Genomic analyses of 10,376 individuals provides comprehensive map of genetic variations, structure and reference haplotypes for Chinese population

Houfeng Zheng (✉ zhenghoufeng@westlake.edu.cn)
Westlake University

Peikuan Cong
Westlake University

Weiyang Bai
Westlake University and Westlake Institute for Advanced Study

Jinchen Li
Central South University

Nan Li
Westlake University

Sirui Gai
Westlake University

Saber Khederzadeh
Westlake University

Yuheng Liu
Westlake University

Mochang Qiu
Jiangxi Medical College

Xiaowei Zhu
Westlake University

Pianpian Zhao
Westlake University

Jiangwei Xia
Westlake University

Shihui Yu
KingMed Diagnostics

Weiwei Zhao
KingMed Diagnostics

Junquan Liu
KingMed Diagnostics

Penglin Guan

Westlake University
Yu Qian
Westlake University
Jianguo Tao
Westlake University
Mengyuan Yang
Westlake University
Geng Tian
Binzhou Medical University
Shuyang Xie
Binzhou Medical University
Keqi Liu
Jiangxi Medical College
Beisha Tang
Central South University

Biological Sciences - Article

Keywords: population genetics, rare variants

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-184446/v1

Abstract

Here, we initiated the Westlake BioBank for Chinese (WBBC) pilot project with 4,535 whole-genome sequencing individuals and 5,481 high-density genotyping individuals. We identified 80.99 million SNPs and INDELs, of which 38.6% are novel. The genetic evidence of Chinese population structure supported the corresponding geographical boundaries of the Qinling-Huaihe Line and Nanling Mountains. 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 6 thousand years ago. In addition, we identified a novel locus (*SNX29*) under selection pressure and confirmed several loci associated with alcohol metabolism and histocompatibility systems. We observed significant selection of genes on epidermal cell differentiation and skin development only in southern Chinese. Finally, the WBBC haplotype panel, which is a population-specific reference panel, yielded substantial improvement of imputation performance in Chinese population for low-frequency and rare variants compared to 1KG Project, and merging EAS individuals to increase the haplotype size of WBBC could improve the performance across all MAF bins. We provided an online imputation server (https://wbbc.westlake.edu.cn/) which could result in higher imputation accuracy compared to the existing panels, especially for lower frequency variants.

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]. 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[9] and Korean[10]. The Singapore SG10K pilot project reported 4,810 whole-genome sequenced samples, including 903 Malays, 1,127 Indians and 2,780 Chinese[11], and the pilot study of the GenomeAsia 100K Project presented a dataset of 1,267 individuals from different countries across Asia[12].

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 (34 administrative divisions, including provinces, municipalities and special administrative regions) could be divided into Northern and Southern area by the geographical barrier of the Qinling-Huaihe Line[13]. 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 differed. 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[14]. Although the genetic structure of north-south differentiation in the Chinese population was consistently observed in previous studies[15-19], clear subgrouping of the population was not always consistent. For example, Cao et al distinguished the Han Chinese into 7 population subgroups[15], while Xu et al clustered the Han Chinese into 3 sub-areas[16].

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[4,20,21]. In addition, our previous study[22] demonstrated that, even with the Haplotype Reference Consortium (HRC) reference panel which contained 64,976 human haplotypes[23], the imputation of Chinese population could not reach the highest accuracy, a population specific reference panel was still needed[22]. 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[24] to characterize the genomic variation and population structure in a large-scale cohort aiming to collect ~100,000 samples with deep phenotypes. Here, the findings of the pilot project of the WBBC from 10,376 samples were described, covering 29 out of 34 administrative divisions of China.

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. Hunan (n = 3,203) and Jiangxi (n = 719) provinces accounted for 87% of all WGS samples. After removing contaminated and duplicated samples, 4,489 unrelated 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 × (Extended Data Fig.1a and Supplementary Table 2). Additionally, 5,841 individuals were genotyped by high-density Illumina Asian Screening Array (ASA) with 73.9M variants. Shandong (n = 2,801) and Jiangxi (n = 1,730) provinces comprised 77.6% of all ASA genotyped samples. In total, we identified 80,993,588 variants after filtration from 104.2 million total raw variants (Ts/Tv = 2.15), including 73,942,614 single-nucleotide variants (SNVs) and 7,050,974 insertions and deletions (INDELs) (Supplementary Information). Of these, 93.3% comprised rare (allele frequency, AF < 0.5%) and low-frequency (AF = 0.5-5%) variants, with the majority of variants being singletons (44.2 million, 54.5%, Fig.1b and Supplementary Table 3). We also 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).

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 genotype from SNP arrays as the reference allele, the heterozygote discordance rate of genotypes was reduced 6-fold from 0.134 to 0.022 at 13.1 × sequencing depth and 4-fold from 0.004 to 0.001 at 25.3 × after genotype refinement (Fig.1c). The non-reference genotype concordance rate extended to 99.88% at 25 × with increasing sequencing depth (Extended Data Fig.2a). 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, whereas it had inconspicuous improvement on homozygote genotype concordance (Extended Data Fig.2b-d).

Novel Variants and Functional Annotation

Comparing the variants with the WBBC and other existing databases, 45,894,245 variants were found not to present in the 1000 Genome Project (1KG)[25], gnomAD[26] and UK10K[6] (Fig.1d). Of these, 45.84 (99.878%) million were rare variants (MAF < 0.005), 45,093 (0.098%) were low frequency variants (0.005 ≤ MAF ≤ 0.05) and 10,836 (0.024%) were common variants (MAF > 0.05). We found 31.27 (38.6%) million novel variants that were not present in dbSNP Build 151[27], including 28,969,267 (92.6%) SNVs and 2,301,842 (7.4%) INDELs (Extended Data Fig.1b). Of these variants, singletons accounted for 83.3%, and 99.95% of the variants (31.26 million) were rare with MAF < 0.005.

To characterize variants with a biological consequence, we annotated all the variants by ANNOVAR tools. As expected, 77,377,200 (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 (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 (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 (Extended Data Fig.3). We predicted about 300,000 deleterious variants by SIFT, PolyPhen-2 or MutationTaster in 4,489 individuals, with the majority of 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. 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. In addition, 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-frameshit indels (Supplementary Table 4 and Supplementary Information). Each genome carried 3.6 ± 2.1 (mean ± SD) pathogenic homozygote variants in Han Chinese population.

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.2a). PC1 and PC2 revealed the main genetic structure of the Chinese population, with PC1 displaying a significant population stratification along the north-south cline, reflecting the geographical locations (Fig.2b). 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, the Han Chinese could be classified into four clusters: North Han (Gansu, Hebei, Heilongjiang, Henan, Inner Mongolia, Jilin, Liaoning, Ningxia, Qinghai, Shaanxi, Shandong, Shanxi and Tianjin) (Fig.2b and Supplementary Information Figure S4), Central Han (Anhui and Jiangsu) (Fig.2b and Supplementary Information Figure S5), where Central Han were closed to North, but embedded in both North and South Han, South Han (Chongqing, Fujian, Guizhou, Hubei, Hunan, Jiangxi, Sichuan, Yunnan and Zhejiang) (Fig.2b and Supplementary Information Figure S6), and Lingnan Han (Guangxi, Guangdong and Hainan) (Fig.2b and Supplementary Information Figure S7).

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.2c). 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 (Extended Data Fig.4 and Supplementary Information). In southern China, South Han and Lingnan Han were clearly distinguished from each other, which was consistent with the PCA results.

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

We calculated pairwise F_{ST} and performed hierarchical clustering for 27 administrative divisions of China and 26 populations of the 1KG (Supplementary Information). The 27 administrative divisions were mainly clustered into three groups and showed an association with geography (Fig.3a and Supplementary Table 6). 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.3a), 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.3a).

Next, we detected the IBD segments with the logarithm of the odds (LOD) score > 3 across individuals in the WBBC[28]. 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.3a and Fig.3b). Besides, the results showed that most Southern provinces shared more IBD segments with Northern provinces than with Lingnan (Fig.3b), just as observed in the Fst analysis (Fig.3a), suggesting that the Han Chinese in South and North shared more common ancestry than South and Lingnan.

We inferred the history of effective population size for the Han Chinese, and the results across the four regions were shown in Fig.3d. 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[11,25,29]. 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.3d).

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[30]. In the result shown in Fig.3c, 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.3c). 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 (Extended Data Fig.7). 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 (Extended Data Fig.7).

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[31] 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*-value = 4.6e-10) and genetic drifts (*p*-value = 2.5e-11), and shared more IBD segments with each other (*p*-value = 1.9e-13) than those from South (Fig.3e). 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

We inferred recent allele frequency changes at SNVs of the Han Chinese population by calculating singleton density score (SDS) using the WGS data. In total, 4,259,171 bi-allelic SNVs and 17,943,790 singletons from 4,395 Han individuals were conducted for the SDS computation. On chromosome 16p, we found novel significant selection signatures in *SNX29* gene (Fig.3f), which encoded the sorting nexin-29 protein and was ubiquitously expressed in the kidney, lymph node, ovary and thyroid gland tissues[32]. In *SNX29* gene, more than 30 SNPs showed strong selection signatures ($p < 5Í10^{-8}$), which indicated significant enrichment of selection in this genomic region. Relatively higher DAF was observed on the top SNP rs75431978 (DAF = 0.176, $p = 1.31Í10^{-15}$) in the Han Chinese population, compared to the values obtained in 1000 Genome Project EUR (DAF = 0.003) and AFR (DAF = 0.002) populations. SNX29 was reported to be a biomarker for vasodilator-responsive Pulmonary Arterial Hypertension[33]. Although the function of the *SNX29* gene remained unknown, it could be considered a biological target of nature selection pressure in the Han population. In addition, we also confirmed several significant natural selection signals at ADH gene clusters (rs1229984, $p = 5.51Í10^{-16}$), the MHC region (rs9380181, $p = 2.04Í10^{-10}$), and BRAP-ALDH2 (rs3782886, $p = 4.29Í10^{-12}$) (Fig.3f, Extended Data Fig.8a, Supplementary Table 9 and Supplementary Information).

We employed the iHS test to identify recent natural signatures of positive selective sweeps in the North, Central South, and Lingnan Han populations[34]. In total, 130 genomic regions with higher |iHS| scores were found in each population (Supplementary Table 10-13). The numbers of overlapping genomic windows of selective sweep regions across the four populations were shown in Extended Data Fig.8b. 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 found the *EDAR* gene in the first three sweep regions in all four subgroups, which have showed the strong signatures of positive selection in East Asians[35-37]. We conducted Gene Ontology (GO) and KEGG pathway analysis for candidate genes in the top 1% genomic regions with signals of recent selection. We observed intriguing enrichment of keratinocyte differentiation, epidermal cell differentiation and skin development in the South and Lingnan Han, which were not present in the North and Central Han populations (Supplementary Information Table S1).

Imputation in the Chinese Population

We evaluated the genotype imputation accuracy of the WBBC, 1KG (Phase 3, v5a)[25], CONVERGE[38], and two combined reference panels (WBBC+EAS and WBBC+1KG) in the Chinese population (Extended Data Fig.9). 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.4a). The two combined panels, WBBC+EAS and WBBC+1KG, almost tied and possessed both the highest Rsq and well-imputed variant counts for variants with a MAF range of 0.2% to 50%, followed by the WBBC, 1KG and CONVERGE (Fig.4a). 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.4a). 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, while the CONVERGE panel imputed the least (Fig.4b). 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.4c), 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 population-diverse haplotypes of the reference panel (Fig.4d). 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.4d). The performance of the WBBC panel was slightly behind the two combined panels, but was superior to the 1KG and CONVERGE (Fig.4d). 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.4e). 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.4f).

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, we found that merging EAS haplotypes into the WBBC could improve the imputation accuracy, while the extra diverse haplotypes of the 1KG could barely contribute to it.

The WBBC Genotype Imputation Server

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 v2 and MINIMAC v4 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.

Discussion

We initiated the Westlake BioBank for Chinese (WBBC) pilot project and performed the whole genome sequencing at 13.9 × coverage of 4,535 individuals from 29 of 34 administrative divisions of China. We described a comprehensive map of the whole genomic variation in the Chinese population (https://wbbc.westlake.edu.cn) and identified 31.27 (38.6%) million novel variants. Together with 5,841 individuals genotyped by high-density Illumina Asian Screening Array (ASA), we have investigated into the structure of Chinese population, and 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 Han, Central Han, South Han and Lingnan Han). The genetic architecture within North Han was more homogeneous than South Han. We found novel significant selection signatures around *SNX29* gene in the Han Chinese, and confirmed several significant natural selection signals at ADH gene clusters and MHC region. We observed enriched positive selective sweeps of keratinocyte differentiation, epidermal cell differentiation and skin development in the South and Lingnan Han. 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.

The genetic structure of a population defines the level and extent of genetic variation within its constituent subpopulations. 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). Our data did not support the classification of seven subgroups in the Han Chinese as reported previously[15]. Shuhua Xu et al showed that the Han Chinese was distinguished with three clusters corresponding roughly to northern Han, central Han and southern Han[16]. 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.

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. The Qinling-Huaihe line is a geographical dividing line between northern China and southern China, which is the boundary between semi-humid warm temperate continental monsoon climate and humid subtropical monsoon climate in China. Most areas of northern China are dry and cold in winter, whereas there are mild in winter and hot and muggy in summer in southern provinces. 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.

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[22,39], we built and evaluated two more combined panels that merged the WBBC with the 1KG and EAS group by the reciprocal imputation approach[40]. 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[25]. The CONVERGE is the largest and population-specific reference panel for Chinese imputation so far. The quality of variants, however, is not very reliable because of the low-coverage sequencing depth[38]. 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[22]. 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. Our finding 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 among the Han Chinese ethnic. We also created a user-friendly website and high-performance genotype imputation server for 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.

Methods

Samples

The WBBC pilot 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). A total of 4,535 individuals were whole genome sequenced and 5,841 individuals were genotyped by high-density Illumina Asian Screening Array (ASA) with 73.9 million variants (Supplementary Table 1). All the participants signed the consent forms. The research program was approved by the Institutional Review Board of the Westlake 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). We performed the whole genome sequencing on Illumina NovaSeq 6000 system (150 bp paired-end reads) following the standard Illumina library construction and instructions 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[41] and GATK4 version 4.1.4.0[42] (Supplementary Information).

Sample filtrations

The sex estimation and confirmation were analyzed by the ratio of sequencing depths aligned to the X chromosome and autosomes[11]. The inferred sex was consistent with the self-reported sex for each sample. The FREEMIX scores were used to estimate DNA contamination by verifyBamID version1.1.3 with –maxDepth 100 –precise –minMapQ 20 –minQ 20 –maxQ 100 and the allele frequencies inferred from our genotyped data[43]. In total, 15 samples with FREEMIX scores > 0.05 were excluded. We identified the duplicates samples by KING version 2.2.4 –duplicate and removed 31 duplicated individuals or MZ twins[44]. Finally, 4,489 samples were retained in the final cohort.

Variant annotations

The functional annotation of variants were performed with the ANNOVAR tool[45]. We annotated the gene name, protein change, location and function for all the variants. The pathogenic or benign of variants were annotated by SIFT[46], PolyPhen-2[47], MutationTaster[48] and ClinVar version 20200728[49].

Genotyping

The 5,841 samples of the WBBC Project and 184 individuals (13.3×-54.7×) sequenced by WGS were genotyped by ASA-750K (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[50] 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 reference allele and our calling variants were test set. After filtration, about 0.5 million common 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 for the low confidence genotypes and missing sites via BEAGLE 5.1 with default settings[51]. We computed the heterozygote disconcordance, non-reference genotype concordance, homozygote genotype concordance, specificity and non-reference sensitivity for the shared variants (Extended Data Fig.10)[52].

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[53], 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[54], we randomly extracted 150 samples from each of the three provinces. Finally, 2,056 Han population individuals, 205 Minority population individuals, and 2,504 1KG individuals were included. We then performed PCA[55], ADMIXTURE[56] 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[50]. 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)[57] 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 by ADMIXTURE version 1.3.0 using default parameters[58]. 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. We further estimated the history of effective population size for four regions using SMC++[29]. 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[25], 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[11,29].

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

We next performed F_{ST} statistics[59], 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[46] 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[60] 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[30]. The genetic drift was represented by a 'drift parameter' in TreeMix, more details were described elsewhere in the study[30]. 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[50].

Calculation of the singleton density score

The Singleton Density Score (SDS) can be applied to infer recent allele frequency changes in the past 2,000-3,000 years by calculating the distance between the nearest singletons on either side of a test-SNP using whole-genome sequence data[61]. In our data set, 73.91 million autosomal variants were identified in 4,395 Han Chinese samples, of which 68,492,157 were bi-allelic SNVs in all individuals. We 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 4,259,171 SNVs and 17,943,790 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 8,790, 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.

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[34,62]. 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. In total, 1,967,791, 1,897,093, 1,981,861 and 1,853,882 biallelic SNVs were obtained in all autosome chromosomes in four Han populations respectively. The SNVs were further filtered by Hardy–Weinberg equilibrium (–hwe 0.000001) and minor allele frequency (–maf 0.01) using the Plink software[50]. 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 filtered the windows with < 20 SNVs[63]. The genes located in the top 1% of windows were considered to be significant regions. The genes or genomic regions were defined within 100 kb of the identified non-overlapping SNVs.

Reference panel construction

The multi-allelic sites were split into bi-allelic sites via the BCFtools norm tool version 1.7[64]. We filtered the variants with –max-missing 0.9 and –hwe 0.000001 using VCFtools version 0.1.13[46]. In total, 508,196 variants were excluded. BEAGLE version 5.1 was used to perform haplotype phasing of all 4,489 samples with default settings. We conducted the haplotype re-phasing with SHAPEIT version 2 (r900) by windows 0.5, states 200 and effective-size 14,269[65]. 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[66], and then samples with the pairwise genetic relationship coefficient > 0.025 were thought to be cryptically related and removed; the variants and samples with a missing call rate > 5% were excluded by Plink version 1.9[50]; the variants deviating from Hardy-Weinberg equilibrium at $p < 10e\text{-}6$ 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 v2 setting the effective-size parameter to 14,269 as the software recommended for the Asian population[67]. 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[68]. 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 ≤ 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[25], and the largest Chinese-specific panel CONVERGE[38], and our own WBBC panel, and two combined panels that merged the WBBC datasets with the 1KG and EAS respectively. The imputation accuracy of these panels was then compared with each other by three different metrics. The design for the entire evaluation was detailed in Extended Data Fig. 9.

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 ≤ 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 8,610 haplotypes for the WBBC. Note that all the manipulations were conducted on chromosome 2[11,69]. 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[40]. 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[64]. Eventually, the WBBC+1KG combined panel consisted of 13,618 haplotypes at 4,450,989 variants, with 917,784 variants shared by both panels. The WBBC+EAS combined panel consisted of 9,618 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[11,69], 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[68]. 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)[68], 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[8]. We obtained average R-square values (Rsq) from Minimac4 info files[7] 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)[69]. 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 Extended Data Fig.10[52].

Genotype imputation server

Using the WBBC Phase 1 WGS data and 1KG Phase 3 data[25], 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[68] 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[64], 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[67,68]. 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).

Data availability

The allele frequencies of all variants and genotype imputation server are 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/). The application forms are required for researchers and the study must conform to the regulations of the Human Genetic Resources Administration of China (HGRAC). Researchers who request access to the raw genetic data must get permission from Ministry of Science and Technology of the People's Republic of China and the Institutional Review Board of the Westlake University.

Declarations

Competing interests

S.Y., W.Z. and J.L. are employee of KingMed Diagnostics.

Author contributions

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

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grants No: 32061143019 and 81871831), 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). 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.

References

1 Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* **19**, 110-124 (2018).

2 Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302-310 (2017).

3 Genetics for all. *Nat. Genet.* **51**, 579 (2019).

4 Martin, A. R. *et al.* 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, U. K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

7 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435-444 (2015).

8 Kowalski, M. H. *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 Nagasaki, M. *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* **6**, 8018 (2015).

10 Jeon, S. *et al.* Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv* **6**, eaaz7835 (2020).

11 Wu, D. *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736-749 e715 (2019).

12 GenomeAsia, K. C. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106-111 (2019).

13 Shi, Y., Li, L., Wang, Y., Chen, J. & Stanley, H. E. A study of Chinese regional hierarchical structure based on surnames. *Physica A* **518**, 169-176 (2019).

14 Xie, G., Lin, Q., Wu, Y. & Hu, Z. The Late Paleolithic industries of southern China (Lingnan region). *Quaternary International* **535**, 21-28 (2020).

15 Cao, Y. *et al.* The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* (2020).

16 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).

17 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).

18 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).

19 Chiang, C. W. K., 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).

20 Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).

21 Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).

22 Bai, W. Y. *et al.* Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Brief. Bioinform.* (2019).

23 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279-1283 (2016).

24 Zhu, X. *et al.* Cohort profile: The Westlake BioBank for Chinese (WBBC) pilot cohort: a prospective study for the late adolescence. *medRxiv*, 2020.2012.2016.20248291 (2020).

25 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

26 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

27 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

28 Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037-2048 (1994).

29 Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303-309 (2017).

30 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

31 Wilcoxin, F. Probability tables for individual comparisons by ranking methods. *Biometrics* **3**, 119-122 (1947).

32 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).

33 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).

34 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

35 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).

36 Tan, J. *et al.* The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum. Genet.* **132**, 1187-1191 (2013).

37 Riddell, J., Basu Mallick, C., Jacobs, G. S., Schoenebeck, J. J. & Headon, D. J. Characterisation of a second gain of function EDAR variant, encoding EDAR380R, in East Asia. *Eur. J. Hum. Genet.* (2020).

38 CONVERGE, c. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588-591 (2015).

39 Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu Rev Genomics Hum Genet* **19**, 73-96 (2018).

40 Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015).

Methods References

41 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

42 Van der Auwera, G. A. *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).

43 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).

44 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

45 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

46 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

47 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).

48 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362 (2014).

49 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).

50 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

51 Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338-348 (2018).

52 Linderman, M. D. *et al.* Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med. Genomics* **7**, 20 (2014).

53 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

54 McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, e1000686 (2009).

55 Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786-792 (1978).

56 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655-1664 (2009).

57 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).

58 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).

59 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358-1370 (1984).

60 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471 (2013).

61 Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).

62 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).

63 Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826-837 (2009).

64 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

65 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).

66 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).

67 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181 (2011).

68 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284-1287 (2016).

69 Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84**, 235-250 (2009).

Figures

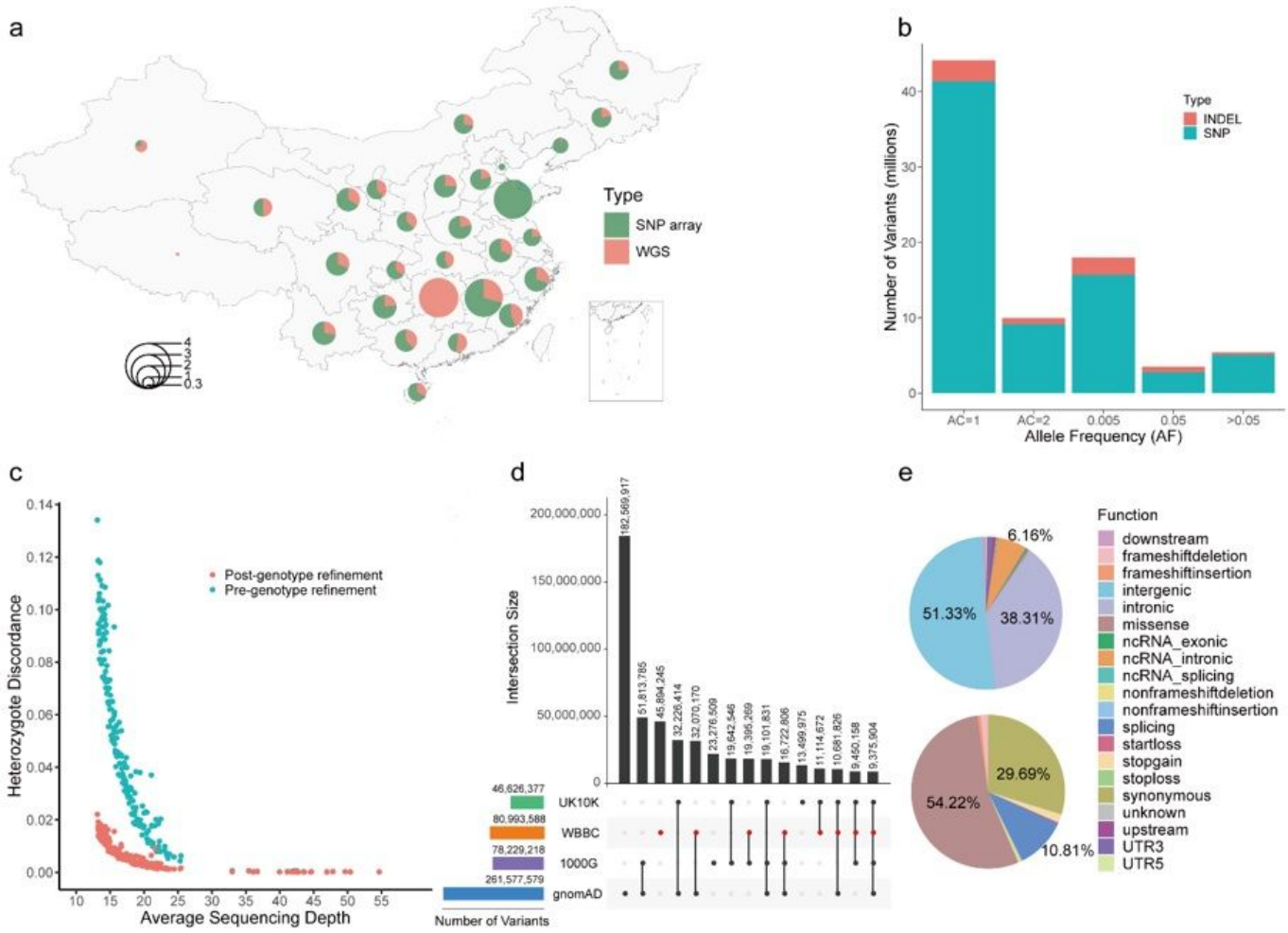

Figure 1

The statistics of samples and variants in the WBBC-cohort. a, Sample distribution and statistics by geography. The pie chart shows the number of samples in each administrative division. 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. The values were converted into Log10. 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 estimated heterozygote discordance rate versus sequencing depth for 184 samples. The red dots indicate the average proportion of post-genotype refinement via Beagle tools and the green dots denote the raw genotype calls. d, 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. e, Functional annotations of all novel variants were absent in dbSNP Build 151. The proportion of each category was filled with a different color. The upper pie chart showed all the 20 classification terms. The bottom pie chart only displayed the variants in the coding and splice regions (10bp from exon-intron boundary). Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

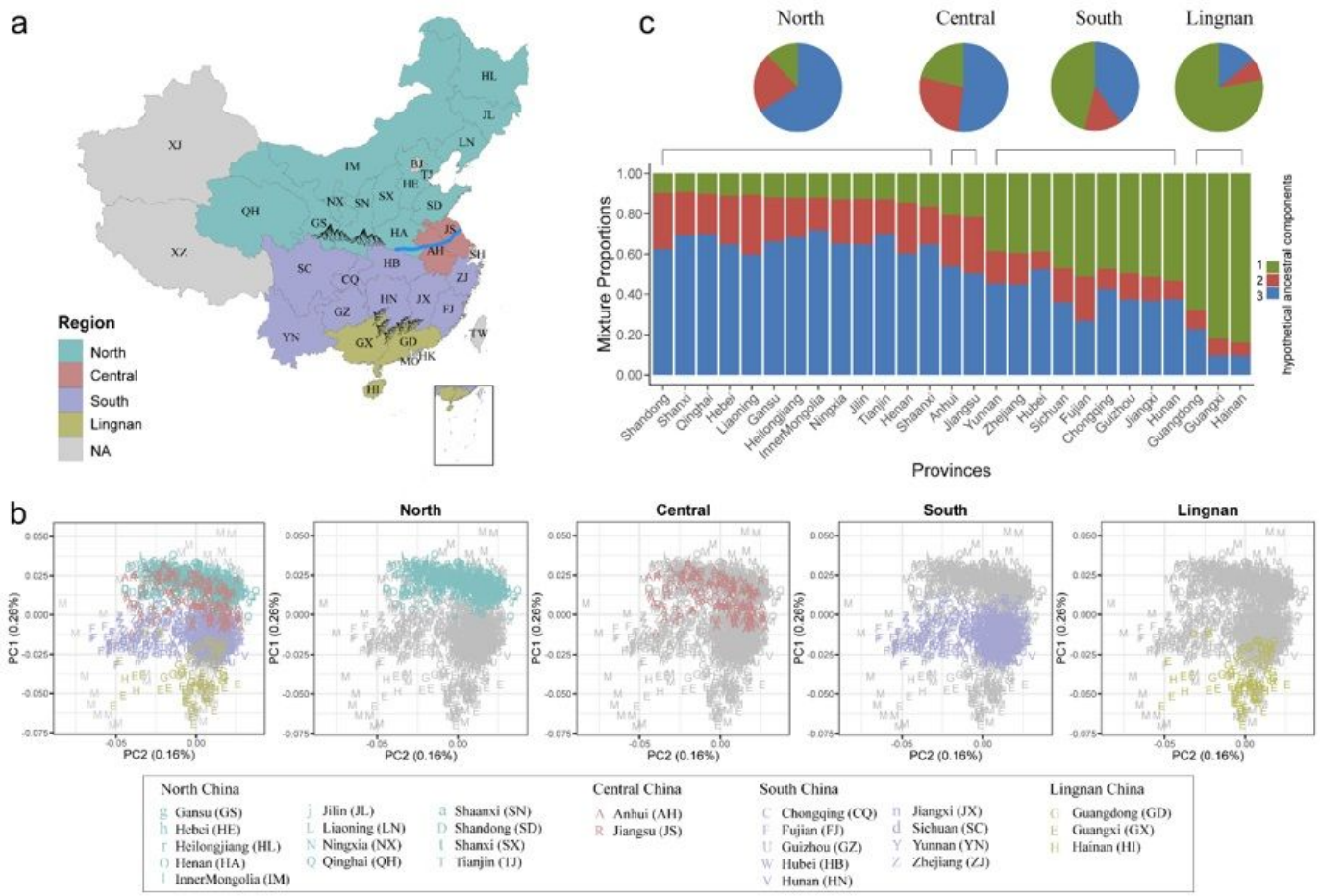

Figure 2

PCA and ADMIXTURE analysis of the Han Chinese populations. 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 regions. The administrative divisions are shown by the distinct letters. Minority people 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. Provinces are arranged by the value of hypothetical ancestral components 1 in each group. The upper pie charts denote the average proportion of components across individuals from the four subgroups. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

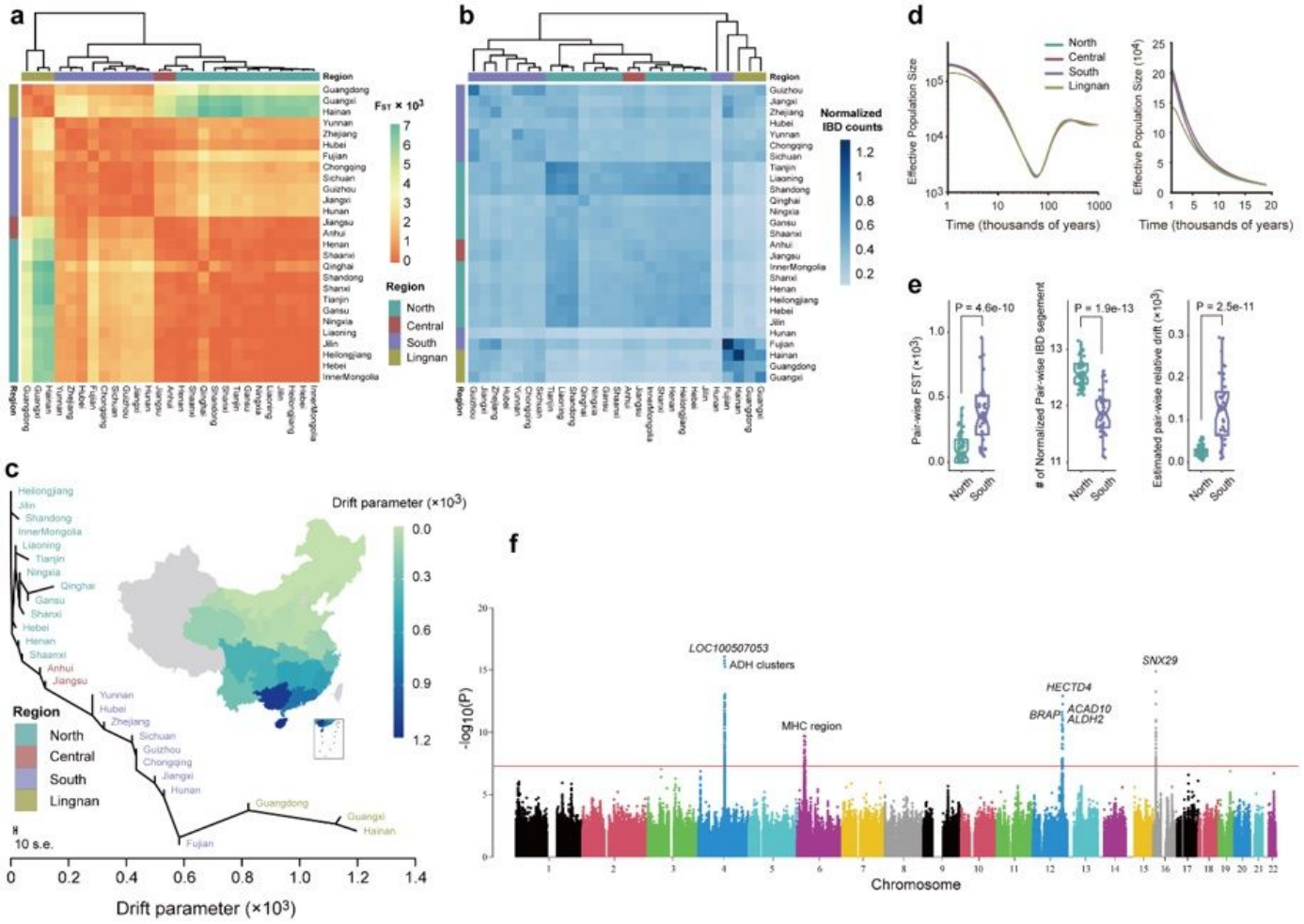

Figure 3

Genetic structure, demographic history and recent selection signatures 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. The hierarchical clustering is implemented by the hclust function in the pheatmap R package. 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. The scale bar shows ten times the average standard error of the entries in the sample covariance matrix for estimating the drift parameter. d, Dynamics of effective population sizes of the Han Chinese in four regions. 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. A generation time of 29 years was used to convert coalescent scaling to calendar time. 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. f, 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). Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

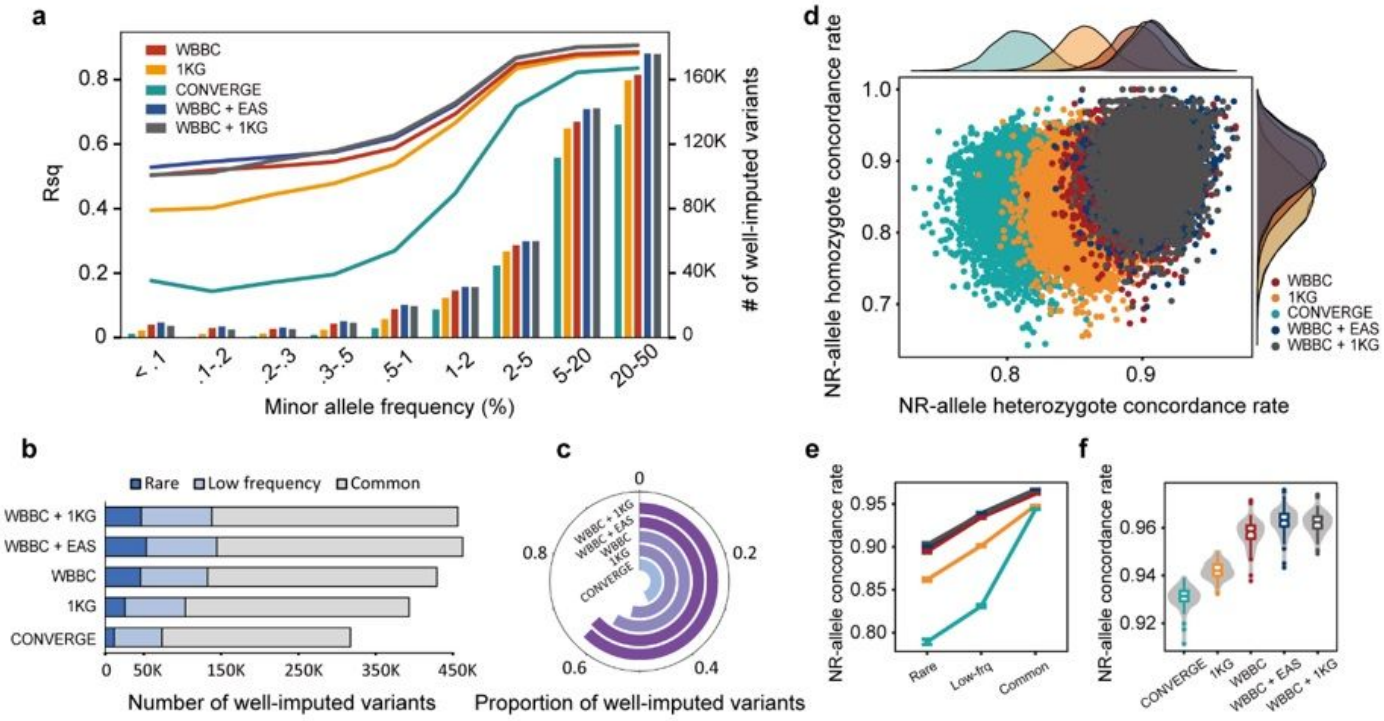

Figure 4

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 each MAF bin for Chinese imputation by five reference panels. b, the cumulative number of well-imputed variants. c, the proportion of well-imputed variants. 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.

Supplementary Files

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

- supplementaryTable.xlsx
- Supplementaryinformation20210129.docx

- ExtendedDataFile.pdf