Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project

Peikuan Cong
Westlake University https://orcid.org/0000-0002-4921-5657
Weiyang Bai
Westlake University
Jinchen Li
Xiangya hospital Central South University
Mengyuan Yang
Westlake University
Saber Khederzadeh
Westlake University
Sirui Gai
Westlake University
Nan Li
Westlake University
Yuheng Liu
Westlake University
Shihui Yu
KingMed Center for Clinical Laboratory Co.
Weiwei Zhao
KingMed Diagnostics
Junquan Liu
KingMed Diagnostics
Yi Sun
KingMed Diagnostics
Xiaowei Zhu
Westlake University
Pianpian Zhao
Westlake University
Jiangwei Xia
Westlake University
Penglin Guan
Westlake University
Yu Qian

Westlake University
Jianguo Tao
Westlake University
Lin Xu
Binzhou Medical University
Geng Tian
Binzhou Medical University
Pingyu wang
Binzhou Medical University
Shu-Yang Xie
Binzhou Medical University https://orcid.org/0000-0002-8090-2180
Mochang Qiu
Jiangxi Medical College
Keqi Liu
Jiangxi Medical College
Beisha Tang
Xiangya Hospital, Central South University
Hou-Feng Zheng (zhenghoufeng@westlake.edu.cn)
Westlake University https://orcid.org/0000-0001-5681-8598

Article

Keywords: genetic resources, complex traits, Westlake BioBank for Chinese pilot project, genomic analyses

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-814288/v1

Version of Record: A version of this preprint was published at Nature Communications on May 26th, 2022. See the published version at https://doi.org/10.1038/s41467-022-30526-x.

Abstract

Imbalance of genetic resources in global population restricts the understanding of complex traits across populations. Here, we initiated the Westlake BioBank for Chinese (WBBC) pilot project with 4,535 whole-genome sequencing (NGS) individuals and 5,841 high-density genotyping individuals. We identified 81.5 million SNPs and INDELs, of which 38.5% are novel. We found that 5.05% of the rare variants in WBBC were common in European population, and some trait-associated common variants in European had much lower allele frequency in Chinese. We provided a population-specific reference panel and an online imputation server (https://wbbc.westlake.edu.cn/) which could yield substantial improvement of imputation performance in Chinese population, especially for low-frequency and rare variants. By analyzing the singleton density of the WGS data, we found novel selection signatures in *SNX29*, *DNAH1* and *WDR1* genes, and the selection of the alcohol metabolism genes (*ADH1A* and *ADH1B*) strengthened from about 4,000 years ago in East Asia. Genetic evidence supported the corresponding geographical boundaries of the Qinling-Huaihe Line and Nanling Mountains, which separated the Han Chinese into subgroups, and we revealed that North Han was more homogeneous than South Han, and the history of effective population size of Lingnan began to deviate from the other three regions from 6,000 years ago. Significant selection of genes on epidermal cell differentiation was only observed in southern Chinese. Genetic divergence from north to south was more noticeable in the ancient than modern populations.

Introduction

Understanding the architecture of the human genome has been a fundamental approach to precision medicine. Over the past decade, great progress has been made to unravel either the genetic basis of complex traits and diseases[1] or the human evolutionary history[2]. The in-depth analysis of global populations with diverse ancestry could improve the understanding of the relationship between genomic variations and human diseases[3]. However, genetic studies exhibited a vast imbalance in global population, with individuals of European descent took up ~ 79% of all genome wide association study (GWAS) participants[3, 4]. Similarly, most of the whole-genome sequencing (WGS) efforts were predominantly conducted on European populations, such as Dutch[5], UK[6] and Icelandic population[7]. Even in larger genomic projects such as the Trans-Omics for Precision Medicine (TOPMed) program, which consisted of ~ 155k participants from > 80 different studies, only 9% of samples were of Asian descent[8, 9]. Therefore, large-scale genomic data are required to understand the genetic basis in Asian population. Recently, some studies have sequenced and analyzed the Asian populations including Japanese[10], Korean[11] and Chinese[12, 13]. The Singapore SG10K pilot project reported 4,810 whole-genome sequenced samples, including 903 Malays, 1,127 Indians and 2,780 Chinese[14] and the pilot study of the GenomeAsia 100K Project presented a dataset of 1,267 individuals from different countries across Asia[15].

Despite the above efforts, the Chinese population was still underrepresented in human genetic studies, which could increase the health disparities if Chinese personal genomes were underserved[16, 17, 18]. China,

as the most populated country, is a multi-ethnic nation, in which the Han Chinese accounts for 90% of the population. Generally, the entire territory of the country includes 34 administrative divisions, including provinces, municipalities and special administrative regions. Our previous study[19] demonstrated that, even with the Haplotype Reference Consortium (HRC) reference panel which contained 64,976 human haplotypes[20], the imputation of Chinese population could not reach the highest accuracy, a population specific reference panel was still needed[19]. Therefore, the genetic study of Chinese population has the potential to benefit ~ 20% of the world population, and provide a comparison to the rest of the world. Thus, we initiated the Westlake BioBank for Chinese (WBBC) project[21] to characterize the genomic variation and population structure in a large-scale cohort aiming to collect ~ 100,000 samples with deep phenotypes. Here, the genomic findings of the pilot project of the WBBC from 10,376 samples were described, covering 29 out of 34 administrative divisions of China.

Results

The WBBC Pilot Dataset and Variants Identified

The WBBC pilot project sampled 10,376 individuals from 29 of 34 administrative divisions of the People's Republic of China (Fig. 1a and Supplementary Table 1). We performed whole-genome sequencing (WGS) in 4,535 individuals on NovaSeq 6000 platform. After removing contaminated and duplicated samples, 4,480 individuals were retained for downstream analyses and statistics. The mean sequencing coverage was 13.9×, which covered 99.77% of the genome, with a range between 9.6× and 65.2× (Supplementary Fig. 1a and Supplementary Table 2). Additionally, 6,025 individuals were genotyped by high-density Illumina Asian Screening Array (ASA), including 184 individuals who were also whole-genome sequenced.

Here, we identified 81,498,995 variants after filtration from 103.96 million total raw variants (Ts/Tv = 2.15), including 74,118,191 single-nucleotide variants (SNVs) and 7,380,804 insertions and deletions (INDELs). Of these variants, which the majority of variants (44.2 million, 54.5%) were singletons (Fig. 1b), 93.3% were rare (allele frequency, AF < 0.005) and low-frequency (AF = 0.005-0.05) variants (Supplementary Table 3). Additionally, C>T, G>A, A>G and T>C constituted the most common SNV substitution types (Supplementary Fig. 1b), and the length of INDELs mainly distributed between -10bp and 10bp (Supplementary Fig. 1c). We provided a database of genetic variations for the Han population in four sub-regions (North, Central, South and Lingnan) (https://wbbc.westlake.edu.cn/genotype.html).

Comparing the variants with the WBBC and other existing databases, 45,696,726 variants were found not to present in the 1000 Genome Project (1KG)[22], gnomAD[23] and UK10K[6] (Fig. 1c). Of these, 45.6 (99.79%) million were rare variants (MAF < 0.005). We also found 31.37 (38.5%) million novel variants that were not present in dbSNP Build 151[24], including 29,015,419 SNVs and 2,353,726 INDELs. Of these novel variants, singletons accounted for 83.3%, and 99.97% of the variants (31.26 million) were rare with MAF < 0.005.

We assessed the SNV variants calling accuracy and sensitivity by comparison with SNP array data in 184 individuals from the whole genome sequencing samples (13.3× - 54.7×). Supposing the SNP array data as the true genotype, the heterozygote discordance rate of genotypes was reduced 6-fold from 0.134 to 0.022 at 13.3× sequencing depth and 4-fold from 0.004 to 0.001 at 25.3× after genotype refinement (Supplementary Fig. 2a). The non-reference genotype concordance rate extended to 99.88% at 25× with increasing sequencing depth (Supplementary Fig. 2b). The non-reference sensitivity and specificity had an effective increase after genotype refinement with BEAGLE from 0.9211 to 0.9924 and from 0.9931 to 0.9999 (Supplementary Fig. 2c, d).

Trait-associated Variants between Populations

We downloaded the public annotation data of human genome-wide association studies (GWAS) from the NHGRI-EBI GWAS Catalog database (https://www.ebi.ac.uk/gwas/) to compare the difference in the rare and low-frequency alleles in 107,124 shared variants between the Chinese (WBBC) and European (EUR). As we know, GWAS usually reported common variants, we found that some trait-associated common variants in European had much lower allele frequency in Chinese. For example, in the shared 17,781 rare variants in WBBC, 61.04% of the variants were common in EUR (Fig. 1d). Regarding to the number of trait-associated variants with different allele frequency, we found that body height, body mass index, blood protein levels, bone mineral density and educational attainment were among the top five (Supplementary Table 4 listed the top 50 phenotypes), suggesting that it should be careful when interpreting and applying the genetic data of these traits between populations. The 4 SNPs with highest difference in derived allele frequency were in the *SLC24A5* (rs1426654_A, WBBC = 0.012 and EUR = 0.997)[25], *SLC45A2* (rs16891982_G, WBBC = 0.003 and EUR = 0.938)[26], *EDAR* (rs3827760_G, WBBC = 0.926 and EUR = 0.011)[27] and *SULT1C4* genes (rs4149433_T, WBBC = 0.884 and EUR = 0.011)[28], which were associated with the traits of body mass index, eye color, hair morphology, and lobe attachment. These biological traits between the Chinese and European were remarkable, which were probably inherited from the ancient ancestors and adapted to local environments in a long period of time.

Further, we compared the difference of allele frequency between WBBC and 1000 Genome Project populations, including East Asian (EAS), South Asian (SAS), Admixed American (AMR), European (EUR) and African (AFR)[22]. A total of 19,778,701 shared variants were extracted and consequently divided into rare (MAF < 0.5%), low-frequency (0.5% ≤ MAF ≤ 5%) and common variants (MAF > 5%) (Fig. 1e and Supplementary Table 5). We found that 6.06% and 5.05% of the rare variants in the WBBC were the common variants in the AFR and EUR, respectively. In terms of the common variants in the WBBC, 6.83% and 5.79% variants were rare variants in the AFR and EUR, respectively. Further, we assessed the 2,020,520 shared variants among East Asians including Chinese (WBBC), Inner Mongolians (IMG)[29], Koreans (KOR)[11], Japanese (JPT) and Vietnamese (KHV)[22] (Supplementary Fig. 1d and Supplementary Table 6). We observed that 1.52% and 0.56 % common variants in the IMG and JPT populations were rare frequency variants in the WBBC. The low-frequency variants in the WBBC displayed diverse frequency in other East Asian population (Supplementary Fig. 1d). Relatively, the variants frequency of the KHV and

KOR populations indicated a minor difference with the WBBC population. Evidently, the East Asian population had a similar frequency spectrum in the common variants (Supplementary Fig. 1d).

Variants Annotation and Individual Genome

To characterize variants with a biological consequence, we annotated all the variants from 4,480 individuals regardless of medical conditions by ANNOVAR tools[30]. As expected, 77,862,672 (95.5%) variants were in intergenic and intronic regions (Supplementary Table 3). The variants in intergenic and intronic regions comprised 89.64% of novel variants (Supplementary Fig. 1e). In coding and splice regions, the missense accounted for 54.22% of the novel variants, while synonymous and splice variants made up 40.5% of the novel variants (Supplementary Fig. 1e). We also found that the missense, stop-gain, frameshift indels and non-frameshift indels variants were markedly increased among rare variants, compared with low frequency and common variants, which were signatures of population expansion and weak purifying selection. We predicted about 300,000 deleterious variants by SIFT, PolyPhen-2 or MutationTaster in 4,480 individuals, with 97.3% of the variants being rare alleles with MAF < 0.5% (Supplementary Table 3). Interestingly, we also identified 1,842 pathogenic or likely pathogenic variants recorded by ClinVar in our dataset regardless of medical conditions. Of these predicted disease-causing variants, 97.4% variants were rare, 1.7% variants were low frequency, and 0.9% were common variants, which arose from selection pressure subjected on these rare variants. The c.315-48 T>C in *FECH* gene were common variant (AF > 0.3) in the Chinese and Asian population, however, the AF was only 0.06 in the CEU population and gnomAD. This deep intronic variant was considered as pathogenic variant and might cause erythropoietic protoporphyria (EPP) in the European patients[31].

We selected 1,151 healthy individuals for the autosomal variants' statistic of a personal genome. On average, an individual carried 2,936,012 SNVs and 191,333 INDELs, including 8,915 missense, 10 stop loss, 70 stop gain and 126 frameshift or non-frameshift indels (Supplementary Table 7). In total, 96.5% of the variants were located in intergenic and intronic regions. For the disease-associated variants predicted in silico, about 1,623 variants were deleterious by SIFT[32], 1,714 variants were probably or possibly damaging by Polyphen2[33], and 8,591 variants were disease-causing by MutationTaster[34]. In these pathogenic and deleterious variants, we observed the higher ratio of heterozygote and non-reference homozygote (Het/Hom) (Supplementary Table 7). The proportions of Het/Hom were also very high in novel SNVs and INDELs variants, which indicated that the majority of novel variants occurred as heterozygotes in the Han Chinese population.

The homozygous pathogenic variants are responsible for recessive Mendelian disorders. To measure the prevalence of pathogenic variants in a healthy individual, we annotated the variants by ClinVar[35]. In total, we identified 757 pathogenic variants in all the healthy population, and in average each individual carried 11 pathogenic variants (het/hom = 2.08) (Supplementary Table 7 and Supplementary Table 8). Each genome carried 3.6 ± 2.1 (mean ± SD) pathogenic homozygote variants in Han Chinese population. Additionally, we found that 19 pathogenic variants existed in all four sub-regions (North, Central, South and Lingnan population) were mainly relevant to the immune system, metabolic, and hearing impairment

diseases, and 16 out of the 19 (84.2%) variants had a higher frequency (MAF > 0.01) in the Han Chinese population (Supplementary Table 8), and lower frequency (MAF < 0.001) in 1000 Genome Project (1KG)[22] and gnomAD database[23].

Whole Genome-wide Singleton Density Score Analysis and Selection Inference

The Singleton Density Score (SDS) can be applied to infer recent allele frequency changes by calculating the distance between the nearest singletons on either side of a test-SNP using whole-genome sequence data[36]. We tried to infer the recent allele frequency changes at SNVs of the Han Chinese population by calculating SDS based on 4,258,941 bi-allelic SNVs and 17,951,337 singletons from 4,334 whole-genome sequenced Han individuals. We found a novel significant selection signatures in *SNX29* gene (Fig. 2a) on chromosome 16p, which encoded the sorting nexin-29 protein and was ubiquitously expressed in the kidney, lymph node, ovary and thyroid gland tissues[37]. More than 30 SNPs on *SNX29* gene exhibited strong selection signatures ($P < 5Í10^{-8}$), which indicated significant enrichment of selection in this genomic region. Relatively higher derived allele frequency (DAF) was observed on the top SNP rs75431978 in the Han Chinese population (DAF = 0.181, $P = 5.54Í10^{-16}$) and East Asian population (DAF = 0.146, Mongolian = 0.18, Korean = 0.187, Japanese = 0.12), compared to the values obtained in 1000 Genome Project SAS (DAF = 0.062), EUR (DAF = 0.003) and AFR (DAF = 0.002) populations (Fig. 2b). SNX29 was reported to be a biomarker for vasodilator-responsive Pulmonary Arterial Hypertension[38] and major mental disorders[39]. We also identified other two novel potential selection signals *DNAH1* rs78947691 on chromosome 3 (DAF = 0.270, $P = 2.65Í10^{-8}$) and *WDR1* rs148629931 (DAF = 0.058, $P = 5.44Í10^{-8}$) on chromosome 4 (Fig. 2a). The top rs78947691 in the intron 16 of *DNAH1* gene and rs148629931 in the upstream of *WDR1* gene have relatively high DAF in the EAS population, comparing with other populations (Fig. 2b). Although these two SNPs were unreported, the polymorphisms in the *DNAH1* gene showed the potential association with male infertility in the Chinese[40, 41], while the variation in the *WDR1* gene were the risk factors for gout development in the Chinese population[42, 43].

We also confirmed several significant natural selection signals at *ADH* gene clusters (rs1229984, $P = 6.07Í10^{-16}$), the MHC region (rs9380181, $p = 6.43Í10^{-11}$), and *ALDH2* (rs671, $P = 1.68Í10^{-14}$) (Fig. 2a). These three selection signature regions have also been identified in the Japanese population[44]. The alcohol-metabolizing enzymes such as the alcohol dehydrogenase (ADH) genes, including *ADH1A*, *ADH1B*, *ADH4*, *ADH5*, *ADH6*, and the aldehyde dehydrogenase (*ALDH2*) gene, had an effective impact on the alcohol metabolism pathway and the consequent alcoholism protective effect, which strongly indicated diverse ethnic-specific alcohol consumption patterns[45, 46, 47, 48, 49]. Similarly, the high derived allele frequency (DAF) in this genomic loci, particularly rs1154413 (*ADH5*), rs4148887 (*ADH4*), rs2156733 (*ADH6*), rs3819197 (*ADH1A*), rs1229984 (*ADH1B*), and rs671 (*ALDH2*) (0.726, 0.734, 0.764, 0.790, 0710 and 0.239, respectively), illustrated corresponding alleles associated with alcoholism in the Han Chinese, when compared with other non-East Asian (Fig. 2b and Supplementary Table 9). Interestingly, we observed a higher-level DAF in these SNVs in South and Lingnan regions compared to the North and Central Han, which reflected the recent regional DAF changes and adaptation in this populous ethnicity

and articulated different drinking habits or specific-alcohol consumption. In addition, we identified other selection signals in chromosome 12, including rs11066280 (P = 1.41Í10^{-13}) in *HECTD4* gene, rs11066015 (P = 2.57Í10^{-12}) in *ACAD10* gene and rs3782886 (P = 4.11Í10^{-12}) in *BRAP* gene, which were limited to EAS ancestry populations. These three genes adjacent to the *ALDH2* gene in chromosome 12q were within a large linkage disequilibrium (LD) block[50], revealing that the region had been under positive selection for a long time.

We estimated the selection coefficient trajectories for *ADH1A* and *ADH1B* genes with a hidden Markov model[51] using the ancient allele frequencies from East Asian ancient individuals (9,500 - 300 BP) and present-day allele frequencies in the WBBC. The derived allele of the SNP rs3819197 was present around the 7,000 year ago, but was very rare for a long time (Fig. 2c). The strength of selection at the *ADH1A* (rs3819197) and *ADH1B* (rs1229984) gene increased around 4,000 years ago. The derived allele (A) of rs671 in *ALDH2* gene was a common variant (MAF = 0.174) and strongly selected in modern East Asian population, yet very rarely in non-East Asian (Fig. 2b), however, the allele A was absent in the East Asian ancient DNA data, suggesting a more recent selection.

Imputation Reference Panel for the Chinese Population

We evaluated the genotype imputation accuracy of the WBBC, 1KG (Phase 3, v5a)[22], CONVERGE[52], and two combined reference panels (WBBC+EAS and WBBC+1KG) in the Chinese population (Supplementary Fig. 3). The results showed that the WBBC panel, with almost fifteen-fold more Chinese samples than the 1KG Project, yielded substantial improvement for imputation for low-frequency and rare variants (Fig. 3a). The two combined panels, WBBC+EAS and WBBC+1KG, almost tied and possessed both the highest Rsq and number of well-imputed variant in the shared sites with a MAF range of 0.2% to 50%, followed by the WBBC, 1KG and CONVERGE (Fig. 3a). For the rare variants with MAF less than 0.2%, WBBC+EAS panel showed the best performance, and the WBBC panel performed roughly the same as the WBBC+1KG (Fig. 3a). This result indicated that merging EAS individuals of the 1KG to increase the haplotype size of the WBBC could improve panel's performance across all MAF bins, but merging the whole 1KG cannot yield more improvement than merging-EAS-only and even not equal to it when the imputed variants were quite rare. Taking all shared variants together, the WBBC+EAS yielded the most well-imputed variants in shared sites, while the CONVERGE panel imputed the least (Fig. 3b). The proportion of imputed variants with Rsq $\geq$ 0.8 for CONVERGE was the only one under 50% across five panels, even it was population-specific to Chinese (Fig. 3c), indicative of the importance of coverage sequencing depth of a reference panel.

To comprehensively evaluate the imputation accuracy for the five panels, we further calculated the non-reference (NR) genotype concordance rate between imputed and genotyped variants by chip array and WGS respectively (imputation vs. chip array and imputation vs. WGS). Two combined panels had the most promising distributions of the NR concordance rates, which were almost coincident with each other, indicating that the NR concordance rates for Chinese imputation could barely benefit from the extra non-East Asian haplotypes of the reference panel (Fig. 3d). Besides, we could know that the peaks of two combined panels in density plots were higher than other panels, indicating that the distributions of NR

concordance rates were more concentrated in the two combined panels (Fig. 3d). The performance of the WBBC panel was slightly behind the two combined panels, but was superior to the 1KG and CONVERGE (Fig. 3d). We also calculated the NR allele concordance rate between the imputed genotypes and the directly sequenced genotypes. Not surprisingly, the two combined panels performed best and were approximately coincident and very closely followed by the WBBC (Fig. 3e). This result suggested that the improvement provided by the EAS and 1KG were unremarkable. Considering all variants together, the WBBC+EAS panel showed the highest NR allele concordance rate, followed by the WBBC+1KG, WBBC, 1KG and CONVERGE (Fig. 3f).

Overall, we employed Rsq, and NR allele concordance rate for both WGS and array genotype to measure the imputation accuracy for the five panels. Our results demonstrated the superiority of the WBBC as a reference panel for Chinese population imputation. Compared to the 1KG and CONVERGE, WBBC panel greatly improved the imputation accuracy, especially for the rare and low-frequency variants. Besides, merging EAS/1KG haplotypes into the WBBC could further improve the imputation accuracy.

To facilitate genotype imputation in Chinese population, we developed an imputation server with user-friendly website interface for public use (https://imputationserver.westlake.edu.cn/). Users can register and create imputation jobs freely by uploading their bgzipped array data (VCF-formatted) to our server under a strict policy of data security. To ensure the integrity of array data for next phasing and imputation, some basic QC should be performed, such as removing mismatched SNPs, monomorphism and duplicate SNPs. The server provided a choice of four reference panels to conduct the imputation, including the WBBC, 1KG Phase3, WBBC combined with EAS, and WBBC combined 1KG Phase3. All panels in both GRCh37 and GRCh38 were built to meet different needs. Besides, service of phasing was also provided in our server for users who cannot afford the corresponding heavy computational load. An email of reminder will be sent to the user when the imputation job is finished, and then user can download the imputed genotype data and the corresponding statistics file with an encrypted link. The SHAPEIT and MINIMAC were employed in our server for phasing and imputation, respectively. More details including the policy of data security, statistics of four reference panels, and the reference manual were specified in our website.

Genetic Evidence Supported the Geographical Boundaries of the Qinling-Huaihe Line and Nanling Mountains

To explore the Chinese population structure, we performed principal component analysis (PCA) on 2,056 Han Chinese individuals and 205 minority individuals from 29 of 34 administrative divisions of China (Fig. 4a). PC1 and PC2 revealed the main genetic structure of the Chinese population, with PC1 displaying a population stratification along the north-south cline, reflecting the geographical locations (Fig. 4b). The genetic difference of the Han population corresponded to the geographical boundaries of the Qinling-Huaihe River Line and Nanling Mountains. Based on the PCA analysis and traditional geographical boundaries in China, the Han Chinese could be classified into four groups: the North Han (Gansu, Hebei, Heilongjiang, Henan, Inner Mongolia, Jilin, Liaoning, Ningxia, Qinghai, Shaanxi, Shandong, Shanxi and

Tianjin) (Supplementary Fig. 4a and Supplementary Fig. 5), the Central Han (Anhui and Jiangsu) (Supplementary Fig. 4b and Supplementary Fig. 6), where Central Han were closed to North, but embedded in both North and South Han, the South Han (Chongqing, Fujian, Guizhou, Hubei, Hunan, Jiangxi, Sichuan, Yunnan and Zhejiang) (Supplementary Fig. 4c and Supplementary Fig. 7), and the Lingnan Han (Guangxi, Guangdong and Hainan) (Supplementary Fig. 4d and Supplementary Fig. 8). PC3 and PC4 displayed no discernable geographical structure and subpopulations (Supplementary Fig. 4e). When the 104 JPT (Japanese in Tokyo, Japan) and 99 KHV (Kinh in Ho Chi Minh City, Vietnam) samples from the 1000 Genomes Projects (1KG) were included, the KHV population formed a cluster overlapping with Lingnan Han, while the JPT population was closer to the North Han Chinese (Supplementary Fig. 4f).

We estimated ancestral composition in the Han Chinese population from 27 provinces using the ADMIXTURE program. The average number of presumed ancestral populations were calculated in each province with the optimal K = 3. When the value of component 1 was sorted, the four regions were arranged from northern to southern China (Fig. 4c). The ancestry fractions of the North Han accounted for about 66% on component 3. The ancestral component of the Central Han was closer to the North Han with 52.1% on component 3, while the admixture components in the South Han were 46.3% on component 1 and 40% on component 3 respectively, which did not show the predominant ancestral components. We found a distinctly higher proportion of component 1 in Lingnan Han, at 78% of ancestry composition compared to other ancestral components. North Han, South Han, and Lingnan Han showed significantly different clusters, while central Han embodied the ancestral components of both northern and southern populations. In southern China, South Han and Lingnan Han were clearly distinguished from each other, which was consistent with the PCA results.

We combined the 1KG data (CHB, CHS, CDX, JPT and KHV) and conducted ADMIXTURE analysis from K = 2 to K = 8 to explore the admixture with Asian population in the Han Chinese and 1KG population (Supplementary Fig. 9). At K = 4, the cross-validation error was the lowest. Based on the outcome of the admixture analysis, individuals from the Guangdong province displayed a complex genetic pattern consisting of Lingnan Han and South Han. Moreover, Fujian province which was obviously consisted of a genetic mosaic of diverse people showed more genetic differentiation compared to the other southern provinces at K = 4 and K = 5. Exploration of structuring patterns also revealed that the Jiangxi and Hunan provinces shared the same components from K = 2 to K = 6 in the South region, corresponding to the adjacent geographical locations, and resulting from known historical migration events. There were no significant components difference among the Northern provinces. Most components of the CHB population were consistent with North Han, except for part of samples originating from South Han, while the CHS population mainly came from South Han. From K = 2 to K = 6, the JPT population and Han Chinese population were classified into different groups, but more close to the North Han at K = 2. KHV population clustered with Lingnan Han at K = 2, whereas from K = 3 onwards, the KHV population and Lingnan Han clustered separately.

We collected 340 published ancient genomes from 8 countries or regions from 40,000 to 300 years ago and 95 representative present-day genomes [Inner Mongolia (IMG), North, Central, South and Lingnan

Han from the WBBC, and CHB, CHS, CDX, JPT and KHV individuals from the 1KG] to reveal the population relationships between modern and ancient individuals in Asia (Fig. 4d). The PCA analysis showed that there was strong genetic divergence of ancient individuals between the Northern (Mongolia and Russia) and Southern area (Taiwan, Thailand and Vietnam). And the ancient samples spread from North to South dispersedly compared to the modern samples. The ancient individuals from North Asia (e.g., Mongolia and Russia) were closer to modern North Han than South Han, while both modern and ancient samples from the Southern area (South, Lingnan, Taiwan, Thailand and Vietnam) were closely clustered together, which was consistently fit well with geographic distribution of the populations. The 88 ancient individuals from the China mainland were mostly close to the modern North Han, and there was population stratification with the modern Southern Chinese population in the PCA analysis, which suggested the human migration and admixture in Northern and Southern China during the long population history of East Asia[53, 54].

Population Structure and Demographic History in Four Sub-regions of the Han Chinese Population

Weir-Cockerham F_{ST} is an allele frequency-based metric to measure the population differentiation due to genetic structure. We calculated pairwise F_{ST} and performed hierarchical clustering for 27 administrative divisions of China and 26 populations of the 1KG. The 27 administrative divisions were mainly clustered into three groups and showed an association with geography (Fig. 5a and Supplementary Table 10). Anhui and Jiangsu provinces, which we designated as Central region of China, were clustered with Northern provinces, indicative of a closer genetic relationship. The other two groups, South and Lingnan, aligned with the regions we designated. Besides, the hierarchical branches suggested that the population differentiation between South and North was smaller than that between the South and Lingnan (Fig. 5a), reflecting the relatively shorter genetic distance. The two most remote regions in geography, North and Lingnan, were also found to have the largest population differentiation (Fig. 5a). Not surprisingly, the pairwise F_{ST} clustering results between the WBBC and 1KG populations showed that the four designated regions were clustered into the East Asian (EAS) group (Supplementary Fig. 10 and Supplementary Table 11). In particular, North and Central were clustered with CHB, while South and Lingnan were clustered with CHS (Supplementary Fig. 10). Using the four 1KG continent-level ancestry groups (AFR, EUR, AMR and SAS) as the Non-Chinese population reference, we further investigated the geographical patterns of F_{ST} in 27 administrative divisions of China. The AFR group showed the largest F_{ST} that ranged from 0.14 to 0.15 (Supplementary Fig. 11b and Supplementary Table 12), indicative of the greatest population differentiation to the WBBC, while the SAS and AMR group yielded the least value (Supplementary Fig. 11a, c). The geographical patterns of F_{ST} across the four sub-regions were similar to each other. On average, the Han Chinese in Northern provinces had the relatively closer genetic structure to the Non-Chinese populations of the 1KG. Interestingly, Qinghai province was conspicuously highlighted in the geographic heatmaps, as its pairwise F_{ST} value was obviously smaller than that of other administrative divisions (Supplementary Fig. 11), indicative of the genetic structure particularity of the Qinghai Han Chinese.

Next, we detected the IBD segments with the logarithm of the odds (LOD) score > 3 across individuals in the WBBC[55]. Unlike the Weir-Cockerham F_{ST}, IBD analysis is a haplotype-based approach to reveal the genetic structure and investigate the common ancestry of populations. The total IBD segment counts in each pair of administrative divisions were normalized by the corresponding sample size. We then performed the hierarchical clustering based on the matrix of normalized pairwise IBD counts. Similar to the results of F_{ST} clustering, 27 administrative divisions were also mainly clustered into three groups, and individuals from Anhui and Jiangsu provinces were clustered in North (Fig. 5a, b). Besides, the results showed that most Southern provinces shared more IBD segments with Northern provinces than with Lingnan (Fig. 5b), just as observed in the Fst analysis (Fig. 5a), suggesting that the Han Chinese in South and North shared more common ancestry than South and Lingnan. The Fujian and Hunan, which we designated as the Southern province, had been found that joined up with Lingnan provinces by the multiple hierarchical branches, indicating that they were more close to Lingnan in ancestry (Fig. 5b), and contiguous to Lingnan geographically (Fig. 5c).

We inferred the history of effective population size for the Han Chinese, and the results across the four regions were shown in Fig. 5d. In the period from 1 million years ago to ~ 6 thousand years ago (kya), the Han Chinese size histories of four regions experienced almost identical dynamics. From 200 kya to ~10 kya, the effective population size experienced a steep decline and then grew rapidly, with the lowest point reached at ~60 kya, which was indicative of a bottleneck, consistent with previous demographic history studies[14, 22, 56]. Around 6 kya, the size histories of the Han Chinese from the Lingnan began to deviate from the other three regions, potentially reflecting the existence of a population substructure within the Lingnan Han Chinese (Fig. 5d).

Using the Han Chinese in the most northern province (Heilongjiang) of China as the reference, we estimated relative genetic drifts and inferred a rooted maximum likelihood tree between 27 administrative divisions by TreeMix software[57]. In the result shown in Fig. 5c, the relative drift of the provinces and municipalities were in line with the geographic location. To gain a better understanding of the result, we further drew a geographic heatmap that suggested a general genetic drift trend from the North to Lingnan, with the drift parameter increasing as the latitude decreased (Fig. 5c). To judge the confidence in the trend and tree topology, we performed ten bootstrap replicates by resampling blocks of SNPs. The trend was repeated in all replicate results (Supplementary Fig. 12). Besides, we found that the tree topology of administrative divisions in Central, South and Lingnan was stable. In the North, however, the tree topology was slightly different across the replicates, indicating that the genetic structures of the Northern administrative divisions were very similar and could not be precisely presented in the tree topology (Supplementary Fig. 12).

Enlightened by the genetic drift estimation results, we further investigated the homogeneity degree in the genetic structure of the Northern and Southern Han Chinese respectively. We performed the Wilcoxon rank-sum tests[58] for Northern and Southern administrative divisions using their respective pairwise F_{ST} values, normalized IBD segments counts and relative drift parameters. The results showed that the Han

Chinese from North had smaller population differentiation (P = 4.6e-10) and genetic drifts (P = 2.5e-11), and shared more IBD segments with each other (P = 1.9e-13) than those from South (Fig. 5e). These results suggested that the genetic structure of the Han Chinese in North was significantly homogeneous than those in South.

Signatures of Recent Positive Selection in Four Sub-regions of the Han Chinese Population

We employed the integrated haplotype score (iHS) test to identify recent natural signatures of positive selective sweeps in the North, Central South, and Lingnan Han populations[59]. The top 1% genomic regions with higher |iHS| scores were found in each population (Supplementary Table 13-16). The numbers of overlapping genomic windows of selective sweep regions across the four populations were shown in supplementary Fig. 13. Only 34 (26%) sweep regions were found in all the four populations. Most regions were shared in two or three of the four subgroups. Averagely, 23.2% of the regions were independent in North, Central and South Han. However, the Lingnan Han had distinctly excess independent sweeps (50, 38.5%), which might be inherited from separate ancestral components, consistent with the conclusion from our demographic history analysis. Importantly, we observed the *EDAR* gene in the first three sweep regions in all four subgroups. *EDAR* is genetic determinant of hair thickness and has been under the strong selection pressure in East Asian[60, 61, 62]. A large genomic region extending for at least 285 kb on the chromosome 7 in the top 10 window regions in all four subgroups contained eight contiguous robust selection genes (*EPHB6*, *KEL*, *LLCFC1*, *MTRNR2L6*, *PRSS1*, *PRSS3P2*, *TRPV5* and *TRPV6*; Supplementary Table 13-16). The signature of positive selection of this region has been described only in European-Americans and not in African-Americans population[63]. Our finding indicates that the interesting candidate region is not population-specific.

We conducted Gene Ontology (GO) and KEGG pathway analysis for candidate genes in the top 1% genomic regions with signals of recent selection by iHS (Supplementary Table 17 and Supplementary Table 18). The terms were selected according to p-value (< 0.05). The results of the GO analysis showed a significant enrichment of positively selected genes for ethanol metabolic process and ethanol oxidation in four sub-regions (Supplementary Table 17), consistent with the selective signatures by whole genome-wide singleton density score (SDS) analysis in the Han Chinese population. We also observed intriguing enrichment of keratinocyte differentiation, epidermal cell differentiation and skin development (*CTSL*, *COL7A1*, *PKP3*, *SEC24B*, *SLITRK6*, *WNT10A*, *KRT10*, *KRT12*, *KRT20* and *KRT23*) in the South and Lingnan Han, which were not present in the North and Central Han populations. The KEGG analysis found that 23 pathways were enriched in the Han population with adjusted p-values < 0.05 (Supplementary Table 18). Of these pathways, Southern individuals displayed significantly enriched terms more than the northern population. Tyrosine metabolism, retinol metabolism and fatty acid degradation were identified in four sub-regions.

Discussion

We initiated the Westlake BioBank for Chinese (WBBC) pilot project and performed the whole genome sequencing of 4,535 individuals from 29 of 34 administrative divisions of China. We provided a comprehensive map of the genomic variations for the Chinese population (https://wbbc.westlake.edu.cn/genotype.html). In addition, we found that the genetic evidence supported the geographical boundaries of the Qinling-Huaihe Line and Nanling Mountains, which separated the Chinese into four sub-regions (North, Central, South and Lingnan). The genetic architecture within North Han was more homogeneous than South Han, and the history of effective population size of Lingnan began to deviate from the other three regions from about 6,000 years ago. Furthermore, we found novel significant selection signatures around *DNAH1*, *WDR1*, and *SNX29* genes in the Han Chinese, and confirmed selection signals at alcohol metabolism genes, and the selection of *ADH1A* and *ADH1B* strengthened from about 4,000 years ago. We observed enriched positive selective sweeps of keratinocyte differentiation, epidermal cell differentiation and skin development in the South and Lingnan Han. Finally, we provided a comprehensive reference panel for genotype imputation for Chinese and Asian population, and an online imputation server (https://imputationserver.westlake.edu.cn/) is publicly available now for genotype imputation.

In our study, we identified several novel significant selection signatures in *DNAH1*, *WDR1*, and *SNX29* genes in the Han Chinese and confirmed several positive selection signatures using the singleton density-based and haplotype-based methods. The ethanol oxidation was the most significant enrichment in the Han Chinese population, which included the *ADH* gene cluster and *ALDH* gene cluster. The strength of selection in *ADH1A* and *ADH1B* gene increased from around 4,000 years ago, which coincided with the period from the late Neolithic to Bronze Age in China. The selection events might be driven by the advances of agricultural and wine production technology, which resulted in prevalence of derived alleles. The major histocompatibility complex (MHC) genes carried the significant selection signatures for adaptive autoimmune and infectious diseases in all populations shaped by natural selection over a long time[64]. The ectodysplasin A1 receptor (*EDAR*) encodes a member of the tumor necrosis factor receptor family, which was involved in the development of hair, teeth and glands[65, 66, 67]. The SNV rs3827760 (NM_022336: c.1109T > C p.Val370Val) in *EDAR* gene had been under strong natural selection in WBBC with a very high allele frequency (0.93). This non-synonymous SNV displayed a higher |iHS| score in four groups (North: 3.08, Central: 3.15, South: 3.15 and Lingnan: 3.65), which indicated a strongest signal of positive selection in this genomic region.

The genetic structure of a population defines the level and extent of genetic variation within its constituent subpopulations. Although the genetic structure of north-south differentiation in the Chinese population was consistently observed in previous studies[12, 13, 68, 69, 70], subgrouping of the administrative divisions was not always consistent. For example, Hubei province was grouped into central in Cao et al[12], while it was clustered into south in Xu et al[68]. Our finding demonstrated that the Han Chinese populations were divided into four sub-regions (North, Central, South and Lingnan), which corresponds to the geographical boundary, the Qinling-Huaihe Line and Nanling Mountains (Five Ridges). The Qinling Mountains are the east-west mountain range that stretch across the south of Gansu and Shaanxi

provinces. The ~ 1,000 kilometers long Huaihe River flows through the south of Henan province and the middle of Anhui and Jiangsu provinces. To some extent, the climate, culture, lifestyle and cuisine between the Northern and Southern regions were different. Lingnan area is the region in the south of Nanling mountains (with five ridges) and the southeast of Yunnan-Guizhou Plateau in southern China, which refers to the administrative divisions of Guandong, Guangxi, Hainan, Hong Kong and Macao[71]. Shuhua Xu et al showed that the Han Chinese was distinguished with three clusters corresponding roughly to northern Han, central Han and southern Han[68]. Notably, the administrative divisions of North Han and Central Han by Xu et al were consistent with our results, however, the southern Han would be accurately separated into South Han and Lingnan Han by the geographical barrier of the Nanling Mountains and Yunnan-Guizhou Plateau, which had been confirmed by our PCA and ADMIXTURE results. Additionally, the genetic architecture within North Han were distinctly homogeneous, while the ancestral components of admixture in South Han were more diverse. Due to the absence of Han samples in seven administrative divisions (Beijing, Shanghai, Tibet, Xinjiang, Taiwan, Hong Kong and Macao), we have not inferred the population structure in these areas. The Qinling-Huaihe line is a boundary between semi-humid warm temperate continental monsoon climate and humid subtropical monsoon climate in China. The enrichment differences of candidate genes on skin development related traits between northern and southern Han Chinese population might be the results of adaptive pressures selection, including the effects of geography, climate and human migration. The epidermis is the outermost layer of the skin, which protects the body against pathogens and ultraviolet radiation, and is under adaptive pressure from sunlight duration and intensity.

Finally, using R-square and NR-allele concordance rate metrics, we evaluated and compared the genotype imputation performance of the WBBC pilot with two existing panels, the 1KG Phase3 and CONVERGE. Besides, given that the haplotype size of a panel and the genetic background between the panel and array are two crucial factors for imputation accuracy[19, 72], we built and evaluated two more combined panels that merged the WBBC with the 1KG and EAS group by the reciprocal imputation approach[73]. The 1KG Project, which consisted of 2,504 individuals from 26 worldwide populations, is the most diverse and commonly used panel for genotype imputation due to its high quality[22]. The CONVERGE is a population-specific reference panel for Chinese imputation but with low-coverage sequencing depth[52]. In our study, the WBBC panel yielded substantial improvement for imputation accuracy for low-frequency and rare variants than these two existing panels. The WBBC + EAS and WBBC + 1KG panels performed better than WBBC panel alone, and the WBBC + EAS panel yielded highest imputation accuracy for rare variants, the most well-imputed variants and the highest proportion of well-imputed variants. This observation was consistent with and further expanded our previous finding that population-specificity between reference panel and the imputed array was reasonably rigorous for the Han Chinese genotype imputation, and the accuracy benefited from the increasing of haplotype size via extra diverse individuals was limited, especially for rare variants[19]. Here, to maximize utilization of the WBBC pilot, we provided a large population-specific Genotype Imputation Server, which included the WBBC, 1KG and the two combined reference panels for Chinese sample imputation.

In summary, we characterized large-scale genomic variations in Chinese population and provided the comprehensive genetic evidence for the geographical boundaries of the Qinling-Huaihe line and Nanling Mountains to divide the Han Chinese population into four subgroups. We elucidated the regional genetic structure and signatures of recent positive selection differences in modern and ancient individuals in East Asia. We also created a user-friendly website and high-performance genotype imputation server for East Asian samples. The online resource would practically be important for the genomic variants filtration of monogenic diseases and consequent association with complex traits in the population genetics field.

Materials And Methods

Study samples

Both Westlake University and Xiangya Hospital contributed to the sample collection. The WBBC pilot project[21] of Westlake University has enrolled 14,726 individuals with diverse traits across 29 of 34 administrative divisions in China (Provinces, Municipalities and Special Administrative Regions), following the regulations of the Human Genetic Resources Administration of China (HGRAC). The Xiangya Hospital contributed another 3,335 samples including 1,973 patients with Parkinson's disease and 1,362 health controls, these samples were included in WBBC later on. Of these samples, a total of 4,535 individuals were whole genome sequenced (WGS) and 6,025 individuals (including 184 WGS samples) were genotyped by high-density Illumina Asian Screening Array (ASA) with 750K variants (Supplementary Table 1). Hunan and Jiangxi provinces accounted for 87% of all WGS samples. Shandong (n = 2,801) and Jiangxi (n = 1,730) provinces comprised 77.6% of all ASA genotyped samples. All the participants signed the consent forms. The research program was approved by the Institutional Review Board of the Westlake University and Xiangya Hospital of Central South University.

Whole genome sequencing and variants calling

Genomic DNA was extracted from peripheral blood samples collected from all the participants using the blood DNA extraction kit (TianGen Biotech, China). Library preparation was performed by NEXTflex Rapid DNA-Seq Kit (Bioo Scientific) following the standard protocol. Whole genome sequencing was conducted on the Illumina NovaSeq 6000 system at the KingMed Diagnostics Co. Ltd. The target depth was ~13× per individual, with about 40 GB sequencing data. Variants calling were conducted on all the samples via BWA version 0.7.17[74] and GATK4 version 4.1.4.0 (Supplementary Fig. 14) [75].

The reads in each lane were aligned to the GRCh38 and GRCh37 human reference genome via the BWA mem tool to produce the SAM files, respectively. We used SAMtools (v1.7) view tool to convert SAM format files into BAM files[76] and GATK4 MergeSamFiles tool to sort and merge multiples lanes data into one bam. The MarkDuplicates was used to mark the PCR duplicates with a REMOVE_DUPLICATES parameter setting of false. BaseRecalibrator generated the recalibration table by the known sites dbSNP and 1000G VCF resources. ApplyBQSR outputted the recalibrated final BAM files for HaplotypeCaller. We first obtained the GVCF file for each sample and combined all the GVCF files into a single VCF files using

GATK4 GenomicsDBImport and GenotypeGVCFs following the suggested pipelines. VariantFiltration was used for filtering excessively heterozygous variants (ExcessHet >54.69) marked with ExcessHet. We calculated the VQSLOD value of each SNV and INDEL variant by setting the max-Gaussians value to 5 and annotating with ReadPosRankSum, MQRankSum, DP, QD, FS and SOR for SNVs, with ReadPosRankSum, DP, QD and FS for INDELs. In the ApplyVQSR filtration, 99.0 was applied for the truth sensitivity level for INDELs and 99.6 for SNPs. All the passed variants were retained for the downstream analyses.

For the X chromosome, we called the genotype in the pseudo-autosomal region (PAR) and non-pseudo-autosomal region (non-PAR), separately. We defined the parameter ploidy with 2 for females and 1 for males in non-PAR, while the parameter was 2 for all individuals in PAR. Then we merged the data together. We only called the genotypes on the Y chromosome for males as the haploid chromosome.

Sample and variant filtrations

The sex was inferred by the ratio of homozygous and heterozygous SNP variant in the chromosome X. Females have homozygote / heterozygote ratio less than 4. In comparison to self-reported sex, 36 samples were not consistent and the mismatch rate was 0.8%. The FREEMIX scores were used to estimate DNA contamination by verifyBamID version1.1.3 with –maxDepth 100 –precise –minMapQ 20 –minQ 20 –maxQ 100 with the allele frequencies inferred from our SNP genotyping array data[77]. In total, 15 samples with FREEMIX scores > 0.05 were excluded. We identified the duplicates samples by KING version 2.2.4 –duplicate with default values and by Plink version 1.9 –genome with the proportion of IBD >0.95 (PI_HAT>0.95)[78] removed 40 duplicated individuals or MZ twins[79]. Finally, 4,480 samples were retained in the final cohort.

In the raw calling set (GRCh38 build), 99,774,562 SNVs and INDELs on autosomes, 4,186,591 SNVs and INDELs on sex chromosomes were identified. Of these, 1,067,527 variants were excluded by ExcessHet. At a truth sensitivity of 99.6% for SNVs and 99% for INDELs, 11,160,919 SNVs and 2,690,120 INDELs were removed. We used the bcftools filter to remove the 2,933,200 SNVs and 2,643,970 INDELs variants closer to INDEL with SnpGap 3 and IndelGap 5. Then we filtered 12,274 INDELs with a length more than 50 bp. Lastly, we excluded 1,947,496 variants with a HWE p value < 10^{-6} by VCFtools (v 0.1.13) [80].

Variant annotations

The functional annotation of variants were performed with the ANNOVAR tool[30]. We annotated the gene name, protein change, location and function for all the variants. The pathogenic or benign of variants were annotated by SIFT[80], PolyPhen-2[33], MutationTaster[34] and ClinVar version 20200728[35].

Genotyping

The 5,841 samples of the WBBC Project and 184 individuals (13.3 ×-54.7 ×, median = 17.1 ×) sequenced by WGS were genotyped by ASA (Asian Screening Array) BeadChip designed for the East Asian

population. The genotype call rates for each sample were more than 95%. We computed the allele frequencies in the Chinese population using 5,841 samples and 484,554 SNP variants passed the filtrations (--geno 0.05 --hwe 0.000001 and --maf 0.01) by Plink version 1.9[78] and were consequently retained for further analyses.

Evaluation of genotype concordance

We applied the sequencing and genotype data from 184 individuals (13.3 × - 54.7 ×) to estimate the whole genome sequencing calling accuracy. The genotype from SNP arrays were considered as the true genotype and our calling variants were test set. We extract the SNP genotyping array sites from 184 whole genome sequencing samples. The variants with calling rates for each sample more than 95% and the frequency more than 1% were retained. Finally, 483,755 variants in autosomes detected by both WGS and SNP array were used to estimate the genotype concordance. We also conducted the LD-based genotype refinement via BEAGLE 5.1 with default settings[81]. We computed the heterozygote disconcordance, non-reference genotype concordance, specificity and non-reference sensitivity for the shared variants (Supplementary Fig. 15) [82].

Calculation of the singleton density score (SDS)

We estimated the kinship of samples by and IBDkin[83] and KING -kinship[79]. In total, 29 samples in one of pair individuals within the degree 3 of relationship or the kinship coefficient more than 0.0442 were removed. Finally, SDS analysis were conducted with 4,334 Han Chinese samples (2,405 normal samples and 1,929 Parkinson's disease samples). We extracted the bi-allelic SNVs in all autosomal variants and filtered the SNVs by Hardy-Weinberg equilibrium ($p < 1\times10^{-6}$). We downloaded the Homo sapiens ancestral annotation information from the Ensembl release-98. SNVs without defined ancestral allele were subsequently removed. Additional SNPs were excluded by MAF <5% and less than 5 individuals for each of the three genotypes. The final data set included 4,258,941 SNVs and 17,951,337 singletons for the SDS computation.

Gamma-shape was estimated with Gravel_CHB as a demographic model for each derived allele frequency (DAF) bin by 0.005 from 0.05 to 0.95. The haplotypes was set to twice the number of individuals. We excluded the centromeres and heterochromatic regions with chromosome boundaries files. The skip boundary missing singletons fraction threshold was 0.5. The raw SDS scores were computed using recommended scripts and standardized within each 1% bin of DAF for each chromosome by calculating z-scores. Two-tailed p-values were converted by whole genome-wide standardized SDS z-scores.

Reference panel construction

The multi-allelic sites were split into bi-allelic sites via the BCFtools norm tool version 1.7[84]. We filtered the variants with --max-missing 0.9 and --hwe 0.000001 using VCFtools version 0.1.13[80]. BEAGLE version 5.1 was used to perform haplotype phasing of all 4,480 samples with default settings. The chromosomes

were divided into chunks of 1Mb with 0.1Mb overlapping. We conducted the haplotype re-phasing with SHAPEIT v2 by windows 0.5, states 200 and effective-size 14,269[85]. Finally, the SHAPEIT haplotypes were converted into VCF format files.

Quality control, pre-phasing and imputation

The rigorous variant-level and sample-level quality control were then performed as following steps: we kept autosome bi-allelic SNPs and calculated genetic relationship matrix across all individuals using variants with MAF > 0.01 by GCTA v1.91[86], and removed 339 cryptically related samples with the pairwise genetic relationship coefficient > 0.025 The variants and samples with a missing call rate > 5% were excluded by Plink version 1.9[78] The variants deviating from Hardy-Weinberg equilibrium at p < 10e-6 or with MAF < 0.01 were also excluded. Finally, 5,679 individuals and 470,279 bi-allelic SNPs on autosomes passed the filters and QC. We pre-phased the array dataset by SHAPEIT setting the effective-size parameter to 14,269 as the software recommended for the Asian population[87]. Imputation was then performed with our own haplotype reference panel, which consisted of 8,978 haplotypes at 34,948,874 SNPs (no singleton), by MINIMAC v4[88]. The length of chunks was set to 20Mb with a 4Mb overlap between contiguous chunks for the imputation. We employed R-square (Rsq) to control the quality of imputed results and filtered out variants with Rsq $\leq$ 0.95.

Reference panel evaluation for imputation in the Chinese population

We evaluated the accuracy of genotype imputation for five reference panels in the Chinese population. These panels included the most widely used panel, the 1KG[22], and the largest Chinese-specific panel CONVERGE[52], and our own WBBC panel, and two combined panels that merged the WBBC datasets with the 1KG and EAS respectively. The 1KG panel was population-diverse while the CONVERGE was Chinese-specific. The sequencing depth of CONVERGE was actually low, only ~1.7×. The imputation accuracy of these panels was then compared with each other by three different metrics (Supplementary Fig. 15).

The 1KG Project reference panel (Phase 3, v5a) was downloaded from the ftp sites (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/), and the CONVERGE Project reference panel was downloaded from the European Variation Archive (http://ftp.ebi.ac.uk/pub/databases/eva/PRJNA289433/). For the 1KG, CONVERGE and WBBC reference panels, we split multi-allelic variants into multiple bi-allelic variants and removed singletons and doubletons (minor allele counts, MAC $\leq$ 2) by using BCFtools. Besides, there were 184 samples that were included in both the WGS and DNA array genotyping for the evaluation purpose. These samples were held-out from the current WBBC panel. Finally, we obtained 3,284,591 variants and 5,008 haplotypes for the 1KG, 1,115,342 variants and 23,340 haplotypes for the CONVERGE, and 2,089,508 variants and 8,592 haplotypes for the WBBC. Note that all the manipulations were conducted on chromosome 2[14, 89]. For two combined reference panels, the WBBC+1KG and WBBC+EAS, we employed the reciprocal imputation approach to implement the combination to preserve maximal variants[73]. The EAS dataset was directly extracted from the 1KG, and sites with MAC equals zero were removed subsequently. We reciprocally

conducted imputation for the WBBC/1KG and WBBC/EAS, and then respectively excluded 2,663 and 2,142 INDELs with incompatible alleles in panels that could fail the next panel-merging. BCFtools was used to finally merge the reference panels[84]. Eventually, the WBBC+1KG combined panel consisted of 13,600 haplotypes at 4,450,989 variants, with 917,784 variants shared by both panels. The WBBC+EAS combined panel consisted of 9,600 haplotypes at 2,411,382 variants, between them, 849,281 variants were shared. We extracted chromosome 2 from our QCed chip array dataset and randomly masked one fifteenth SNPs[14, 89]. A total of 5,679 individuals were included and 2,600 SNPs were masked for the next evaluation.

We transformed the format of five panels into M3VCF and performed genotype imputation by jointly using Minimac3/4[88]. The length of chunks for imputation was set to 20MB with 4MB overlapped between contiguous chunks. The accuracy of different reference panels was evaluated by three metrics. In the first one, the estimated value of the squared correlation between imputed genotypes and true, unobserved genotypes (i.e., R-square)[88], was calculated based on the imputed dosage and produced with the imputation results by Minimac4. This value was also the most commonly used metric. In this study, an imputed variant with the Rsq $\geq$ 0.8 was considered as 'well-imputed'. For the comparison purpose, we extracted 729,958 imputed variants that were shared by the five panels. The variants were then grouped into nine MAF bins (< 0.1%, 0.1%-0.2%, 0.2%-0.3%, 0.3%-0.5%, 0.5%-1%, 1%-2%, 2%-5%, 5%-20% and 20%-50%) to differentiate the detailed imputation performance for variants with different MAF, especially for low-frequency and rare variants, which are usually difficult to impute accurately. We obtained average R-square values (Rsq) from Minimac4 info files and counted the well-imputed variants in each MAF bin. The second metric was non-reference allele (NR-allele) concordance. The variants that had been masked in the beginning were imputed by different panels. We then calculated the NR-allele concordance between imputed genotypes and the original ones in chip array for each individual (Imputed vs. Array)[89]. To gain a better understanding of the distribution of the genotype concordance, we separated the NR alleles into homozygote and heterozygote. The third metric was similar to the second, but the NR-allele concordance was calculated between imputed genotypes and WGS genotypes by the samples that we hold-out (Imputed vs. WGS). The definition of concordance and corresponding formula was specified in supplementary Fig. 15[82].

Genotype imputation server

Using the WBBC Phase 1 WGS data and 1KG Phase 3 data[22], we developed a genotype imputation server for public use. We included the WBBC and 1KG reference panel in the server and re-constructed two combined panels, the WBBC+EAS and WBBC+1KG. All panels were built in both GRCh37 and GRCh38 version, and singletons were excluded. MINIMAC v3[88] was used here to build genotype data in the M3VCF format to save the computational memory. We developed the pipeline in Python and Shell, and employed MySQL for the management of data. For the VCF-formatted array data uploaded by users, validity of data would be checked first. Before the actual imputation, there were some basic filtering steps conducted by BCFtools[84], including removing all mismatched SNPs, monomorphism, and duplicate SNPs. The 1KG

was used here as the allele reference. The next phasing and imputation were performed using SHAPEIT v2 and MINIMAC v4[87, 88]. We specified a policy of data security to protect the user's data across the entire interaction process with the server. Also, we wrote a help manual and illustrated all processes of our pipeline to facilitate users. Detailed information could be found in our website (https://wbbc.westlake.edu.cn).

PCA, ADMIXTURE and effective population size inference

We removed the variants in imputed dataset by Rsq ≤ 0.95, and merged it with our sequencing dataset by GATK v4.1.4.0[90], resulting in 9,996 individuals and 2,016,533 bi-allelic SNPs. We further merged the WBBC dataset with the 1KG Project. After filtering SNPs by MAF ≤ 0.01, a total of 1,857,766 bi-allelic SNPs with 100% call rate were left for subsequent analysis. We noted that the participants of the WBBC Project mainly came from three provinces of China, including Jiangxi (23.8%), Shandong (26%) and Hunan (31.4%). To avoid the potential bias of oversampling certain provinces[91], we randomly extracted 150 samples from each of the three provinces. Finally, 2,056 Han population individuals, 205 Minority population individuals (Tujia, Zhuang, Yi and Mongolian etc), and 2,504 1KG individuals were included. We then performed PCA[92], ADMIXTURE[93] and inference of effective population size. Note that the minority population individuals were held-out from each province group.

We excluded the SNPs with HWE p value < $1x10^{-6}$, MAF < 0.05 and genotype missing > 0.05 using the Plink software[78]. Then we performed the linkage disequilibrium based SNP pruning with –indep-pairwise 50 10 0.5. The final data sets had 338,275 bi-allelic SNPs for PCA and ADMIXTURE analyses. We used the smartpca command from the software EIGENSOFT (v6.1.4)[94] and calculated the components for the first ten PCs. PC1 and PC2 were selected for the genetic diversity comparison, which were plotted by in-house R scripts.

ADMIXTURE analysis were conducted with 2,056 Han individuals, 103 CHB (Han Chinese in Beijing, China), 105 CHS (Han Chinese South, China), 93 CDX (Chinese Dai in Xishuangbanna, China), 104 JPT (Japanese in Tokyo, Japan) and 99 KHV (Kinh in Ho Chi Minh City, Vietnam) individuals from combined dataset by ADMIXTURE version 1.3.0 using default parameters[95]. To obtain the optimal K value, we analyzed the ADMIXTURE with 10 random seeds for each K ranging from 2 to 8. The default 5-fold cross-validation procedure was carried out to estimate prediction errors. The K value with the highest log-likelihood was selected as the most probable model.

For the ancient pattern projection of the East Asia, the published genome data of 340 ancient individuals (40,000 – 300 BP) from Eight countries or regions, including Russia (n = 66)[54, 96, 97, 98], Mongolia (n = 99)[54, 99, 100], China mainland (n = 88)[53, 54, 101, 102, 103], Taiwan (n = 46)[54], Japan (61)[54, 104, 105], Laos (n = 3) [105], Thailand (n = 9)[105, 106] and Vietnam (n = 19)[105, 106] were downloaded from the Allen Ancient DNA Resource https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable genotypes-present-day-and-ancient-dna-data, version 44.3). Moreover, the 95 representative whole-genome sequences of the modern individuals were selected in the WBBC and 1KG, which were comprising of the

Inner Mongolian (IMG, n = 5), North Han (n = 10), Central Han (n = 10), South Han (n = 10) and Lingnan Han (n = 10) from the WBBC, and 10 samples of each CHB, CHS, CDX, JPT and KHV from the 1KG[22]. Principal component analysis (PCA) was performed by the smartPCA program in the EIGENSOFT v.6.1.4[94].

We further estimated the history of effective population size for four regions using SMC++[56]. Using the ancestral components analyzed by ADMITURE with K = 4, we designated 10 most representative samples with the high sequence-depth as the distinguished lineage sample for each region. We followed the suggestion of SMC++ authors and masked all low-complexity regions of the genome using the 1KG Phase3 supported data[22], and kept all left bi-allelic SNPs for next analysis. For each region, we repeated SMC++ 10 times according to each distinguished lineage sample. The combined results were used to form the composite likelihood for the final estimation. The per-generation mutation rate was set at 1.25e-8 and a generation time of 29 years was used to convert coalescent scaling to calendar time[14, 56].

F_{ST} statistics, IBD analysis and genetic drift estimation

We next performed F_{ST} statistics[107], genetic drift estimation and identity-by-descent (IBD) analysis. We calculated weighted Weir-Cockerham F_{ST} estimates for each pair of the WBBC provinces and 1KG populations using VCFtools v0.1.13[80] based on 1,857,766 bi-allelic SNPs. The window size was set to 50,000 and step size to 5,000. We built F_{ST} values matrix and performed hierarchical clustering with it using complete-linkage method implemented in the *hclust* function in the pheatmap package in R.

The IBD analysis was based on haplotypes of individuals. The genome-wide IBD segments were identified for all pairwise Han Chinese from 27 administrative divisions of China using Refined IBD software[108] with default settings. We built the IBD counts matrix for each pair of administrative divisions. Given that the sample size of 27 administrative divisions were different, we normalized the total IBD counts by sample size. For the IBD segment counts within administrative divisions (for example, province 'A'), $IBD_{normalized\ counts\ of\ A} = IBD_{total\ counts\ of\ A} / comb(N_A)$, where *comb* was the combination function in math and N_A was the sample size of province 'A'. For the IBD segment counts between two administrative divisions (for example, province 'A' and 'B'), $IBD_{normalized\ counts\ of\ A\ vs.\ B} = IBD_{total\ counts\ of\ A\ vs.\ B} / N_A * N_B$, where N_A and N_B were the sample size of province 'A' and 'B' respectively. The hierarchical clustering was then performed based on the matrix by using the same method as F_{ST} clustering.

We computed relative genetic drift estimates between each province using TreeMix v1.13 with default settings on the same SNPs as the F_{ST} analysis used[57]. The genetic drift was represented by a 'drift parameter' in TreeMix, more details were described elsewhere in the study[57]. A maximum likelihood tree for the Han Chinese population from 27 administrative divisions was then plotted. Note that the Heilongjiang province, which was located in the northern most part of China, was set as the reference point. For judging the confidence in our tree topology, ten bootstrap replicates were generated by setting the -bootstrap -k flag ranging from 10 to 100 (step-size = 10) to resample blocks of contiguous SNPs for

drift parameter estimation. Plink version 1.9 was used in this part to calculate allele counts of SNPs for reformatting of input data that the software required[78].

Calculation of iHS values

To detect the genomic signatures of recent positive selection, we computed the integrated haplotype score (iHS) using the R package rehh v3.1.0[59, 109]. The data from 2,860 North, 148 Central, 5,274 South and 92 Lingnan Han Chinese individuals were extracted from the imputed and phased files. Averagely, 1,925,157 biallelic SNVs were obtained in all autosome chromosomes in four Han populations. The SNVs were further filtered by Hardy–Weinberg equilibrium (–hwe 0.000001) and minor allele frequency (–maf 0.01) using the Plink software[78]. The ancestral allele of SNVs were defined by the data downloaded from Ensembl release-98. We removed SNVs without an ancestral allele state.

In total, 1,725,164 SNVs in North population, 1,712,580 SNVs in Central population, 1,720,051 SNVs in South population and 1,685,839 SNVs in Lingnan population passed quality control and were retained for statistical analysis. We performed iHS statistics independently for the population. The absolute values of the iHS scores were taken to analyze the data. We calculated the fraction of SNVs with |iHS| > 2 in 200 kb non-overlapping genomic windows ($N_{|iHS|>2}$ / N_{total}) and excluded the regions with < 20 SNVs[110]. The genes or genomic regions were defined within 100 kb of the identified non-overlapping SNVs. We performed the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for the adaptive candidate genes using R package clusterProfiler v3.16.0[111].

Declarations

Acknowledgments

We thankfully acknowledge Kangyong Hu from the Westlake University Supercomputer Center (WLSC) for the computational supports. We would like to thank Novogene Co., Ltd for their support and assistance in the genotyping of the study samples. This study was supported by the National Natural Science Foundation of China (Grants No: 32061143019 and 81871831), and by the Westlake Biobank for Chinese (WBBC) funds from the Westlake University, and by the National Key Plan for Scientific Research and Development of China (Grants No: 2016YFC1306000).

Author contributions: H.-F.Z. conceptualized and designed the study. P.-K.C., W.-Y.B., M.-Y.Y. and S.K. conducted the data analysis. S.-H.Y., W.-W.Z. and J.-Q.L. conducted the whole sequencing experiments. B.-S.T. and J.-C.L. provided the whole sequencing data from Hunan province. X.-W.Z., P.-P.Z., J.-W.X., P.-L.G., J.-G.T., Y.Q., G.T., S.-Y.X., L.X., M.-C.Q., and K.-Q.L. contributed to the collection of study samples. Y.-H.L., W.-Y.B., P.-K.C., S.-R.G. and N.L. designed the online website resource. H.-F.Z., P.-K.C. and W.-Y.B. drafted the manuscript, H.-F.Z., B.-S.T., J.-C.L. and S.K. reviewed and edited manuscript. All authors contributed, discussed and approved manuscript.

Competing interests: S.-H.Y., W.-W.Z. and J.-Q.L. are employee of KingMed Diagnostics Co., Ltd. The other authors have no conflict of interest to declare.

Data availability: All data are available within the main text and the supplementary materials. The allele frequency of all variants and genotype imputation server are freely available via the website (https://wbbc.westlake.edu.cn). Raw sequencing data have been deposited to the CNGB Sequence Archive (CNSA) of China National GeneBank (CNGBdb) with accession number (CNP0001516) (https://db.cngb.org/cnsa/).

References

1. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* **19**, 110–124 (2018).
2. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
3. Genetics for all. *Nat Genet* **51**, 579 (2019).
4. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
5. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818–825 (2014).
6. Consortium UK, *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
7. Gudbjartsson DF, *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435–444 (2015).
8. Kowalski MH, *et al.* Use of > 100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).
9. Taliun D, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
10. Nagasaki M, *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* **6**, 8018 (2015).
11. Jeon S, *et al.* Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv* **6**, eaaz7835 (2020).
12. Cao Y, *et al.* The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res*, (2020).
13. Chiang CWK, Mangul S, Robles C, Sankararaman S. A Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol Biol Evol* **35**, 2736–2750 (2018).

14. Wu D, *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736–749 e715 (2019).
15. GenomeAsia KC. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
16. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).
17. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
18. Jones KM, *et al.* Complicated legacies: The human genome at 20. *Science* **371**, 564–569 (2021).
19. Bai WY, Zhu XW, Cong PK, Zhang XJ, Richards JB, Zheng HF. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Brief Bioinform*, (2019).
20. McCarthy S, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
21. Zhu XW, *et al.* Cohort profile: the Westlake BioBank for Chinese (WBBC) pilot project. *BMJ Open* **11**, e045564 (2021).
22. Genomes Project C, *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. Lek M, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
24. Sherry ST, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
25. Adhikari K, *et al.* A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat Commun* **10**, 358 (2019).
26. Morgan MD, *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat Commun* **9**, 5271 (2018).
27. Endo C, *et al.* Genome-wide association study in Japanese females identifies fifteen novel skin-related trait associations. *Sci Rep* **8**, 8974 (2018).
28. Adhikari K, *et al.* A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun* **6**, 7500 (2015).
29. Yoo SK, *et al.* NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med* **11**, 64 (2019).
30. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
31. Chiara M, *et al.* Targeted resequencing of FECH locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism. *Genet Med* **22**, 35–43 (2020).
32. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–1081 (2009).

33. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet Chap. 7*, Unit7 20 (2013).
34. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361–362 (2014).
35. Landrum MJ, *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
36. Field Y, *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
37. Fagerberg L, *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**, 397–406 (2014).
38. Thayer T, *et al.* Sorting Nexin 29 (SNX29) as a Novel Biomarker for Vasoresponsive Pulmonary Arterial Hypertension. *Am J Respir Crit Care Med* **201**, A4397-A4397 (2020).
39. Chen JH, *et al.* SNX29, a new susceptibility gene shared with major mental disorders in Han Chinese population. *World J Biol Psychiatry*, 1–9 (2020).
40. Yang X, *et al.* Associations between DNAH1 gene polymorphisms and male infertility: A retrospective study. *Medicine (Baltimore)* **97**, e13493 (2018).
41. Sha Y, *et al.* DNAH1 gene mutations and their potential association with dysplasia of the sperm fibrous sheath and infertility in the Han Chinese population. *Fertil Steril* **107**, 1312–1318 e1312 (2017).
42. Lan B, *et al.* WDR1 and CLNK gene polymorphisms correlate with serum glucose and high-density lipoprotein levels in Tibetan gout patients. *Rheumatol Int* **36**, 405–412 (2016).
43. Liu LJ, *et al.* Genetic variation in WDR1 is associated with gout risk and gout-related metabolic indices in the Han Chinese population. *Genet Mol Res* **15**, (2016).
44. Okada Y, *et al.* Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun* **9**, 1631 (2018).
45. Edenberg HJ. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res Health* **30**, 5–13 (2007).
46. Ehlers CL, Liang T, Gizer IR. ADH and ALDH polymorphisms and alcohol dependence in Mexican and Native Americans. *Am J Drug Alcohol Abuse* **38**, 389–394 (2012).
47. Choi IG, *et al.* Scanning of genetic effects of alcohol metabolism gene (ADH1B and ADH1C) polymorphisms on the risk of alcoholism. *Hum Mutat* **26**, 224–234 (2005).
48. Druesne-Pecollo N, *et al.* Alcohol and genetic polymorphisms: effect on risk of alcohol-related cancer. *Lancet Oncol* **10**, 173–180 (2009).
49. Bierut LJ, *et al.* ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Mol Psychiatry* **17**, 445–450 (2012).
50. Levy D, *et al.* Genome-wide association study of blood pressure and hypertension. *Nat Genet* **41**, 677–687 (2009).

51. Mathieson I, McVean G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984 (2013).
52. CONVERGE c. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
53. Yang MA, *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288 (2020).
54. Wang CC, *et al.* Genomic insights into the formation of human populations in East Asia. *Nature* **591**, 413–419 (2021).
55. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
56. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**, 303–309 (2017).
57. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).
58. Wilcoxin F. Probability tables for individual comparisons by ranking methods. *Biometrics* **3**, 119–122 (1947).
59. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).
60. Mou C, *et al.* Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Hum Mutat* **29**, 1405–1411 (2008).
61. Tan J, Yang Y, Tang K, Sabeti PC, Jin L, Wang S. The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum Genet* **132**, 1187–1191 (2013).
62. Riddell J, Basu Mallick C, Jacobs GS, Schoenebeck JJ, Headon DJ. Characterisation of a second gain of function EDAR variant, encoding EDAR380R, in East Asia. *Eur J Hum Genet*, (2020).
63. Akey JM, *et al.* Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**, e286 (2004).
64. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. *Ann Hum Genet* **65**, 1–26 (2001).
65. Schmidt-Ullrich R, Aebischer T, Hulsken J, Birchmeier W, Klemm U, Scheidereit C. Requirement of NF-kappaB/Rel for the development of hair follicles and other epidermal appendices. *Development* **128**, 3843–3853 (2001).
66. Fujimoto A, *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet* **17**, 835–843 (2008).
67. Fujimoto A, *et al.* A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Hum Genet* **124**, 179–185 (2008).
68. Xu S, *et al.* Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet* **85**, 762–774 (2009).

69. Chen J, *et al.* Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet* **85**, 775–785 (2009).
70. Liu S, *et al.* Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **175**, 347–359 e314 (2018).
71. Xie G, Lin Q, Wu Y, Hu Z. The Late Paleolithic industries of southern China (Lingnan region). *Quaternary International* **535**, 21–28 (2020).
72. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels. *Annu Rev Genomics Hum Genet* **19**, 73–96 (2018).
73. Huang J, *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015).
74. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
75. Van der Auwera GA, *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11–11 10 33 (2013).
76. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
77. Jun G, *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839–848 (2012).
78. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
79. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
80. Danecek P, *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
81. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338–348 (2018).
82. Linderman MD, *et al.* Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics* **7**, 20 (2014).
83. Zhou Y, Browning SR, Browning BL. IBDkin: fast estimation of kinship coefficients from identity by descent segments. *Bioinformatics* **36**, 4519–4520 (2020).
84. Li H, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5–6 (2013).
86. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).

87. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2011).
88. Das S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
89. Huang L, *et al.* Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**, 235–250 (2009).
90. McKenna A, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
91. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**, e1000686 (2009).
92. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
93. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
94. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).
95. Patterson N, *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
96. Sikora M, *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).
97. Sikora M, *et al.* Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **358**, 659–662 (2017).
98. Yu H, *et al.* Paleolithic to Bronze Age Siberians Reveal Connections with First Americans and across Eurasia. *Cell* **181**, 1232–1245 e1220 (2020).
99. Damgaard PB, *et al.* 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018).
100. Jeong C, *et al.* Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proc Natl Acad Sci U S A* **115**, E11248-E11255 (2018).
101. Ning C, *et al.* Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan. *Curr Biol* **29**, 2526–2532 e2524 (2019).
102. Ning C, *et al.* Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat Commun* **11**, 2700 (2020).
103. Yang MA, *et al.* 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. *Curr Biol* **27**, 3202–3208 e3209 (2017).
104. Kanzawa-Kiriyama H, *et al.* A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. *J Hum Genet* **62**, 213–221 (2017).
105. McColl H, *et al.* The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).

106. Lipson M, *et al.* Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92–95 (2018).
107. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).
108. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
109. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour* **17**, 78–90 (2017).
110. Pickrell JK, *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* **19**, 826–837 (2009).
111. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

Figures

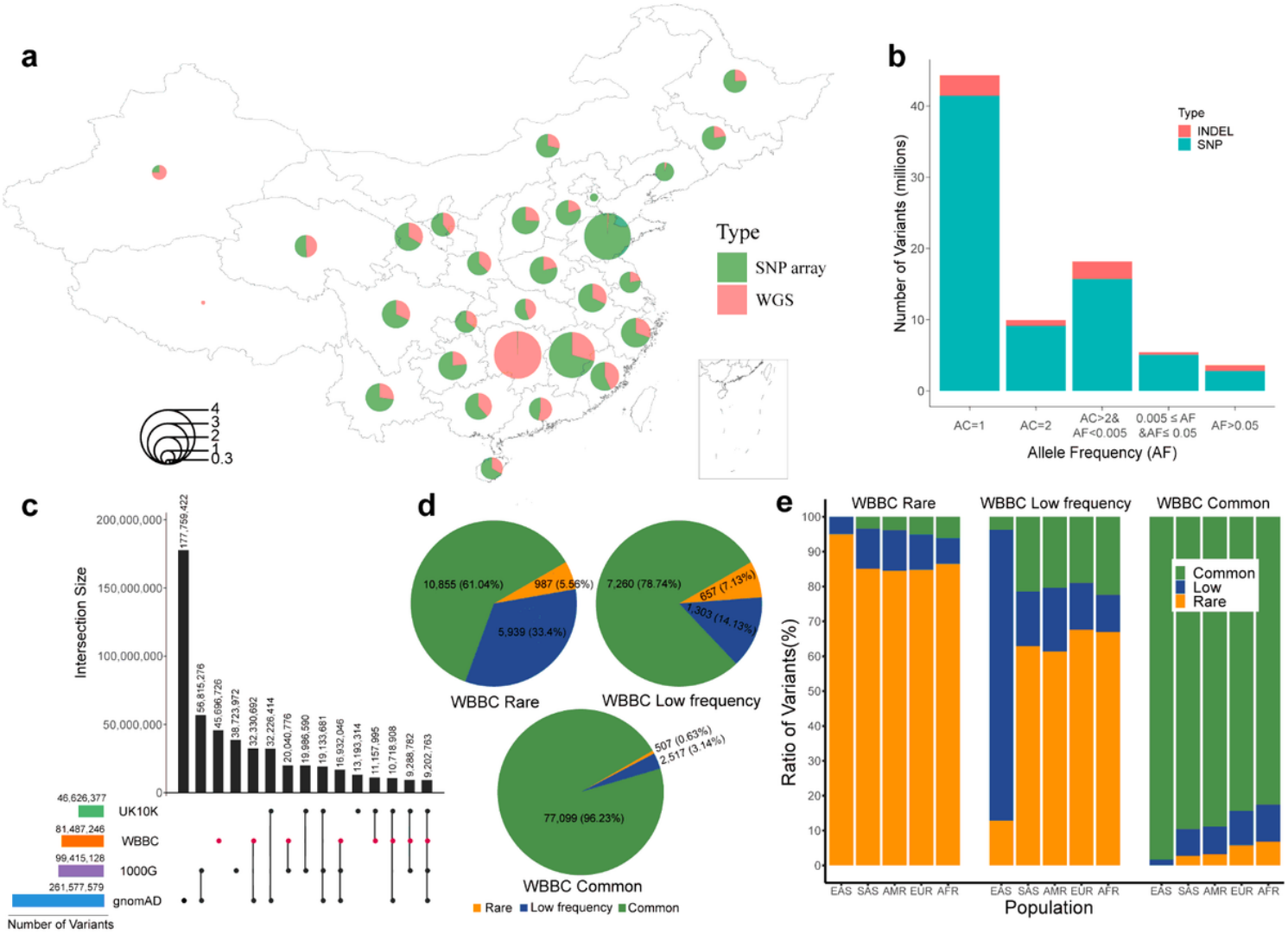

Figure 1

The statistics of samples and variants in the WBBC. a Sample distribution and statistics by geography. The proportion of samples sequenced by whole-genome sequencing (WGS) and those genotyped by high-density Illumina Asian Screening Array (ASA) were marked in red and green, respectively. b The number of SNV and INDEL variants identified in the WBBC cohort in six frequency bins: AC = 1, AC = 2, AC > 2 & AF < 0.005, 0.005 ≤ AF ≤ 0.05, and AF > 0.05. c The number of variants in 22 autosomes and X chromosome in the WBBC, 1000 Genome Project (1000G), gnomAD, and UK10K datasets. The horizontal bar plot shows the total number of variants in each of the four datasets. The individual dots and connected dots indicate each dataset and a combination of two or more datasets, respectively. Each vertical bar represents the number of variants in each dataset or overlapping variants in those datasets. d Comparison of allele frequency in shared variants between WBBC and EUR populations. e The difference of allele frequency between WBBC and 1000 Genome Project populations in rare (MAF < 0.5%), low-frequency (0.5% ≤ MAF ≤ 5%) and common variants (MAF > 5%).

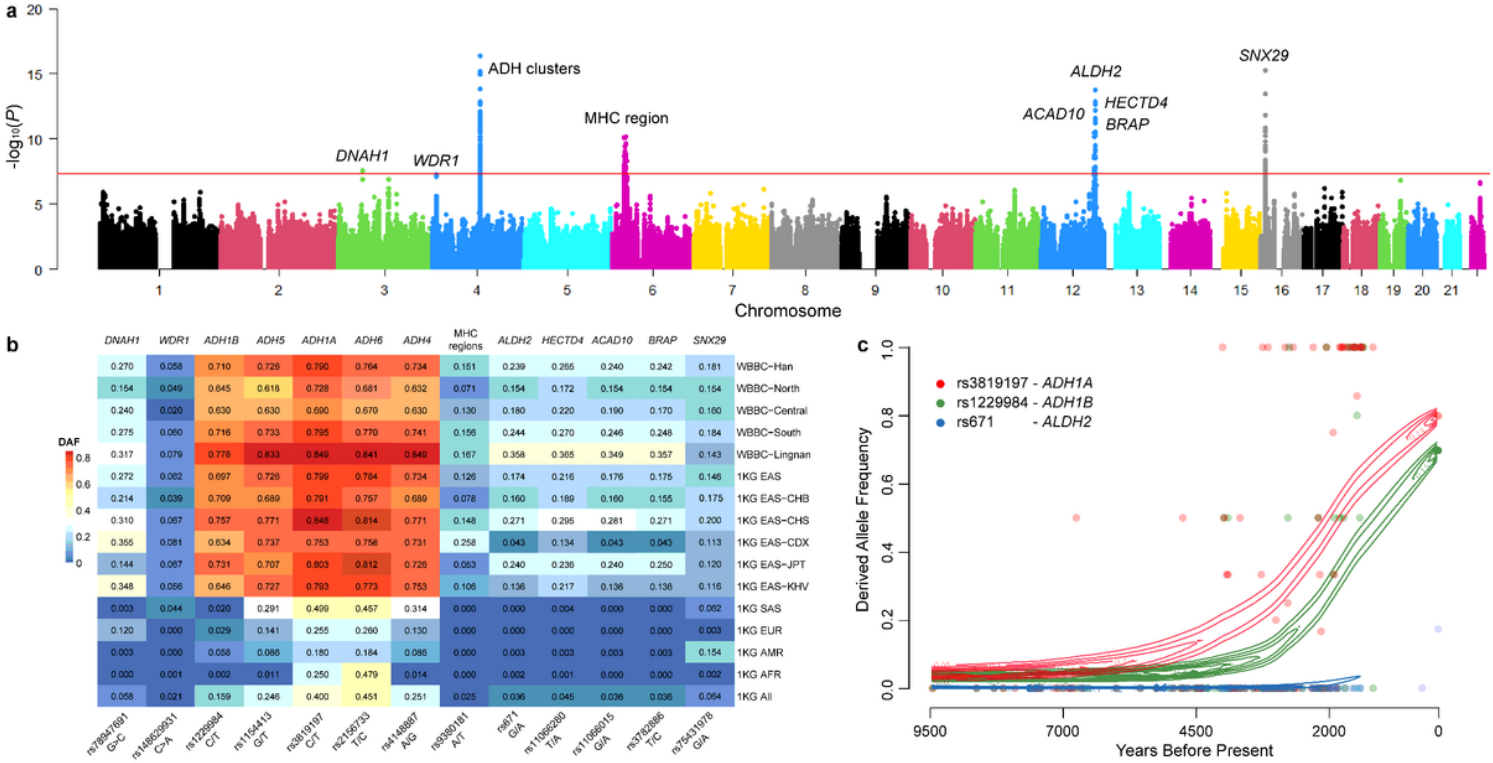

Figure 2

Whole-genome-wide recent selection signatures of the Han Chinese population by singleton density score (SDS) analysis. a Manhattan plot of the natural selection signatures from the WGS data of the Han Chinese individuals. The y-axis represents the -log10 (P) of the two-tailed p-values for standardized SDS z-scores. The horizontal red line indicates the significance threshold (p < 5 × 10-8). b The derived allele frequency (DAF) of SNVs with significant selection signatures for different populations. The WBBC-Han is all the Han Chinese individuals sequenced by whole-genome sequencing (WGS) in the WBBC cohort. North, Central, South, and Lingnan are the four Han subgroups. EAS, SAS, EUR, AMR and AFR come from

the 1000 Genome Project (1KG). c The inferred allele frequency trajectory for the derived alleles at rs3819197, rs1229984 and rs671 over the past 9,500 years from the ancient individuals of East Asia. The dot indicates the allele frequency in each generation (25 years/generation).

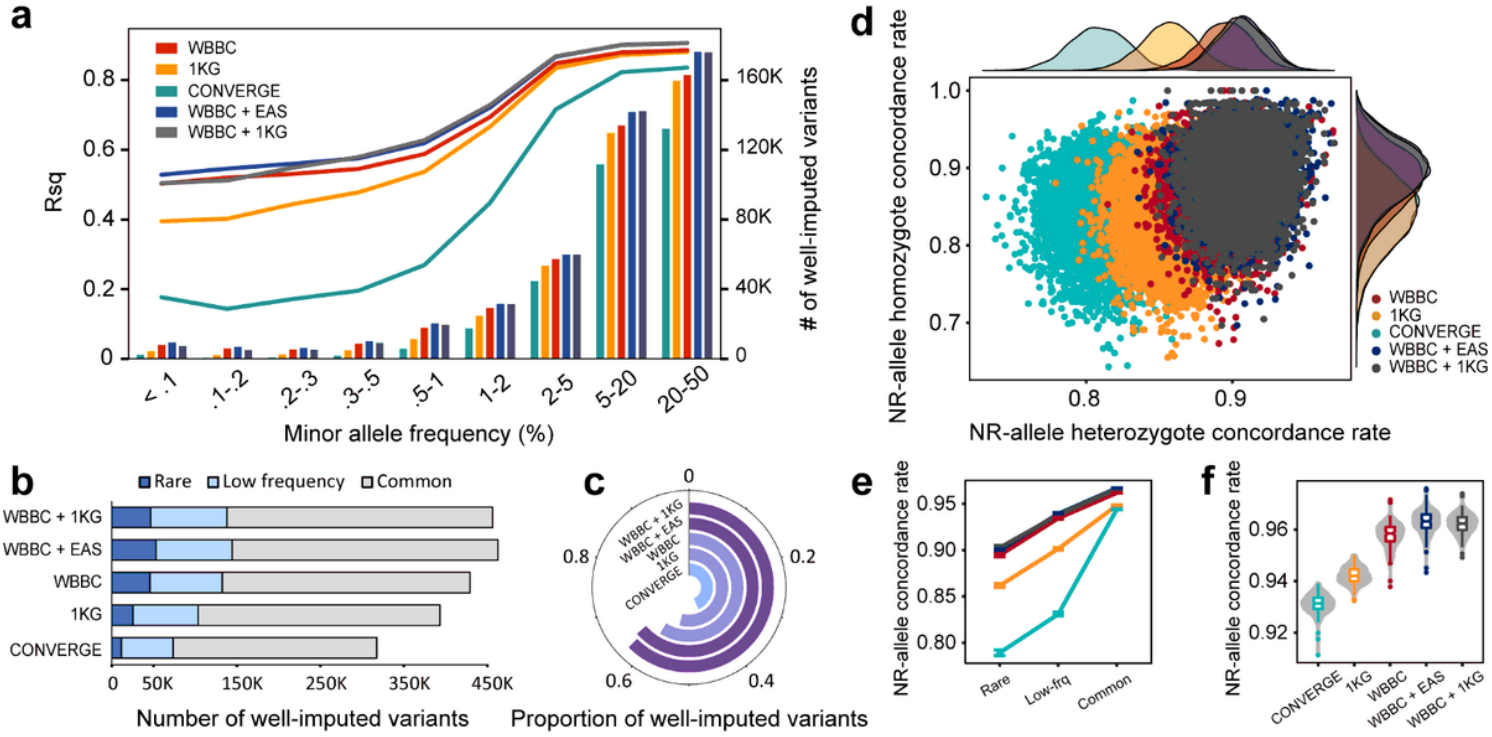

Figure 3

Imputation performance of five reference panels in the Han Chinese. a The average R-square (Rsq) and number of well-imputed (Rsq ≥ 0.8) variants in shared sites of five reference panels (729,958 SNPs). All shared variants were grouped into nine MAF bins. b and c The cumulative number and proportion of well-imputed variants in shared sites of five panels, there were 729,958 shared SNPs in total. d Non-reference allele (NR-allele) concordance rate distribution (imputed variants vs. array variants). Each dot represents an individual. The plots on the top and right are the corresponding density distributions. e and f The NR-allele genotype concordance rate for rare, low-frequency, and common variants and overall variants (imputed variants vs. WGS variants). The 1KG means 1000G Phase3 and EAS means East Asian group in 1000G Phase 3. All imputations were conducted on chromosome 2.

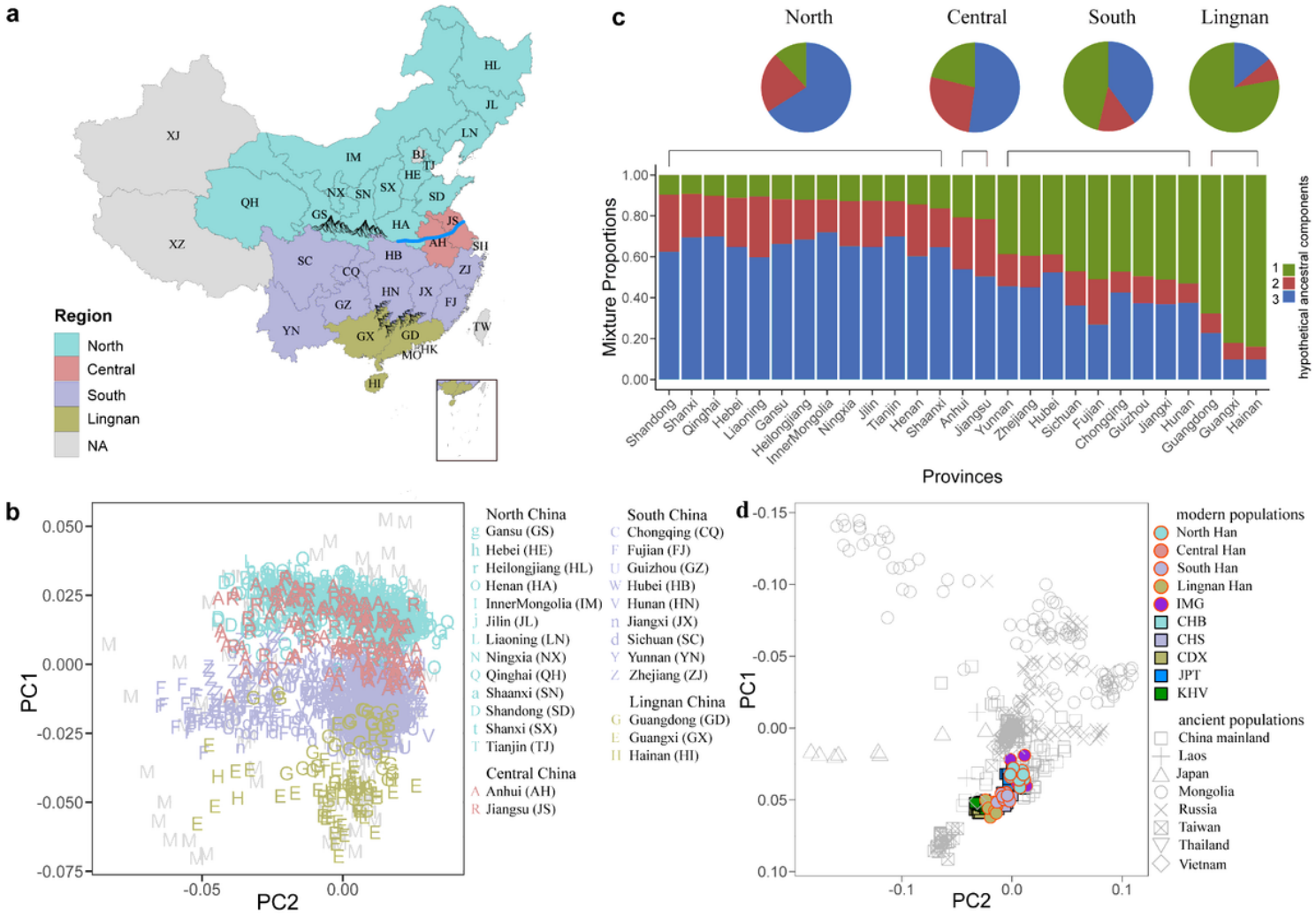

Figure 4

PCA and ADMIXTURE analysis of the Han Chinese populations and East Asian. a A map of the People's Republic of China showing its 34 administrative divisions. "NA" indicates that the Han Chinese samples were not recruited from that region. The Qinling-Huaihe River line lies in central China, while the Nanling Mountains are in southern China. b Principal Component Analysis (PCA) of the Han and Minority Chinese individuals from four sub-regions. The administrative divisions are shown by the distinct letters. Minority individuals are marked with "M". The Han Chinese populations can be classified into four subgroups: North Han (cyan color), Central Han (dark-red color), South Han (purple color), and Lingnan Han (golden color). c ADMIXTURE analysis of 2,056 Han Chinese individuals from 27 administrative divisions for the optimal K value = 3. Each vertical bar represents the average proportion of ancestral components in the regions. The length of each color indicates the percentage of inferred ancestry components from ancestral populations. The upper pie charts denote the average proportion of components across individuals from the four subgroups. d Plots of the first two principal components for modern and ancient East Asian individuals.

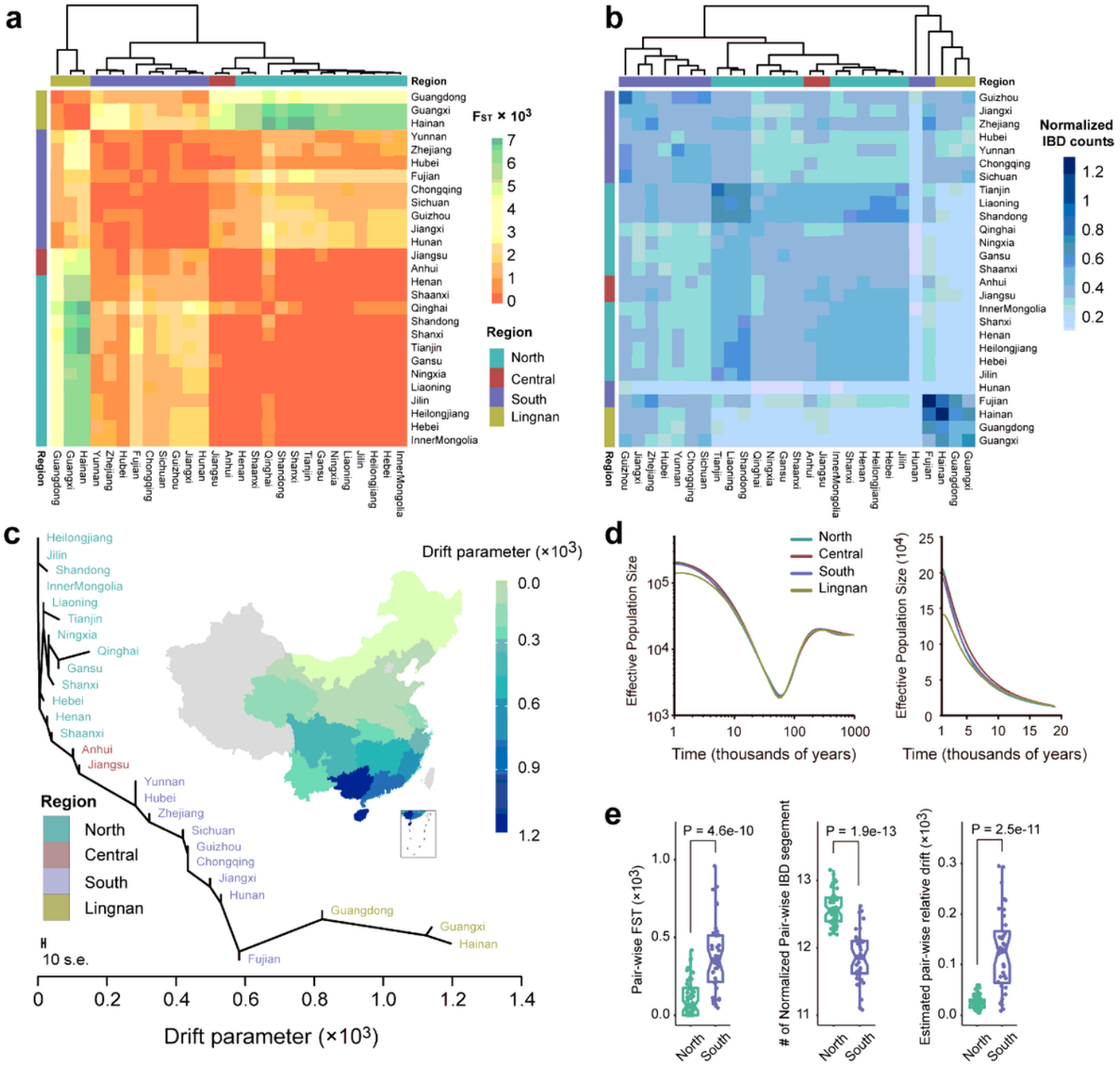

Figure 5

FST, IBD, genetic drift, and effective population size of the Han Chinese populations. a A heatmap of pairwise FST between any two of the 27 administrative divisions in China. The bars on the top and left show the classification of administrative divisions in the four regions. b A heatmap of pairwise IBD segments count between administrative divisions in China. The number of IBD segments is normalized by the sample size of each province. c A maximum likelihood tree of the Han Chinese in 27 administrative divisions. The plot is rooted in the northernmost province, and the x-axis represents estimated genetic drift. All administrative divisions in the tree are colored by different regions. d Dynamics of effective population sizes of the Han Chinese in four regions. The x-axis means the thousands of years before

present. The left panel shows the results on a log-log scale from 1 million to 1,000 years ago and the right panel shows the results on a linear scale over the past 20,000 years. e Wilcoxon rank-sum test results for the FST (left panel), normalized IBD segments (middle panel), and relative genetic drift (right panel) between pairwise Northern provinces and pairwise Southern provinces.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryinformationFigures.docx
- supplementaryTables.xlsx