5**). Among these haplotypes, H231 contained the largest number of selected alleles compared to H98 and H540, and H231 had the highest frequency in the network (**Figure 5**). Furthermore,

H98 and H540 showed a striking geographic distribution. In particular, haplotypes H98 and H540 were present only in non-Africans while haplotypes H265, H361, H402 were found mainly in populations of African descent (**Figure 5**; **Table S41**). Our analyses also revealed that different haplotypes carried distinct combinations of selected alleles. For example, the derived alleles at *rs*13147967 and *rs*1898831 occurred on several haplotypes (e.g. H98, H231, and H540) that were widespread in Eurasian populations, while the derived alleles at *rs*2083287 and *rs*6811185 existed on haplotypes (e.g., H361, H402) that were more common in populations of African descent (**Figure 5; Table S41**).

At *CCL2*, H9, H11, H19, and H29 haplotypes occurred at high frequency in global populations, except for H9 which was absent in East Asia (**Figure 5; Tables S37-S39)**. Notably, H9 carried the ancestral *rs*4589 C-allele, which has been inferred to be under strong selection in Africans. Haplotype H11 contained both the *rs*4589 C-allele and the derived *rs*13900 T-allele, which have significant *s* estimates in East Asian populations (**Table S42**). Comparatively, H29 carried the derived *rs*2857657 G-allele and the derived *rs*13900 T-allele; the highest frequency of this haplotype occurred in Europe (**Figure 5)**.

Finally, at *SLC11A1*, the common H334 and H484 haplotypes were found primarily in non-African populations (**Figure 5; Table S43**). Specifically, H334 occurred at low frequency (~2%) in populations of African descent (**Table S26; Table S27**), while it was present at a much higher frequency in non-Africans (i.e., ~40%, ~29% and ~29% in East Asians, Europeans, and South Asians, respectively; **Tables S37-S39**). H484 was observed at relatively high frequency in Europeans (~60%) but at much lower frequency in populations of South Asian and African descent (~20%; **Tables S37-S39**); H484 was absent in East Asian populations. Furthermore, H334 and H484 harbored multiple alleles with significant *s* estimates (**Figure 5**; **Tables S37-S39**). Likewise, haplotypes specific to populations of African descent (such as H454 and H551) carried alleles with significant *s* values (**Table S43**). These African-specific haplotypes also exhibited reticulations—represented by loops in the haplotype network—which indicate past recombination (**Figure 5**).

In silico **functional analysis**

Given that gene expression differences are known to contribute to disease susceptibility and/or severity [84–86], we analyzed the common SNPs that were inferred to be adaptive using the SNP2TFBS tool [87], which predicts whether or not polymorphisms of interest occur in transcription factor binding sites (TFBSs) [87, 88]. This analysis identified two SNPs at *IL-4* (rs56174361 and rs56408050) which were inferred to affect the binding of the ERG2 transcription factor (TF). Furthermore, one SNP at *CCL2* (rs2857657)—located in a TFBS—is predicted to affect the binding of the SOX10 and ZNF263 transcription factors (**Table 1**). At *SLC11A1*, *rs*2279014 and *rs*2279015 are predicted to alter the binding

of the CEBPA and ERG1, respectively (**Table 1**). In contrast, none of the SNPs at *TLR2* overlapped with known TFBSs in the SNP2TFBS database.

Genotype-phenotype analysis

Because we hypothesized that loci associated with immune-mediated diseases could also influence severe COVID-19 outcomes, we focused our analyses on previously identified candidate loci associated with immune-related illnesses in laboratory-confirmed SARS-CoV-2-postitive individuals in two GEN-COVID cohorts. First, we examined comorbidities known to increase the risk of severe COVID-19 along with genotypes from our four genes in a cohort of 1,141 SARS-CoV-2-positive individuals. Using a linear mixed model approach, we detected a significant association between the missense *rs*139227237 mutation in the coding region of *TLR2* and organ transplant status ($P < 0.03$; **Table S44**). Notably, the *rs*139227237 C-allele [p.(Phe217Ser)] was present in SARS-CoV-2-infected individuals who received an organ transplant and absent in SARS-CoV-2-positive individuals without a transplant history. This finding is particularly relevant given that organ transplant recipients are known to be at higher risk for severe COVID-19 [89, 90]. We also detected a marginally significant association between *rs*139227237 and *rs*5743704 [p.(Pro631His)], another missense mutation in *TLR2*, and the incidence of cardiac infarction ($P=0.056$; **Table S45**). Although this result was not significant after statistical correction, it is notable that the *rs*139227237 C-allele and the *rs*5743704 A-allele were present in individuals who had a cardiac infarction while these alleles were absent in individuals who had not experienced this cardiac event. Additionally, we did not observe any associations between genetic variation at the other immunity genes and comorbidity in this dataset (**Tables S44 and S45**).

We further explored the effect of the coding region on COVID-19 outcomes in an independent dataset. Specifically, we analyzed whole exome sequencing data from a separate cohort of 2,920 SARS-CoV-2-positive individuals from GEN-COVID. These participants were classified by degrees of COVID-19 severity (*i.e.*, not hospitalized, hospitalized with no oxygen therapy, hospitalized with oxygen therapy, hospitalized with CPAP/BiPAP, hospitalized with intubation, and deceased). When we examined these phenotypes as a binary trait (*i.e.*, not-hospitalized versus the more serious outcomes), we detected a significant association at *rs*5743709 in *TLR2* [p.(Ala781Ala)] ($P < 0.035$; **Table S46**). In addition, when we treated levels of COVID-19 severity as quantitative phenotypes (*i.e.*, as graded levels of COVID-19 severity), we observed a significant association between the *rs*5743709 A-allele and COVID-19 severity ($P < 0.008$; **Table S47**) using a mixed linear regression model; this linear regression analysis also revealed a small to medium effect size for this variant (**Table S47**). We did not detect any significant associations in the other immunity genes. Lastly, we found that our candidate risk alleles (at *rs*139227237 and *rs*5743709)

generally occurred at low frequency in Eurasian populations (between ~0.05% and ~1%) but were absent in populations of African descent in the 1000 Genomes Project (**Figure 6**; **Table S2**).

Discussion

Evidence for positive selection

Our genetic analyses uncovered strong evidence for positive selection at *IL-4*, *TLR2*, *CCL2*, and *SLC11A1* in African and non-African populations. Specifically, we detected significant deviations from neutrality, extended haplotype homozygosity around either derived or ancestral alleles, high levels of population differentiation, and significant *s* estimates at multiple loci at these genes. Taken together, these analyses led to the identification of 72 common SNPs that were inferred to be under selection. Furthermore, several of these SNPs were found on different common haplotypes, suggesting these haplotypes could have risen to high frequency in populations due to multiple-locus soft sweeps [91, 92]. Under this model, no single favored mutation sweeps to fixation [93, 94]; instead beneficial alleles at different nucleotide sites increase in frequency simultaneously undergoing incomplete selective sweeps. In this scenario, multiple alleles (and haplotypes) can co-exist at intermediate frequency, resulting in strongly positive summary statistics in populations, which we observed at *IL-4* (in Africans and non-Africans), *TLR2* (in non-Africans) and *CCL2* (in both Africans and non-Africans).

Among these three genes, some of the most striking genetic patterns arguably occurred in or around *IL-4* which harbored alleles that were inferred targets of selection, including those shared with Neanderthal/Denisovan species—a finding not previously reported in the literature. This result prompts new questions about whether archaic and modern humans displayed similar phenotypic traits due to their shared alleles and whether these alleles were subject to selection in archaic humans. Furthermore, we observed extreme population differentiation primarily at archaic variation at *IL-4* and a pattern of divergent haplotypes. Sakagami et al. (2004), one of the few evolutionary studies of *IL-4*, reported finding divergent haplotypes in their dataset and proposed that this pattern of variation was likely the result of selection in response to different environmental pressures in populations [95]. Overall, our results and those from this previous study highlight the complex nature of adaptive evolution at this gene.

In comparison, the markedly negative *H* values and extensive EHH mainly around derived alleles at *TLR2* in African populations suggest that this gene likely experienced a classic (or hard) selective sweep in Africa. Similar signals of positive selection have been identified at other loci, such as *IFN*-γ on Chromosome 12 [96] and the *LCT* region on Chromosome 2 associated with lactase persistence [77, 97, 98]. Likewise, the pronounced negative *H* values at *SLC11A1* and the long-range haplotype structure around this gene in non-African populations suggest *SLC11A1* underwent a hard selective sweep, leading to the dominance of the H231 haplotype in these groups. Non-Africans also possessed high-frequency selected

variants that occurred at low frequency or were absent in populations of African descent, suggesting local (or regional) adaptation outside of Africa. Although African and African-descended populations exhibited significantly negative *H* statistics at *SLC11A1*, they lacked the extensive haplotype structure observed in non-Africans. One possible explanation for this pattern is that classic selective sweeps at this gene in ancestral Africans occurred much earlier in time than those in non-Africans, decreasing the likelihood of observing long-range EHH in African populations. In the end, it is feasible that a skew in the site frequency spectrum indicative of positive selection may be present in modern human populations, but long-range haplotype structure may be absent [99]. In addition, we found haplotypes carrying selected variants that were specific to populations of African descent, likely reflecting local (or regional) adaptation. As the primary interface with the environment, it is not surprising that immunity genes have been subject to local pathogen-driven selection over evolutionary time.

Functional effects of common adaptive loci

Using the SNP2TFBS tool, we identified four inferred targets of selection that were predicted to alter the binding of multiple transcription factors, potentially affecting gene expression. These transcription factors included EGR2 (rs56174361 and rs56408050 in *IL-4*); SOX10 and ZNF263 (*rs*2857657 in *CCL2*); EGR1 and C/EBPα (*rs*2279015 and rs2279014, respectively, in *SLC11A1*). Previous studies have shown that EGR2 can influence susceptibility to autoimmune eye disorders [100]. Although the function of ZNF263 is not well-understood, this transcription factor has been implicated in basic cellular processes, such as transcriptional repression [101]. SOX10 has been correlated with rare genetic disorders, such as Waardenburg Syndrome [102] and Hirschsprung Disease [103]. Additionally, previous studies have reported that ERG1 regulates cell growth, proliferation, and apoptosis, as well as fibrotic gene expression [104, 105]. Finally, the C/EBPα transcription factor recognizes the CCAAT motif in the promoters of target genes. Collectively, these findings suggest that adaptive loci identified in our study could contribute to gene expression regulation, conferring a fitness advantage.

While computational approaches can be powerful for predicting the impact of nucleotide variation on transcription factor binding [106–112], it is important to recognize their current limitations. In particular, only a small fraction of transcription factors in the genome has been functionally characterized [113]. As a result, not all of the TFBSs, containing our SNPs of interest, have known corresponding transcription factors, limiting our ability to infer function *in silico*. Therefore, future *in vitro* and *in vivo* experiments will be needed to determine the extent to which our loci influence gene expression.

Relationship between positive selection and risk loci

Interestingly, we found that several common SNPs—previously implicated in respiratory disease risk—were inferred to be under selection. For example, studies have reported that the *rs*13900 C-allele located in the 3′-UTR of *CCL2* was correlated with a decreased level of cytokine production [72, 114], predisposing individuals to infectious disease [115–117]. Our analyses showed that the *rs*13900 T-allele was under selection. Furthermore, *rs*1898830 G-allele in *TLR2* and rs4586 C-allele in *CCL2*, each identified as targets of selection in our study, have been previously implicated in respiratory disease [65, 69–71]. More explicitly, a recent genetic analysis showed that the frequency of the derived *rs*1898830 G-allele was significantly lower in TB patients compared to controls in the Chinese Han population—a finding that was replicated in the Tibetan population [71]. In a separate study, the *rs*4586 T-allele at *CCL2* also was significantly associated with an increased risk for chronic obstructive pulmonary disease, which is characterized by airflow obstruction, chronic inflammation, and a progressive decline in lung function)[65]. Thus, the *rs*13900 T-allele, *rs*1898830 G allele, and the *rs*4586 C allele, all inferred to be under selection, are not associated with increased disease risk and may in fact be protective.

Conversely, we also found that alleles under selection were themselves correlated with increased disease susceptibility. For example, a case-control study of individuals from Japan showed that the derived *rs*2279014 T-allele in the 3'-UTR of *SLC11A1* was associated with protection from *Mycobacterium avium* complex, a predominant contributor to disease burden in the region [118, 119]. However, our results showed selection on the ancestral *rs*2279014 C-allele (rather than the *rs*2279014 T-allele) in East Asian samples, including the Japanese. Furthermore, several studies have reported that the ancestral *rs*2070874 T-allele at *IL-4* was correlated with severe respiratory disease and an exacerbated immune response [57, 66–68]. However, our results indicated selection favoring the ancestral *rs*2070874 T-allele in populations of East Asian and African descent (in contrast, we detected selection on the derived *rs*2070874 C-allele in European and South Asian populations). While it may seem counter-intuitive for alleles contributing to disease susceptibility to be favored in populations, this pattern may be explained by antagonistic pleiotropy—a phenomenon in which a gene or mutation affects multiple traits related to fitness and is antagonistic when a beneficial change in one trait is linked to a disadvantageous change in another within a given environment [120, 121]. Thus, these selected alleles could have been beneficial in one context but costly in another context, particularly as selective pressures change over time. Explicitly, further functional and evolutionary studies are needed to explore this possibility.

Genotype-phenotype associations

Our analysis of SARS-CoV-2-positive cases from the GEN-COVID Multicenter Study uncovered low-frequency polymorphisms in the *TLR2* coding region that influence COVID-19 and COVID-19-related

outcomes. These findings are consistent with other studies that have reported association between host polymorphisms, particularly in genes related to immunodeficiency or inflammasomes, and severe COVID-19 (Elhabyan et al., 2020). For example, mRNA expression levels of Toll-like receptor genes, including *TLR2*, have been associated with disease severity in SARS-CoV-2-infected individuals [19, 39]. Furthermore, rare mutations in the toll-like receptor 7 (TLR-7) gene—which is involved in the innate antiviral immune response—were correlated with severe COVID-19 in patients who were admitted to the ICU in Italy [122]. The authors also found that these mutations suppress the production of interferon, predisposing individuals to severe disease [122]. Similarly, a separate case–control study identified rare missense mutations at *TLR7* in patients with life-threatening COVID-19 [123]. In addition, rare mutations in other immunity genes (e.g., *TLR3*, *IL12RB1*, *TBK1*, and *IFNGR2*) have been argued to play a role in the onset of severe COVID-19 symptoms [122–126]. In contrast, a recent large-scale, multiethnic study of >500,000 individuals found no significant associations between protein-coding variants (ascertained through whole exome sequencing) and COVID-19 severity after applying a stringent Bonferroni correction [127]. However, given this study design, we argue that this analysis would likely not detect any variants with small to medium effect sizes, overlooking novel associations. In contrast, our study focused on alleles that were previously implicated in immune-mediated diseases, allowing us to identify variants with small pleiotropic effects that would otherwise be missed in broader genome-wide analyses. Nonetheless, we recognize that our findings require replication in independent cohorts and/or further functional validation to determine the biological relevance of the variants identified in our study (as well as the ones reported in prior genetic association analyses).

The candidate loci identified in our study have also been implicated in other immune-related diseases in other analyses. For example, data have indicated that the *rs*139227237 [p.(Phe217Ser)] plays a role in abnormal innate immunity signaling and the pathophysiology of myelodysplastic syndromes [128]. Furthermore, the synonymous *rs*5743709 [p.(Ala781Ala)] SNP has been associated with the risk of tuberculosis [129]. Although *rs*5743709 does not alter the amino acid sequence, it is plausible that this SNP could be in strong LD with a functional variant (or exist in a regulatory motif involved in splicing) that directly influences disease risk. Given the overlap in symptoms between severe COVID-19 and other acute illnesses [130, 131], it is possible that the same genetic mutations could affect multiple immune-related diseases (i.e., these alleles could be pleiotropic). Indeed, the pleiotropic effects of variants on disease have been reported in several association studies of complex traits, including disease susceptibility [132–134].

Finally, in contrast to common variants, we found that the low-frequency candidate loci associated with severe COVID-19 were not directly under selection. Given the low frequency of these variants outside of Africa and their absence in our African samples, it is likely these polymorphisms arose more recently following the out-of-Africa migration of modern humans.

Concluding Remarks

In summary, we uncovered evidence of natural selection and observed the presence of trans-species polymorphisms in modern populations. Furthermore, we identified low-frequency alleles in *TLR2* coding region that are correlated with COVID-19 severity. Although these mutations likely explain a small fraction of COVID-19 variability, a more detailed study of this genetic variation—and other low-frequency variants of small to medium effect—could lead to new insights into the pathophysiology of severe COVID-19. Given the persistence of COVID-19, there is a growing need to identify additional predictive biomarkers of risk for severe disease. In addition to acute disease, long COVID, which is characterized by lingering symptoms affecting multiple organ systems, remains a poorly understood and understudied consequence of SARS-CoV-2 infection. Expanding genetic studies to identify variants associated with post-viral syndromes, such as long COVID, may reveal important pathways underlying chronic immune dysregulation and recovery.

As humans continue to transform their environments, emerging respiratory diseases and their global spread will be one of the grand challenges that countries around the world will face over the coming decades. To minimize the impact of these respiratory illnesses, novel statistical methods and a concerted effort to include diverse populations in genomic studies will be needed. Such approaches will be essential for identifying additional risk alleles that influence disease susceptibility and severity in different populations. Ultimately, the findings from these studies will deepen our understanding of the genetic architecture underlying differential infectious disease risk and help guide the development of effective therapeutic interventions.

Methods

Population samples

We analyzed nucleotide and haplotype variation in four immune-related genes: *IL-4* (8,698 base pairs [bps]) on chromosome 5, *TLR2* (22,623 bps) on chromosome 4, *CCL2* (2,293 bps) on chromosome 17, and *SLC11A1* (14,926 bps) on chromosome 2 in 2,039 unrelated individuals from the 1000 Genomes Project (Phase 3). These individuals originated from Africa, South Asia, Europe and East Asia [135]. In this dataset, Africa consisted of 99 Esan (ESN, Nigeria), 113 Mandinka (GWD, Gambia), 54 Mende (MSL, Sierra Leone), 107 Yoruba (YRI, Nigeria), 99 Luhya (LWK, Kenya), 60 African Americans (ASW, United States), and 96 Barbadians (ACB, Barbados). Throughout this paper, the term "African populations" refers to indigenous African populations only while the term "African and African-descended populations" refers to indigenous Africans and African Americans/Afro-Caribbean populations, respectively. In addition, we periodically refer to these populations collectively as populations of African descent. South Asia included

84 Bengali (BEB, Bangladesh), 103 Gurjarati (GIH, Gujarati Indian recruited in Houston, Texas), 99 Indian Telugu (ITU, Indian Telugu recruited in the United Kingdom), 93 Punjabi (PJL, Pakistan), and 99 Tamil (STU, Sri Lanka). Europe consisted of 90 Great British (GBR, Great Britain), 99 Finnish (FIN, Finland), 107 Iberian (IBS, Spain), and 104 Toscani Italians (TSI, Italy). East Asia encompassed 93 Dai (CDX, China), 103 Han (CHB, Han Chinese in Beijing, China), 105 Han (CHS, Han Chinese in South China), 104 Japanese (JPT, Japan), and 99 Kinh (KHV, Vietnam). Individual chromosomes for the Altai Neanderthal genome were downloaded from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/. Individual chromosomes for the Denisovan genome were downloaded from http://cdna.eva.mpg.de/denisova/VCF/human/.

To investigate the genetic architecture of COVID-19 severity, we examined genetic and clinical data from GEN-COVID, a multicenter network of more than 40 hospitals and affiliated partners that collected high-quality samples and data across Italy for COVID-19 research under the direction of Dr. Alesandra Renieri. More specifically, we analyzed variants genotyped with the Illumina Global Screening Array-24 v3.0 + Multi-Disease BeadChip together with traits related to COVID-19 severity in a cohort of 1,141 SARS-CoV-2-infected participants. In addition, we analyzed whole-exome sequencing data and graded levels of COVID-19 severity in a different cohort of 2,920 SARS-CoV-2-positive individuals from GEN-COVID. The whole exome data were generated using the Illumina NovaSeq6000 System (Illumina, San Diego, CA, USA) with ≥ 20 x coverage. Variant calling was also performed using the GATK best practice guidelines. Notably, there is no overlap of SARS-CoV-2-positive individuals between the genotyping and whole exome sequencing datasets.

Nucleotide variation

We extracted variant calls (from vcf files; build GRCh37 human assembly) for the *IL-4, TLR2, CCL2,* and *SLC11A1* genes in the 1000 Genomes using `vcftools` [136]. The genomic coordinates (*i.e.*, the start and end positions) for these genes were obtained from the NCBI database. The minor allele frequency (MAF) was defined as the second most frequent allele at a given site in pooled populations (*i.e.*, 2N=4078). SNPs were also broadly classified as common or rare based on the MAF at a given site. More explicitly, SNPs with an MAF ≥ 5% were classified as common, while SNPs with an MAF < 5% were classified as rare.

Tests of neutrality

To determine if *IL-4, TLR2, CCL2,* and *SLC11A1* evolved neutrally, we calculated two summary statistics, Fay and Wu's H (H)[137] and Tajima's D (D_T)[138], for each population using DnaSP [75]. These statistical tests analyze different aspects of the site frequency spectrum [139], providing information

about the evolutionary forces that have influenced patterns of polymorphic variation in populations. We also simulated theoretical H and D_T values within a coalescent framework under different demographic models [140–145] using the `msprime` software [146, 147]. For African populations, we simulated a 2–, 4–, 6–, 8–, and 10– fold increase in population size (from an initial population size of 7,300 individuals) starting at 148,000 years ago (ya) until the present (2,000 replicates per simulation).

For non-Africans, we simulated demographic models involving a population bottleneck (consisting of 1,861 individuals) 51,000 years ago, which corresponds to the approximate time anatomically modern humans migrated out of Africa [143, 148, 149], followed by a 20–, 40–, 60–, 80–, and 100– fold increase in population size starting at 23,000 years ago until the present (2,000 replicates per simulation)[140, 141, 143–145, 150]. For all simulations, we used a mutation rate of 2.5×10^{-8}, recombination rate of 1×10^{-8}, and a generation time of 28 years [143–145]. We also incorporated bidirectional migration between Africans and non-Africans at ~30,000 ya and ~20,000 ya using the number of migrants given in Gravel et al. [143]. Finally, we assessed the statistical significance of observed H and D_T values by comparing them to a distribution of theoretical values. In cases where observed values were either negative or positive in populations, we tested for significance in one direction (*i.e.*, a one-tailed test, $P < 0.05$). In cases where observed values were a mix of positive and negative numbers in populations, we tested for significant difference in both directions (*i.e.*, a two-tailed test, $P < 0.05$).

Long-range haplotype homozygosity

We characterized the length of haplotype homozygosity on Chromosomes 2 (*SLC11A1*), 4 (*TLR2*), 5 (*IL-4*), and 17 (*CCL2*) for each population using the *i*HS statistic calculated with the `selscan` and `Polaris` packages [76, 142, 151]. To identify outlier *i*HS values, the unstandardized scores for >600,000 SNPs across the ~240 Mb region of Chromosome 2, >520,000 SNPs across the ~190 Mb region of Chromosome 4, >470,000 SNPs across the ~180 Mb region of Chromosome 5, and >180,000 SNPs across the ~80 Mb region of Chromosome 17 in each population were normalized using the norm program implemented in `selscan` [76]. Standardized | *i*HS | > 2 represents the top 5% of | iHS | scores. We also calculated nS_L, another haplotype-based statistic, to estimate the length of haplotype homozygosity in populations using selscan [152]. Unlike the *i*HS, this statistic measures haplotype lengths based on the number of segregating sites in a sample and does not depend on the recombination rate [152], making it robust to recombination rate variation. Furthermore, nS_L can detect selective sweeps from standing variation and incomplete selective sweeps [152]. The resulting nS_L statistics were normalized using the norm software [76]. By convention, selscan reports positive statistics when the derived allele (coded as "1") is under selection and negative statistics when the ancestral allele (coded as "0") is favored [76, 153–155], which is opposite to the orientation described in Voight *et al.* [142].

Finally, we quantified the decay of identity of haplotypes with distance by calculating the extended haplotype homozygosity (EHH) statistic with `selscan` and `Polaris` [76, 78, 151] using loci with extreme | iHS | and/or | nS_L | scores identified above as core SNPs. If we observed outlier standardized | iHS | and/or | nS_L | statistics along with long-range EHH for a given SNP, we classified that SNP as a target of selection.

Selection coefficient estimates

The selection coefficient (s) quantifies the fitness effects of mutations that arise in populations, providing insight into the impact of selection at the genetic level. We estimated s for alleles with an MAF ≥ 5% in our immunity genes using the CLUES2 algorithm implemented in the CLUES2 Companion package [79, 81]. CLUES2 incorporates ancestral recombination graphs (ARGs)—which capture the history of genomic regions in a sample of sequences—along with importance sampling of ARGs to estimate s in different time periods [79]. Thus, CLUES2 utilizes the full set of allele histories in a sample of sequences to estimate s [79]. We calculated s at different time depths (i.e., 1,000, 500, and 200 generations ago, which correspond to 28,000, 14,000, and 5,600 ya assuming a generation time of 28 years) in each gene using CLUES2 Companion [81]. Because selection intensity can vary over time [79], we aimed to capture time points at which selection coefficients might change. SNP loci with significant s estimates (i.e., -log(p) > 1.3 or $P < 0.05$) at any time depth were classified as targets of positive selection, complementing the | iHS |, | nS_L |, and EHH results.

Population differentiation

To measure the degree of genetic divergence among populations, we calculated among-population F_{ST} at individual polymorphic sites across each gene using the Weir and Cockerham method in vcftools [136, 156]. The observed F_{ST} estimates for SNPs in the *IL-4*, *TLR2*, *CCL2*, and *SLC11A1* genes were then compared against an empirical distribution of F_{ST} values derived from 8.1 million genome-wide SNPs that were randomly chosen to identify outlier values (≥ 95th percentile of the distribution). This F_{ST} empirical distribution was generated using the same 21 populations that were used to calculate the observed values. Based on this approach, we determined that $F_{ST} > 0.286$ represented the top 95th percentile of empirical values.

Haplotype inferences and relationships

Fully phased haplotypes from the 1000 Genomes Project were extracted for 2,039 individuals using `vcftools` [136]. A custom script was applied to the already phased sequence data to identify unique haplotypes along with the count of each haplotype in the dataset. Next, the genealogical relationships among haplotypes were reconstructed using the median-joining method implemented in the `Network v. 5.0`

software [82]. For each gene, the resulting visual tree for each gene was the one with the least number of changes among all possible trees [157].

In silico functional analysis

We applied the SNP2TFBS tool to polymorphisms that were inferred to be under selection [87]. This tool assesses whether or not SNPs of interest map to transcription factor binding sites (TFBSs) and outputs a list of SNPs that overlap with TFBSs along with the transcription factors that interact with them [87].

Genetic association analyses

We examined genotypic and clinical data from laboratory-confirmed SARS-CoV-2-postitive individuals enrolled in the GEN-COVID Multicenter Study in Italy. The GEN-COVID cohorts consisted of SARS-CoV-2-infected individuals from whom DNA and detailed clinical information were collected through a consortium of more than 40 Italian Hospitals, 16 Continuity Assistance Special Units, and eight Departments of Preventive Medicine for host genetic analysis.

In our first GEN-COVID cohort of 1,141 SARS-CoV-2-positive individuals, we analyzed comorbidities (specifically, cancer, chronic kidney, chronic pulmonary, transplant, heart failure, diabetes, immunocompromised, asthma, hypertension, cardiac infarction, stroke, and ICU admission status) known to increase risk for severe COVID-19 along with genotypes from our four genes. This dataset consisted of 648 males and 493 females ranging in age from 18 to 99 years. Prior to testing for any genetic associations, we generated a genetic relationship matrix (GRM) between pairs of individuals using genome-wide SNPs in the 1,141 SARS-CoV-2-positive individuals to account for population structure and familial relatedness with the GTCA software [158]. We then used a linear mixed model approach implemented in GCTA [158] to test for associations between genetic variation and traits related to severe COVID-19 in our cohort (sex and age were included in this analysis as covariates). Statistical significance was determined after applying: 1) the saddle point approximation (SPA) method to correct for unbalanced samples sizes in comparative groups; and 2) the Benjamini-Hochberg method to correct for multiple testing, mitigating p-value inflation.

Additionally, we analyzed whole exome sequencing data in a second GEN-COVID cohort comprised of 2,920 SARS-CoV-2-infected individuals (1,730 males and 1,189 females ranging in age from a few months to 99 years). These 2,920 SARS-CoV-2-positive cases were also classified by degrees of COVID-19 severity; specifically, score 0=not hospitalized (N=499), 1=hospitalized with no oxygen therapy (N=406), 2=hospitalized with oxygen therapy (N=946), 3=hospitalized with CPAP/BiPAP intervention (N=668), 4=hospitalized with intubation (N=207), and 5=deceased (N=194). Severe COVID-19 in the current study is defined as the stage of illness requiring hospitalization or resulting in death. For this

analysis, we generated a new GRM using the whole exome sequencing data from our 2,920 SARS-CoV-2-positive cases. We then employed linear mixed model approaches implemented in `GCTA` [158] to examine the effects of exonic variants on COVID-19 severity. More specifically, we analyzed phenotypes as binary and quantitative traits using logistic and linear methods, respectively. In both analyses, the GRM was incorporated as a random variable to account for population structure and familial relatedness; sex and age were included as covariates. Statistical significance was determined after applying SPA and Benjamini-Hochberg corrections. It is important to note that there was no overlap of individuals between the first and second cohorts of SARS-CoV-2-infected individuals.

Acknowledgements

We thank the Center for Computational Biology and Bioinformatics (CCBB) at Howard University and the Center for Advanced Research Computing at the University of Southern California for providing computational resources essential to this study. We are also grateful to Professor Alessandra Renieri at the University of Siena (Italy) for providing the genotype–phenotype data used in our analyses.

Funding

This work was supported by funds from the National Geographic Society [grant HJ–116ER–17] to C.N.C, the U.S. National Science Foundation (NSF) grant IOS-1355034, Howard University College of Medicine, and the District of Columbia Center for AIDS Research, an NIH funded program [P30AI117970] to T.H., as well as U.S. NSF grant BCS–2221924 and U.S. NSF grant BCS–2221920 to M.C.C.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

All authors contributed to the study conception and design. Material preparation and analyses were performed by Christopher N. Cross, Alessandro Lisi, Faith C. Simmonds, and Michael C. Campbell. The first draft of the manuscript was written by Christopher N. Cross and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability

The genetic diversity data underlying this article are available in the 1000 Genomes Consortium Project at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The clinical datasets analyzed for the current study are available from GEN-COVID upon reasonable request.

Abbreviations

ACB: African Caribbeans in Barbados

ACE2: Angiotensin-Converting Enzyme 2

ANPEP: Alanyl Aminopeptidase

APC: Antigen-Presenting Cell

ARG: Ancestral Recombination Graph

ASW: African Ancestry in Southwest US

BEB: Bengali in Bangladesh

bp: Base Pair

CCL2: C-C Motif Chemokine Ligand 2

CCL3: C-C Motif Chemokine Ligand 3

CCBB: Center for Computational Biology and Bioinformatics

CDX: Chinese Dai in Xishuangbanna, China

CHB: Han Chinese in Beijing, China

CHS: Southern Han Chinese

CLUES2: Composite Likelihood for Estimating Selection, version 2

COVID-19: Coronavirus Disease 2019

CPAP: Continuous Positive Airway Pressure

CTSB: Cathepsin B

CTSL: Cathepsin L

CXCL10: C-X-C Motif Chemokine Ligand 10

DPP4: Dipeptidyl Peptidase-4

DNA: Deoxyribonucleic Acid

D_T: Tajima's D Statistic

EHH: Extended Haplotype Homozygosity

ER: Endoplasmic Reticulum

ESN: Esan in Nigeria

FIN: Finnish in Finland

F_{ST}: Fixation Index

FURIN: Furin Protease

GBR: British in England and Scotland

GEN-COVID: Genetic and Clinical Data Collection for COVID-19

GIH: Gujarati Indian in Houston, Texas

GRCh37: Genome Reference Consortium Human Build 37

GRM: Genetic Relationship Matrix

GWD: Gambian in Western Division, The Gambia

H: Fay and Wu's H Statistic

HBV: Hepatitis B Virus

HCV: Hepatitis C Virus

HIV: Human Immunodeficiency Virus

HLA: Human Leukocyte Antigen

IBD: Inflammatory Bowel Disease

IBS: Iberian population in Spain

ICU: Intensive Care Unit

IFN-α: Interferon-alpha

IFN-γ: Interferon-gamma

iHS: Integrated Haplotype Score

IL: Interleukin

IL-4: Interleukin-4

IL-6: Interleukin-6

IL-10: Interleukin-10

ITU: Indian Telugu in the UK

JPT: Japanese in Tokyo, Japan

KHV: Kinh in Ho Chi Minh City, Vietnam

LD: Linkage Disequilibrium

LWK: Luhya in Webuye, Kenya

MAF: Minor Allele Frequency

MSL: Mende in Sierra Leone

mRNA: Messenger RNA

nS_L: Number of Segregating Sites by Length (haplotype-based test)

ORF: Open Reading Frame

PCA: Principal Component Analysis

PJL: Punjabi in Lahore, Pakistan

RNA: Ribonucleic Acid

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2

SLC11A1: Solute Carrier Family 11 Member 1

SNP: Single Nucleotide Polymorphism

SPA: Saddle Point Approximation

STU: Sri Lankan Tamil in the UK

TB: Tuberculosis

TF: Transcription Factor

TFBS: Transcription Factor Binding Site

TLR: Toll-like Receptor

TLR2: Toll-like Receptor 2

TMPRSS2: Transmembrane Protease, Serine 2

TMPRSS11D: Transmembrane Protease, Serine 11D

TSI: Toscani in Italy

UTR: Untranslated Region

VCF: Variant Call Format

YRI: Yoruba in Ibadan, Nigeria

References

1. Mallah SI, Ghorab OK, Al-Salmi S, Abdellatif OS, Tharmaratnam T, Iskandar MA, et al. COVID-19: breaking down a global health crisis. Annals of Clinical Microbiology and Antimicrobials. 2021;20.
2. Herman H, Gharini PPR, Dinarti LK. Clinical cardiac manifestations in patients with coronavirus disease 2019 (COVID-19). Journal of thee Medical Sciences (Berkala Ilmu Kedokteran). 2020;52.
3. Basu-Ray I, Almaddah N k., Adeboye A, Soos MP. Cardiac Manifestations Of Coronavirus (COVID-19). In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022.
4. Grolmusz VK, Bozsik A, Papp J, Patócs A. Germline Genetic Variants of Viral Entry and Innate Immunity May Influence Susceptibility to SARS-CoV-2 Infection: Toward a Polygenic Risk Score for Risk Stratification. Front Immunol. 2021;12:653489.
5. Musa HH, Musa TH, Musa IH, Musa IH, Ranciaro A, Campbell MC. Addressing Africa's pandemic puzzle: Perspectives on COVID-19 transmission and mortality in sub-Saharan Africa. Int J Infect Dis. 2021;102:483–8.
6. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature. 2021;600:472–7.
7. Kucher AN, Babushkina NP, Sleptcov AA, Nazarenko MS. Genetic Control of Human Infection with SARS-CoV-2. Russian Journal of Genetics. 2021;57:627–41.
8. Zhang C, Verma A, Feng Y, Melo MCDR, McQuillan M, Hansen M, et al. Impact of natural selection on global patterns of genetic variation, and association with clinical phenotypes, at genes involved in SARS-CoV-2 infection. Res Sq. 2021. https://doi.org/10.21203/rs.3.rs-673011/v1.
9. Wang L, Balmat TJ, Antonia AL, Constantine FJ, Henao R, Burke TW, et al. An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. Genome Med. 2021;13:83.
10. Consortium MGS, Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520–62.
11. Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, et al. New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC Med. 2020;18:216.
12. De Sanctis JB, García AH, Moreno D, Hajduch M. Coronavirus infection: An immunologists' perspective. Scand J Immunol. 2021;93:e13043.

13. Zhang C, Verma A, Feng Y, Melo MCR, McQuillan M, Hansen M, et al. Impact of natural selection on global patterns of genetic variation and association with clinical phenotypes at genes involved in SARS-CoV-2 infection. Proc Natl Acad Sci U S A. 2022;119:e2123000119.
14. Byun J, Han Y, Walsh KM, Park AS, Bondy ML, Amos CI. Shared genomic architecture between COVID-19 severity and numerous clinical and physiologic parameters revealed by LD score regression analysis. Sci Rep. 2022;12:1891.
15. Langton DJ, Bourke SC, Lie BA, Reiff G, Natu S, Darlay R, et al. The influence of HLA genotype on the severity of COVID-19 infection. Hladnikia . 2021;98:14–22.
16. Fricke-Galindo I, Falfán-Valencia R. Genetics Insight for COVID-19 Susceptibility and Severity: A Review. Front Immunol. 2021;12:622176.
17. Del Valle DM, Kim-Schulze S, Huang H-H, Beckmann ND, Nirenberg S, Wang B, et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. Nat Med. 2020;26:1636–43.
18. Velavan TP, Pallerla SR, Rüter J, Augustin Y, Kremsner PG, Krishna S, et al. Host genetic factors determining COVID-19 susceptibility and severity. EBioMedicine. 2021;72:103629.
19. Zheng M, Karki R, Williams EP, Yang D, Fitzpatrick E, Vogel P, et al. TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. Nat Immunol. 2021;22:829–38.
20. Raza RZ, Abbasi SW. An Evolutionary Insight Into the Heterogeneous Severity Pattern of the SARS-CoV-2 Infection. Frontiers in Genetics. 2022;13.
21. Wu P, Ding L, Li X, Liu S, Cheng F, He Q, et al. Trans-ethnic genome-wide association study of severe COVID-19. Commun Biol. 2021;4:1034.
22. Group TSC-19 G, The Severe Covid-19 GWAS Group. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. New England Journal of Medicine. 2020;383:1522–34.
23. Khan S, Shafiei MS, Longoria C, Schoggins JW, Savani RC, Zaki H. SARS-CoV-2 spike protein induces inflammation via TLR2-dependent activation of the NF-κB pathway. Elife. 2021;10.
24. Han H, Ma Q, Li C, Liu R, Zhao L, Wang W, et al. Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors. Emerg Microbes Infect. 2020;9:1123–30.
25. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591:92–8.
26. Totura AL, Whitmore A, Agnihothram S, Schäfer A, Katze MG, Heise MT, et al. Toll-like receptor 3 signaling via TRIF contributes to a protective innate immune response to severe acute respiratory syndrome Coronavirus infection. MBio. 2015;6:e00638-15.
27. Zguro K, Fallerini C, Fava F, Furini S, Renieri A. Host genetic basis of COVID-19: from methodologies to genes. Eur J Hum Genet. 2022. https://doi.org/10.1038/s41431-022-01121-x.
28. Saba AA, Nur J, Alam MS, Howlader ZH, Islam LN, Nabi AHMN. Missense variant rs75603675 within TMPRSS2 gene is associated with the increased risk of severe form of COVID-19. Gene Rep. 2024;37:102039.
29. Loktionov AV, Kobzeva KA, Karpenko AR, Sergeeva VA, Orlov YL, Bushueva OY. GWAS-significant loci and severe COVID-19: analysis of associations, link with thromboinflammation syndrome, gene-gene, and gene-environmental interactions. Front Genet. 2024;15:1434681.
30. Kim GJ, Elnaggar JH, Varnado M, Feehan AK, Tauzier D, Rose R, et al. A bioinformatic analysis of T-cell epitope diversity in SARS-CoV-2 variants: association with COVID-19 clinical severity in the United States population. Front Immunol. 2024;15:1357731.
31. Sun Z, Pan L, Tian A, Chen P. Critically-ill COVID-19 susceptibility gene CCR3 shows natural selection in sub-Saharan Africans. Infect Genet Evol. 2024;121:105594.
32. Badla BA, Hanifa MS, Jain R, Naofal ME, Halabi N, Yaslam S, et al. Genetic determinants of severe COVID-19 in young Asian and Middle Eastern patients: a case series. Sci Rep. 2023;13:20294.
33. Araújo A, Sgorlon G, Aguiar LE, Cidrão MHMC, Teixeira KS, Villalobos Salcedo JM, et al. Influence of polymorphic variations of IFNL, HLA, and IL-6 genes in severe cases of COVID-19. Exp Biol Med (Maywood). 2023;248:787–97.

34. Butler-Laporte G, Povysil G, Kosmicki JA, Cirulli ET, Drivas T, Furini S, et al. Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative. PLoS Genet. 2022;18:e1010367.
35. Rizvi S, Rizvi SMS, Raza ST, Abbas M, Fatima K, Zaidi ZH, et al. Implication of single nucleotide polymorphisms in Interleukin-10 gene (rs1800896 and rs1800872) with severity of COVID-19. Egypt J Med Hum Genet. 2022;23:145.
36. Degenhardt F, Ellinghaus D, Juzenas S, Lerga-Jaso J, Wendorff M, Maya-Miles D, et al. Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Hum Mol Genet. 2022;31:3945–66.
37. Namkoong H, Edahiro R, Takano T, Nishihara H, Shirai Y, Sonehara K, et al. DOCK2 is involved in the host genetics and biology of severe COVID-19. Nature. 2022;609:754–60.
38. Balanovsky O, Petrushenko V, Mirzaev K, Abdullaev S, Gorin I, Chernevskiy D, et al. Variation of genomic sites associated with severe covid-19 across populations: Global and national patterns. Pharmgenomics Pers Med. 2021;14:1391–402.
39. Schimke LF, Marques AHC, Baiocchi GC, de Souza Prado CA, Fonseca DLM, Freire PP, et al. Severe COVID-19 Shares a Common Neutrophil Activation Signature with Other Acute Inflammatory States. Cells. 2022;11.
40. Zaid Y, Doré É, Dubuc I, Archambault A-S, Flamand O, Laviolette M, et al. Chemokines and eicosanoids fuel the hyperinflammation within the lungs of patients with severe COVID-19. J Allergy Clin Immunol. 2021;148:368-380.e3.
41. Ramasamy S, Subbian S. Critical Determinants of Cytokine Storm and Type I Interferon Response in COVID-19 Pathogenesis. Clin Microbiol Rev. 2021;34.
42. Tang L, Yin Z, Hu Y, Mei H. Controlling Cytokine Storm Is Vital in COVID-19. Front Immunol. 2020;11:570993.
43. Chen L-D, Zhang Z-Y, Wei X-J, Cai Y-Q, Yao W-Z, Wang M-H, et al. Association between cytokine profiles and lung injury in COVID-19 pneumonia. Respir Res. 2020;21:201.
44. Liu Y, Zhang C, Huang F, Yang Y, Wang F, Yuan J, et al. Elevated plasma levels of selective cytokines in COVID-19 patients reflect viral load and lung injury. Natl Sci Rev. 2020;7:1003–11.
45. van Blokland IV, Lanting P, Ori APS, Vonk JM, Warmerdam RCA, Herkert JC, et al. Using symptom-based case predictions to identify host genetic factors that contribute to COVID-19 susceptibility. PLoS One. 2021;16:e0255402.
46. Al Khatib HA, Benslimane FM, Elbashir IE, Coyle PV, Al Maslamani MA, Al-Khal A, et al. Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients With Variable Disease Severities. Front Cell Infect Microbiol. 2020;10:575613.
47. Smatti MK, Al-Sarraj YA, Albagha O, Yassine HM. Host Genetic Variants Potentially Associated With SARS-CoV-2: A Multi-Population Analysis. Front Genet. 2020;11:578523.
48. Ben-Ali M, Barbouche M-R, Bousnina S, Chabbou A, Dellagi K. Toll-like receptor 2 Arg677Trp polymorphism is associated with susceptibility to tuberculosis in Tunisian patients. Clin Diagn Lab Immunol. 2004;11:625–6.
49. Meilang Q, Zhang Y, Zhang J, Zhao Y, Tian C, Huang J, et al. Polymorphisms in the SLC11A1 gene and tuberculosis risk: a meta-analysis update. Int J Tuberc Lung Dis. 2012;16:437–46.
50. Liu Y, Zhao E, Zhu L, Zhang D, Wang Z. 3'UTR polymorphisms in NRAMP1 are associated with the susceptibility to pulmonary tuberculosis: A MOOSE-compliant meta-analysis. Medicine. 2019;98:e15955.
51. Stagas MK, Papaetis GS, Orphanidou D, Kostopoulos C, Syriou S, Reczko M, et al. Polymorphisms of the NRAMP1 gene: Distribution and susceptibility to the development of pulmonary tuberculosis in the Greek population. Medical Science Monitor. 2011;17:PH1–6.
52. Ogus AC, Yoldas B, Ozdemir T, Uguz A, Olcen S, Keser I, et al. The Arg753GLn polymorphism of the human toll-like receptor 2 gene in tuberculosis disease. Eur Respir J. 2004;23:219–23.
53. Saleh MA, Ramadan MM, Arram EO. Toll-like receptor-2 Arg753Gln and Arg677Trp polymorphisms and susceptibility to pulmonary and peritoneal tuberculosis. APMIS. 2017;125:558–64.

54. Pabst S, Yenice V, Lennarz M, Tuleta I, Nickenig G, Gillissen A, et al. Toll-Like Receptor 2 Gene Polymorphisms Arg677Trp and Arg753Gln in Chronic Obstructive Pulmonary Disease. Lung. 2009;187:173–8.
55. Sookaromdee P, Wiwanitkit V, Joob B. Genetic polymorphisms namely TLR2 Gene Arg753gln and Arg677trp polymorphisms and tuberculosis: Comparison on the effects of molecular weight alterations. Biomedical and Biotechnology Research Journal (BBRJ). 2019;3:129.
56. Mukherjee S, Huda S, Sinha Babu SP. Toll-like receptor polymorphism in host immune response to infectious diseases: A review. Scand J Immunol. 2019;90:e12771.
57. Patarčić I, Gelemanović A, Kirin M, Kolčić I, Theodoratou E, Baillie KJ, et al. The role of host genetic factors in respiratory tract infectious diseases: systematic review, meta-analyses and field synopsis. Sci Rep. 2015;5:16119.
58. Cortez A, Muxfeldt E. Monocyte chemoattractant protein-1 and hypertension: An overview. Hipertens Riesgo Vasc. 2022;39:14–23.
59. Sabah-Özcan S, Gürel G. The polymorphism rs4696480 in the TLR2 gene is associated with psoriasis patients in the Turkish population. Immunol Lett. 2019;211:28–32.
60. Xu Z, Li J, Yang H, Jiang L, Zhou X, Huang Y, et al. Association of CCL2 gene variants with osteoarthritis. Arch Med Res. 2019;50:86–90.
61. Bank S, Skytt Andersen P, Burisch J, Pedersen N, Roug S, Galsgaard J, et al. Polymorphisms in the inflammatory pathway genes TLR2, TLR4, TLR9, LY96, NFKBIA, NFKB1, TNFA, TNFRSF1A, IL6R, IL10, IL23R, PTPN22, and PPARG are associated with susceptibility of inflammatory bowel disease in a Danish cohort. PLoS One. 2014;9:e98815.
62. Simula ER, Jasemi S, Cossu D, Manca PC, Sanna D, Scarpa F, et al. The genetic landscape of systemic rheumatic diseases: A comprehensive multigene-panel study identifying key gene polymorphisms. Pharmaceuticals (Basel). 2024;17.
63. Zecevic M, Kotur N, Ristivojevic B, Gasic V, Skodric-Trifunovic V, Stjepanovic M, et al. Genome-wide association study of COVID-19 outcomes reveals novel host genetic risk loci in the Serbian population. Front Genet. 2022;13:911010.
64. Ghafouri-Fard S, Noroozi R, Vafaee R, Branicki W, Pośpiech E, Pyrc K, et al. Effects of host genetic variations on response to, susceptibility and severity of respiratory infections. Biomed Pharmacother. 2020;128:110296.
65. Lin C, Wang Z, Shen L, Yi G, Li M, Li D. Genetic Variants, Circulating Level of MCP1 with Risk of Chronic Obstructive Pulmonary Disease: A Case-Control Study. Pharmgenomics Pers Med. 2021;14:561–7.
66. Abhimanyu, Bose M, Jha P, Indian Genome Variation Consortium. Footprints of genetic susceptibility to pulmonary tuberculosis: cytokine gene variants in north Indians. Indian J Med Res. 2012;135:763–70.
67. Paim AAO, Lopes-Ribeiro Á, Daian e Silva DS, Andrade LAF, Moraes TFS, Barbosa-Stancioli EF, et al. Will a little change do you good? A putative role of polymorphisms in COVID-19. Immunology Letters. 2021;235:9–14.
68. Amat F, Louha M, Benet M, Guiddir T, Bourgoin-Heck M, Saint-Pierre P, et al. The IL-4 rs2070874 polymorphism may be associated with the severity of recurrent viral-induced wheeze. Pediatr Pulmonol. 2017;52:1435–42.
69. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323:1239–42.
70. Eldar-Yedidia Y, Hillel M, Cohen A, Bar-Meir M, Freier-Dror Y, Schlesinger Y. Association of toll-like receptors polymorphism and intrauterine transmission of cytomegalovirus. PLoS One. 2017;12:e0189921.
71. Wu S, Liu X, Chen L, Wang Y, Zhang M, Wang M, et al. Polymorphisms of TLR2, TLR4 and TOLLIP and tuberculosis in two independent studies. Biosci Rep. 2020;40.

72. Celik SK, Genc GC, Dursun A. A bioinformatic approach to investigating cytokine genes and their receptor variants in relation to COVID-19 progression. International Journal of Immunogenetics. 2021;48:211–8.
73. Lin, C., Wang Z., Shen L., Yi G., Li M., Li D. Genetic Variants, Circulating Level of MCP1 with Risk of Chronic Obstructive Pulmonary Disease: A Case-Control Study. Pharmacogenomics and personalized medicine. 2021;14:561–7.
74. Sapkota BR, Hijikata M, Matsushita I, Tanaka G, Ieki R, Kobayashi N, et al. Association of SLC11A1 (NRAMP1) polymorphisms with pulmonary Mycobacterium avium complex infection. Hum Immunol. 2012;73:529–36.
75. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34:3299–302.
76. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31:2824–7.
77. Lisi A, Campbell MC. Polaris: Polarization of ancestral and derived polymorphic alleles for inferences of extended haplotype homozygosity in human populations. Bioinformatics. 2025;41.
78. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7.
79. Vaughn AH, Nielsen R. Fast and accurate estimation of selection coefficients and allele histories from ancient and modern DNA. Mol Biol Evol. 2024;41.
80. Nielsen R, Vaughn AH, Deng Y. Inference and applications of ancestral recombination graphs. Nat Rev Genet. 2025;26:47–58.
81. Lisi A, Campbell MC. CLUES2 Companion: Computational pipelines to estimate, visualize, and date selection on multi-locus sites. bioRxiv. 2025. https://doi.org/10.1101/2025.06.10.658975.
82. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16:37–48.
83. Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol. 2001;16:37–45.
84. D'Antonio M, Nguyen JP, Arthur TD, Matsui H, COVID-19 Host Genetics Initiative, D'Antonio-Chronowska A, et al. SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues. Cell Rep. 2021;37:110020.
85. Li J, Liu Y, Kim T, Min R, Zhang Z. Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. PLoS Computational Biology. 2010;6:e1000910.
86. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun. 2021;12:5647.
87. Kumar S, Ambrosini G, Bucher P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res. 2017;45:D139–44.
88. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42 Database issue:D980-5.
89. Marinelli T, Ferreira VH, Ierullo M, Ku T, Lilly L, Kim SJ, et al. Prospective clinical, virologic, and immunologic assessment of COVID-19 in transplant recipients. Transplantation. 2021;105:2175–83.
90. Vinson AJ, Agarwal G, Dai R, Anzalone AJ, Lee SB, French E, et al. COVID-19 in solid organ transplantation: Results of the National COVID Cohort Collaborative. Transplant Direct. 2021;7:e775.
91. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28:659–69.

92. Hermisson J, Pennings PS. Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation.
93. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010;20:R208-15.
94. Hermisson J, Pfaffelhuber P. The pattern of genetic hitchhiking under recurrent mutation. Electronic Journal of Probability. 2008;13 none.
95. Sakagami T, Witherspoon DJ, Nakajima T, Jinnai N, Wooding S, Jorde LB, et al. Local adaptation and population differentiation at the interleukin 13 and interleukin 4 loci. Genes Immun. 2004;5:389–97.
96. Campbell MC, Smith LT, Harvey J. Population genetic evidence for positive and purifying selection acting at the human IFN-γ locus in Africa. Genes Immun. 2019;20:143–57.
97. Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, et al. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet. 2014;94:496–510.
98. Campbell MC, Ranciaro A. Human adaptation, demography and cattle domestication: an overview of the complexity of lactase persistence in Africa. Hum Mol Genet. 2021;30:R98–109.
99. Lohmueller KE, Bustamante CD, Clark AG. Detecting directional selection in the presence of recent admixture in African-Americans. Genetics. 2011;187:823–35.
100. Li W, He S, Tan J, Li N, Zhao C, Wang X, et al. Transcription factor EGR2 alleviates autoimmune uveitis via activation of GDF15 to modulate the retinal microglial phenotype. Proc Natl Acad Sci U S A. 2024;121:e2316161121.
101. Frietze S, Lan X, Jin VX, Farnham PJ. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. J Biol Chem. 2010;285:1393–403.
102. Ma J, Zhang Z, Jiang H-C, Sun H, Ming C, Zhao L-P, et al. A novel dominant mutation in the SOX10 gene in a Chinese family with Waardenburg syndrome type II. Mol Med Rep. 2019;19:1775–80.
103. Amiel J, Lyonnet S. Hirschsprung disease, associated syndromes, and genetics: a review. J Med Genet. 2001;38:729–39.
104. Bhattacharyya S, Fang F, Tourtellotte W, Varga J. Egr-1: new conductor for the tissue repair orchestra directs harmony (regeneration) or cacophony (fibrosis). J Pathol. 2013;229:286–97.
105. Aliperti V, Sgueglia G, Aniello F, Vitale E, Fucci L, Donizetti A. Identification, Characterization, and Regulatory Mechanisms of a Novel EGR1 Splicing Isoform. Int J Mol Sci. 2019;20.
106. Povysil G, Hochreiter S. IBD Sharing between Africans, Neandertals, and Denisovans. Genome Biol Evol. 2016;8:3406–16.
107. Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, et al. The timing and spatiotemporal patterning of Neanderthal disappearance. Nature. 2014;512:306–9.
108. Hohenlohe PA, Phillips PC, Cresko WA. USING POPULATION GENOMICS TO DETECT SELECTION IN NATURAL POPULATIONS: KEY CONCEPTS AND METHODOLOGICAL CONSIDERATIONS. Int J Plant Sci. 2010;171:1059–71.
109. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 2012;8:e1003011.
110. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96:986–91.
111. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta. 2014;1842:1910–22.
112. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015;16:359–71.
113. Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation. Cell. 2016;166:538–54.
114. Pham M-HT, Bonello GB, Castiblanco J, Le T, Sigala J, He W, et al. The rs1024611 regulatory region polymorphism is associated with CCL2 allelic expression imbalance. PLoS One. 2012;7:e49498.

115. Balamayooran G, Batra S, Balamayooran T, Cai S, Jeyaseelan S. Monocyte chemoattractant protein 1 regulates pulmonary host defense via neutrophil recruitment during Escherichia coli infection. Infect Immun. 2013;81:4323–4323.
116. Gomes RN, Teixeira-Cunha MGA, Figueiredo RT, Almeida PE, Alves SC, Bozza PT, et al. Bacterial clearance in septic mice is modulated by MCP-1/CCL2 and nitric oxide. Shock. 2013;39:63–9.
117. Cruickshank SM, Deschoolmeester ML, Svensson M, Howell G, Bazakou A, Logunova L, et al. Rapid Dendritic Cell Mobilisation to the Large Intestinal Epithelium is Associated with Resistance to Trichuris muris Infection. The Journal of Immunology. 2009;182:3055–62.
118. Namkoong H, Kurashima A, Morimoto K, Hoshino Y, Hasegawa N, Ato M, et al. Epidemiology of pulmonary nontuberculous Mycobacterial disease, japan. Emerg Infect Dis. 2016;22:1116–7.
119. Schildkraut JA, Gallagher J, Morimoto K, Lange C, Haworth C, Floto RA, et al. Epidemiology of nontuberculous mycobacterial pulmonary disease in Europe and Japan by Delphi estimation. Respir Med. 2020;173:106164.
120. Carter AJR, Nguyen AQ. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. BMC Med Genet. 2011;12:160.
121. Byars SG, Voskarides K. Antagonistic pleiotropy in human disease. J Mol Evol. 2020;88:12–25.
122. van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, et al. Presence of genetic variants among young men with severe COVID-19. JAMA. 2020;324:663–73.
123. Fallerini C, Daga S, Mantovani S, Benetti E, Picchiotti N, Francisci D, et al. Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study. Elife. 2021;10.
124. Zhang S-Y, Zhang Q, Casanova J-L, Su HC, COVID Team. Severe COVID-19 in the young and healthy: monogenic inborn errors of immunity? Nat Rev Immunol. 2020;20:455–6.
125. Asano T, Boisson B, Onodi F, Matuozzo D, Moncada-Velez M, Maglorius Renkilaraj MRL, et al. X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. Sci Immunol. 2021;6.
126. Mantovani S, Daga S, Fallerini C, Baldassarri M, Benetti E, Picchiotti N, et al. Rare variants in Toll-like receptor 7 results in functional impairment and downregulation of cytokine-mediated signaling in COVID-19 patients. Genes Immun. 2022;23:51–6.
127. Kosmicki JA, Horowitz JE, Banerjee N, Lanche R, Marcketta A, Maxwell E, et al. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am J Hum Genet. 2021;108:1350–5.
128. Wei Y, Dimicoli S, Bueso-Ramos C, Chen R, Yang H, Neuberg D, et al. Toll-like receptor alterations in myelodysplastic syndrome. Leukemia. 2013;27:1832–40.
129. Krieger E, Vissichelli N, Leichtle S, Kashioris M, Sabo R, Brophy D, et al. Immunological determinants of clinical outcomes in COVID-19: A quantitative perspective. arXiv [q-bio.CB]. 2020.
130. Brosnahan SB, Jonkman AH, Kugler MC, Munger JS, Kaufman DA. COVID-19 and respiratory system disorders: Current knowledge, future clinical and translational research questions. Arterioscler Thromb Vasc Biol. 2020;40:2586–97.
131. Dheda K, Perumal T, Moultrie H, Perumal R, Esmail A, Scott AJ, et al. The intersecting pandemics of tuberculosis and COVID-19: population-level and patient-level impact, clinical presentation, and corrective interventions. Lancet Respir Med. 2022;10:603–22.
132. Bone WP, Siewert KM, Jha A, Klarin D, Damrauer SM, VA Million Veteran Program, et al. Multi-trait association studies discover pleiotropic loci between Alzheimer's disease and cardiometabolic traits. Alzheimers Res Ther. 2021;13:34.
133. Zhang X, Lucas AM, Veturi Y, Drivas TG, Bone WP, Verma A, et al. Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders. Nat Commun. 2022;13:3428.
134. Xu Q, Yang C, Pei Y-F. Editorial: Genetic pleiotropy in complex traits and diseases. Front Genet. 2022;13:897383.

135. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
136. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
137. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155:1405–13.
138. Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol. 1984;1:269–85.
139. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. Genetics. 2013;195:181–93.
140. Cox MP, Morales DA, Woerner AE, Sozanski J, Wall JD, Hammer MF. Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. PLoS One. 2009;4:e6366.
141. Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics. 2004;166:351–72.
142. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.
143. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 2011;108:11983–8.
144. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proc Natl Acad Sci U S A. 2016;113:5652–7.
145. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367:eaay5012.
146. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022;220.
147. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 2016;12:e1004842.
148. Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. The peopling of the African continent and the diaspora into the new world. Curr Opin Genet Dev. 2014;29:120–32.
149. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008;9:403–33.
150. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences. 2005;102:18508–13.
151. Lisi A, Campbell MC. Polaris: Polarization of ancestral and derived polymorphic alleles for inferences of extended haplotype homozygosity in human populations. bioRxiv. 2024.
152. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31:1275–91.
153. Roca-Umbert A, Caro-Consuegra R, Londono-Correa D, Rodriguez-Lozano GF, Vicente R, Bosch E. Understanding signatures of positive natural selection in human zinc transporter genes. Sci Rep. 2022;12:4320.
154. Shihabi M, Lukic B, Cubric-Curik V, Brajkovic V, Oršanić M, Ugarković D, et al. Identification of selection signals on the X-chromosome in East Adriatic sheep: A new complementary approach. Front Genet. 2022;13:887582.
155. Szpiech ZA. Selscan 2.0: Scanning for sweeps in unphased data. Bioinformatics. 2024;40.
156. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38:1358–70.
157. Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution. 2005;59:2312–23.

158. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

Table Legend

Table 1. Inferred targets of selection and their predicted functional impact. The SNPs identified as adaptive loci are given in the Table. These loci were inferred to be targets of selection based on | *i*HS |, | nS_L |, EHH, and CLUES2 results. For each of these selected sites, we listed the rs identifier, the genic region in which SNPs are located, the ancestral allele, the allele believed to be the target(s) of selection, the population in which the selected allele(s) was found, and the known transcription factor that binds to the transcription factor binding site (TFBS) where the SNP occurred; "—" means that a given SNP was not found in a known TFBS according to the SNP2TFBS database (last accessed 08/14/2025). The full population names for the abbreviations in the Table can be found in the Methods section.

Figure Legends

Figure 1: Gene structure and the distribution of polymorphisms across genic regions. This figure illustrates the chromosomal location and gene structure of *IL-4*, *TLR2*, *CCL2*, and *SLC11A1* along with the positions and *rs* identifiers of common SNPs previously associated with respiratory disease. Bar plots to the right of each gene diagram show the proportion of polymorphisms located in the 5′-untranslated region (5′-UTR), introns, exons, and the 3′-untranslated region (3′-UTR) of each immune-related gene (when applicable).

Figure 2: Signatures of positive selection and allele frequencies patterns at rs2279014 in a subset of global populations. Panels A-D show Manhattan plots of standardized | *i*HS | statistics on Chromosome 2 in exemplar populations from the 1000 Genomes Project. The horizontal dashed line indicates the threshold for identifying outlier *i*HS statistics. Red dots indicate the derived allele at SNP loci while the green dots denote ancestral alleles at SNPs across *SLC11A1*; we also listed their corresponding rs identifiers. Panels E-H show the decay of extended haplotype homozygosity (EHH) surrounding the functional *rs*2279014 SNP. The x-axis is the distance in megabases (Mb) upstream (negative values) and downstream (positive values) from the *rs*2279014 core site on the forward strand; the *y*-axis is the probability that two chromosomes are homozygous at all SNPs for the interval from the core site to distance *x*. Blue and red shading indicates the decay of homozygosity on chromosomes carrying the ancestral and derived alleles, respectively, from the core site. The frequencies of the ancestral and derived alleles at *rs*2279014 are given for each population. The additional *i*HS and EHH plots for other populations are provided in the Supplementary Materials.

Figure 3: Signatures of positive selection and allele frequency patterns at rs2279015 in a subset of populations. Panels A-D illustrate Manhattan plots of standardized *i*HS statistics on Chromosome 2 in exemplar populations from the 1000 Genomes Project. The horizontal dashed lines represent the cut-off for extreme | *i*HS | scores (| *i*HS | = 2). Red dots denote the derived alleles at *SLC11A1* SNPs while the green dots indicate ancestral alleles at SNPs in the same gene. Panels E-H display EHH around the functional *rs*2279015 SNP (core site) in exemplar populations. Specifically, these plots show the decay of identity of haplotypes on chromosomes carrying the derived allele (red shading) versus the ancestral allele (blue shading) at *rs*2279015 with increasing distance from the core site. The x-axis shows the distance of homozygosity from the core SNP (at position zero), with negative numbers indicating upstream regions and positive values signifying downstream regions on the forward strand. The EHH probabilities are given on the y-axis. The frequencies of the ancestral and derived alleles at *rs*2279015 are provided for each population. Additional *i*HS and EHH plots for other populations are available in the Supplementary

Materials.

Figure 4: Estimated selection coefficients for SNP loci in *IL-4* in global populations. Panels A-D, Panels E-H, and Panels I-L show *s* estimates for SNPs in a subset of global populations together with the rs identifiers for SNPs that exhibit significant s values ($P < 0.05$) across *IL-4* at three different time depths (1,000 generations ago, 500 generations ago, and 200 generations ago, respectively). Each dot represents an estimated *s* statistic, and the vertical bars indicate the confidence intervals around the estimate. Negative *s* estimates indicate selection on ancestral alleles while positive *s* values signify selection on derived alleles. The color of each dot indicates the level of significance associated with the *s* statistics in addition to the asterisks.

Figure 5: Haplotype genealogies of global variation across *IL-4*, *TLR2*, *CCL2*, and *SLC11A1*. Haplotype relationships were reconstructed using the median-joining algorithm implemented in `Network 5.0` [82]. The resulting phylogeny consists of haplotypes (or nodes), and the size of each haplotype is proportional to the number of chromosomes with that given haplotype. The colors within the nodes (or haplotypes) indicate the frequency of that haplotype in continental populations. Specifically, blue represents African and African-descended populations (AFR); pink signifies South Asian populations (SAS); red denotes European populations (EUR); green indicates East Asian populations (EAS). The lines connect the nodes in the shortest possible tree. High-frequency haplotypes carrying alleles inferred to be under selection are shaded in orange, and the corresponding haplotype identifiers (e.g., H69, H98) are listed near them. The genomic coordinates and rs identifiers for selected SNPs on these common haplotypes are given in Table S40 (*IL-4*), Table S41 (*TLR2*), Table S42 (*CCL2*), and Table S43 (*SLC11A1*).

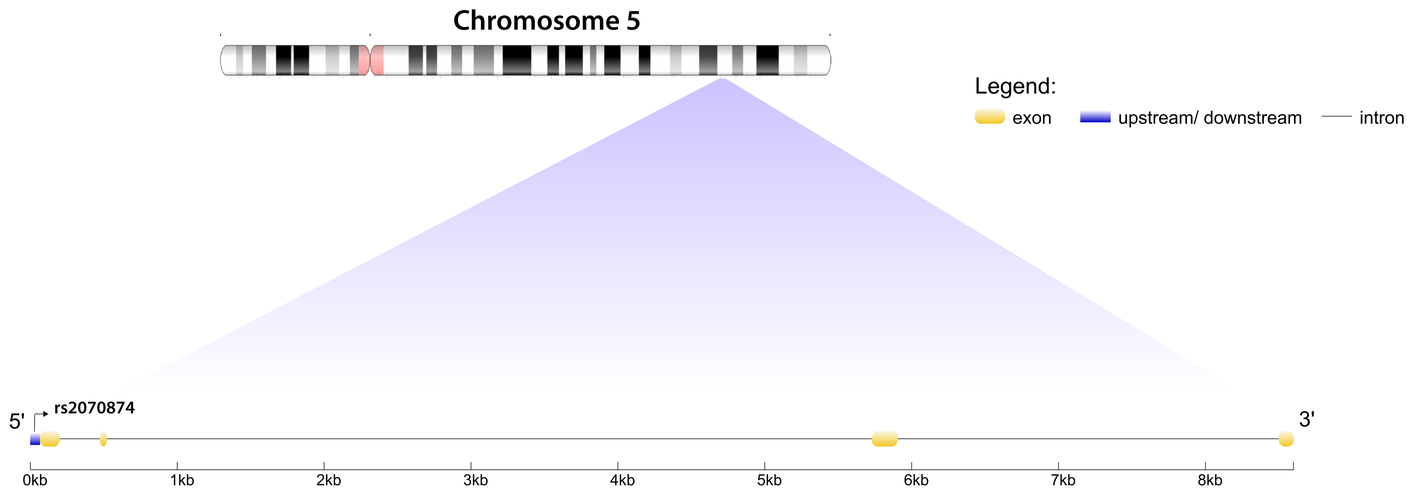
Figure 6. Geographic distribution and the frequency of candidate risk alleles for severe COVID-19. This map displays the geographic distribution of the rs5743709 A-allele and the rs139227237 C-allele in global populations from the 1000 Genomes Project. Particularly, the red shading shows the distribution of the *rs*139227237 C-allele (in the European Great British [GBR], Finnish [FIN], Tuscan Italians [TSI]) while blue shading shows the distribution of the rs5743709 A-allele (in the East Asian Japanese [JPT]). Darker shading reflects a higher allele frequency (~1.1%) while lighter shading indicates a lower allele frequency (~0.5%).

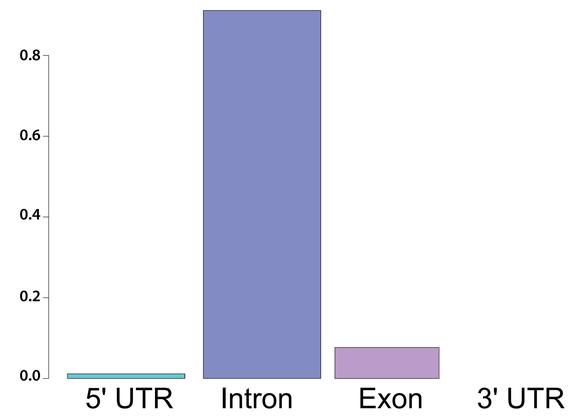
Table 1— Inferred targets of selection and their predicted functional impact

Gene	rsID	Genic region	Ancestral allele	Allele(s) under selection	Population where selected allele was identified	Transcription factor	Neandertal/Denisovan species
	rs2070874	5'-UTR	T	C/T	CDX, PJL, STU,	—	Not shared
	rs2243251	exon	A	G	ESN	—	Not shared
	rs734244	intron	C	C/T	CDX, CHB, CHS, PJL, STU,	—	Not shared
	rs2243253	intron	C	T	GWD	—	Not shared
	rs56174361	intron	A	A/G	GBR, GWD, ESN, FIN, MSL, STU, TSI, YRI	ERG2	Not shared
	rs58349986	intron	A	A/G	ACB, CDX, CHB, CHS, GBR, GWD, JPT, KHV, LWK, MSL, TSI	—	**Shared**
	rs199693275	intron	A	A/G	CHB, CHS, ESN, GWD, JPT, LWK, MSL, TSI, YRI	—	Not shared
	rs56408050	intron	A	A/G	BEB, CHB, CHS, ESN, FIN, GBR, IBS, ITU, JPT, KHV, PJL, STU, TSI, YRI	ERG2	Not shared
IL-4	rs2227284	intron	T	T/G	BEB, ESN, GBR, GIH, GWD, IBS, ITU, PJL, STU, TSI	—	Not shared
	rs2227282	intron	C	C/G	BEB, GBR, GIH, GWD, IBS, ITU, PJL STU, TSI, YRI	—	Not shared
	rs2243263	intron	G	G/C	KHV, MSL, PJL,	—	Not shared
	rs2243266	intron	G	G/A	BEB, CDX, CHS, GWD, PJL, STU, TSI	—	Not shared
	rs2243267	intron	G	G/C	CDX, GIH, GWD, PJL	—	Not shared
	rs2243268	intron	A	A/T	GBR, GWD, LWK, MSL	—	Not shared
	rs2243270	intron	G	G/A	BEB, CDX, GBR, IBS, TSI	—	**Shared**
	rs2243282	intron	C	C/A	BEB, CHB, CHS, IBS, KHV, LWK, MSL, PJL	—	**Shared**
	rs2243283	intron	C	C/G	ACB, GWD, LWK, MSL, PJL,	—	Not shared
	rs2243284	intron	G	G/A	CDX, GBR, GIH, IBS, ITU, PJL, TSI	—	**Shared**
	rs2243285	intron	G	G	ACB, GWD	—	Not shared
	rs2243288	intron	A	A/G	CDX, CHB, GBR, IBS, ITU, JPT, PJL, TSI	—	**Shared**
	rs2243289	intron	A	A/G	CDX, CHS, GIH, ITU, MSL, PJL, STU,	—	Not shared
	rs2243290	intron	C	C/A	CDX, CHS, GIH, IBS, KHV, MSL, STU, TSI,	—	**Shared**
	rs7696323	5'-UTR	C	T	LWK	—	Not shared
	rs55714645	5'-UTR	G	A	ACB, CDX, LWK	—	Not shared
	rs4696480	5'-UTR	T	T	CDX, CHS, GIH, ITU, PJL, STU	—	Not shared
	rs2083287	5'-UTR	A	G	ACB, ASW, ESN, GWD, LWK, YRI	—	Not shared
	rs1898830	5'-UTR	A	G	ACB, ASW, ESN, GIH, GWD, IBS, PJL, STU, PJL	—	Not shared
	rs13147967	5'-UTR	G	G/A	GBR, MSL, GIH	—	Not shared
	rs6811309	5'-UTR	G	A	LWK, PJL	—	Not shared
	rs6811715	5'-UTR	G	G	GIH, ITU, PJL, STU	—	Not shared
	rs11099894	5'-UTR	C	T	ACB, ASW, CHS, ESN, GIH, GWD, IBS, KHV, PJL, STU, TSI	—	Not shared
	rs148248740	5'-UTR	T	C	CDX	—	Not shared
	rs58425222	5'-UTR	G	C	ACB,	—	Not shared
	rs10222800	5'-UTR	G	A	CHS, GBR, TSI, YRI	—	Not shared
TLR2	rs7675718	5'-UTR	G	A	ACB, ASW, CHS, ESN, GIH, GWD, IBS, PJL, STU, TSI	—	Not shared
	rs28872182	5'-UTR	C	A	ACB	—	Not shared
	rs6822309	5'-UTR	T	A	LWK	—	Not shared
	rs1898831	5'-UTR	G	A	STU	—	Not shared

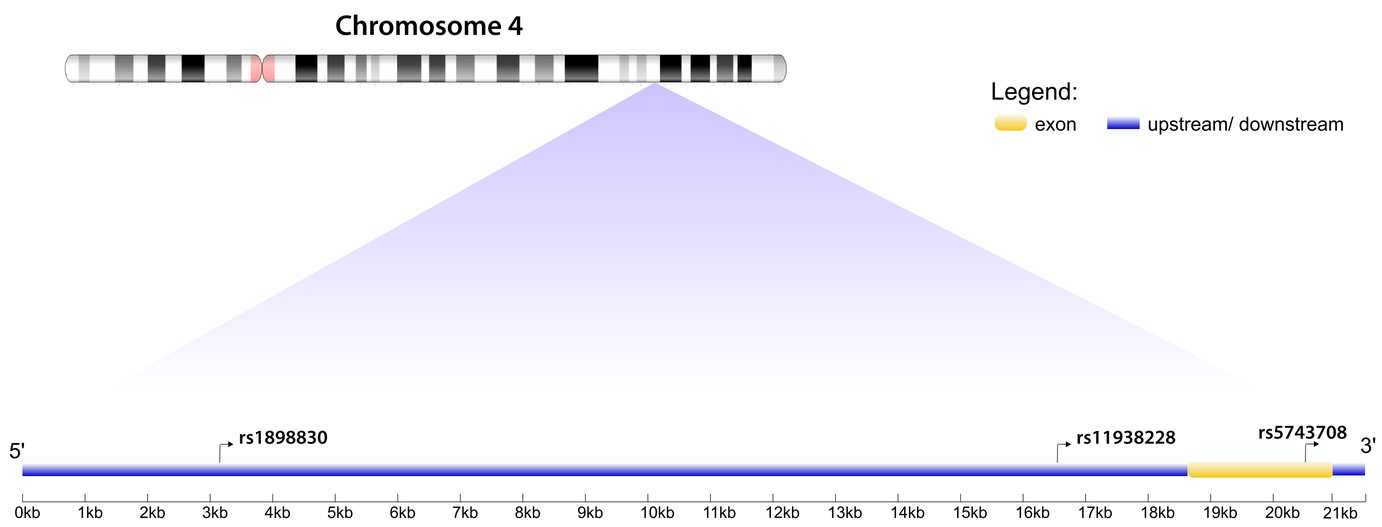
	rs1898832	5'-UTR	T	C	JPT, LWK	—	Not shared
	rs74341183	5'-UTR	A	G	TSI	—	**Shared**
	rs7694512	5'-UTR	G	T	ACB, ASW, CHB, CHS, ESN, GIH, GWD, IBS, PJL, TSI	—	Not shared
	rs6811185	5'-UTR	A	G	ESN, GWD, STU, YRI	—	Not shared
	rs4696483	5'-UTR	T	C	BEB, GIH, STU, TSI, YRI	—	Not shared
	rs13115230	5'-UTR	A	T	ACB, ASW, ESN, GWD, IBS, KHV, PJL	—	Not shared
	rs13123230	5'-UTR	A	G	ACB, ESN, PJL, TSI	—	Not shared
	rs12510187	5'-UTR	G	A	ACB, ASW, CHS, ESN, GBR, GWD, IBS, PJL, STU, TSI	—	Not shared
	rs11938228	5'-UTR	C	A	ACB, ASW, CHS, ESN, GBR, GWD, PJL	—	Not shared
	rs1439164	5'-UTR	T	C	ACB, ASW, BEB, STU	—	Not shared
	rs13105517	5'-UTR	G	A	ASW, CHS, GBR, GWD, IBS, TSI, PJL	—	Not shared
	rs62323856	5'-UTR	T	C	IBS, KHV	—	**Shared**
	rs72954762	5'-UTR	T	C	CDX, ESN, GWD, MSL	—	Not shared
	rs3804099	5'-UTR	C	T	PJL	—	Not shared
	rs3804100	exon	T	T	GWD	—	Not shared
CCL2	rs2857657	intron	C	G	BEB, KHV, STU	SOX10, ZNF263	Not shared
	rs4586	exon	C	C	CHB, CHS, JPT, MSL, YRI	—	Not shared
	rs13900	3'-UTR	C	T	CHS, JPT, LWK	—	Not shared
	rs17228967	intron	T	T	ACB	—	Not shared
	rs2276631	exon	C	C	CHS, GIH, IBS	—	Not shared
	rs137892936	intron	G	G	CHS, GIH, GWD, PJL	—	Not shared
	rs3731865	intron	G	G	BEB, CHS, GIH, YRI	—	Not shared
	rs3731863	intron	C	C	BEB, CHB, CHS, GIH, JPT, PJL	—	Not shared
	rs2290708	intron	C	C	CHB, CHS, FIN, GIH,	—	Not shared
	rs17221959	exon	C	C	ACB, CHB, CHS, GIH, JPT, TSI	—	Not shared
SLC11A1	rs3816560	intron	T	T	CHB, CHS, GIH, JPT, PJL, YRI	—	Not shared
	rs2695343	intron	A	G	ACB, BEB, CDX, CHB, CHS, FIN, GBR, GIH, IBS, ITU, JPT, KHV, LWK, PJL, STU, TSI	—	Not shared
	rs17228995	intron	C	C	ASW, BEB, CHB, CHS, GIH, JPY, STU	—	Not shared
	rs2279015	intron	A	G	ASW, BEB, CDX, CHB, CHS, GBR, GIH, GWD, IBS, ITU, JPT, KHV, LWK, PJL, STU	ERG1	Not shared
	rs17235409	exon	G	G	CDX, CHB, CHS, ITU, KHV	—	Not shared
	rs1059823	3'-UTR	G	A	ASW, BEB, CDX, CHB, CHS, FIN, GBR, GIH, IBS, ITU, JPT, KHV, PJL, STU, TSI	—	Not shared
	rs17229016	3'-UTR	A	G	ASW, CDX, CHB, CHS, FIN, GIH, JPT, KHV	—	Not shared
	rs13062	3'-UTR	C	C	ACB, ASW, BEB, CDX, CHB, CHS, FIN, GWD, GIH, ITU, JPT, KHV, LWK, PJL, STU, ESN, TSI, YRI	—	Not shared
	rs2279014	3'-UTR	C	C/T	BEB, CDX, CHB, CHS, FIN, GIH, GBR, ITU, JPT, KHV, MSL, PJL, STU, TSI, YRI	C/EBPα	Not shared

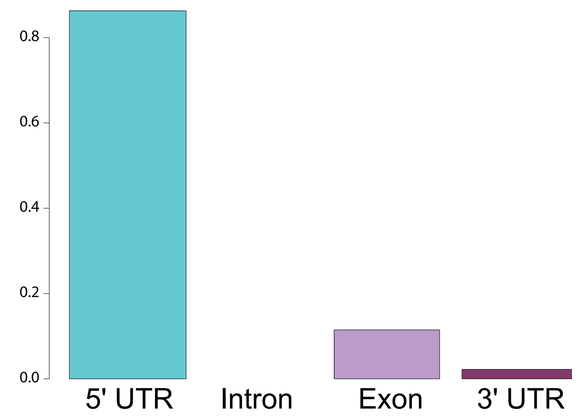
A. *IL-4*



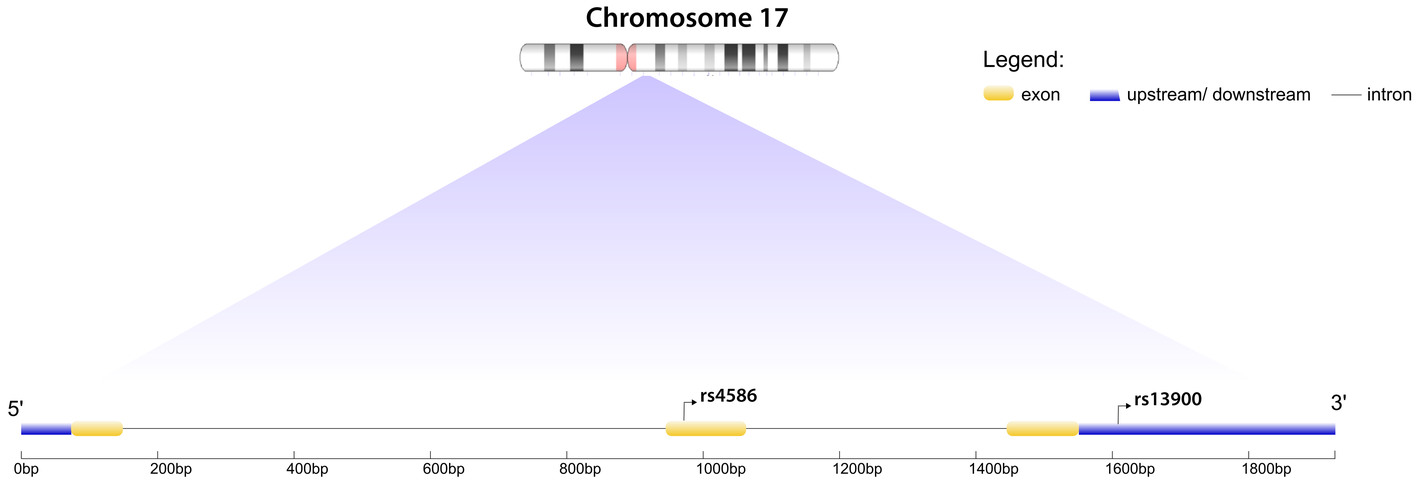


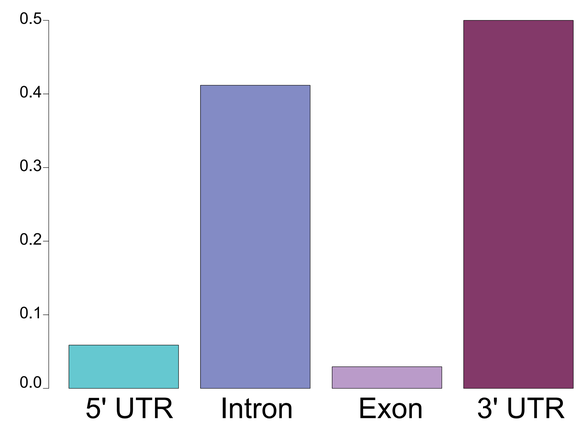
B. *TLR2*



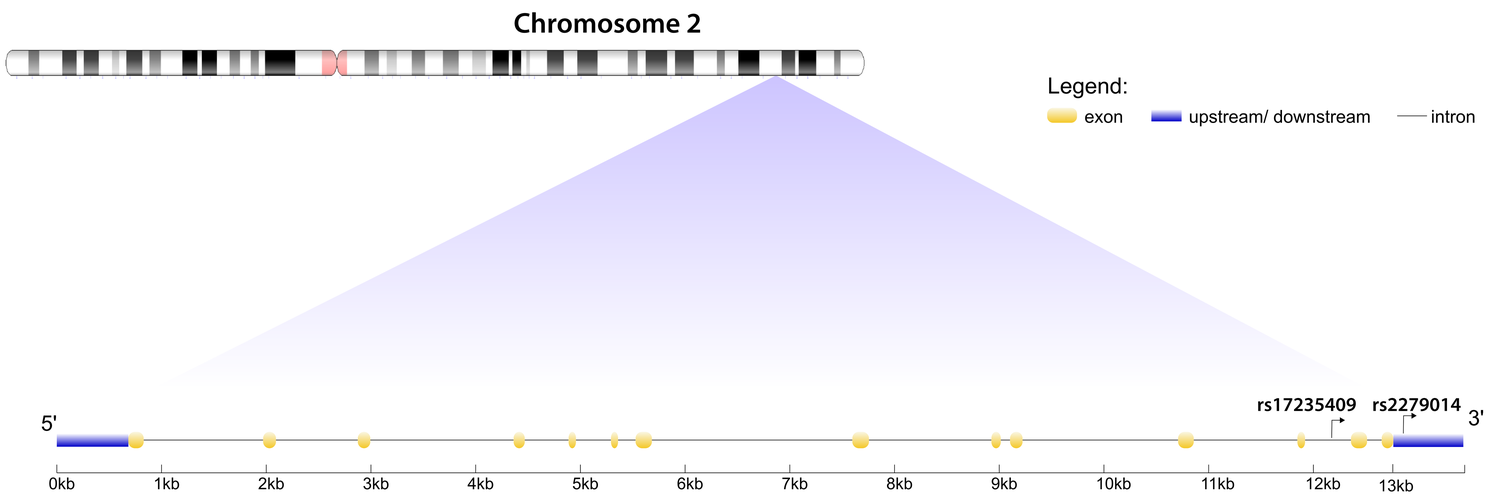


C. *CCL2*





D. *SLC11A1*



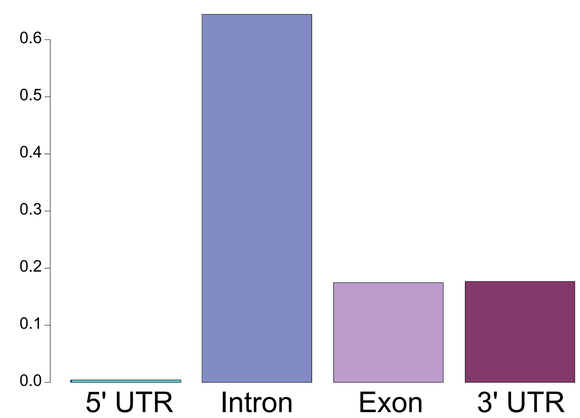


Figure 1

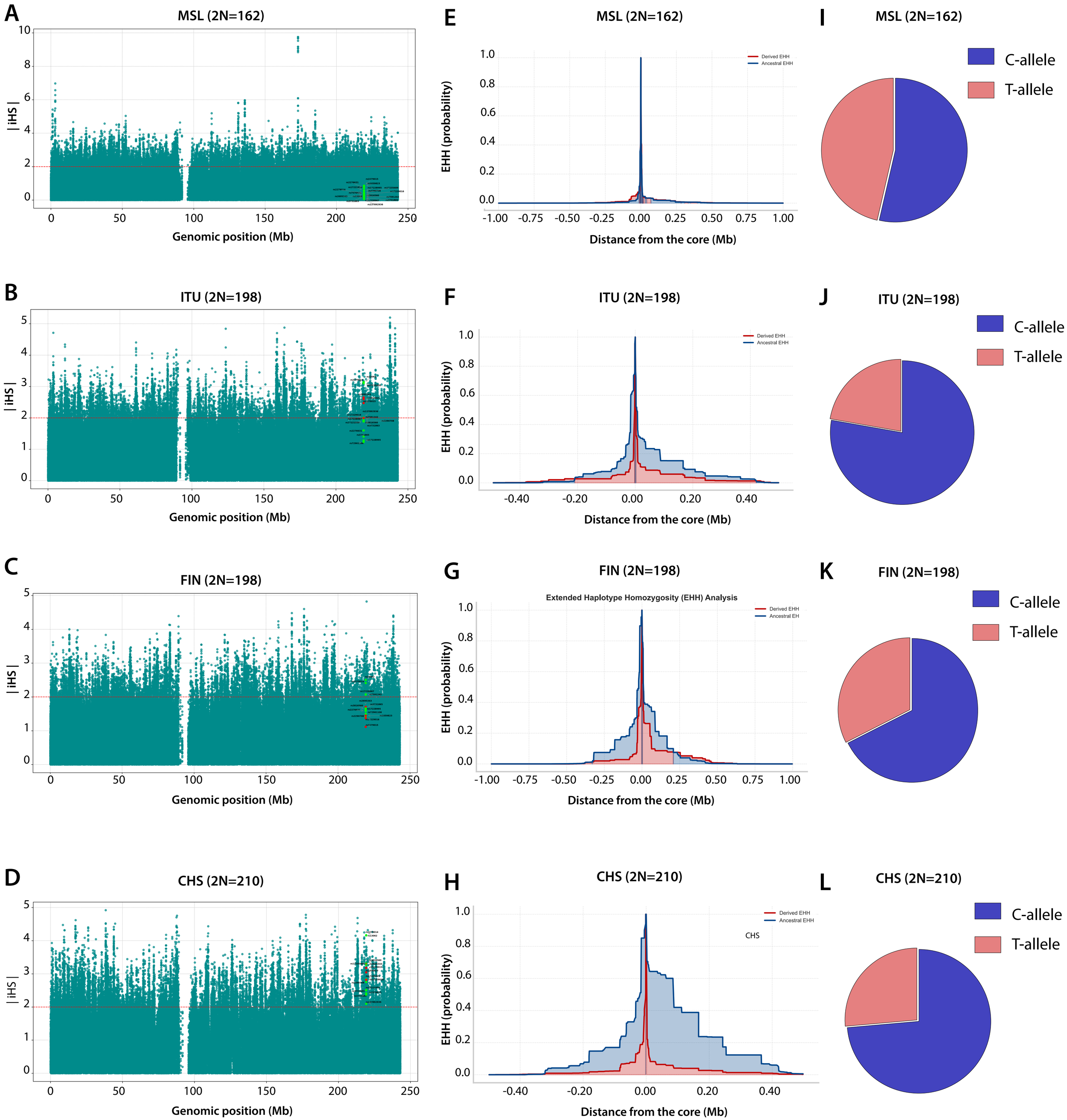

A
MSL (2N=162)
| iHS |
Genomic position (Mb)
B
ITU (2N=198)
| iHS |
Genomic position (Mb)
C
FIN (2N=198)
| iHS |
Genomic position (Mb)
D
CHS (2N=210)
| iHS |
Genomic position (Mb)
E
MSL (2N=162)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
F
ITU (2N=198)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
G
FIN (2N=198)
Extended Haplotype Homozygosity (EHH) Analysis
EHH (probability)
Derived EHH
Ancestral EH
Distance from the core (Mb)
H
CHS (2N=210)
EHH (probability)
Derived EHH
Ancestral EHH
CHS
Distance from the core (Mb)
I
MSL (2N=162)
C-allele
T-allele
J
ITU (2N=198)
C-allele
T-allele
K
FIN (2N=198)
C-allele
T-allele
L
CHS (2N=210)
C-allele
T-allele

Figure 2

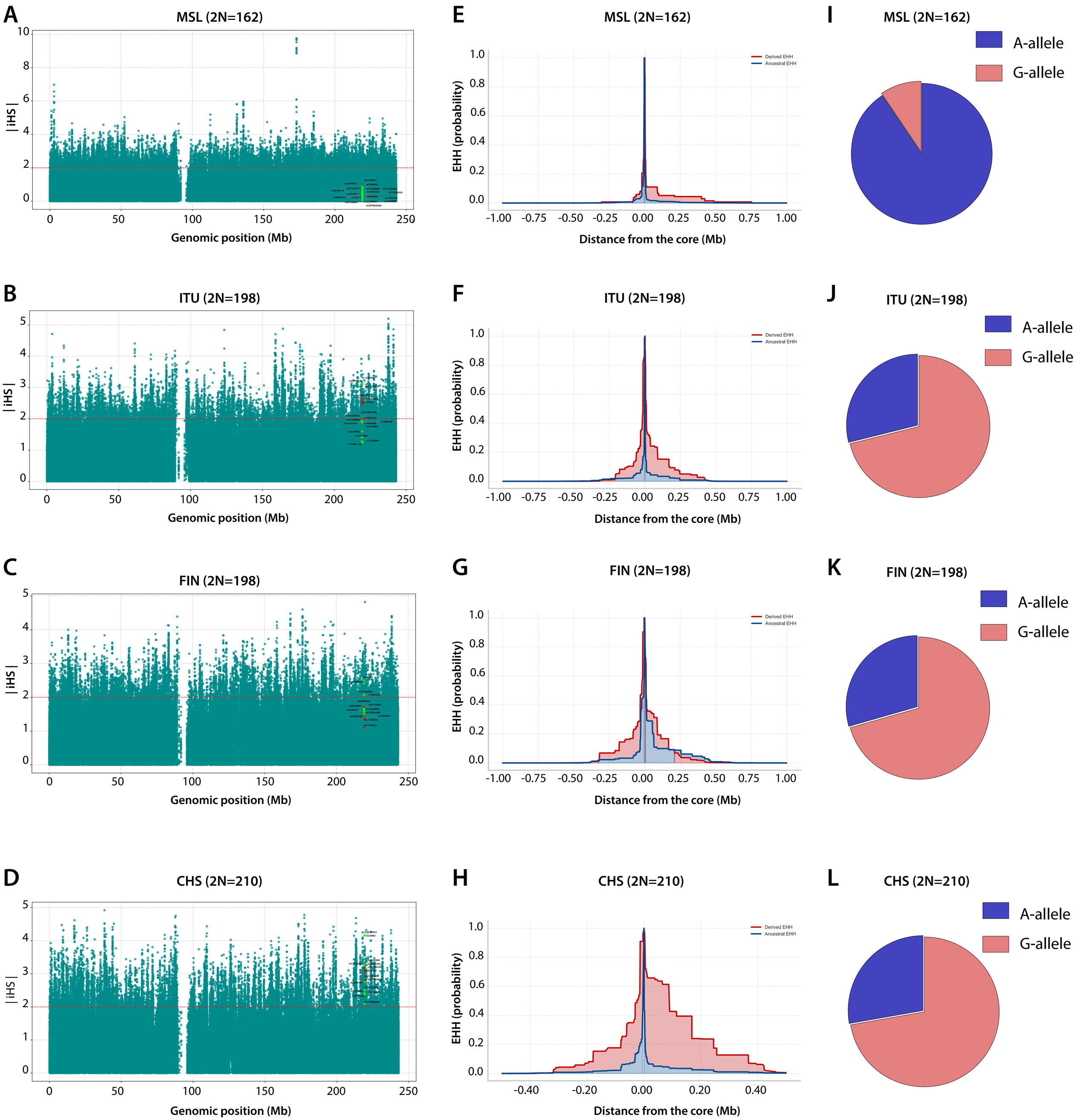

A
MSL (2N=162)
| iHS |
Genomic position (Mb)
B
ITU (2N=198)
| iHS |
Genomic position (Mb)
C
FIN (2N=198)
| iHS |
Genomic position (Mb)
D
CHS (2N=210)
| iHS |
Genomic position (Mb)
E
MSL (2N=162)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
F
ITU (2N=198)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
G
FIN (2N=198)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
H
CHS (2N=210)
EHH (probability)
Derived EHH
Ancestral EHH
Distance from the core (Mb)
I
MSL (2N=162)
A-allele
G-allele
J
ITU (2N=198)
A-allele
G-allele
K
FIN (2N=198)
A-allele
G-allele
L
CHS (2N=210)
A-allele
G-allele

Figure 3

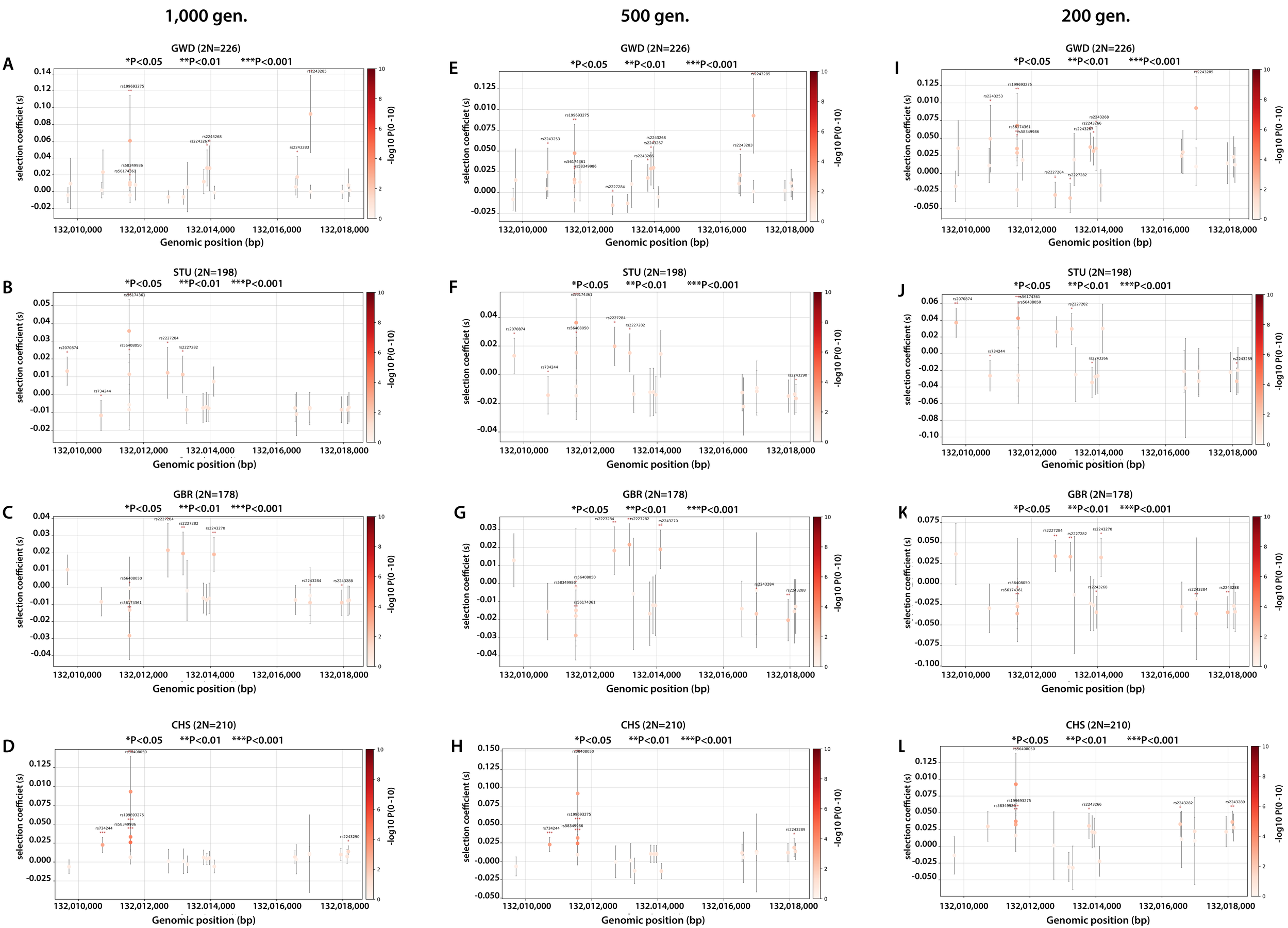

1,000 gen.
500 gen.
200 gen.
A
B
C
D
E
F
G
H
I
J
K
L
GWD (2N=226)
STU (2N=198)
GBR (2N=178)
CHS (2N=210)
*P<0.05 **P<0.01 ***P<0.001
selection coefficiet (s)
Genomic position (bp)
-log10 P(0 -10)
132,010,000
132,012,000
132,014,000
132,016,000
132,018,000
rs199693275
rs2243285
rs2243268
rs2243267
rs2243283
rs58349986
rs56174361
rs2243253
rs2243266
rs2227284
rs2227282
rs2070874
rs56408050
rs734244
rs2243270
rs2243284
rs2243288
rs2243289
rs2243290

Figure 4

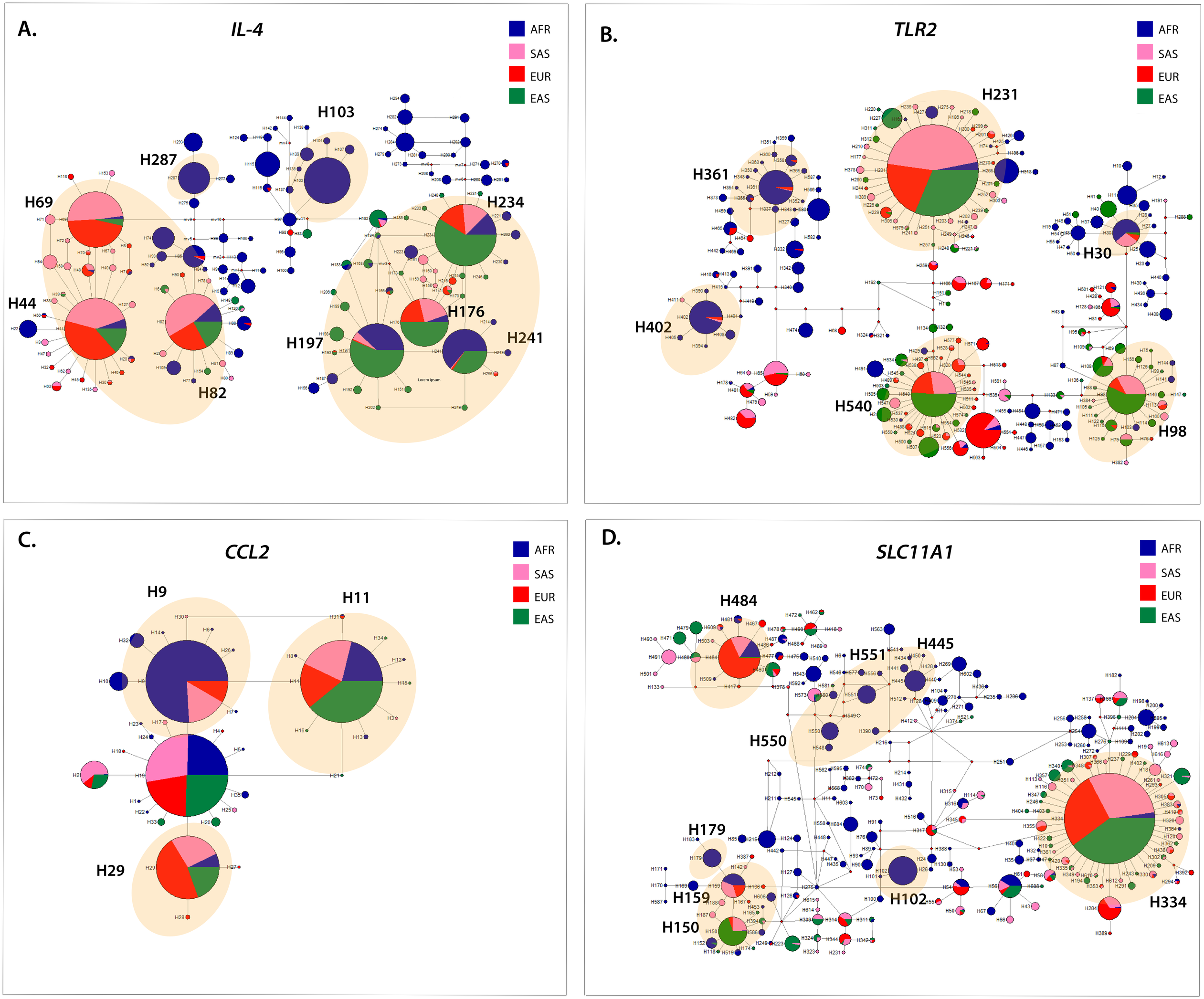

A.
IL-4
AFR
SAS
EUR
EAS
H103
H287
H69
H234
H44
H176
H197
H241
H82
B.
TLR2
AFR
SAS
EUR
EAS
H231
H361
H30
H402
H540
H98
C.
CCL2
AFR
SAS
EUR
EAS
H9
H11
H29
D.
SLC11A1
AFR
SAS
EUR
EAS
H484
H445
H551
H550
H179
H159
H150
H102
H334

Figure 5

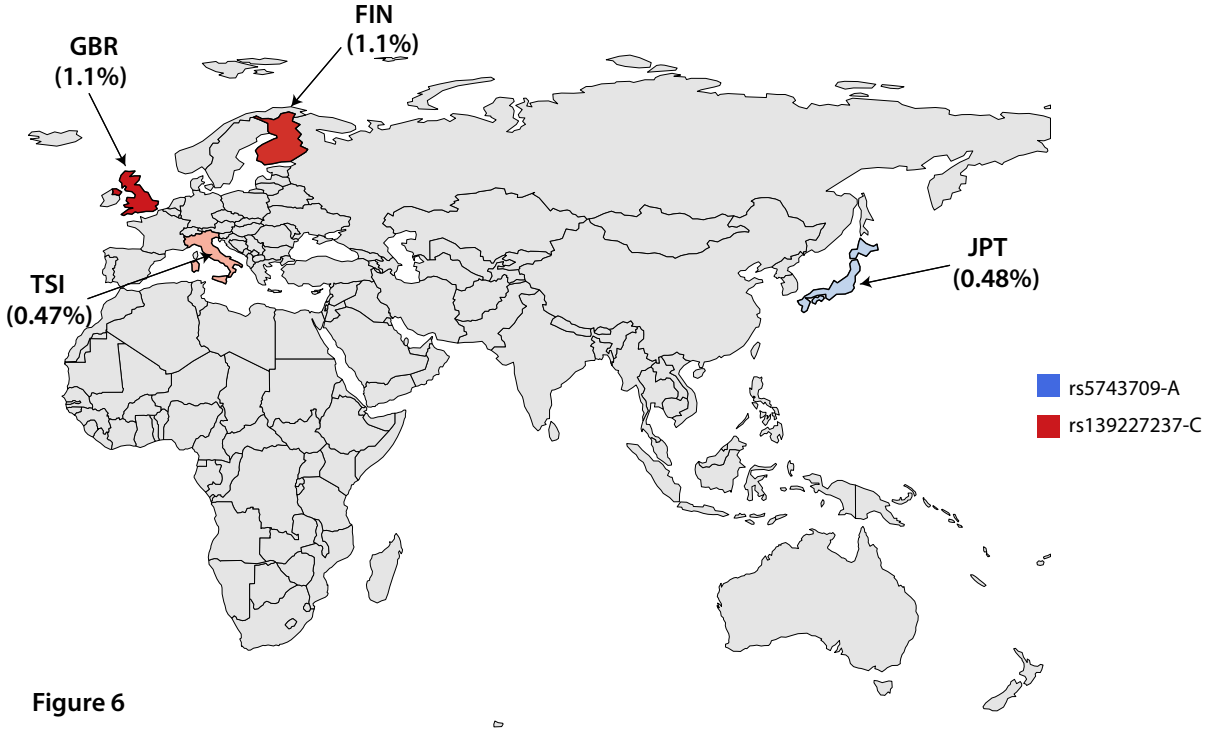


Figure 6