Genome-wide detection of natural selection in African Americans pre-and post-admixture

Wenfei Jin, Shuhua Xu, Haifeng Wang, et al.

Genome Res. published online November 29, 2011
Access the most recent version at doi:10.1101/gr.124784.111

A Genome-Wide Detection of Natural Selection in African Americans Pre-and Post-Admixture

Wenfei Jin,[1] *Shuhua Xu,*[1,*] *Haifeng Wang,*[3] *Yongguo Yu,*[4] *Yiping Shen,*[2,5] *Bailin Wu,*[2,5] *Li Jin,*[1,2,*]

[1] Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;

[2] Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China;

[3] Chinese National Human Genome Center at Shanghai 201203, China;

[4] Shanghai Children's Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai 200127, China;

[5] Children's Hospital Boston, Harvard Medical School, Boston, MA 02115, USA.

* To whom correspondence should be addressed. E-mail: xushua@picb.ac.cn (S.X.); lijin.fudan@gmail.com (L.J.)

Running title: Natural Selection in African Americans

Keywords: Natural Selection; Admixture; African Americans; SNP; IPA; Malaria

Abstract

It is particularly meaningful to investigate natural selection in African Americans (AfA) due to the high mortality their African ancestry has experienced in history. In this study, we examined 491,526 autosomal SNPs genotyped in 5,210 individuals and conducted a genome-wide search for selection signals in 1,890 AfA. Several genomic regions showing excess of African or European ancestry, which were thought as the footprints of selection since population admixture, were detected based on a commonly used approach. However, we also developed a new strategy to detect natural selection both pre-and post-admixture by reconstructing an ancestral African population (AAF) from inferred African components of ancestry in AfA and comparing it with indigenous African populations (IAF). Interestingly, many selection-candidate genes identified by the new approach were associated with AfA specific high-risk diseases such as prostate cancer and hypertension, suggesting an important role these disease-related genes might have played in adapting to new environment. *CD36* and *HBB*, whose mutations confer a degree of protection against malaria, were also located in the highly differentiated regions between AAF and IAF. Further analysis showed that the frequencies of alleles protecting against malaria in AAF were lower than that in IAF, which consists with the relaxed selection pressure of malaria in the New World. There is no overlap between the top candidate genes detected by the two approaches, indicating the different environmental pressures AfA experienced pre-and post-population-admixture. We suggest that the new approach is reasonably powerful and can also be applied to other admixed populations such as Latinos and Uyghurs.

Supplementary material is available online at http://www.genome.org

Introduction

Although the vast majority of human genetic variations evolve neutrally, some parts, however, have been shaped by natural selection (Kimura 2003; Balaresque et al. 2007). Studies on recent natural selection have led to the discovery of genes showing population differences in adapting to pathogens, diet, climate and other environmental challenges. These discoveries have greatly enriched our understanding about the origins and also the evolutionary history of human species, identified many genes with important biological functions, and will lead to further elucidation of the genetic basis of some human diseases (Balaresque et al. 2007; Nielsen et al. 2007; Sabeti et al. 2007; Hancock et al. 2008; Akey 2009). The recent availability of high-density SNPs has provided essential resources for genome-wide detection of natural selection, especially in ethnically well-defined populations with little admixture (Akey 2009; Barreiro et al. 2008; Pickrell et al. 2009; Sabeti et al. 2007). Although there have been several studies on recently admixed populations (Tang et al. 2007; Basu et al. 2008; Bryc et al. 2010), no study so far has particularly investigated the locus-specific population differentiation between the ancestral components of the admixed population and its ancestral parental population, which might reflected the natural selection since the two split. In this case, we used African Americans, a well-studied admixed population, as the object of our study.

African Americans (AfA) are residents of the United States with partial recent Sub-Saharan African ancestry. The majority of them are descendents of the Africans, probably 500,000 to 650,000 in number, who were forcibly brought to North America during the Middle Passage (Thomas 1999; Zakharia et al. 2009). However, many of those captive Africans died either during the Atlantic shipment to America due to the severe

conditions, or soon upon their arrival in the New World as a result of exposure to foreign pathogens and/or the poor living conditions. Although the exact amount of life lost in this process remains a mystery, it may equal or exceed the actual amount enslaved (Stannard 1993). Therefore, the high mortality of the AfA in the whole slavery era could be attributed to the overwhelming environmental challenges. These persistent selection pressures might make the frequencies of 'beneficial' alleles increase continually, which should lead to higher population differentiation between African components of ancestry in AfA and indigenous Africans at these loci, in contrast to those evolved neutrally.

Meanwhile, before Africans and Europeans migrated to the New World, their ancestral parental populations had evolved independently in distinct environments for tens of thousands of years (Basu et al. 2008). The completely new environment in the New World constituted a challenge for both populations and the populations of their admixture. Therefore, it is very likely that some genomic regions in AfA show excess of a particular ancestry as a result of selection pressures after the population admixture (Tang et al. 2007; Basu et al. 2008). For example, there were various studies in detecting signatures of selection in AfA by examining admixture proportion using a small number of available loci (WORKMAN et al. 1963; Reed 1969; Blumberg and Hesser 1971; Adams and Ward 1973; Long 1991). Recently, Bryc *et al.* (2010) identified three autosomal regions showing excessive or reduced African ancestry (by at least three standard deviation of the mean) as natural selection candidates based on approximately 500K SNPs genotyped in 365 AfA.

Here, we analyzed the genotype data from approximately 500,000 autosomal SNPs shared by 1,890 AfA and 3,320 non-AfA for detecting the signatures of selection in AfA that were classified into pre-admixture and post-admixture according to the different

environments they experienced (Figure S1). Firstly, we detected the natural selections that were more likely to occur after admixture by examining the genome-wide distribution of ancestry in AfA, Then we developed a new strategy by reconstructing an ancestral African population (AAF) from inferred African components of ancestry in AfA and comparing it with indigenous African populations (IAF) (Figure S2), which reflect natural selection since the African ancestors of AfA left Africa (including both pre-and post-admixture). Many candidate genes identified by our new approach could explain the challenges that AfA and their ancestry had experienced. Thus we suggest that our new approach is reasonably powerful and can also be applied to the studies of other admixed populations.

Results

African and European ancestries in AfA

The populations from West Africa and Europe, who had undoubtedly contributed to the current gene pool of AfA, were considered as ancestral parental populations of AfA in this study. However, it was practically difficult to select proper populations as genetic donor to the gene pool of AfA since the extant populations in West Africa and Europe might not necessarily represent those 300 years ago, given the possible influence of genetic drift, selection, and demographic history. In this study, we chose the samples of YRI and CEU to respectively represent those who contributed to the formation of AfA, largely due to the availability of phased genetic data for these two populations. Besides, an exploratory analysis including Human Genome Diversity Project (HGDP) populations (Li et al. 2008) also suggested that YRI and CEU are better choice for ancestral parental populations of

AfA than the other populations with available genome-wide data (see Supplemental Text and Figure S3).

FRAPPE (Tang et al. 2005) was applied to the genome-wide high-density SNPs data (by taking K = 2), and the estimated European contribution to AfA was 21.65% at population level (Figure S4, S5). STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) analyses yielded virtually identical results. Based on the thinned data with 341,672 SNPs, the European contribution to the 1,890 AfA was estimated to be 21.61% (Figure S3D). These estimations were essentially consistent with those in previous studies (Bryc et al. 2010; Smith et al. 2004; Xu et al. 2007), although datasets analyzed were different.

Identification of genomic regions with biased ancestry in AfA

To estimate the distribution of genetic contribution of European and African ancestry to AfA across the genome, we used haplotypes of 88 CEU and 88 YRI (the phased parents of trios, presumably unrelated, from HapMap3 data) to represent their ancestral parental populations. The identical monomorphic SNPs in CEU and YRI samples were removed for they could not provide valuable information in the local ancestry inference. HAPMIX (Price et al. 2009) was employed to estimate the locus-specific genetic contributions of the ancestral parental populations to AfA (see Methods) by taking 21.65% as the European contribution, as estimated by FRAPPE. Based on the likelihood given by HAPMIX, generation since admixture (λ) was estimated to be $\lambda = 7$ (essentially hybrid isolation model), which was similar to those based on other AfA datasets (Smith et al. 2004; Price et al. 2009). Detailed analysis showed that λ values for most individuals ranged from 1 to 12, and could be explained by continuous-gene-flow (CGF) model (see Supplemental Text and Figure S6). The locus-specific European ancestry proportion across the genome of AfA

was estimated to be 21.68% ± 0.75% (mean ± SD). The standard deviation (SD) of locus-specific ancestral genetic contributions in this study is lower than those in any previous studies (WORKMAN et al. 1963; Reed 1969; Blumberg and Hesser 1971; Tang et al. 2007; Basu et al. 2008; Bryc et al. 2010), which was not beyond our expectation since we used a much larger sample size.

The genomic regions showing excessive or reduced ancestry in admixed population are likely to be signatures of natural selection (Tang et al. 2007; Basu et al. 2008; Bryc et al. 2010; Oleksyk et al. 2010). The loci showing strong deviation of European ancestry (3 SDs above or below the genome-wide average) were therefore identified as candidates of natural selection in this study. Four regions (2p22, 3q13, 6q26, 16q21) with excessive European influence and two regions (1p36, 2q37) with excessive African influence were observed in AfA genomes (Figure 1). Each of the six regions was significantly different from the genome-wide average of ancestral contributions ($P < 2.2\times10^{-16}$, t test). The detailed annotations of the six candidate regions are presented in Table 1. Most genomic regions showing ancestry deviation can be replicated by the analysis with LAMP-ANC (see Supplemental Text and Figure S7), although it has lower accuracy than HAPMIX based on our simulated data.

A close examination of SNPs in the six regions showing biased ancestry revealed neither significant deviation from Hardy-Weinberg expectation nor unusual fraction of missing data, suggesting that the genotyping quality is unlikely the cause of the bias. In addition, the six regions did not overlap with previously reported long-ranged LD blocks including inversions which may confound genome-wide scans for selection signals in admixed populations (Price et al. 2008), although three short inversions have been found in

one European individual in 1p36. Although the estimated locus-specific ancestral contributions across the genome are generally consistent with the study by Bryc *et al.* (2010), the three regions showing biased ancestry identified by them were only in moderate excess of African or European ancestry in this study, possibly due to the different samples and/or sample sizes between the two studies. Our estimations were also generally consistent with the genome-wide distribution of ancestry calculated in previous admixture mappings (Reich et al. 2005; Kao et al. 2008).

The region showing the strongest bias is located on 1p36, where African ancestry is over-represented. In this region, *IGSF21* and *AKR7A2* are located next to the SNPs that show strongest signals, respectively. *IGSF21* belongs to immunoglobin superfamily, while *AKR7A2* is involved in the detoxification of aldehydes and ketones, and is implicated in various cancers such as pancreatic cancer (Praml et al. 2008; Cui et al. 2009). Another region with over-represented African ancestry is 2q37, in which *PDCD1* is involved in signaling of immune system and various diseases. The region showing highest excessive European ancestry is located on 3q13 which only harbors *LSAMP*, a candidate of tumor suppressor gene in human osteosarcomas, and is associated with coronary artery disease (Kresse et al. 2009; Yen et al. 2009). Interestingly, *EIF2AK2*, involved in influenza infection pathways (McAllister et al. 2010; Pereira et al. 2010), is located ~200 Kb away from the peak (2p22) showing the second highest excessive European ancestry, which suggested the possible difference between African and European in resisting to influenza. *PACRG*, located in 6q26, is associated with Parkinson's disease. However, the region in 16q21 showing excessive European ancestry did not contain any genes or known function elements.

African/European components of ancestry in AfA

The segments of African ancestry and those of the European ancestry in AfA, inferred by HAPMIX, were collectively referred to as African components of ancestry and European components of ancestry, respectively, each of which could be considered as an ancestral African population (AAF) or an ancestral European population (AEU) residing in America before admixture. The inference of AAF and AEU are credible given the accuracy of HAPMIX that is over 98% based on simulated data (see Supplemental Text).

Then population differentiation between AEU and each putative European parental population was calculated, with CEU showing the lowest F_{ST} with AEU ($F_{ST\ [AEU\text{-}CEU]}$ = 0.0005) among all putative parental populations. When 2,648 Caucasian from GWAS data (referred to as CAU-GWAS) were considered, $F_{ST\ [AEU\text{-}CAU\text{-}GWAS]}$ was 0.0006, which was the second lowest among all values. Among all indigenous African populations (IAF), YRI showed the lowest F_{ST} with AAF ($F_{ST\ [AAF\text{-}YRI]}$ = 0.0007). When the observed F_{ST}'s were compared with those simulated under neutrality (see Methods), different pattern emerged between European ancestry and African ancestry. For European ancestry, the observed F_{ST} between AEU and CEU (0.0005) was lower than that simulated (simulated $F_{ST\ [AEU\text{-}CEU]}$ = 0.0006), and the genome-wide distribution of observed locus-specific F_{ST} did not deviate much from those simulated (P = 0.042, Kolmogorov–Smirnov test).

However, for African ancestry, the observed F_{ST} between AAF and YRI (0.0007) was higher than that simulated (simulated $F_{ST\ [AAF\text{-}YRI]}$ = 0.0006), and genome-wide distribution of locus-specific $F_{ST\ [AAF\text{-}YRI]}$ was significantly different from those simulated ($P < 2.2\times10^{-16}$, Kolmogorov–Smirnov test). In particular, Q-Q plot between the observed and simulated locus-specific $F_{ST\ [AAF\text{-}YRI]}$ showed an enrichment of SNPs with high F_{ST} at

the tail of the observed locus-specific F_{ST} (Figure S8), suggesting a possible role of natural selection. The results remain essentially unchanged when bottlenecks for African ancestry was assumed in our simulations.

Identification of regions highly differentiated between AAF and African

Since the African immigrants left for America, they might have experienced a completely different population history compared with indigenous Africans. The high mortality of AfA during the slavery era (Meltzer 1993; Stannard 1993; Thomas 1999) suggested a possible presence of strong selective pressures. According to the theory of neutrality, F_{ST} is largely influenced by demographic history which affect all loci similarly (Weir and Cockerham 1984; Kimura 2003), if not equally. By contrast, the force of positive selection acts in a locus-specific manner and tends to increase F_{ST} (Nielsen 2005), which has been widely used to detect the positive selection in various studies (Akey et al. 2002; Oleksyk et al. 2010). In this study, we calculated the locus-specific F_{ST} between AAF and YRI (Figure S9). However, loci with a very low minor allele frequency (MAF) may be subjected to sampling error and statistical error, therefore, the SNPs with MAF < 0.05 in AAF or YRI were removed from further analyses.

The genome-wide distribution of F_{ST} between AAF and YRI for 401,599 autosomal SNPs were presented in Figure 2A. In spite of the low differentiation between AAF and YRI, a substantial proportion of SNPs were located at the right tail of the distribution. The functional annotations of the most differentiated SNP clusters (99.99^{th} percentile; F_{ST}>0.0452) were listed in Table 2. Although it is reported that F_{ST} of individual-marker was too variable (Weir et al. 2005), the four most significant regions (7q21, 6p21-22, 1q22 and 11p15) carrying multiple highly differentiated SNPs should be

indicative. The 7q21 region harbors two genes: *CD36* and *SEMA3C*. *CD36* directly mediates cytoadherence of *Plasmodium falciparum* parasitized erythrocytes and it binds long chain fatty acids and may function in the transport and/or as a regulator of fatty acid transport (Oquendo et al. 1989; Baruch et al. 1996; Erdman et al. 2009). It was reported that *CD36* has been subjected to positive selection for malaria or some unknown selection pressures (Aitman et al. 2000; Omi et al. 2002; Omi et al. 2003; Fry et al. 2009). *SEMA3C*, induced by *ADAMTS1*, promotes the migration of cancer cells (Esselens et al. 2010). The 6p21-22 region, harboring human major histocompatibility complex (MHC), also showed an over-representation of African ancestry in this study (although less than 3SD) and has been reported under selection in various studies (Garrigan and Hedrick 2003; Tang et al. 2007). *MUC1*, located in 1q22, is involved in signaling by PDGF and serves a protective function by binding to pathogens (Davila et al. 2010; Li et al. 2010). *HBB* and *HBD*, located in 11p15, have been subjected to balancing selections because their mutations protected against malaria according to numerous studies (Ashley-Koch et al. 2000; Wood et al. 2005).

Ingenuity pathway analysis (IPA) was particularly helpful in exploring the function and pathways of the selection-candidate genes in the context of higher-order cellular and molecular mechanisms. The 402 SNPs with the highest F_{ST} (99.90^{th} percentile; F_{ST} >0.0287) were subjected to IPA analysis, whose results showed that genes involved in metabolic diseases were the most significantly enriched ($P = 1.51\times10^{-16}$) among all function classes, followed by the genes involved in endocrine system disorder ($P = 2.23\times10^{-16}$), immunological diseases ($P = 9.30\times10^{-12}$) and genetic disorder ($P = 5.67\times10^{-11}$). Antigen presentation pathway was most significantly enriched ($P = 1.95\times10^{-4}$), followed

by allograft rejection signaling ($P = 4.69\times10^{-3}$), Graft-versus-host disease signaling ($P = 4.69\times10^{-3}$) and autoimmune thyroid diseases signaling ($P = 5.35\times10^{-3}$). All of the four aforementioned pathways are related to immune system, which might reflect a great environmental differentiation between Sub-Saharan Africa and North America. We also conducted IPA on 4,011 SNPs showing the highest F_{ST} (99.00^{th} percentile; $F_{ST} > 0.0162$), which yielded results similar to that with $F_{ST} > 0.0287$ (99.90^{th} percentile). And two additional pathways emerged: IL-9 signaling pathway ($P = 8.01\times10^{-3}$) and EGF signaling pathway ($P = 6.38\times10^{-3}$).

Reconstituted African Americans and its difference with AfA

We then compared the AfA genome with that of reconstituted African American (rAfA) using genotypes of YRI and CEU. The rationale behind is that the former have been subjected to possible natural selection, while the latter have not. In particular, the allele frequencies of rAfA were estimated at each locus using YRI and CEU for the given level of admixture (21.68%) assuming no natural selection after admixture. The candidates identified under selection in this way could avoid potential errors introduced in the inference of ancestry. The genome-wide distribution of F_{ST} between AfA and rAfA was shown in Figure 2B. Overall 81% of the SNPs showing highest difference between AAF and IAF could be validated in the comparison between AfA and rAfA. In particular, these main F_{ST} peaks between AAF and YRI are essentially the same as those between AfA and rAfA, which indicated that almost all the selection signals identified by comparing AfA and rAfA were originated from AAF or IAF. We also used CAU-GWAS instead of CEU to construct rAfA, with the genomic distribution of F_{ST} similar to that using CEU (Figure S10).

Next, we reconstructed a rAfA population using a set of putative parental populations. First, we constructed an African parental population of AfA (APP) using 64% Yoruba, 19% Mandenka and 14% Bantu according to previous reported ancestral contribution to AAF (Zakharia et al. 2009). Then we used genotypes of the APP and CAU-GWAS to reconstruct rAfA. We obtained the genome-wide distribution of F_{ST} between AfA and rAfA (Figure S11). Although the population differentiation between this rAfA and AfA is higher than that based on two aforementioned rAfAs, the genome-wide distribution of F_{ST} is similar to that using only two pure parental populations.

Further evidences for positive selection in AAF and African

Since the force of positive selection acts in a locus-specific manner and tends to increase F_{ST} (Nielsen 2005), we hypothesized that positive selection preferentially acted upon functionally important loci over the others in the genome, thus leading to an enrichment of functional SNPs in the high F_{ST} bin (Barreiro et al. 2008). Here, we investigated the enrichment of different SNP classes among the high F_{ST} bin (top 1st percentile among all SNPs, $F_{ST} > 0.0164$) between AAF and YRI.

The SNPs, based on their location and function relative to the genes, were classified into nongenic, genic, intronic, 3'UTR, 5'UTR, synonymous, non-synonymous, coding, transcriptonic, near-gene-3 and near-gene-5 according to UCSC annotation. We found that F_{ST} distributions of each SNP category were not significantly different from those of non-genic SNPs. However, the proportion of genic SNPs among high F_{ST} bin (top 1st percentile among all SNPs, $F_{ST} > 0.0164$) is significantly higher than that of nongenic SNPs (χ^2 test, P=0.046; Figure 3). Notably, this excess is particularly marked for transcriptonic SNPs (χ^2 test, P=0.004; Figure 3). The proportion of synonymous SNPs in

high F_{ST} bin was 1.22-fold higher than the expectation under neutrality, which could attribute to the linkage disequilibrium of those SNPs with loci under selection (a phenomenon known as hitchhiking). Similar observations could be made when the thresholds for high F_{ST} bin were set at top 5% (F_{ST} >0.0083) or top 0.1% (F_{ST} >0.0304) (Figure S12, S13), and the conclusions still held when the SNPs with MAF >0.05 in both YRI and AAF were examined (Figure S14). The significant enrichment of high F_{ST} loci in the SNP categories with genetic functions supports the presence of positive selection either in AAF or in YRI, or both.

Discussion

Admixed populations such as African Americans provide a unique opportunity to study very recent natural selection, as their genomes are donated by long-diverged continental ancestries and may have been subjected to novel environmental challenges. The first strategy, detecting excessive or decreased ancestry contribution from its ancestral parental populations, has been used in several recent studies (Tang et al. 2007; Basu et al. 2008; Bryc et al. 2010). However, natural selections before admixture cannot be detected by this approach because the distribution of ancestry across the admixed genomes would not be affected by such selections. Therefore, we developed a new strategy, which examined selection since the African ancestry left for America (including selection both pre-and post-admixture). The candidate genes under selection identified by our new strategy were also confirmed in an analysis of allele frequence differentiation between rAfA and AfA (although the two analyses are not completely independent).

Above all, we have identified six regions showing excess of African or European ancestry using the first strategy. The highest ancestry deviation among all regions showing

excess of African and European ancestry is <0.026, which is lower than the values of any previous studies on the admixed populations in the New World (WORKMAN et al. 1963; Reed 1969; Blumberg and Hesser 1971; Tang et al. 2007; Basu et al. 2008; Bryc et al. 2010). For example, the African ancestry deviation even exceeds 0.14 in Puerto Rican according to a study by Tang *et al.* (2007), which is more than five-fold higher than that in this study. We attribute the low ancestry deviation to the much larger sample size, denser markers, more accurate phased data using trios, and more powerful statistical methods used in this study. Based on the maximum deviation regions, we estimated that the highest selection coefficient is approximate to be 0.002 (assuming 12 generations since admixture under CGF model). In fact, the real selection coefficient could be much lower than 0.002 considering the statistics and evolutionary noises. These results are reasonable considering the fact that there is much lower mortality rate of AfA in recent 200 years compared with that during Atlantic slave trade.

Secondly, we identified a large number of genes with highly differentiated allele frequencies between AAF and YRI using our new approach. These genes do not overlap with those identified by the former approach, suggesting the different environmental pressures AfA experienced before and after population admixture. IPA Analysis of SNPs with high population differentiation between AAF and YRI showed that genes involved in metabolic diseases ($P = 1.51\times10^{-16}$), endocrine system disorder ($P = 2.23 \times10^{-16}$), immunological diseases ($P = 9.30\times10^{-12}$) and genetic disorder ($P = 5.67\times10^{-11}$) were significantly enriched. Especially, we found that genes such as *PSCA*, *ZP4*, *AKAP12* were associated with AfA specific high-risk diseases such as hypertension and prostate cancer (Smith and O'Brien 2005; Goran 2008). Five genes (*CD36*, *HBB*, *HBD*, *HLA-B*, *HLA-DR*),

whose mutations protect against malaria, are also located in the highly differentiated regions between AAF and IAF.

Compared with Caucasians, African Americans have higher mortality rate for all cancers combined and for most major cancers (Jemal et al. 2006), as well as higher risk of obesity-related disease such as diabetes, hypertension and prostate cancer (Goran 2008). Interestingly, many genes located in selection candidate regions identified by the novel approach are associated with AfA ethnic high-risk diseases such as hypertension, prostate cancer and systemic sclerosis. Especially, One of the most significantly differentiated SNPs (rs2294008; F_{ST} =0.04561) between AAF and YRI, located in 8q24, is a missense mutation c.57T>C (p.Met1Thr) in *PSCA* and was reported to be associated with gastric and bladder cancer (Sakamoto et al. 2008; Matsuo et al. 2009; Wu et al. 2009). Many studies also reported that multiple loci in 8q24 were associated with prostate cancer in AfA (Freedman et al. 2006; Al Olama et al. 2009; Yeager et al. 2009). We proposed a hypothesis that most of the genes associated with AfA ethnic diseases may have played an important role in AfA's adaptation to local environment and thus show higher population differentiation between AAF and IAF. Further analysis of the 8q24 region would provide new insights into the etiology and evolutionary history of these cancers.

Among the selection candidate genes detected by genome-wide locus specific F_{ST} between AAF and YRI, five genes (*CD36*, *HBB*, *HBD*, *HLA-B*, *HLA-DR*) have been reported subjected to natural selection probably due to malaria (Kwiatkowski 1999; Kwiatkowski 2005). Because of the strong selection pressure of malaria, loss-of-function or abnormality of these genes was supposed to increase the survival rate of individual living in Africa. Some mutations in these genes have reached much higher frequencies in

Africans compared with that in areas of low incidence of malaria. However, these mutations that defend malaria could become disadvantage in AfA because the malaria was no longer a strong selection pressure in North America and these mutations could even lead to morbidity or mortality (Platt et al. 1994). We hypothesize that frequencies of these mutations would have decreased in AAF compared with those in indigenous African due to their disadvantage in AfA.

Next, we examined this hypothesis in the empirical data. Since the functional mutations in these genes were not genotyped in this study, we examined the SNPs strongly linked (linkage disequilibrium) with these mutations instead of these mutations themselves. It is well known that rs3211938 is a nonsense mutation c.1389T>G (p.Tyr325X) in *CD36* and has been subjected to natural selection because of malaria or some other environmental factors in African (Aitman et al. 2000; Omi et al. 2003; Erdman et al. 2009; Fry et al. 2009). We found three SNPs that are highly differentiated between AAF and YRI (Table S1) in 7q21 are strongly linked with rs3211938 (each with r^2 >0.4 in YRI). Interestingly, we did observe that the frequencies of the alleles linked with rs3211938 (G), which is the derived allele, were much lower in AAF compared with that in YRI (Table S1). Another example, rs334 is a missense mutation c.70A>T (p.Glu7Val) in *HBB*, which leads to sickle cell anemia [MIM 603903] (Ashley-Koch et al. 2000; Winichagoon et al. 2000; Wood et al. 2005), one of the most well-studied genetic disorder. rs7952293, one of the SNPs showing high F_{ST} between AAF and YRI, was strongly linked with rs334 (r^2 = 0.237 in YRI). In particular, the haplotype constructed by rs7952293(A) and rs334(T) accounted for 86.67% of the haplotypes containing rs334(T) in YRI. We observed that the frequency of rs7952293(A) in AAF (0.2261) was lower compared with

that in YRI (0.3172), which also supports the hypothesis that frequencies of alleles protecting against malaria in AAF are lower than those in indigenous African. The other three genes were not examined using the same procedure because the frequencies of mutations on these genes are too low to find strong-linked representative SNPs.

Our study takes advantage of both large sample size and high-density genome-wide data. However, our analysis demonstrated that the maximum deviation showing excess of African or European ancestry was small (<2.6%). Detecting such weak selection signals in admixed population such as AfA is a big challenge, which needs large sample size as in this study, or even larger sample size, to distinguish the real signals from the ancestry deviation caused by genetic drift and sampling error. Therefore, we propose that any study in the future trying to detect such weak selection signals in AfA or other recently admixed populations should collect at least thousands of samples. With the new strategy, we detected a lot of genes associated with AfA specific high-risk diseases such as hypertension and prostate cancer, and we also detected five genes whose mutations are against malaria. This new approach is powerful in detecting natural selection both before and after the establishment of AfA and can be applied to other admixed populations such as Latinos in the New world and the Uyghurs in Asia (Xu et al. 2008; Xu and Jin 2008; Xu et al. 2009).

Methods

Data assemble and quality control

The genotypic data were obtained from International HapMap Project (HapMap; http://www.hapmap.org), Human Genome Diversity Project (HGDP;

http://www.cephb.fr/en/hgdp), and Illumina iControlDB (http://www.illumina.com), respectively. The combined data set processed with PEAS v1.0 (Xu et al. 2010) includes the genotypic data of 588 HapMap samples (87 ASW, 167 YRI, 165 CEU, 85 CHD, 84 CHB) (Frazer et al. 2007), 300 HGDP samples (156 indigenous European, 102 indigenous African and 42 Amerindian) (Li et al. 2008), 2,161 AfA and 3,294 Caucasians (referred to as CAU-GWAS) from iControlDB genotyped by Illumina 550K Beadarray. The samples from iControlDB have passed Illumina's rigorous quality control and have been used as controls in five genome-wide association studies.

ASW from HapMap and AfA from iControlDB were merged into one AfA population, and Yoruba from HGDP was collected from Nigeria and therefore was merged with YRI from HapMap in subsequent analyses. The following samples were removed from the data set: (1) the relatives based on sample information or PLINK (Purcell et al. 2007) result (IBD >0.2), (2) individuals identified as outliers of each population (except AfA) based on the top ten principal components of PCA analysis (SD >6), (3) AfA individuals with >2% Native-American/East-Asian, (4) AfA individuals with >99% European contribution, which are likely to be descendents of individuals of European ancestry, (5) AfA individuals with >99% African contribution, in which recent African immigrants could not be practically identified (Bryc et al. 2010).

These filtered samples described above were merged and the SNPs sharing reference SNP ID (rs) and vendor-specified strands were kept in combined data. Then, the data set was further filtered for individuals with >10% missing genotypes and SNPs with >10% missing data, as well as Hardy-Weinberg disequilibrium ($P < 2\times10^{-6}$) within each population except AfA. The final data set comprised of 503,694 SNPs (491,526 autosomal

SNPs) shared by 5,210 individuals from 21 population groups, with total genotyping call rate 99.74%.

Populations and samples

Overall, 1,890 unrelated AfA samples with ignorable ancestry outside of Africa and Europe were studied, among which 1,838 individuals were downloaded from iControlDB, and another 52 individuals were from the International HapMap Project (Altshuler et al. 2010). In addition, 113 YRI from HapMap (Altshuler et al. 2010) and 102 indigenous Africans in 7 different groups collected from HGDP (Li et al. 2008) were merged, representing the extant Africans, while 113 CEU from HapMap (Altshuler et al. 2010) were merged with 156 Europeans in 8 different groups collected from HGDP, representing the extant Europeans. In total 2,648 CAU-GWAS samples from iControlDB were taken as another representation of the extant Europeans. In addition, 84 CHB and 85 CHD represented populations from East Asia, and 24 pure Amerindians from HGDP represented Native American.

Population genetic analysis

In order to reduce the linkage disequilibrium (LD) between markers, those with r^2 >0.5 were removed, calculated in sliding window of 50 SNPs and shifted every five SNPs (see Supplemental Text). This process reduced the original dataset to 341,672 autosomal SNPs. Based on these thinned markers, principle component analysis (PCA) was performed at the individual level using *smartpca*, from the package EIGENSOFT (Patterson et al. 2006). Individual ancestry proportion was estimated using FRAPPE, which implements an expectation-maximization (EM) algorithm (Tang et al. 2005). FRAPPE was run on all 491,526 SNPs with 10,000 iterations by setting K from 2 to 4. We

also ran STRUCTURE (Falush et al. 2003) on SNPs with inter-marker distance more than 1M (see Supplemental Text). Genetic difference between populations was measured using F_{ST} following Weir and Cockerham (Weir and Cockerham 1984), which accounts for differences in the sample size in each population. The locus-specific F_{ST} between any two populations was also calculated using the same formula.

Locus-specific ancestry inference

Various methods and software have been developed for inferring locus-specific ancestry based on high density SNPs data, such as ANCESTRYMAP (Patterson et al. 2004), SABER (Tang et al. 2006), LAMP and LAMP-ANC (Sankararaman et al. 2008b), uSWITCH and uSWITCH-ANC (Sankararaman et al. 2008a), HAPAA (Sundquist et al. 2008) and HAPMIX (Price et al. 2009). A simple analysis showed that HAPMIX outperformed other methods with implemented software based on our simulated data (see Supplemental Text). Therefore, HAPMIX was used to infer locus-specific ancestry in African Americans in this study. We also used LAMP-ANC to do a similar analysis since it also performs very well.

The phased data of HapMap 3 were downloaded from HapMap website (Altshuler et al. 2010). We used haplotypes of 88 CEU and 88 YRI (all from trio samples) representing the European and African ancestral populations, respectively in the subsequent analysis. The mean European ancestry proportion in AfA (θ), which is a required input parameter for HAPMIX, was based on the estimation of FRAPPE. Generation since admixture (λ) with the largest likelihood was taken as its estimation. By running HAPMIX in diploid mode, we obtained the haplotypes and ancestry segments for each AfA individual. Then we reconstructed an ancestral African population (AAF) and an ancestral European

population (AEU) using inferred chromosomal segments of African ancestry and those of European ancestry in AfA, respectively. In brief, we constructed AAF using only those chromosomal segments with pure African ancestry. For each given SNP, allele frequency of AAF was calculated based on the available genotypes with African ancestry across all AfA. This procedure was also applied to the construction of AEU using chromosomal segments with pure European ancestry.

Simulation of AfA and its parental populations

Under selective neutrality, genetic drift of the admixed population and its parental populations contribute to the variation of ancestry proportion and population differentiation among loci (Weir and Cockerham 1984; Long 1991). Therefore, we performed an extensive forward-time simulation to explore the potential impact of genetic drift on the locus-specific population differentiations between AAF and YRI, as well as that between AEU and CEU. In this simulation, recombination was introduced according to the genetic map adapted from HapMap (release #22) (Frazer et al. 2007), and mutation was ignored given the short history of AfA. The effective population sizes (N_e) of each population was obtained from the HapMap website (Altshuler et al. 2010). In particular, N_e for African, European and AfA were set to 17,094, 11,418 and 17,094, respectively. Continuous-gene-flow (CGF) model (Figure S15) was used based on previous studies (Pfaff et al. 2001; Price et al. 2009).

The aforementioned phased data of 88 YRI and 88 CEU were taken respectively as the genotypes of common African ancestry and common European ancestry before population admixture. We simulated individuals of AfA by constructing their genomes from a mosaic of haploid YRI and haploid CEU genomes. The generations since admixture

(λ) was set to 12 according to CGF model based on previous reports (Pfaff et al. 2001; Price et al. 2009), which was also supported by our observations (Figure S6). The gene flow (α) that African ancestry received from European each generation was calculated by $\alpha = 1- (m_1)^{1/t}$, in which m_1 represents the mean proportion of European ancestry in AfA. Both African and European parental populations evolved 12 generations simultaneously. Finally, genotypes of 113 European, 113 African and 1,890 AfA, as well as their SNP-specific ancestral status, were output to match the sample sizes of the data. Based on the primary simulation, we performed extended simulations by setting a bottleneck event in the first generation during which N_e for AfA were reduced to 8,000, 5,000, 3,000, 2,000 and 1,000, respectively.

Function annotations and ingenuity pathway analysis (IPA)

Genomic regions that deviated from genome-wide ancestral contributions or with extremely high F_{ST} were annotated based on HapMap website (Thorisson et al. 2005). The SNPs showing substantial population differentiation were interrogated for network and functional interrelatedness using the Ingenuity Pathway Analysis (IPA) version 8.5 software tools. This software searches for information on genes in Ingenuity Pathways Knowledge Base, a repository of molecular interactions, regulatory events, gene-to-phenotype associations, and chemical knowledge, all collected from the full text of the peer-reviewed life sciences literatures. With IPA, we can analyze data in the context of molecular mechanisms, identify key mechanistic differences between subpopulations, and further relate molecular events to higher-order cellular and disease processes.

Statistical analysis

All statistical computation and graphics were performed using R version 2.9 (Ihaka and Gentleman 1996). Kolmogorov–Smirnov tests were performed to compare the empirical distributions of locus-specific F_{ST} with those simulated. And Chi-squared (χ^2 test) tests were performed to test the over-representation of each SNP category compared with nongenic SNPs among the high F_{ST} bin.

Acknowledgements

SX was supported by the National Science Foundation of China (30971577, 31171218), Shanghai Rising-Star Program (11QA1407600), and Science Foundation of The Chinese Academy of Sciences (KSCX2-EW-Q-1-11, KSCX2-EW-R-01-05, KSCX2-EW-J-15-05). LJ was supported by the National Science Foundation of China (30890034, 30625016) and the Science and Technology Commission of Shanghai Municipality (09540704300). SX is a Max-Planck Independent Junior Research Group Leader and a member of CAS Youth Innovation Promotion Association. SX also gratefully acknowledges the support of K.C.Wong Education Foundation, Hong Kong. BW is a senior author of Harvard group. This work was also supported by the MoST International Cooperation Base of China.

Authors' contributions

S.X. and **L.J.** conceived and designed the study. **H.W.** collected genotype data from the Illumina iControl Database (iControlDB). **W.J.** performed data analysis, with contribution from **S.X.**. **Y.Y.**, **Y.S.** and **B.W.** performed IPA analysis. **S.X.** and **W.J.** interpreted the data. **W.J.**, **S.X.** and **L.J.** wrote the paper. All authors read and approved the final manuscript. All authors declare that no competing financial interests exist.

1

2

References

Adams, J. and R.H. Ward. 1973. Admixture studies and the detection of selection. *Science* 180: 1137-1143.

Aitman, T.J., L.D. Cooper, P.J. Norsworthy, F.N. Wahid, J.K. Gray, B.R. Curtis, P.M. McKeigue, D. Kwiatkowski, B.M. Greenwood, R.W. Snow et al. 2000. Malaria susceptibility and CD36 mutation. *Nature* 405: 1015-1016.

Akey, J.M. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* 19: 711-722.

Akey, J.M., G. Zhang, K. Zhang, L. Jin, and M.D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12: 1805-1814.

Al Olama, A.A., Z. Kote-Jarai, G.G. Giles, M. Guy, J. Morrison, G. Severi, D.A. Leongamornlert, M. Tymrakiewicz, S. Jhavar, E. Saunders et al. 2009. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* 41: 1058-1060.

Altshuler, D.M., R.A. Gibbs, L. Peltonen, D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, L. Peltonen et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58.

Ashley-Koch, A., Q. Yang, and R.S. Olney. 2000. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. *Am J Epidemiol* 151: 839-845.

Balaresque, P.L., S.J. Ballereau, and M.A. Jobling. 2007. Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* 16 (R2): R134-139.

Barreiro, L.B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* 40: 340-345.

Baruch, D.I., J.A. Gormely, C. Ma, R.J. Howard, and B.L. Pasloske. 1996. Plasmodium falciparum erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule 1. *Proc Natl Acad Sci U S A* 93: 3497-3502.

Basu, A., H. Tang, X. Zhu, C.C. Gu, C. Hanis, E. Boerwinkle, and N. Risch. 2008. Genome-wide distribution of ancestry in Mexican Americans. *Hum Genet* 124: 207-214.

Blumberg, B.S. and J.E. Hesser. 1971. Loci differentially affected by selection in two American black populations. *Proc Natl Acad Sci U S A* 68: 2554-2558.

Bryc, K., A. Auton, M.R. Nelson, J.R. Oksenberg, S.L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S.A. Tishkoff et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* 107: 786-791.

Cui, Y., M. Tian, M. Zong, M. Teng, Y. Chen, J. Lu, J. Jiang, X. Liu, and J. Han. 2009. Proteomic analysis of pancreatic ductal adenocarcinoma compared with normal adjacent pancreatic tissue and pancreatic benign cystadenoma. *Pancreatology* 9: 89-98.

Davila, S., F.E. Froeling, A. Tan, C. Bonnard, G.J. Boland, H. Snippe, M.L. Hibberd, and M. Seielstad. 2010. New genetic associations detected in a host response study to hepatitis B vaccine. *Genes Immun* 11: 232-238.

Erdman, L.K., G. Cosio, A.J. Helmers, D. Gowda, S. Grinstein, and K.C. Kain. 2009. CD36 and TLR Interactions in Inflammation and Phagocytosis: Implications for Malaria. *J Immunol 183: 6452-6459*.

Esselens, C., J. Malapeira, N. Colome, C. Casal, J.C. Rodriguez-Manzaneque, F. Canals, and J. Arribas. 2010. The cleavage of semaphorin 3C induced by ADAMTS1 promotes cell migration. *J Biol Chem* 285: 2463-2473.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567-1587.

Frazer, K.A. D.G. Ballinger D.R. Cox D.A. Hinds L.L. Stuve R.A. Gibbs J.W. Belmont A. Boudreau P. Hardenbol S.M. Leal et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

Freedman, M.L., C.A. Haiman, N. Patterson, G.J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R.G. Steen, K. Ardlie, E.M. John et al. 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A* 103: 14068-14073.

Fry, A.E., A. Ghansa, K.S. Small, A. Palma, S. Auburn, M. Diakite, A. Green, S. Campino, Y.Y. Teo, T.G. Clark et al. 2009. Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Hum Mol Genet* 18: 2683-2692.

Garrigan, D. and P.W. Hedrick. 2003. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution* 57: 1707-1722.

Goran, M.I. 2008. Ethnic-specific pathways to obesity-related disease: the Hispanic vs. African-American paradox. *Obesity (Silver Spring)* 16: 2561-2565.

Hancock, A.M., D.B. Witonsky, A.S. Gordon, G. Eshel, J.K. Pritchard, G. Coop, and A. Di Rienzo. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet* 4: e32.

Ihaka, R. and R. Gentleman. 1996. R: A language for data analysis and graphics. *J Comput Graph Statist* 5: 299-314.

Jemal, A., R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, and M.J. Thun. 2006. Cancer statistics, 2006. *CA Cancer J Clin* 56: 106-130.

Kao, W.H., M.J. Klag, L.A. Meoni, D. Reich, Y. Berthier-Schaad, M. Li, J. Coresh, N. Patterson, A. Tandon, N.R. Powe et al. 2008. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nat Genet* 40: 1185-1192.

Kimura, M. 2003. The neutral theory of molecular evolution. *Cambridge University Press, Cambridge, United Kingdom.*

Kresse, S.H., H.O. Ohnstad, E.B. Paulsen, B. Bjerkehagen, K. Szuhai, M. Serra, K.L. Schaefer, O. Myklebost, and L.A. Meza-Zepeda. 2009. LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. *Genes Chromosomes Cancer* 48: 679-93.

Kwiatkowski, D. 1999. The molecular genetic approach to malarial pathogenesis and immunity. *Parassitologia* 41: 233-240.

Kwiatkowski, D.P. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* 77: 171-192.

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza et al. 2008. Worldwide human

relationships inferred from genome-wide patterns of variation. *Science* 319: 1100-1104.

Li, Y., D.L. Dinwiddie, K.S. Harrod, Y. Jiang, and K.C. Kim. 2010. Anti-inflammatory effect of MUC1 during respiratory syncytial virus infection of lung epithelial cells in vitro. *Am J Physiol Lung Cell Mol Physiol* 298: L558-563.

Long, J.C. 1991. The genetic structure of admixed populations. *Genetics* 127: 417-428.

Matsuo, K., K. Tajima, T. Suzuki, T. Kawase, M. Watanabe, K. Shitara, K. Misawa, S. Ito, A. Sawaki, K. Muro et al. 2009. Association of prostate stem cell antigen gene polymorphisms with the risk of stomach cancer in Japanese. *Int J Cancer* 125: 1961-1964.

McAllister, C.S., A.M. Toth, P. Zhang, P. Devaux, R. Cattaneo, and C.E. Samuel. 2010. Mechanisms of protein kinase PKR-mediated amplification of beta interferon induction by C protein-deficient measles virus. *J Virol* 84: 380-386.

Meltzer , M. 1993. Slavery: A World History. *Da Capo Press*, New York.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu Rev Genet* 39: 197-218.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A.G. Clark. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* 8: 857-868.

Oleksyk, T.K., M.W. Smith, and S.J. O'Brien. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci* 365: 185-205.

Omi, K., J. Ohashi, J. Patarapotikul, H. Hananantachai, I. Naka, S. Looareesuwan, and K. Tokunaga. 2002. Fcgamma receptor IIA and IIIB polymorphisms are associated with susceptibility to cerebral malaria. *Parasitol Int* 51: 361-366.

Omi, K., J. Ohashi, J. Patarapotikul, H. Hananantachai, I. Naka, S. Looareesuwan, and K. Tokunaga. 2003. CD36 polymorphism is associated with protection from cerebral malaria. *Am J Hum Genet* 72: 364-374.

Oquendo, P., E. Hundt, J. Lawler, and B. Seed. 1989. CD36 directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. *Cell* 58: 95-101.

Patterson, N., N. Hattangadi, B. Lane, K.E. Lohmueller, D.A. Hafler, J.R. Oksenberg, S.L. Hauser, M.W. Smith, S.J. O'Brien, D. Altshuler et al. 2004. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 74: 979-1000.

Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. *PLoS Genet* 2: e190.

Pereira, R.M., K.L. Teixeira, V. Barreto-de-Souza, T.C. Calegari-Silva, L.D. De-Melo, D.C. Soares, D.C. Bou-Habib, A.M. Silva, E.M. Saraiva, and U.G. Lopes. 2010. Novel role for the double-stranded RNA-activated protein kinase PKR: modulation of macrophage infection by the protozoan parasite Leishmania. *Faseb J* 24: 617-626.

Pfaff, C.L., E.J. Parra, C. Bonilla, K. Hiester, P.M. McKeigue, M.I. Kamboh, R.G. Hutchinson, R.E. Ferrell, E. Boerwinkle, and M.D. Shriver. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet* 68: 198-207.

Platt, O.S., D.J. Brambilla, W.F. Rosse, P.F. Milner, O. Castro, M.H. Steinberg, and P.P. Klug. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* 330: 1639-1644.

Praml, C., W. Schulz, A. Claas, J. Mollenhauer, A. Poustka, R. Ackermann, M. Schwab, and K.O. Henrich. 2008. Genetic variation of Aflatoxin B1 aldehyde reductase genes (AFAR) in human tumour cells. *Cancer Lett* 272: 160-166.

Price, A.L., A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5: e1000519.

Price, A.L., M.E. Weale, N. Patterson, S.R. Myers, A.C. Need, K.V. Shianna, D. Ge, J.I. Rotter, E. Torres, K.D. Taylor et al. 2008. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 83: 132-135.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575.

Reed, T.E. 1969. Caucasian genes in American Negroes. *Science* 165: 762-768.

Reich, D., N. Patterson, P.L. De Jager, G.J. McDonald, A. Waliszewska, A. Tandon, R.R. Lincoln, C. DeLoa, S.A. Fruhan, and P. Cabre. 2005. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genet* 37: 1113-1118.

Sabeti, P.C. P. Varilly B. Fry J. Lohmueller E. Hostetter C. Cotsapas X. Xie E.H. Byrne S.A. McCarroll R. Gaudet et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-918.

Sakamoto, H., K. Yoshimura, N. Saeki, H. Katai, T. Shimoda, Y. Matsuno, D. Saito, H. Sugimura, F. Tanioka, S. Kato et al. 2008. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat Genet* 40: 730-740.

Sankararaman, S., G. Kimmel, E. Halperin, and M.I. Jordan. 2008. On the inference of ancestries in admixed populations. *Genome Res* 18: 668-675.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin. 2008. Estimating local ancestry in admixed populations. *Am J Hum Genet* 82: 290-303.

Smith, M.W. and S.J. O'Brien. 2005. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet* 6: 623-632.

Smith, M.W., N. Patterson, J.A. Lautenberger, A.L. Truelove, G.J. McDonald, A. Waliszewska, B.D. Kessing, M.J. Malasky, C. Scafe, and E. Le. 2004. A high-density admixture map for disease gene discovery in african americans. *The Am J Hum Genet* 74: 1001-1013.

Stannard, D. 1993. American Holocaust. *Oxford University Press*, Oxford, United Kingdom.

Sundquist, A., E. Fratkin, C.B. Do, and S. Batzoglou. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res* 18: 676-682.

Tang, H., S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E.G. Burchard, and N.J. Risch. 2007. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* 81: 626-633.

Tang, H., M. Coram, P. Wang, X. Zhu, and N. Risch. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 79: 1-12.

Tang, H., J. Peng, P. Wang, and N.J. Risch. 2005. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology* 28: 289-301.

Thomas, H. 1999. The Slave Trade: The Story of the Atlantic Slave Trade. *Simon & Schuster* 1440-1870.

Thorisson, G.A., A.V. Smith, L. Krishnan, and L.D. Stein. 2005. The international HapMap project web site. *Genome Res* 15: 1592-1593.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 15: 1468-1476.

Weir, B.S. and C.C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.

Winichagoon, P., S. Fucharoen, P. Chen, and P. Wasi. 2000. Genetic factors affecting clinical severity in beta-thalassemia syndromes. *J Pediatr Hematol Oncol* 22: 573-580.

Wood, E.T., D.A. Stover, M. Slatkin, M.W. Nachman, and M.F. Hammer. 2005. The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. *Am J Hum Genet* 77: 637-642.

WORKMAN, P.L., B.S. BLUMBERG, and A.J. COOPER. 1963. SELECTION, GENE MIGRATION AND POLYMORPHIC STABILITY IN A U. S. WHITE AND NEGRO POPULATION. *Am J Hum Genet* 15: 429-437.

Wu, X., Y. Ye, L.A. Kiemeney, P. Sulem, T. Rafnar, G. Matullo, D. Seminara, T. Yoshida, N. Saeki, A.S. Andrew et al. 2009. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat Genet* 41: 991-995.

Yeager, M., N. Chatterjee, J. Ciampa, K.B. Jacobs, J. Gonzalez-Bosquet, R.B. Hayes, P. Kraft, S. Wacholder, N. Orr, S. Berndt et al. 2009. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* 41: 1055-1057.

Yen, C.C., W.M. Chen, T.H. Chen, W.Y. Chen, P.C. Chen, H.J. Chiou, G.Y. Hung, H.T. Wu, C.J. Wei, and C.Y. Shiau. 2009. Identification of chromosomal aberrations associated with disease progression and a novel 3q13. 31 deletion involving LSAMP gene in osteosarcoma. Int J Oncol 35: 775-788.

Zakharia, F., A. Basu, D. Absher, T.L. Assimes, A.S. Go, M.A. Hlatky, C. Iribarren, J.W. Knowles, J. Li, B. Narasimhan et al. 2009. Characterizing the admixed African ancestry of African Americans. *Genome Biol* 10: R141.

Tables

Table 1. Regions Showing Excess of European or African Ancestry.

Regions	Position	Excess ancestry	Size (bp)	SNPs	Highest deviation	Genes	Pathways	Related diseases
1p36	chr1:17409539..21604321	African	4194783	489	0.0253	*AKR7A2**, *IGSF21*, *DDOST**, HTR6 *et al.*	Diabetes pathways, signaling by GPCR, metabolism of amino acids	Diabetes, pancreatic cancer
2q37	chr2:241750403..242568618	African	818216	16	0.0231	*SEPT2**, *HDLBP**, *PDCD*1*, *FARP2 et al.*	Signaling in immune system, Axon guidance, metabolism of nucleotides	Bladder cancer, lung cancer, coronary atherosclerosis
2p22	chr2:37451925..37508581	European	56657	9	0.0230	*QPCT*, (*EIF2AK2** 222kb)	Influenza infection	Influenza infection
3q13	chr3:116930811..118313302	European	1382492	216	0.0253	*LSAMP**	Homophilic adhesion	Osteosarcoma
6q26	chr6:163653158..163653428	European	271	2	0.0225	*PACRG**	Mediate proteasomal degradation	Juvenile Parkinson's disease
16q21	chr16:61214438..61242497	European	28060	9	0.0229	NA	NA	NA

NA: not available. * Denotes genes associated with diseases. Genes in parentheses are strong candidates out of the chromosome location but closest.

Table 2. Regions with highly differentiated allele frequency between AAF and YRI (F_{ST} >0.0452).

Regions or SNPs	Position	Size (bp)	SNPs	Highest F_{ST}	Genes	Pathways	Related disease
1p21	chr1:100125058..100183875	58817	2	0.0562	*AGL**	Metabolism of carbohydrate	Glycogen storage disease
1q22	chr1:153401959..153464086	62127	4	0.0692	*THBS3**, *MUC1**, *MTX1*, *TRIM46*, *KRTCAP2*	Signaling by PDGF	Stomach cancer, breast cancer, osteosarcoma
rs12094201	chr1:236509336	1	1	0.0561	(*ZP4** 389kb)	NA	Hypertension, Non-alcoholic fatty liver
rs7642575	chr3: 31400165	1	1	0.0453	(*STT3B*, *OSBPL10** 149 kb)	NA	Peripheral arterial disease
6p21-p22	chr6:26554684..33961049	7406365	11	0.0711	*HLA-B**, *HLA-C*, *EHMT2**, *HLA-DPA1**, *HLA-DRB5*, *EHM*, *BTN3A3*, *et al*	Signaling by GPCG, signaling in immune system, HIV infection, Diabetes pathway	HIV, Crohn's disease, rheumatoid arthritis, juvenile idiopathic arthritis, colorectal cancer, systemic sclerosis
6q25	chr6:151555551..151569258	13707	2	0.0545	(*AKAP12** 40kb)	Cell growth	Hypertension, hemorrhagic stroke
rs10499542	chr7: 22235870		1	0.04606	*RAPGEF5**	GTP/GDP-regulation	Thyroid stimulating hormone
7q21	chr7:79768487..80482597	714110	10	0.0946	*CD36**, *SEMA3C*	Metabolism of lipids and lipoprotein	Metabolic syndrome, malaria
8q24	chr8:143754039..143758933	4894	2	0.04679	*PSCA**	NA	Prostate cancer, bladder cancer, gastric cancer
11p15	chr11:5034229..5421456	387227	3	0.0617	*HBB**, *HBD**, *HBE1**, *HBG2*, *OR51I1*, *et al*	Signaling by GPCR	Sickle cell disease, beta-thalassemia, malaria
rs4883422	chr12:7189594	1	1	0.04721	*CLSTN3*	NA	NA
rs6491096	chr13:25488362	1	1	0.04716	*ATP8A2*	NA	NA
rs1075875	chr16: 47595721	1	1	0.0766	(*CBLN1* 277kb)	NA	NA
rs6015945	chr20:59319574	1	1	0.0627	*CDH4**	Cell junction organization	Alzheimer's Disease

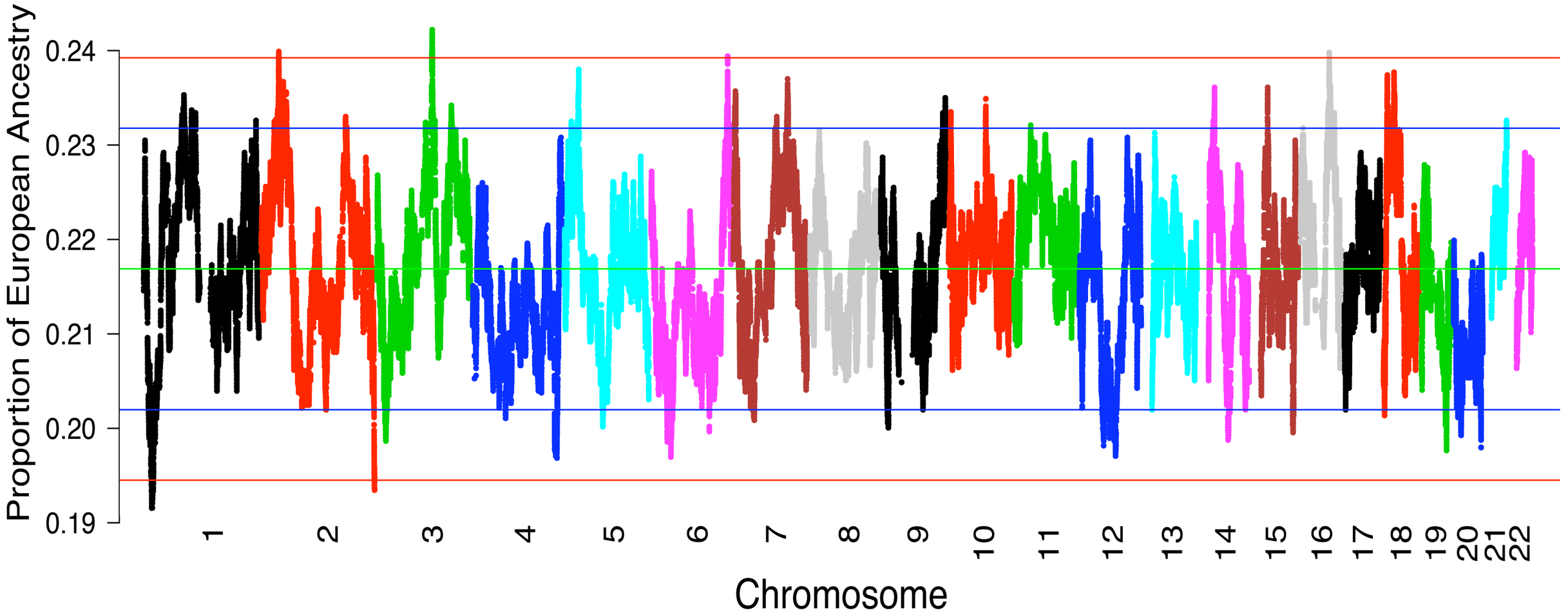
NA: not available. * Denotes the genes associated with diseases. Genes in parentheses are strong candidates out of the chromosome location but closest.

FIGURE LEGENDS

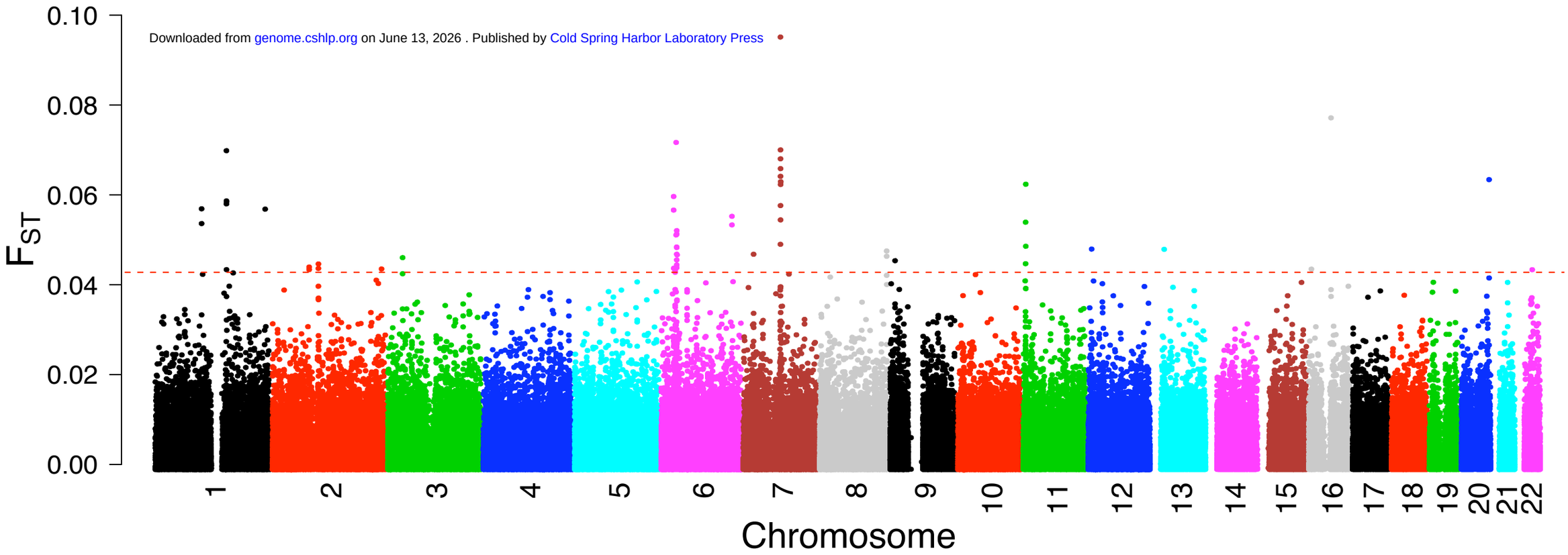
Figure 1. Genome-wide Distribution of European Ancestral Contributions. Mean European ancestral contribution across 1,890 African American individuals at each SNP. Green line is the estimated genome-wide mean European ancestral contribution (21.68%). Blue bands indicate +2 and -2 SDs from the mean ancestral contribution and red Bands indicate +3 and -3 SDs from the mean ancestral contribution.

Figure 2. Genomic Distribution of F_{ST} between AAF and YRI (A) and Genomic Distribution of F_{ST} between African American and rAfA (B). The dashed red horizontal line indicates the cutoff threshold (99.99^{th} percentile). Locus-specific F_{ST} between YRI and CEU were calculated when MAF >0.05 in both populations. The rAfA was constituted according to the ancestry proportion of CEU and YRI under neutrality.

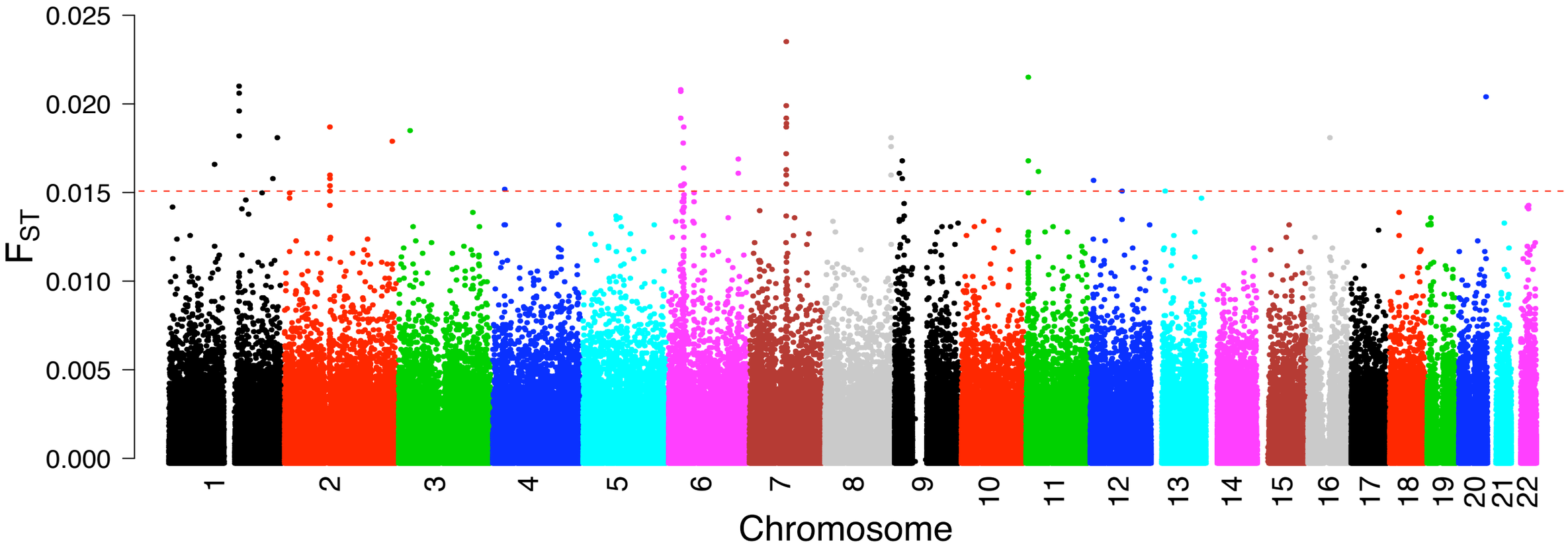
Figure 3. Enrichment of high F_{ST} loci for different SNP categories. Observed excess of high F_{ST} loci in different SNP classes, with respect to non-genic class, among high F_{ST} bin (99^{th} percentile; F_{ST} >0.0164). The values on the bar are p-values of χ^2 tests. “NS” stands for “not significant”.

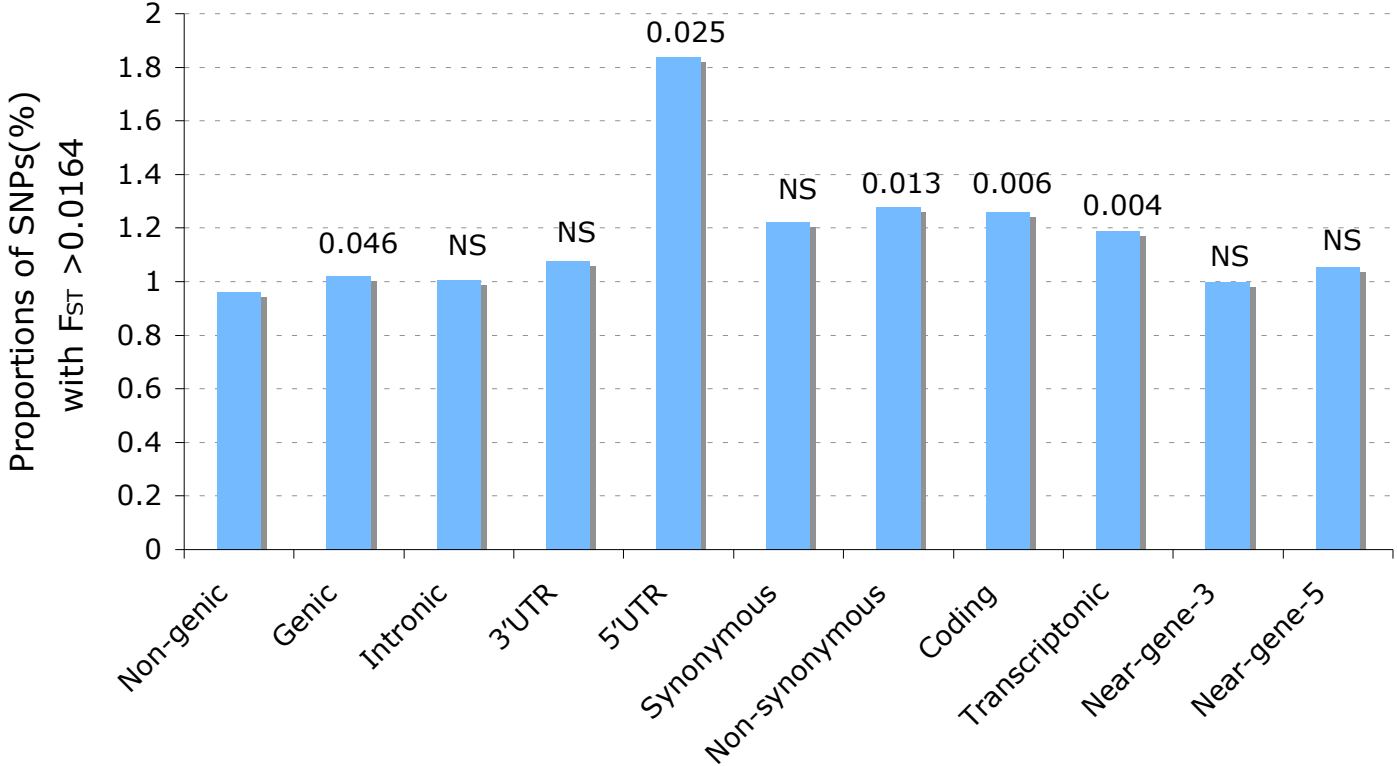

Proportion of European Ancestry
0.24
0.23
0.22
0.21
0.20
0.19
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

A



B




Proportions of SNPs(%) with FST >0.0164
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.046
NS
NS
0.025
NS
0.013
0.006
0.004
NS
NS
Non-genic
Genic
Intronic
3'UTR
5'UTR
Synonymous
Non-synonymous
Coding
Transcriptonic
Near-gene-3
Near-gene-5