Evolutionary Genetics of Coronary Heart Disease

Keyue Ding, PhD; Iftikhar J. Kullo, MD

Nothing in biology makes sense except in the light of evolution.1

—Theodosius Dobzhansky, 1973

No biological problem is solved until both the proximate and the evolutionary causation has been elucidated. Furthermore, the study of evolutionary causes is as legitimate a part of biology as is the study of the usually physico-chemical proximate causes.2

—Ernst W. Mayr, 1982

Susceptibility to common diseases such as coronary heart disease (CHD) may in part reflect historical or evolutionary legacies,3,4 and interest in studying evolutionary biology to gain novel insights into human health and disease is increasing. The evolutionary history of the human species may provide valuable insights into the origin of common diseases beyond what is possible by investigating only the most immediate or “proximal” causes of disease. The potential role of evolutionary biology in explaining disease causation was highlighted by Williams and Nesse1 and is often referred to as darwinian medicine. Although the relevance of an evolutionary perspective may vary depending on the disease under study, a strong argument could be made for studying the evolutionary genetics of CHD, a leading cause of human morbidity and death.

Within the last decade, several important advances have made it possible to study “modern” diseases from an evolutionary perspective. The Human Genome Project5 provided a reference human genome, and the subsequent International HapMap Projects6,7 described genetic variations (mostly single nucleotide polymorphisms [SNPs]) among individuals and the patterns of variation across the genome. Both projects provide the raw material to study natural selection in the human genome, as reviewed in recent reports.8–10 In addition, sequencing of multiple other genomes, including those of primates, provides a framework for generating important insights into the origin and expression of human diseases.4 Such an avenue of investigation may help answer why, compared with their closest relative, the chimpanzee, humans are susceptible to CHD,11–13 why CHD has assumed epidemic proportions,14 and why substantial differences in susceptibility to CHD are present between ethnic groups.15 Genetic factors underlying both “complex” and mendelian diseases may be affected by natural selection16; conversely, regions of the human genome under natural selection17 are likely to harbor functional loci (gene coding and regulatory regions18) that may influence disease susceptibility. Inferences about natural selection may therefore help identify disease susceptibility loci in humans and facilitate disease mapping studies.19

In this review, we discuss the current hypotheses and knowledge about the evolutionary genetics of several risk factors for CHD, including hypertension, dyslipidemia, and the metabolic syndrome. Several candidate genes in pathways of blood pressure, glucose and lipid metabolism, blood coagulation, and inflammation that may be under natural selection are enumerated. An evolutionary perspective might explain why contemporary humans are at high risk for CHD and also helps to better understand variation in disease susceptibility.

Natural Selection

Natural selection is the inevitable consequence of inherited variation in fitness, which is defined as the relative ability of an organism to survive and transmit its genes to the next generation.20 Under natural selection, individuals with advantageous or “adaptive” traits tend to be more successful in producing offspring, and those with nonadvantageous traits will be selected against. Positive selection can increase the prevalence of adaptive traits with a genetic basis. CHD has not been subject to direct selective pressure because it often manifests in middle age and beyond and is unlikely to have an effect on reproductive success.21 However, evolutionary genetics provides a framework for understanding how susceptibility alleles (see the glossary in the online-only Data Supplement) for CHD may have been affected by natural selection. Nesse22 proposed several possible explanations why previously protective alleles may be maladaptive in the present environment. Selection will tend to maintain the frequency of genes that increase reproductive success even if these genes have other effects that increase disease susceptibility in older age.23 For example, genetically determined high cytokine response levels may be associated with adverse cardiovascular outcomes in older individuals but may increase reproductive success in young age by conferring resistance to fatal infectious diseases24; a time lag between environmental change and adaptation by natural selection may also increase disease susceptibility (eg, the proatherogenic environment is relatively recent in human history); natural selection may not

From the Division of Cardiovascular Diseases, Mayo Clinic, Rochester, Minn.
The online-only Data Supplement is available with this article at http://circ.ahajournals.org/cgi/content/full/119/3/459/DC1.
Reprint requests to Iftikhar J. Kullo, MD, Division of Cardiovascular Diseases, Mayo Clinic, 200 First St SW, Rochester MN 55905. E-mail kullo.iftikhar@mayo.edu
(Circulation. 2009;119:459-467.)
© 2009 American Heart Association, Inc.
Circulation is available at http://circ.ahajournals.org
DOI: 10.1161/CIRCULATIONAHA.108.809970
Methods for Detecting Signals of Natural Selection

Natural selection takes places at the levels of transmitted genotypes and occurs on varying time scales. Several approaches have been developed to detect signals of natural selection.28 Broadly, 2 types of selective forces have shaped the evolution of species: purifying selection, which favors conservation of existing phenotypes, and positive selection, which promotes the emergence of new phenotypes. Both purifying selection and positive selection leave distinctive signatures in the form of patterns of genetic variation that can be detected through statistical tests of genotype or sequence data. If a gene has an important role in organismal homeostasis, then changes such as nonsynonymous mutations will be selected against and eliminated both between species and within species and result in an increase in the proportion of low-frequency variants within species. Positive selection typically manifests as rapid divergence of functional sites between species and a reduction in polymorphic variation within species.

In his landmark “neutral” theory of molecular evolution, Kimura27 proposed that most of the DNA sequence substitution observed both within and between species has no effect on the fitness of the individual organism and that most evolutionary change is the result of genetic drift acting on neutral alleles (ie, genes under no selective pressure). The neutral theory is widely used as a “null model” to identify signatures of natural selection in genes under selective neutrality.19,31,32 Deviations from this null hypothesis are indicative of selection, and several tests have been designed to detect such deviation.29 Sabeti et al30 have reviewed in detail statistical tests to detect positive natural selection based on distinctive patterns of genetic variation in the genome.

Broadly, these statistical tests can be classified into 4 categories discussed briefly below (more information about these tests can be obtained in several excellent reviews16,17,19,31,32). Comparison of DNA sequence between species, especially for genetic variants that alter protein function, can be used to detect positive selection that occurred several million years ago. More recent selection can be gleaned from human population genetic data as alterations in levels of nucleotide diversity (or the allele frequency spectrum), linkage disequilibrium (LD), and haplotype structure. Most methods applicable to population genetic data make certain assumptions about the demographic history of populations (a constant population size and no population structure),19 whereas methods comparing species are relatively robust to demographic factors.33 To tease out the demographic effects, one needs to carry out tests of neutrality that account for the demographic history of the population (see Demography and Recombination Versus Selection below). Examples of several methods that have been used to detect signatures of natural selection across the human leukocyte antigen locus are provided in Table I of the online-only Data Supplement.

Tests Based on Divergence of Species

Over a prolonged period, positive selection can increase the fixation rate of beneficial function-altering mutations (eg, substitutions that change amino acids [ie, nonsynonymous substitutions]). A simple way of detecting selection from comparative genomic data is to calculate the ratio between the rate of nonsynonymous substitutions and the rate of synonymous substitutions (ie, dN/dS).34 This ratio provides a means of detecting selective pressure at the protein level: dN/dS = 1 for no selection, dN/dS < 1 for purifying selection, and dN/dS > 1 for positive selection. Two caveats should be noted for dN/dS-based tests. First, purifying or positive selection may selectively influence synonymous or nonsynonymous residues within a single gene based on the functions of individual residues at the protein, RNA, or regulatory level. Second, this method is not easily applied to cis-regulatory regions, many of which are putative and not validated.35 However, several studies have shown that conserved noncoding regions in mammals bear the signature of natural selection.36–38

Tests Based on Allele Frequency Spectrum

The allele frequency spectrum represents a summary of the allele frequencies of various mutations in a sample. Natural selection affects the distribution of alleles within populations, and several tests (eg, Tajima’s D, Fu and Li’s F, Fu and Li’s D, and Fay and Wu’s H) have been developed to detect selection based on the allele frequency spectrum.16,17,19 Selection on 1 allele changes the frequency of other alleles that are in LD with it (ie, when the frequency of an allele increases in a population, the frequency of the linked variants also increases). This effect, also called hitchhiking or a selective sweep, is the process by which a new advantageous mutation eliminates or reduces variation in linked neutral sites as it increases in frequency in the population (Figure 1).

Tests Based on Population Differentiation (FST)

Natural selection may increase the degree of differentiation among geographically separate populations that are subject to
distinct environmental and/or cultural pressures. The signature can arise only when populations are at least partially isolated reproductively (e.g., after the major human migrations out of Africa ~50,000 to 70,000 years ago). This is best summarized by the statistic FST, which measures the variance in allele frequency among populations relative to total variation in the entire population (Figure 2). Local adaptation appears to have featured in recent human evolutionary history, and screening the whole genome for local adaptation may be a means of identifying susceptibility genes for diseases for which prevalence varies as a function of ethnicity. For example, we calculated FST for 15,559 common SNPs in 416 candidate genes in 6 causal pathways for CHD and found significant population differentiation for 9 genes.

Population differentiation leading to varying frequency of gene variants across populations has made admixture mapping of disease susceptibility genes possible. In an admixed group such as blacks in the United States, one can classify the genome into sections that come from their African or European ancestors. Genomic regions in which individuals with a particular complex disease tend to have an unusually high proportion of ancestry from either Europeans or Africans may harbor a disease risk variant that influences disease susceptibility.

**Tests Based on LD**

These tests rely on the LD structure of a given locus in the genome. Levels of LD will increase in selected regions if the rise in frequency of the beneficial mutation occurs faster than the recombination rate. Haplotypes containing the beneficial allele are conserved over significantly longer distances than other nearby haplotypes with similar frequencies. Thus, positive selection manifests as a haplotype at high frequency with homozygosity that extends over large regions and can be detected with the extended haplotype homozygosity test. Figure 3 illustrates extended haplotype homozygosity for the derived allele of the SNP rs2032582 in the ATP-binding cassette, subfamily B, member 1 gene (ABCB1, also known as the multidrug resistance gene) using haplotype bifurcation diagrams. This gene is associated with export of chemical substance from within cells to the extracellular space and may have been under selective pressure related to xenobiotic toxin exposure. The derived allele in Europeans and Chinese demonstrated clear long-range LD as a predominance of 1 thick branch.

**Demography and Recombination Versus Selection**

It should be noted that various demographic scenarios can mimic the effects of natural selection on genetic variation. The challenge for population genetics-based methods is to separate a signature of natural selection from the confounding effects of population demographic history (e.g., population bottlenecks, expansions, and subdivision). For example, population expansion can mimic the effect of a selective sweep (increased proportion of low-frequency variants), whereas a population bottleneck can mirror the pattern seen under balancing selection (an excess of intermediate-
frequency variants). Population isolation can lead to significantly less genetic diversity than humanity as a whole and extended haplotypes over several centimorgans. Thus, it is necessary to tease out the effect of population demographic history if significant deviations from evolutionary neutrality are observed. To do this, one needs to carry out a test of neutrality that accounts for the demographic history of the population. For example, a locus of interest can be compared with the pattern of genetic diversity across the genome. In contrast to demographic factors, the effects of selection will be locus specific. Tests of evolutionary neutrality also need to account for the recombination rate in the region under study because recombination is a determinant of haplotype structure given the negative correlation between recombination rates and extent of LD.

**Evolutionary Models of the Genetic Architecture of CHD**

The genetic architecture of CHD refers to the number of genetic polymorphisms that affect the risk of CHD, the distribution of their allele frequencies and effect sizes, and their genetic mode of action (additive, dominant and/or epistasis, and pleiotropy). Any evolutionary change such as mutation, genetic drift, and natural selection may affect genetic architecture. An evolutionary genetics perspective provides a better understanding of the number and frequency of susceptibility alleles for CHD and thereby helps determine optimal strategies to detect such alleles. Theoretical models for the allelic structure of rare mendelian disease and common complex diseases have been described.

Natural selection has contrasting effects on the genetic architecture of rare mendelian diseases and common complex diseases. The causal alleles for many dominant mendelian diseases are generally new mutations (ie, derived alleles) for which selective pressure exists against their propagation (ie, purifying selection). Examples include familial hypercholesterolemia (eg, mutations in the low-density lipoprotein receptor gene [LDLR]), long-QT syndrome (eg, the calcium channel, voltage-dependent, L type, α1C subunit gene [CACNA1C]), and hypertrophic cardiomyopathy (eg, myosin, heavy chain 7, cardiac muscle, β [MYH7]). However, natural selection may operate in different directions in the heterozygous and homozygous forms of recessive mendelian diseases. Two classic examples, sickle cell disease and cystic fibrosis, illustrate the selective advantage for the heterozygote against malaria and diarrhea illness, respectively, whereas the homozygotes with the disease generally die before they reach reproductive age.

In contrast, the allelic spectrum of common complex diseases includes both common and rare variants. Recent genome-wide association studies have discovered many susceptibility variants of modest effect size that are prevalent in the general population, indicating that the variants were either neutral or advantageous in the past. Lifestyle changes such as culture and environment (eg, climate and diet) would have led to specific selective pressures as humans spread out of Africa. Present-day human beings live in a markedly different environment from their ancestors, so the common complex diseases might result from a mismatch between the ancestral alleles and current environments (ie, the ancestral-susceptibility model for complex diseases).

In other words, ancestral alleles reflect adaptations to the lifestyle of ancient human populations but are deleterious in the “modern” environment (Figure 4). The rare variant–common disease hypothesis posits that much of the disease susceptibility is due to many rare variants at different loci, each conferring a moderate increase in risk. These alleles are not as deleterious for reproductive fitness as are rare mutations in mendelian disease and therefore are not eliminated by strong purifying selection. Detection of such rare alleles will require sequencing of the genome or candidate genes because conventional genome-wide association studies are not powered to detect such alleles.

**Candidate Genes for CHD With Evidence for Natural Selection**

Two major challenges in studying the genetic basis and evolutionary genetics of CHD are its phenotypic complexity and the presence of multiple causal factors. Although considerable overlap exists among various CHD phenotypes, the underlying pathophysiology may be quite different. For example, myocardial infarction often results from rupture of a vulnerable atherosclerotic plaque, and risk factors and their interactions that determine plaque vulnerability may differ from those that influence coronary atherosclerotic burden as measured by the presence and quantity of coronary artery calcium. Epidemiological research over the last several decades has led to the discovery of several risk factors for CHD, including elevated lipid levels, diabetes mellitus, smoking, hypertension, inflammation, oxidative stress, and increased coagulability. In this section, we enumerate genes in causal pathways of atherosclerosis, including blood pressure regulation, lipoprotein and glucose metabolism, coagulation, and inflammation, that may be subject to various degrees of selective pressures resulting from climatic and dietary changes and host response to pathogens (Table). Candidate genes in these pathways may influence susceptibility to cardiovascular disease; we briefly discuss a few examples of such genes. It needs to be emphasized that for the loci identified as being under selection, further work is needed to address potential confounding caused by population history and structure and to establish definitive evidence of selection, mechanism of selection, and functional effects of the allelic variants under selection.
Differential susceptibility to hypertension is likely in part due to explain hypertension in the framework of evolution. Weder has summarized hypotheses that have been put forth to our history of adaptation to climatic changes. For example, the sodium hypothesis posits that sodium-conserving mechanisms conferred a survival advantage among our ancestors in the hot and humid climate of Africa but may lead to hypertension in temperate climates. Thus, sodium-conserving genotypes in US blacks may predispose to excessive sodium retention and sodium-sensitive hypertension in their current milieu. The ancestral allele of 4 is associated with CHD...
sodium-conserving alleles) show strong latitudinal gradients in allele frequency (clines) and are more prevalent in Africans than in populations from northern Europe, in whom signatures of positive selection (eg, high levels of LD and low haplotype diversity) are noted for the derived allele. The genetic patterns observed in candidate genes affecting sodium handling could have implications for the diagnosis and treatment of hypertension in geographically and ethnically defined populations around the world.60

The geographic distribution of the A(-6)G variant in the promoter region of the human angiotensinogen (AGT) gene suggests that the G(-6) variant has been selectively advantageous outside Africa.60 The G(-6) variant, the derived allele, is present at higher frequency in Asians and Europeans than in Africans,60 and evidence exists of a selective sweep in the vicinity of the polymorphism because haplotypes carrying the derived G(-6) allele showed elevated levels of LD in non-African populations. Genetic drift is not a likely explanation because frequencies of other AGT alleles are not similarly affected.60 A population bottleneck followed by population growth can result in relatively high LD. However, the differences in LD are pronounced around the A(-6)G polymorphisms but not in the entire gene.60,85 Other examples of correlation between latitude and “heat-adapted” alleles include variants in CYP3A5,61 GNB3,62 ADRB2,62 and SCNNIA62 (online-only Data Supplement Table II). Among these alleles, the variant 825T in GNB3 may account for a significant portion of worldwide variation in blood pressure.62

Lipid Metabolism

Humans, in whom cholesterol metabolism has diverged in a variety of ways from that of many distant mammals such as rodents and dogs, are susceptible to hypercholesterolemia given the present-day “atherogenic” diet.86 Several selective pressures, including climatic and dietary changes, may have influenced lipoprotein metabolism. Higher serum cholesterol may have been advantageous during the rapid increase in human size during human evolution and for its role in steroid hormone synthesis. For these reasons, the ε4 allele of the apolipoprotein E (APOE) gene may have been advantageous in the past, but it is maladaptive in the setting of nutritional abundance typical of the developed world, leading to increased risk of CHD.87,88 Consequently, the derived ε3 allele is adaptive in the modern environment.65,89

Two genes that play a key role in cholesterol homeostasis, LDLR and proprotein convertase subtilisin/kexin type 9 (PCSK9), a newly discovered regulator of LDLR, may have been under selection. LDLR has been shown to be differentially expressed among mammals,47,90 and an anthropoid primate-specific sequence element (a novel sterol regulatory element) that accounts for the strong activity characterizing human and other anthropoid primates has been identified.66 Species-specific regulation of LDLR in primates may have evolved as a result of varying availability of dietary cholesterol in the past.66 A study68 of the evolutionary genetics of PCSK9 revealed that it may have been subject to various forms of selection, both ancient and recent. Thus, phylogenetic analysis of the dS/dD ratio in PCSK9 in hominoids, Old World monkeys, and New World monkeys revealed evidence of ancient purifying selection, although the functional carboxyl-terminal domain appeared to be under positive selection in many branches across the phylogeny. In addition, a signature of recent positive selection was noted in 2 common nonsynonymous variants (rs505151 and rs562556) associated with higher plasma LDL cholesterol levels in the general population.69

Metabolic Syndrome and Diabetes Mellitus

The metabolic syndrome is characterized by clustering of several risk factors, including increased abdominal fat (obesity), atherogenic dyslipidemia, hypertension, and abnormal glucose metabolism.91 All of these risk factors increase the risk of atherosclerosis and CHD.92 In an influential paper published in 1962, Neel93 introduced the “thrifty-gene” hypothesis to explain the predisposition of certain ethnic groups to obesity and diabetes in the framework of evolution. He postulated that certain genetic variants in humans have evolved to maximize metabolic efficiency, lipid storage, and food searching behavior and that, in times of abundance, these variants predispose their carriers to diseases caused by excess nutritional intake. In the case of thrifty genes, the modern environment promotes the development of the metabolic syndrome.94

A recent study suggested that evolutionary pressures resulting from climatic changes may have played an important role in shaping variation in genes in metabolic pathways.82 Calpain-10 (CAPN10) encodes a member of the calpain-like cysteine protease family that is involved in the regulation of blood glucose levels.95 Population genetic analyses of SNP 44 (rs2975760) reveal a significant deficit of variation in the haplotypes carrying the derived allele,72,73 suggesting that it was quickly driven to high frequency by positive selection. The derived allele at SNP 44 is protective against diabetes mellitus, providing support for the ancestral-susceptibility model of common diseases.96,97 Another example of an ancestral allele predisposing to disease is the T allele of the SNP rs7903146 in the gene for transcription factor 7-like 2 (TCF7L2), which is strongly associated with type II diabetes mellitus.74 Positive selection has driven the haplotypes with the derived allele at this locus to near fixation in East Asians.74

Coagulation

Genes in the coagulation pathway may have been under selective pressure given the potential survival advantage from a robust clotting mechanism in the ancestral environment, possibly by enhancing hemostasis in the setting of bleeding from injury (eg, during hunting or after childbirth).35 The factor V (F5) Leiden mutation, which increases the risk of venous thromboembolism, is found at a high frequency (~5%) in European Americans. The mutation increases circulating thrombin generation and may predispose to atherosclerotic vascular disease.98 However, no evidence for natural selection has been noted so far in F5, and the relatively high prevalence of the Leiden mutation remains unexplained. Evidence for positive selection has been noted in the regulatory variant of the factor VII gene (F7), and the high-expression variant is common in Singaporean Chinese.75

Downloaded from http://circ.ahajournals.org/ by guest on January 15, 2018
The high-expression variant increases risk of CHD, and the selective forces responsible for its increased frequency are unknown. Reduced nucleotide diversity at the factor IX gene (F9) suggests recent positive selection or background selection, but how such selection might have affected reproductive fitness is not clear.

Inflammation

It is well known that the host response to pathogens has been under selective pressure in the past, exemplified by signatures of selection in the human leukocyte antigen locus (online-only Data Supplement Table I). A heightened immune response that developed to fight pathogens may increase the risk of CHD. Recent work has highlighted the role of the inflammatory response in atherosclerosis, and plasma levels of cytokines and the corresponding receptors are associated with CHD risk. Several genes in the inflammation pathway (see Table) have been reported to have evidence for positive selection.

Future Work

Several aspects of evolutionary genetics of CHD merit further investigation. It has yet to be proved whether allelic variants associated with CHD and identified as being under natural selection are associated with functional effects that affect fitness. As a result, mechanistic links between signals of natural selection and CHD have not been fully delineated. How allele frequency shifts caused by natural selection shape physiological and morphological characteristics and thereby CHD risk needs further study. For example, the 5T allele in the promoter region of MMP3 may have increased in frequency in European Americans because of the associated improvement in arterial elasticity and a consequent lowering of the risk of CHD. Second, evolutionary properties of genetic variation in regulatory regions need to be investigated because such variants may, in turn, lead to variation in gene expression, providing a substrate for evolution. Indeed, many of the genetic differences between humans and primates may be due to changes in the regulation of genes rather than differences in gene function. Third, a need exists to develop computational tests that are sensitive to deviation from evolutionary neutrality while being robust to population admixture and help predict which variants in humans may be associated with disease. Such investigation could provide novel insights into the genetic epidemiology and pathophysiology of atherosclerotic vascular disease, including CHD, and potentially open new avenues for prevention and treatment.

Conclusions

Genetic variants in pathways underlying complex disease such as inflammation and energy metabolism are a substrate for natural selection. Evolutionary hypotheses and models such as the thrifty-gene hypothesis and the ancestral-allele susceptibility model have been proposed to explain the epidemiology of complex diseases in the context of evolution. Increasing genetic evidence supports these hypotheses, motivating further exploration of the link between natural selection and common diseases. With the availability of comparative genomics and population genetic data, it should be possible to determine which genes, and what proportion of variants, have been affected by natural selection. This will allow further exploration into the molecular nature of adaptation and help predict which variants in humans may be associated with disease. Such investigation could provide novel insights into the genetic epidemiology and pathophysiology of atherosclerotic vascular disease, including CHD, and potentially open new avenues for prevention and treatment.

Acknowledgment

We would like to thank Pardis Sabeti for helpful comments.

Disclosures

None.

References


Evolutionary Genetics of Coronary Heart Disease
Keyue Ding and Iftikhar J. Kullo

_Circulation_. 2009;119:459-467
doi: 10.1161/CIRCULATIONAHA.108.809970
_Circulation_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/119/3/459

Data Supplement (unedited) at:
http://circ.ahajournals.org/content/suppl/2009/01/26/119.3.459.DC1
http://circ.ahajournals.org/content/suppl/2009/01/29/119.3.459.DC2

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in _Circulation_ can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to _Circulation_ is online at:
http://circ.ahajournals.org/subscriptions/
SUPPLEMENTARY MATERIALS

Glossary

Allele: one of the different forms of a gene, occupying the same locus (position) on a chromosome.

Allele / haplotype frequency: a measure of the relative frequency of an allele or a haplotype at a genetic locus in a population.

Ancestral and derived alleles: are inferred by comparing the allele observed at any given human polymorphic site with its orthologous position in the chimpanzee (the species closest to human). The allele found in chimpanzee is inferred to be ancestral, and the alternative allele is derived.

Background selection: the elimination of neutral polymorphisms as a result of the negative selection of deleterious mutations at linked sites.

Balancing selection: A form of natural selection in which multiple alleles are maintained in a population due to a heterozygote advantage.

dN/dS ratio: The ratio of the number of nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS).

Extended haplotype homozygosity: is the haplotype homozygosity between a distance x and the specified core region for a chromosome population carrying a single core haplotype. Haplotype homozygosity is an effective measure of linkage disequilibrium between two or more markers. It is calculated as \[ \frac{\sum(p_i^2) - 1/n}{1 - 1/n} \] with \( p_i \) being the relative haplotype frequency and \( n \) the sample size.

Haplotype: a combination of alleles at multiple linked loci that are transmitted together.

Linkage disequilibrium: the correlation among alleles from different loci.

Locus: the position of a gene, DNA marker or genetic marker on a chromosome.

Positive selection: a form of natural selection in which a novel variant has a selective advantage over others, and consequently rises in frequency.

Purifying selection: A form of natural selection in which a novel variant that has a selective disadvantage with respect to others, tends to remain at low frequencies or be removed.

Selective sweep: The reduction or elimination of neutral genetic variation in DNA adjacent to a beneficial mutation, as the result of recent and strong positive selection.
### Supplementary Table 1. Signatures of natural selection across the human leukocyte antigen (HLA) loci

<table>
<thead>
<tr>
<th>Loci</th>
<th>Test</th>
<th>Signature</th>
<th>Type of natural selection</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA (class I and II)</td>
<td>Divergence</td>
<td>$d_N$ significantly exceeds $d_S$</td>
<td>Positive selection has acted to enhance the rate of nonsynonymous substitution in the peptide-binding region codons</td>
<td>3</td>
</tr>
<tr>
<td>HLA (class I and DQB1)</td>
<td>Frequency spectrum</td>
<td>Overall-excess of intermediate-frequency alleles</td>
<td>Balancing selection has resulted in heterozygous individuals being able to respond to a broader range of parasite antigens</td>
<td>4</td>
</tr>
<tr>
<td>HLA (class I and DRB1 and DQB1)</td>
<td>Population differentiation</td>
<td>Increase in population differentiation ($F_{ST}$)</td>
<td>Directional selection has favored different alleles in distinct populations</td>
<td>4</td>
</tr>
<tr>
<td>HLA (class I and II)</td>
<td>LD</td>
<td>LD was lower in Africans and Europeans and higher in Oceania and north and South America</td>
<td>Positive selection has led to increased magnitude of LD</td>
<td>4</td>
</tr>
</tbody>
</table>

Abbreviations: Ref, references; LD, linkage disequilibrium
**Supplementary Table 2.** Derived alleles in candidate genes for hypertension with evidence of positive selection

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Gene</th>
<th>Derived alleles under selection*</th>
<th>Population</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGT</td>
<td>Angiotensinogen</td>
<td>A(-6)→G</td>
<td>Non-African</td>
<td>5</td>
</tr>
<tr>
<td>CYP3A5</td>
<td>Cytochrome P450, family, subfamily A, polypeptide 5</td>
<td>CYP3A5*3</td>
<td>European and Han Chinese</td>
<td>6</td>
</tr>
<tr>
<td>GNB3</td>
<td>G protein β-3 subunit</td>
<td>C825→T</td>
<td>Non-African</td>
<td>7</td>
</tr>
<tr>
<td>ADRB2</td>
<td>Adrenergic β-2 receptor</td>
<td>G47→A; G79→C</td>
<td>Non-African</td>
<td>7</td>
</tr>
<tr>
<td>SCNN1A</td>
<td>Sodium channel, non-voltage-gated 1</td>
<td>A(-173)→G; A(-946)→G</td>
<td>Non-African</td>
<td>7</td>
</tr>
</tbody>
</table>

*Alleles in bold indicate derived alleles under positive selection.

Abbreviations: Ref, references; AGT, angiotensinogen; CYP3A5, cytochrome P450, family 3, subfamily A, polypeptide 5; GNB3, guanine nucleotide binding protein (G protein), beta polypeptide 3; ADRB2, β-2-adrenergic receptor; SCNN1A, sodium channel, nonvoltage-gated 1 alpha.
Supplementary References


