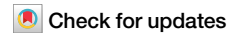


<https://doi.org/10.1038/s42003-025-07760-2>

Genetic diversity and dietary adaptations of the Central Plains Han Chinese population in East Asia



Xiaoyang Qiao^{1,7}, Jianxiang Shi^{1,7}, Hongen Xu^{1,7}, Kai Liu^{2,7}, Youwei Pu^{1,7}, Xia Xue¹, Wangshan Zheng², Yongbo Guo², Hao Ma³, Chuan-Chao Wang³, Habtom K. Bitsue¹, Xiaoyu Xu¹, Shanshan Wang¹, Jingru Zhao¹, Xiangqian Guo⁴, Xinyue Hou¹, Xinwei Wang¹, Lei Peng¹, Zan Qiu¹, Bing Su², Wenxue Tang⁵✉, Yaoxi He²✉, Jiancheng Guo⁶✉ & Zhaohui Yang¹✉

The Central Plains Han Chinese (CPHC) is the typical agricultural population of East Asia. Investigating the genome of the CPHC is crucial to understanding the genetic structure and adaptation of the modern humans in East Asia. Here, we perform whole genome sequencing of 492 CPHC individuals and obtained 22.65 million SNPs, 4.26 million INDELs and 41,959 SVs. We found the CPHC has a higher level of genetic diversity and the glycolipid metabolic genes show strong selection signals, e.g. *LONP2*, *FADS2*, *FGF21* and *SLC19A2*. Ancient DNA analyses suggest that the domestication of crops, which drove the emergence of the candidate mutations. Notably, East Asian-specific SVs, e.g., DEL_21699 (*LINC01749*) and DEL_38406 (*FAM102A*) may be associated with the high prevalence of esophageal squamous carcinoma and primary angle-closure glaucoma. Our results provide an important genetic resource and show that dietary adaptations play an important role in phenotypic evolution in East Asian populations.

Modern humans originated in Africa and then migrated around the world. East Asia is a key region for the study of migration and evolution of modern humans, and the Y-chromosome haplotype study of various East Asian populations suggests a south-north migration that led to the settlement of mainland East Asia^{1,2}. As humans migrated northwards, population expansion and environmental and dietary selection pressures led to changes in the structure of the genomes of East Asian populations³. Previous studies on the Han Chinese in East Asia have provided a wealth of genetic evidence for the complex migratory history and disease susceptibility of the populations^{4–6}. However, not enough attention has been paid to key areas particularly the Central Plains of China.

The Central Plains usually referred to as Henan Province in China, which lies in the middle and lower reaches of the Yellow River basin, is one of the largest agricultural regions in East Asia and even globally^{7,8}. Historically, more than 20 Chinese dynasties chose to build their capitals here (includes the famous the Xia Dynasty, Shang Dynasty and the Song

Dynasty), and it is considered the birthplace of Chinese civilization. Studying the population structure of the Central Plains populations and their subsistence patterns is crucial to understanding the formation of modern humans in East Asia. In addition, structural variation (SV), which consists of insertions, deletions, etc. of long fragments of sequences over 50 bp, is the other major form of variation in the genome. Previous studies have shown that SVs play a crucial role in human phenotypes, adaptation and disease⁹. But to date, investigations of SV in East Asian populations have been limited. Therefore, in-depth studies of the Han Chinese in the Central Plains using SNPs and SVs may provide additional genetic evidence for the population history and phenotypic adaptations of East Asian populations, particularly for adaptation to crop domestication, which has led to changes in the diet of East Asian populations.

In the current study, we collected 492 Central Plains Han Chinese (CPHC) individuals and performed whole genome sequencing. Population

¹Tianjian Laboratory of Advanced Biomedical Sciences, Academy of Medical Science, Zhengzhou University, Zhengzhou, China. ²State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ³State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China. ⁴Zhongyuan Intelligent Medical Laboratory, School of Basic Medical Sciences, Henan University, Kaifeng, China. ⁵The Research and Application Center of Precision Medicine, Departments of Otolaryngology, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, China. ⁶The Research and Application Center of Precision Medicine, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, China. ⁷These authors contributed equally: Xiaoyang Qiao, Jianxiang Shi, Hongen Xu, Kai Liu, Youwei Pu. ✉e-mail: twx@zzu.edu.cn; heyaoxi@mail.kiz.ac.cn; gjc@zzu.edu.cn; yangzh@zzu.edu.cn

structure analysis indicates that the CPHC has a more concentrated set of genetic components and a higher level of genetic diversity than the other populations in East Asia (CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan) in the 1000 Genomes Project (1KGP). We identified a lot of candidate variants (13,035 SNPs and 26 SVs) with higher allele frequencies in the CPHC. In particular, the glycolipid metabolic genes show strong selection signals in CPHC, for example, the *LONP2*, *FADS2*, *CCDC171*, *SULT1C4*, *SLAH1*, *ASXL2* genes and the high-frequency SVs INS_27791 (*FGF21*), INS_03097 (*SLC19A2*). In addition, East Asia-specific structural variants DEL_21699 (*LINC01749*) and DEL_38406 (*FAM102A*) may be associated with esophageal squamous carcinoma and primary angle-closure glaucoma (PACG), respectively. In conclusion, this study demonstrates the unique genetic structure and dietary adaptations of the CPHC, and the result will provide important clues for the study of population structure and phenotypic adaptation of East Asian populations.

Results

Genome sequencing of 492 CPHC individuals and variants annotation

We sampled a total of 492 Han Chinese individuals from 16 cities in Henan Province, China (Fig. 1a). The individuals were between 30–50 years old, with a gender ratio of 1.56:1 (300 female and 192 male). Whole genome sequencing was performed on the Illumina sequencing platform and quantified on the Bioanalyzer 2100 with a paired-end read length of 2×150 bp. The mean NGS coverage was 25.54x, with a range ~20.2–42.2x (Fig. 1b). A total of 22.65 million SNPs and 4.26 million INDELs have been identified using the standard GATK process and a rigorous filtering process (“Methods”). The transition-to-transversion ratio (Ts/Tv) of all these variants is 2.12, indicating the accuracy of our SNP calling¹⁰. A total of 2,417,290 single nucleotide variants (~10.67% of the current study, compared to dbSNP Build 156) were novel variants (Fig. 1c); 151,763 SNPs (0.67%) located in coding and splice regions (including 71,359 missense mutations); 14,847,845 SNPs (65.55%) located in intronic region and 144,514 SNPs

(6.38%) in intergenic region (Fig. 1d; Supplementary Fig. 1); In addition, 17,441,403 SNPs (~77% of the current study) were minor allele frequency variants (AF < 0.01). The total number of these variants is much higher than other Han Chinese databases, e.g. ~1.9 times the total number of CHB and ~2.2 times the total number of CHS in 1KGP (Supplementary Fig. 2).

More importantly, we used Paragraph (a graph-based SVs genotyping tool)¹¹ to type SVs in the CPHC genome and obtained a total of 19,042 DELs and 22,917 INSs with the SV length ranging from 50 to 4000 bp. Next, we used the VEP and RepeatMasker tools to annotate all variants^{12,13}. Of these, 18,546 SVs (44.2%) were located in genomic regulatory regions, 340 SVs (0.81%) in coding regions and 17,916 SVs (42.7%) in intronic regions (Fig. 1e). We also identified SINE-associated SVs in various genomic functional regions (Fig. 1f and g). For Deletions, 2457 (42.4%) were located in intronic regions; 442 (7.63%) in 5'/3'-UTR regions; 386 (6.66%) in non-coding RNAs regions and 93 (1.61%) in up/downstream regions. For Insertions, 2326 (48.68%) were located in intronic regions; 605 (12.66%) in 5'/3'-UTR regions; 416 (8.7%) in non-coding RNAs regions and 65 (1.36%) in up/downstream regions. In addition, minor allele frequency calculations for all variants found 80% of the SVs were rare variants (AF < 0.01), which may be specific to CPHC populations.

These results indicate that many SNPs and SVs are present in the CPHC genome, suggesting a rich genetic diversity that is a great complement to the variant resources of East Asian populations. Further analysis of these variants will be important for understanding the genetic structure, phenotypic adaptation, and disease risk of the CPHC.

The CPHC have a higher level of genetic diversity than other populations in East Asia

To investigate the genetic relationship between the CPHC and other populations, we performed principal component analysis (PCA) using SNPs and SVs generated from the CPHC and public databases (1KGP and HGDP), respectively. When we conducted the PCA of global populations, East Asians, Europeans and African populations showed significant genetic differences, and the CPHC clustered closely with East Asian populations

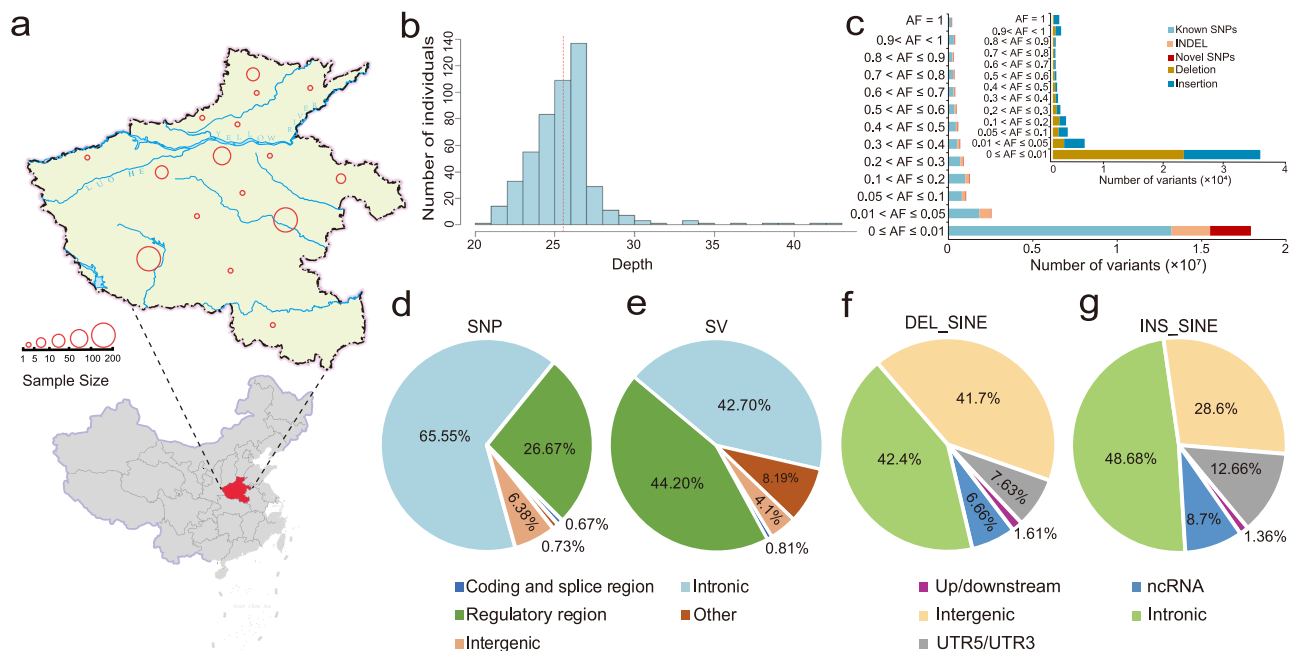


Fig. 1 | Sample distribution and WGS data statistics in the CPHC. a Distribution of samples in the CPHC. The red highlight shows the geographical location of Henan Province in China. Red circles represent the native places of the samples. The map was sourced from an approved standard map service (<http://bzdt.ch.mnr.gov.cn>) endorsed by the Ministry of National Resources of the People's Republic of China (GS(2019)1653, GS(2019)3266 and GS(2019)3333). **b** The distribution of WGS

genomic coverage. The red line refers to the mean coverage of the samples. **c** Numbers of variants detected in different bins of allele frequencies. Variants were classified as known or novel based on public resources (1KGP and dbSNP). **d–g** Functional annotation of SNPs and SVs in the CPHC. The scale of each category is filled with a different color.

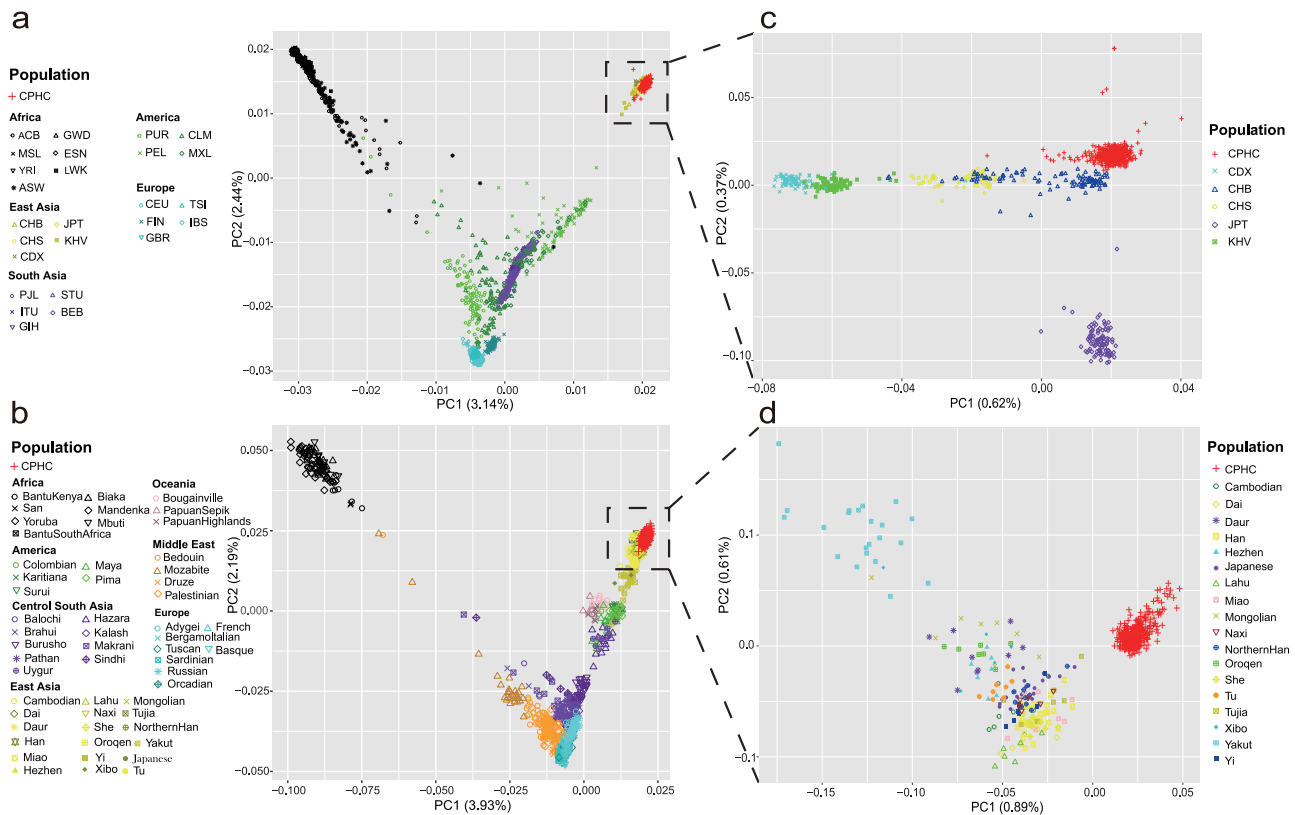


Fig. 2 | Principal component analysis (PCA) of SNPs and SVs in the CPHC and other populations. a The CPHC and all populations in the 1KGP. **b** The CPHC and other East Asian populations in the 1KGP. **c** The CPHC and all populations in the

HGDP. **d** The CPHC and other East Asian populations in the HGDP. The population abbreviations are given in Supplementary Data 4.

(Fig. 2a, b). It is worth noting that the SV analyses suggest that the Native Americans in the HGDP (Colombian, Maya, Karitiana, Pima and Surui) are genetically closer to East Asian populations than modern American populations in the 1KGP (PUR, Puerto Ricans from Puerto Rico; CLM, Colombians from Medellin, Colombia; PEL, Peruvians from Lima, Peru; MXL, Mende in Sierra Leone) (Fig. 2a, b). This may indicate a closer genetic relationship between East Asian populations and Native Americans. We further investigated the genetic relationship between East Asian populations. The results of SNP-based PC1 showed that populations in southern East Asia (KHV, Kinh in Ho Chi Minh City, Vietnam; CDX, Chinese Dai in Xishuangbanna, China) were genetically more closely, whereas CHB and CHS showed mixed Han Chinese population characteristics, but the CPHC individuals showed a more concentrated genetic component. The levels of PC2 showed genetic differences between the Japanese (JPT) and most of the East Asian populations (Fig. 2c). From the perspective of SVs, similarly, a separate genetic component for CPHC was also shown, and the inclusion of CPHC makes the genetic position of the Han Chinese population more prominent (Fig. 2d).

We next inferred the rooted maximum likelihood tree (Fig. 3a) between the CPHC and the 26 world populations in the 1KGP using YRI as the outgroup with TreeMix and MEGA^{4,15}. The CDX and KHV are clustered in the same branch, while the northern East Asian populations (CHS, CHB, CPHC and JPT) are clustered in another branch. CPHC is closer to CHB and together with CHB and CHS forms the Han population, distinct from JPT. We further calculated the F_{ST} ¹⁶ between the populations to indicate the degree of genetic differentiation between the northern, central and southern Han populations and the YRI. The result showed that the F_{ST} between the CPHC and YRI were greater than CHB and CHS with YRI ($F_{ST}^{(CPHC-YRI)} = 0.179$; $F_{ST}^{(CHB-YRI)} = 0.170$; $F_{ST}^{(CHS-YRI)} = 0.169$), implies a greater genetic difference between CPHC and YRI. Meanwhile, we found that CPHC is less genetically differentiated from CHB compared to CHS

($F_{ST}^{(CPHC-CHS)} = 0.00234$; $F_{ST}^{(CPHC-CHB)} = 0.00068$), suggesting that the CPHC is genetically more closely to the northern Chinese population (Supplementary Fig. 3).

To further investigate the genetic composition of the major populations in East Asia, we performed admixture analysis on the CPHC and other populations from the 1KGP. Each population is assumed to be descended from an arbitrary number of ancestral components, denoted by K ¹⁷. After rigorous screening for relatedness and linkage disequilibrium (Methods), the admixture results showed that from $K = 4$, East Asia is dominated by the genetic components of the northern (yellow) and southern (purple) populations (Supplementary Fig. 4). Further analysis of the Cross-validation result shows that $K = 5$ best matches the East Asian populations' genetic structure (Supplementary Fig. 5). The CPHC contains predominantly the main genetic components of the northern East Asian populations, whereas the southern CDX and KHV populations contain predominantly the genetic components of the southern East Asian populations, and the CHB and CHS share the genetic components of the northern and southern East Asian populations. JPT provides a new genetic component (Fig. 3b), which is also in agreement with the PCA of the East Asian populations described above.

Finally, we investigated the genetic diversity and the effective population size of CPHC. LD Decay was performed between the CPHC and other East Asian populations in the 1KGP, using CEU and YRI as control populations. In general, faster LD decay indicates greater genetic diversity¹⁸. The LD decay rate of the CPHC is significantly higher than that of CHB, CHS and JPT (Fig. 3c). This implies a richer genetic diversity in the CPHC. To further investigate the effective population size of the East Asians, we performed MSMC analysis on the above populations¹⁹. The results indicated that non-African populations experienced a dramatic population decline around 100,000 years ago, corresponding to the temperature drop during the Last Glacial Maximum²⁰. On the other hand, there was a rapid increase

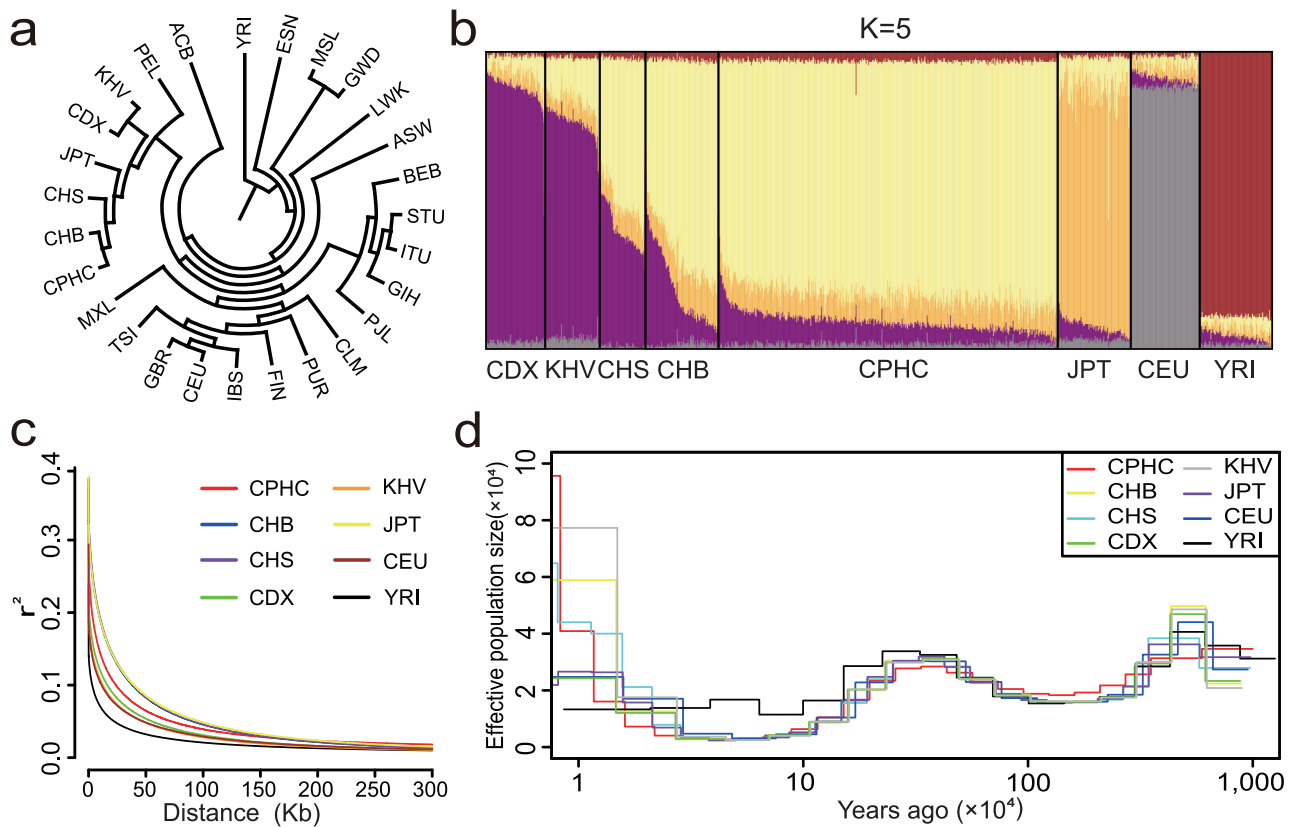


Fig. 3 | Population structure and genetic diversity analysis. **a** The maximum likelihood tree of the CPHC with other populations in the 1KGP. **b** ADMIXTURE analysis of the CPHC with other populations in 1KGP. Different colors represent different ancestry components (Yellow: Northern Asian population; Purple:

Southern Asian population; Orange: JPT; Grey: CEU; Brown: YRI). **c** The pattern of LD decay between the CPHC and the other populations in the 1KGP. **d** Dynamics of effective population sizes of the CPHC and other world populations in 1KGP. The population abbreviations are given in Supplementary Data 4.

in the effective population size of the CPHC around 10,000 years ago (Fig. 3d). This period coincides with the emergence of an early agricultural civilization in China, which later developed gradually in the central plains near the Yellow River basin with the northern dry farming system represented by wheat and millet^{21,22}.

Collectively, this evidence suggests that the CPHC have a rich genetic diversity with some genetic differences between them and other populations in East Asia, which may suggest CPHC can provide more genetic information in the Han population of East Asia.

Genome-wide detection of selective signals in the CPHC

The above analyses demonstrate the unique genetic components of the CPHC, which may be related to the genetic adaptation of modern humans following their migration to East Asia. To further investigate the genetic components of the CPHC, we performed F_{ST} analysis between the CPHC (agricultural group) and 81 hunter-gatherer groups (JPZ) in Southeast Asia²³ and identified a total of 380 genes (13,035 SNPs) showing significant genetic differentiation (Fig. 4a). These SNPs were also tested using XP-EHH, a long-range haplotype statistical method to identify alleles segregating in a population that has undergone selection²⁴. We observed 12,535 highly differentiated SNPs ($F_{ST} \text{ (CPHC-JPZ)} > 0.1$) in 361 genes that only selected in the CPHC (Supplementary Data 1). We took the top 0.1% regions, which contained a total of 9 SNPs with high derived allele frequencies ($F_{ST} \text{ (CPHC-JPZ)} > 0.51$). The 9 SNPs are located in the *LONP2* (4 SNPs), *SIAH1* (1 SNPs), *TDRD10* (2 SNPs) and *SHE* (2 SNPs) gene regions.

LONP2 is located on chromosome 16 and encodes an ATP-dependent peroxidase whose overexpression promotes the synthesis of polyunsaturated fatty acids (PUFA), which is one of the essential fatty acids and plays an important role in human lipid metabolism^{25,26}. Four SNPs rs12446160, rs6500380, rs8052293 and rs111915588 are all found mainly in

East Asia and are most common in CPHC ($DAF > 0.97$) (Supplementary Data 1). First, rs12446160 ($F_{ST} \text{ (CPHC-JPZ)} = 0.579$; XP-EHH $\text{ (CPHC-JPZ)} = 3.620$) is a cis-eQTL for *LONP2*, which can promote its gene expression²⁷. In addition, LocusZoom indicates that the variant rs6500380 ($F_{ST} \text{ (CPHC-JPZ)} = 0.634$; XP-EHH $\text{ (CPHC-JPZ)} = 3.254$) highly linked with rs8052293 ($R^2 = 1$), rs111915588 ($R^2 = 0.929$) and an adjacent intronic variant rs1982782 in *SIAH1* ($F_{ST} \text{ (CPHC-JPZ)} = 0.638$; XP-EHH $\text{ (CPHC-JPZ)} = 2.964$, $R^2 = 0.916$; Fig. 4b) and as cis-eQTL for another lipid metabolism gene *SIAH1*, which can promote *SIAH1* expression to enhance ubiquitination of Scp2 and facilitate intracellular transport of cholesterol and fatty acids, reducing the incidence of various metabolic diseases such as atherosclerosis, diabetes and obesity^{27,28}. This suggests that some specific lipid metabolism genes are under natural selection in the CPHC. Also, two SNPs rs12128408 ($F_{ST} \text{ (CPHC-JPZ)} = 0.526$; XP-EHH $\text{ (CPHC-JPZ)} = -1.732$), rs12136771 ($F_{ST} \text{ (CPHC-JPZ)} = 0.514$; XP-EHH $\text{ (CPHC-JPZ)} = -0.413$) of *TDRD10* (encodes a protein with Tudor domains) and two SNPs rs6684921 ($F_{ST} \text{ (CPHC-JPZ)} = 0.515$; XP-EHH $\text{ (CPHC-JPZ)} = -0.206$), rs12118018 ($F_{ST} \text{ (CPHC-JPZ)} = 0.514$; XP-EHH $\text{ (CPHC-JPZ)} = -1.905$) of *SHE* (encodes a protein with Src homology 2 domain) were also distributed at high frequency ($DAF > 0.8$) in the CPHC (Supplementary Data 1). Previous study found these four variants are all cis-eQTL for *SHE*²⁷. Meanwhile, *TDRD10* and *SHE* have been reported to be associated with the immune component, e.g. serum levels of total protein, C-reactive protein and *IL6* receptor^{29–31}. This suggests that the immune response is another important adaptive phenotype in CPHC. However, the selection signals for XP-EHH did not show the same significance as the lipid metabolism genes.

In addition to the variants described above that are specific to the CPHC, 19 genes were also found in other two Han population studies (Westlake BioBank and the NyuWa Genome)^{4,32} (Supplementary Data 2). Remarkably, there have 8 genes (*FADS2*, *FADS3*, *CCDC171*, *SULT1C4*,

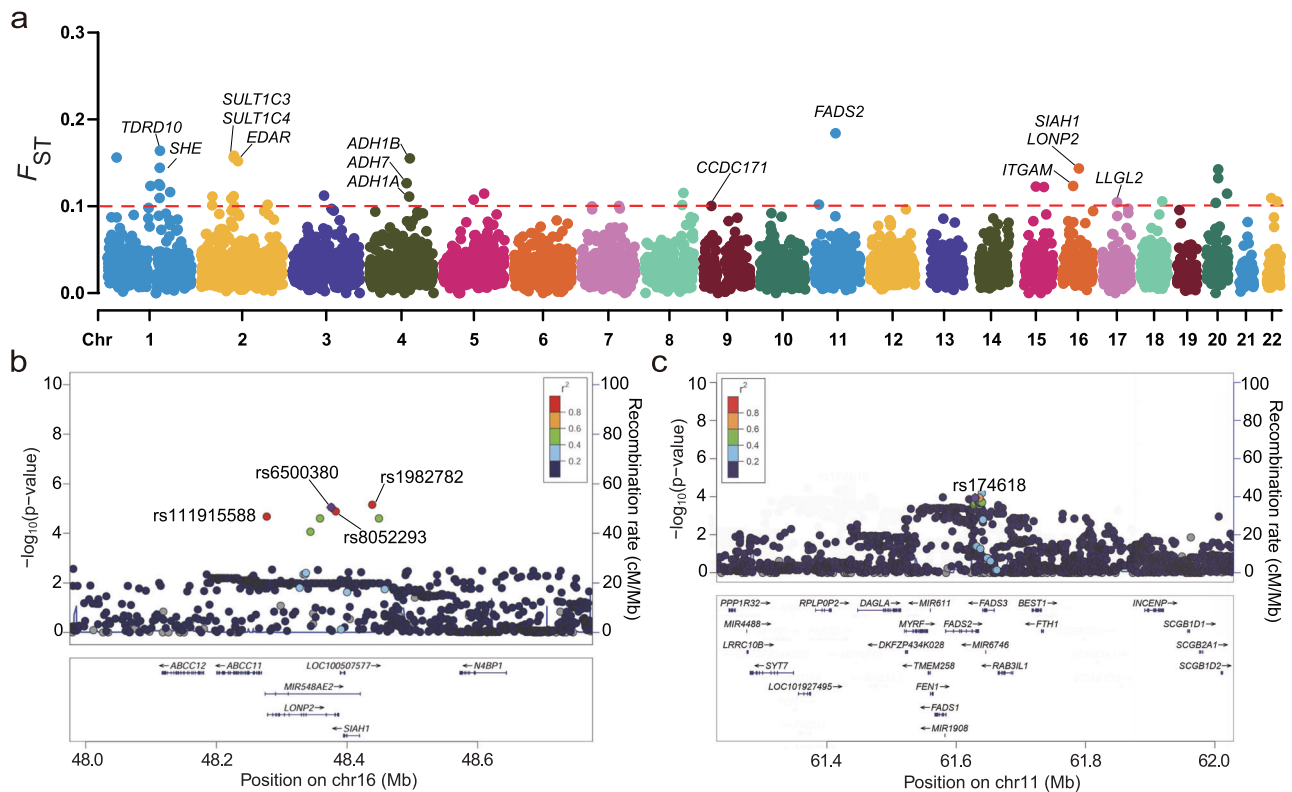


Fig. 4 | Genome-wide selection signatures of the CPHC. a Manhattan plot of the signatures of natural selection from the WGS data of the CPHC. The y-axis represents the F_{ST} values. Significance threshold indicated by horizontal red line

($F_{ST} > 0.1$). **b, c** LocusZoom plots of the rs6500380 and rs174618. The rs6500380 and rs174618 are represented by the purple circle. Linkage disequilibrium r^2 values were calculated using the East Asian population from the 1KGP.

ASXL2, ADH1B, ADH7 and ADH1A) are associated to human glycolipid metabolism. While the other 11 genes (*RAB31L1*, *RAPGEF2*, *CCDC138*, *EDAR*, *RANBP2*, *GCC2*, *KIF3C*, *LIMS1*, *SULT1C3*, *SULT1C2* and *C4orf45*) have a variety of functions, including play a role in the human immune system, angiogenesis, the development of sweat glands and the development of the nervous system^{33–39}.

For these 8 glycolipid metabolic genes, the *FADS* genes (*FADS2* and *FADS3*) encode a desaturase and can also catalyze the biosynthesis of polyunsaturated fatty acids (PUFA)⁴⁰. These two genes are located next to each other at 6,172 bp. We have identified a total of 14 SNPs in *FADS2* and 3 SNPs in *FADS3* (Supplementary Data 2). In particular, the regulatory SNPs rs174618 (associated with PUFA concentrations; $p < 0.05$) in *FADS2* ($F_{ST}(\text{CPHC-JPZ}) = 0.506$; XP-EHH_{(CPHC-JPZ)} = 5.609) and the two regulatory SNPs rs174450 ($F_{ST}(\text{CPHC-JPZ}) = 0.312$; XP-EHH_{(CPHC-JPZ)} = 4.578) and rs174455 ($F_{ST}(\text{CPHC-JPZ}) = 0.294$; XP-EHH_{(CPHC-JPZ)} = 4.118) in *FADS3* are all cis-eQTL for *FADS2*, suggesting a possible synergistic role for these variants in the modification of polyunsaturated fatty acid concentrations in human serum^{27,41,42} (Fig. 4c; Supplementary Data 2). Next, *SULT1C4* and *ASXL2* are involved in secreting insulin. 9 SNPs were found in the gene *SULT1C4*, which is associated with flavanone metabolism and may promote insulin secretion. Among them, a missense mutation rs1402467 ($F_{ST}(\text{CPHC-JPZ}) = 0.269$; XP-EHH_{(CPHC-JPZ)} = 4.706) can regulate insulin secret and then influence glucose metabolism in humans⁴³. 14 SNPs in is an ETP family gene *ASXL2*, which activates adipogenic genes, regulates insulin secretion and plays an important role in adipogenesis and glycolysis in humans⁴⁴. In addition, 31 SNPs showed selective signals in a fat storage and metabolic regulation gene *CCDC171*. The missense mutation rs1539172 ($F_{ST}(\text{CPHC-JPZ}) = 0.101$; XP-EHH_{(CPHC-JPZ)} = 2.457) is a cis-eQTL for *CCDC171* and strongly associated with body mass index and obesity-related diseases (such as hypertension and diabetes)^{27,31}. Besides, *ADH1B* (8 SNPs), *ADH1A* (4 SNPs) and *ADH7* (6 SNPs) are all alcohol metabolism genes, especially the well-known ethanol dehydrogenase gene *ADH1B*. The}}}}}

mutation rs1229984 (Arg47His, $F_{ST}(\text{CPHC-JPZ}) = 0.313$; XP-EHH_{(CPHC-JPZ)} = 2.905) results in higher ethanol dehydrogenase activity and increased alcohol metabolism capacity⁴⁵. The above analyses showed that glycolipid metabolism is one of the most important phenotypic characteristics of CPHC, suggesting that it is a key adaptive trait in East Asian populations.}

In addition to this, several genes that have been proven to be selective in East Asia were also observed. For example, *OCA2* (42 SNPs; $F_{ST}(\text{CPHC-JPZ}) > 0.1$) and *PAH* (21 SNPs; $F_{ST}(\text{CPHC-JPZ}) > 0.1$) are all associated with lighter skin color^{46–49}. *EDAR* (40 SNPs; $F_{ST}(\text{CPHC-JPZ}) > 0.1$), a gene associated with the development of hair straightness³⁹. *SLC35F3* (126 SNPs; $F_{ST}(\text{CPHC-JPZ}) > 0.1$) and *SLC35F4* (21 SNPs; $F_{ST}(\text{CPHC-JPZ}) > 0.1$) associated with vitamin B1 metabolism⁵⁰ (Supplementary Data 3).

Taken together, the analyses of the selective signals of the SNPs in the CPHC suggest that the CPHC has a number of specific and high-frequency mutations, especially SNPs related to glycolipid metabolism, suggesting that the CPHC may undergo adaptive traits as a result of environmental and dietary changes during migration.

Glycolipid metabolism-associated mutations may occur as a result of dietary adaptations due to the expansion of agriculture

The Central Plains have been an important source of crops in East Asia and even globally, and several of the important glycolipid metabolism-associated genes mentioned above show strong selection signals in the CPHC. Therefore, we further investigated whether the occurrence of these specific mutations could have a relationship with the agricultural domestication. We analyzed 21 typical mutations in these genes (*LONP2*, *FADS2*, *FADS3*, *CCDC171*, *SULT1C4*, *SIAH1*, *ASXL2*, *ADH1B*, *ADH1A* and *ADH7*) in 14 ancient DNA samples from the Yellow River basin^{51,52}. We infer the possible time of the candidate SNP emergence based on the age of the ancient DNA samples and the number of typing of its derived allele. First, these SNPs were found to be widespread in the ancient human population of the Yellow River basin (Table 1). The occurrence of some SNPs derived alleles can be dated to

6,000 BP (the Xiaowu site in Lingbao City, Henan Province), such as the derived allele A of rs12446160 (*LONP2*; totalCov ≥ 2 in 9/14 samples) and the derived allele T of rs174618 (*FADS2*; totalCov ≥ 2 in 12/14 samples). The time coincides with the cultivation of crops (such as millet) in the Central Plains⁵¹. Other typical mutations in genes associated with glycolipid metabolism (*CCDC171*, *SULT1C4*, *SIAH1* and *ASXL2*) showed similar results, expect two SNPs in *FADS3* (Table 1). However, it's also worth noting that derived alleles of alcohol metabolism variants associated with the domestication of rice were not widespread in ancient humans from the Yellow River basin. A well-known missense variant rs1229984 in *ADH1B*, which is associated with alcohol metabolism, is predominantly ancestral C (totalCov ≥ 2) in 7/14 ancient DNA samples (Xiaowu, Lingbao; Dahecun, Zhengzhou; Pingliangtai, Huaiyang; Haojiatai, Luohe; and Niecun, Jiaozuo) and derived T (totalCov ≥ 2) is only 3/14 ancient DNA samples (Wadian, Yuzhou; Niecun, Jiaozuo). Previous studies have shown that the rs1229984_T allele appeared around 7,000 years ago, coinciding with the beginning of the domestication of rice in the Yangtze River basin⁴⁵. Two other alcohol metabolism genes, *ADH1A* (rs1813977, rs1039151 and rs3819197) and *ADH7* (rs284784 and rs1827567) showed similar results to *ADH1B* (Table 1). Therefore, the derived mutations of the *ADH* genes did not reach high frequencies in the Central Plains during the Late Neolithic.

In summary, the ancient DNA analysis not only suggests that the generation of these mutations in *LONP2*, *FADS2*, *CCDC171*, *SULT1C4*, *SIAH1* and *ASXL2* in the Central Plains may be related to the agricultural domestication and dietary adaptation in the Yellow River basin humans but also suggests that the genetic basis of dietary adaptations may be different in the major crop regions of the Yellow River (wheat and millet dominated) and Yangtze (rice dominated) basins.

Specific structural variants contribute to the genetic diversity of the populations of East Asia

We used Paragraph to genotype the SVs in the CPHC¹¹, identified and manually checked a total of 26 SVs (14 deletions and 12 insertions) with large frequency differences between the CPHC and other populations in the world. Of these, 9 SVs (4 deletions and 5 insertions) are specific to the CPHC (AF: 0.203–0.516). The other 17 SVs (10 deletions and 7 insertions) are with high frequency (0.4 higher than the average frequency in the other populations around the world) in the CPHC (AF: 0.528–0.924) (Fig. 5a, b and Table 2).

Among the 9 East Asian population-specific deletions, DEL_21699 is a 170 bp deletion located in *LINC01749*, with frequency 0.222 in the CPHC (African frequency is 0 and European is 0.018 in HGDP) (Fig. 5c). *LINC01749* encodes a class of non-protein-coding RNA transcripts greater than 200 nucleotides, has an important role in esophageal squamous cell carcinoma⁵³. We performed the linkage disequilibrium (LD) of its upstream and downstream 10 kb regions and observed a highly linked SNP rs6010839 near the DEL_21699 ($R^2 = 0.96$), an eQTL for the esophageal squamous cell carcinoma risk gene *TCFL5*^{27,54}. The Central Plains of China is known to be a region with a high incidence of esophageal squamous cell carcinoma^{55,56}. We have also seen the distribution of the DEL_21699 in other groups with a high incidence of esophageal squamous cell carcinoma (e.g., Hazara in north-eastern Iran, Russians in Eastern Europe and Japanese) (Fig. 5c)⁵⁷. This suggests that DEL_21699 may be involved in the development of esophageal squamous cell carcinoma. DEL_38406 is a 92 bp deletion located in intron 10 of the *FAM102A*, with frequency is 0.223 in the CPHC (African and Europeans is 0 in HGDP) (Fig. 5d). Previous study reported that *FAM102A* plays an important role in Primary angle-closure glaucoma (PACG), a predominant type of glaucoma in East Asian populations^{58,59}. We found a highly linked variant rs2417064 ($R^2 = 0.93$), an eQTL for *FAM102A*²⁷. This suggests that DEL_38406 may have potential function in PACG.

Next, we also identified 17 high-frequency SVs (10 deletions and 7 insertions) in the CPHC. INS_27791 is a 639 bp insertion (CPHC = 0.924; HGDP.Han = 0.875; HGDP non-East Asian populations = 0.441–0.508), located on upstream of the *FGF21*, which is a metabolic hormone synthesized by several organs that regulates glucose sensitivity, and specifically

encodes a protein that stimulates glucose uptake in adipose tissue⁶⁰. LD analysis showed that INS_27791 highly linked with some intergenic variants, including an *FGF21* eQTL SNP rs4801773 ($R^2 = 1$)^{27,31}. Therefore, INS_27791 may have an effect on *FGF21* expression to alleviate diet-induced insulin resistance and regulate hepatic gluconeogenesis. INS_03097 is a 337 bp insertion (CPHC = 0.751; HGDP.Han = 0.579; HGDP non-East Asian populations = 0.018–0.159), located in intron 4 of the *SLC19A2*, encodes a thiamine transporter. Its deficiency affects intracellular thiamine transport, and this impairment affects insulin secretion and glucose metabolism, which in turn increases the risk of diabetes in people on a high starch diet⁶¹. Therefore, some of the high SVs found in CPHC may also have contributed to dietary adaptation in East Asian populations.

In addition to the typical examples listed above, the remaining 22 SVs (7 unique and 15 high frequencies in the CPHC) located in the intergenic regions, e.g. INS_18720 119 bp, INS_42051 56 bp and INS_05228 77 bp. Some SVs in the intronic regions of the gene, e.g. DEL_29932 located in *PPP2R2B* (2,592 bp); DEL_24082 located in *ENPP7P2* (72 bp); INS_21702 located in *DPEPI* (61 bp) and INS_46471 located in *OSMR-DT* (96 bp) (Table 2). However, very little is known about the phenotypes that are associated with these structural variants and genes.

In general, the results suggest the presence of some high-frequency or specific SVs in the CPHC, which may play an important role in East Asia specific disease genesis and dietary adaptation.

Discussion

The central plains of China, which lie in the middle and lower reaches of the Yellow River basin, are one of the core areas of agricultural domestication in East Asia. Genomic studies of populations in the Central Plains are critical to understanding the genetic structure of East Asian populations and genetic adaptations resulting from agricultural expansion. In the current study, whole-genome sequencing data revealed the unique genetic structure of the CPHC, indicating its rich genetic diversity. The new discoveries of high-frequency mutations and structural variants are an important genetic resource for East Asian populations.

This PCA result suggests that the CPHC has some more concentrated genetic components. In comparison with CHS, CPHC has a closer genetic affinity with CHB, suggesting that it has more genetic characteristics of northern populations. The PCA results of SV also showed that Native Americans are closer to East Asian populations, further confirming this result from the perspective of SV (Fig. 2b). This finding is consistent with the results of previous analyses of the SNPs^{62–65}. However, the Han Chinese in the HGDP are mixed with Yi, Hui, and other East Asian populations, which is inconsistent with the PCA results of SNPs (Fig. 2c, d) and may be as the low number of certain populations in the HGDP and the insufficient number of SVs. Therefore, this study is also an addition to previous databases of Han Chinese data.

The ADMIXTURE and LD Decay analysis showed a higher genetic diversity (Fig. 3b, c). $K = 5$ best represents the number of ancestral components present in East Asian populations, with three ancestral components observed (Fig. 3b; Supplementary Fig. 5). These three ancestral components correspond to populations from southern East Asia (CDX and KHV), Han Chinese populations (CPHC, CHB and CHS) and the Japanese population (JPT). Compared to $K = 4$, a new ancestral component emerges in the Japanese population, leading to its separation from other East Asian populations. These results may be related to the relatively isolated island geography of Japan. Long-term genetic drift and relatively independent populations may have led to the development of genetic characteristics in the Japanese population.

The effective population size analysis shows a rapid increase in the population of the CPHC around 10,000 years ago, probably as a result of the beginning of crop cultivation and the availability of a stable source of food⁶⁶. This implies that crop domestication may have facilitated dietary adaptation in East Asian populations and have contributed to the population increase of East Asia. Early human diets were predominantly hunter-gatherer, rich in animal protein and fiber. However, with the expansion of agriculture, the

Table 1 | The allele types of some typical mutations associated with dietary metabolism in 14 ancient DNA samples from the Yellow River Basin

SNP	Pos (GRCh38)	Region	AA	DA	GENE	DAF (CPHC)	XW-MIR18 (6056 ± 119 BP)	WGM35 (5180 ± 124 BP)	DM28 (4400 ± 30BP)	WD- WTH16 (4038 ± 53 BP)	PLTM310 (3740 ± 30 BP)	PLTM311 (3710 ± 30 BP)	HJTM107 (3860 ± 30 BP)	PLTM313 (3620 ± 30 BP)	HJTM109 (3600 ± 30 BP)	JXNTM2 (3550- 3050 BP)	JXNTM23 (3127 ± 54 BP)	LGM79 (2259 ± 79 BP)	LGM41 (2250- 2050 BP)	HJTM115 (2236 ± 79 BP)
rs12446160	16:48324457	Intron	G	A	LONP2	0.579	A(11)	A(4)	A(3)	A(13)	A(3)	0	A(3)	0	A(4)	A(4)	0	0	0	A(4)
rs6500380	16:48341866	Intron	G	A	LONP2	0.972	A(6)	0	A(8)	A(16)	A(3)	A(6)	A(3)	A(4)	0	A(6)	A(3)	A(4)	A(5)	A(7)
rs8052293	16:48348611	Intron	G	A	LONP2	0.973	A(8)	A(3)	A(6)	A(15)	A(8)	A(4)	A(4)	A(10)	0	0	0	A(4)	A(6)	A(8)
rs111915588	16:48276823	Intron	G	A	LONP2	0.968	A(4)	0	0	A(3)	0	A(3)	0	0	0	0	0	0	0	A(4)
rs1982782	16:48404573	Intron	G	A	SAH1	0.973	A(6)	A(2)	0	A(11)	A(4)	A(4)	A(1)	A(4)	A(1)	A(2)	A(1)	A(3)	A(2)	A(2)
rs1744618	11:61861850	Intron	C	T	FADS2	0.992	T(6)	T(2)	0	T(10)	T(11)	T(7)	T(2)	T(5)	0	T(3)	T(2) C(1)	T(1)	T(2)	T(8)
rs174450	11:61874070	Intron	G	T	FADS3	0.743	T(4) G(4)	0	G(2)	G(3)	G(1)	T(4)	T(2) G(2)	T(2)	0	T(2)	T(2) G(1)	T(3)	T(1)	G(1)
rs174455	11:61888645	Intron	G	A	FADS3	0.749	G(3) A(2)	G(1)	G(11)	G(10)	G(2)	A(5)	0	A(8)	A(2)	A(2)	G(1)	A(1)	A(1)	G(5) A(3)
rs1402467	2:108378352	Missense	G	C	SULT1C4	0.929	C(6)	C(2)	0	C(5) G(3)	C(2)	C(3)	G(1)	C(4)	C(2)	0	C(2)	C(1)	G(2)	C(3)
rs6747116	2:25746860	Intron	T	C	ASXL2	0.859	C(9)	C(1)	C(6)	C(4)	C(1)	C(3) T(1)	C(1)	C(5)	C(1)	C(4) T(2)	C(2)	0	C(3)	C(3)
rs11897652	2:25767112	Intron	C	T	ASXL2	0.807	T(6)	0	0	T(2)	T(1)	T(1)	0	0	0	C(1)	T(3)	0	0	T(1)
rs1539172	9:15784633	Missense	G	A	CCDC171	0.678	A(7)	A(2)	0	A(2) G(2)	G(1)	A(3)	0	A(3)	0	A(1) G(2)	0	A(2)	A(3) G(1)	A(2)
rs1229984	4:98318162	Missense	C	T	ADH1B	0.657	C(6)	0	C(2)	T(3) C(1)	C(3)	T(1) C(1)	T(1) C(2)	C(5)	0	C(2)	C(3)	T(4)	T(2) C(1)	C(1)
rs35333426	4:99302879	Intron	G	A	ADH1B	0.706	G(6)	G(1)	G(1)	G(7) A(4)	G(2)	A(5)	A(1)	G(4)	0	G(1)	0	A(1)	0	A(1)
rs1042026	4:99307309	Intron	T	C	ADH1B	0.711	T(6)	T(2)	0	T(3)	T(3) C(2)	C(5)	0	T(3)	C(1)	T(2)	0	C(2)	0	C(2)
rs1813977	4:98234994	Upstream	A	G	ADH1A	0.703	A(9)	A(1)	0	A(2) G(2)	A(2)	G(2)	G(2)	A(2)	G(1)	0	0	G(3)	G(1)	G(5)
rs1039151	4:98271625	Intron	G	T	ADH1A	0.691	G(5)	G(1)	0	0	0	0	0	G(3)	0	G(1)	T(1)	T(1)	T(1)	T(2)
rs3819197	4:98279352	Intron	C	T	ADH1A	0.688	C(7)	0	C(2)	C(3) T(3)	C(1)	T(3)	0	C(3)	T(1)	C(3)	C(1) T(1)	T(1)	T(2)	T(2)
rs284784	4:98414717	Intron	C	A	ADH7	0.544	C(3)	0	C(2)	C(5)	C(1)	C(3)	0	C(5) T(1)	C(1)	C(1)	C(1)	A(2)	C(1)	C(1)
rs1827567	4:98414945	Intron	G	A	ADH7	0.544	G(1)	0	0	G(4)	A(1)	G(2) A(1)	0	G(1)	0	G(1)	G(1)	A(1)	0	G(1)
rs284793	4:98409137	Intron	A	G	ADH7	0.564	G(3)	A(1)	A(1)	G(9) A(3)	G(1) A(4)	G(7)	G(1) A(2)	G(4) A(1)	0	0	G(2)	A(5)	A(1)	G(4) A(2)

Alleles in the table represent allele types identified in ancient DNA samples.
The numbers in parentheses represent the number of reads that matched ancestral or derived alleles.
The 0 indicates that no genotype data are available.
AA ancestral allele, DA derived allele, DAF derived allele frequency.

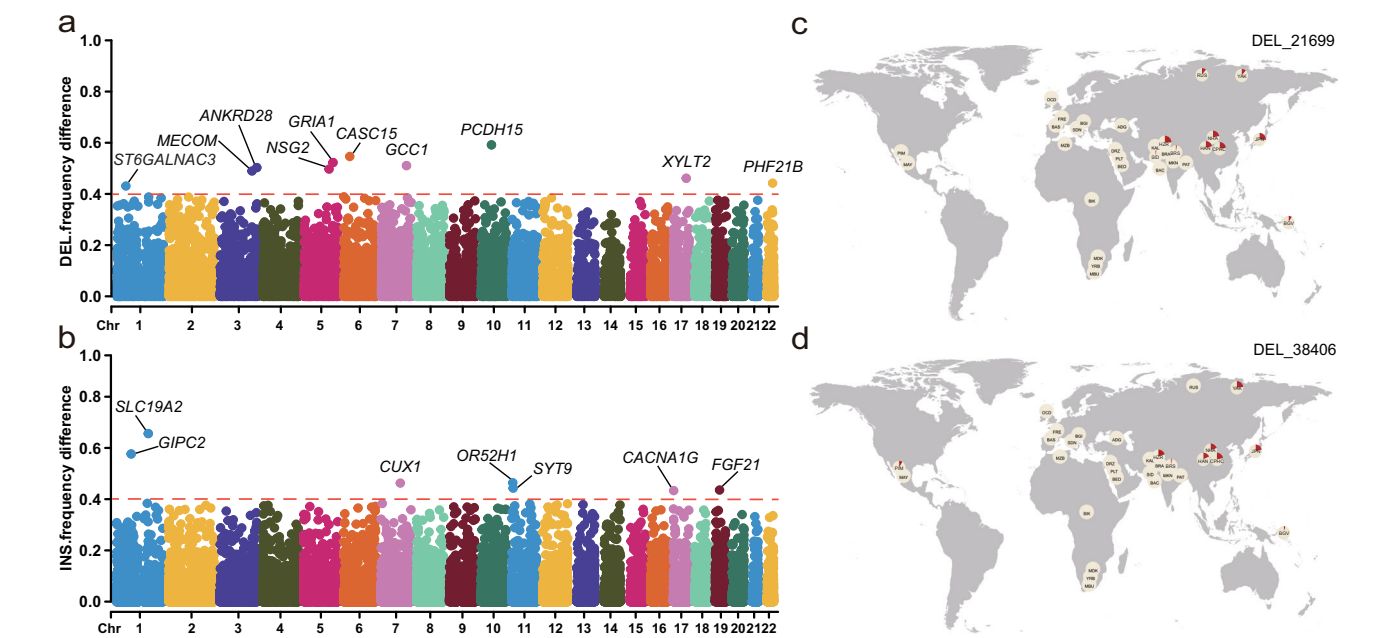


Fig. 5 | Specific and high-frequency SVs in the CPHC. Manhattan plot of Deletions (a) and Insertions (b) in the CPHC. The frequency distribution map of DEL_21699 (c) and DEL_38406 (d). The population abbreviations are given in Supplementary Data 4.

Table 2 | Specific and high-frequency structural variants in the CPHC

SV ID	Chr:Pos (GRCh38)	Length (bp)	AF					Gene
			CPHC	East Asia	Central South Asia	Europe	Africa	
Specific SVs								
DEL_21699	20:63085519	170	0.222	0.212	0.031	0.018	0	LINC01749
DEL_38406	9:127961761	92	0.223	0.218	0.031	0	0	FAM102A
DEL_24082	3:75452121	72	0.284	0.244	0.011	0.004	0	ENPP7P2
INS_42051	4:19588797	56	0.257	0.205	0.019	0	0	–
DEL_29932	5:147015151	2592	0.203	0.187	0.017	0.004	0	PPP2R2B
INS_46471	5:38821455	96	0.222	0.193	0.017	0.007	0.035	OSMR-DT
INS_05228	10:9295452	77	0.211	0.121	0.022	0.007	0	–
INS_18720	15:79492084	119	0.317	0.241	0.022	0	0	–
INS_21702	16:89640547	61	0.516	0.516	0.053	0.011	0	DPEP1
High-frequency SVs								
INS_27791	19:48758867	639	0.924	0.921	0.441	0.508	0.471	FGF21
INS_03097	1:169473717	331	0.751	0.647	0.159	0.011	0.018	SLC19A2
DEL_01111	1:76303034	78	0.635	0.632	0.301	0.234	0.011	ST6GALNAC3
INS_02070	1:78141369	319	0.814	0.745	0.296	0.250	0.106	GIPC2
DEL_23618	3:15803663	1566	0.528	0.511	0.185	0.088	0.176	ANKRD28
DEL_24930	3:169550812	279	0.734	0.661	0.293	0.189	0.176	MECOM
DEL_29988	5:153606868	92	0.763	0.674	0.401	0.088	0.091	GRIA1
DEL_30151	5:174213439	93	0.587	0.562	0.221	0.066	0.261	NSG2
DEL_30693	6:22050711	3406	0.677	0.492	0.080	0.029	0.267	CASC15
DEL_34494	7:127574915	3011	0.655	0.629	0.279	0.281	0.011	GCC1
INS_54902	7:102216655	65	0.811	0.752	0.378	0.395	0.213	CUX1
DEL_03573	10:54133924	448	0.869	0.858	0.326	0.310	0.119	PCDH15
INS_08244	11:5549455	100	0.846	0.834	0.413	0.293	0.523	OR52H1
INS_08261	11:7331215	330	0.909	0.931	0.528	0.389	0.300	SYT9
DEL_14351	17:50350142	384	0.885	0.824	0.586	0.459	0.375	XYLT2
INS_23258	17:50588357	111	0.841	0.820	0.506	0.361	0.477	CACNA1G
DEL_23080	22:44904523	342	0.677	0.583	0.251	0.197	0.182	PHF21B

human diet gradually shifted to a starch-rich plant-based diet, which may have exerted significant selective pressure on insulin secretion, synthesis of long-chain polyunsaturated fatty acids and other metabolic processes^{3,40}. In the current study, we found several of the dietary metabolism-associated genes (*LONP2*, *FADS2*, *FADS3*, *SIAH1*, *SULT1C4*, *CCDC171*, *ASXL2*, *ADH1A*, *ADH1B* and *ADH7*) show strong selection signals in the CPHC. These genes may be important for the dietary evolutionary history of East Asian populations. *e.g.* positive selection signaling for *LONP2* and *FADS2*, which have been implicated in promoting polyunsaturated fatty acid synthesis in humans. The increased capacity for endogenous polyunsaturated fatty acid synthesis is also a consequence of the shift from a carnivorous diet to a diet rich in starch crops^{25,40}. Evidence from ancient DNA suggests that signals of metabolic selection in the CPHC may have coincided with the expansion of agriculture in the region. For example, the occurrence of derived allele of 10 SNPs in 6 candidate genes (*LONP2*, *FADS2*, *SIAH1*, *SULT1C4*, *CCDC171* and *ASXL2*), date back to approximately 6,000 BP, coincides with the cultivation of millet in the Central Plains^{22,51}. This suggests that the high frequency of metabolism-related mutations may be the result of CPHC dietary adaptation and play an important role in glycolipid metabolism in East Asian populations.

The Yangtze River basin is an important area for rice domestication, unlike the Yellow River basin where millet and wheat are the main crops⁶⁷. Selection signals for the *ADH* genes (*e.g.* *ADH1A*, *ADH1B*, *ADH4*, *ADH5*, *ADH6* and *ADH7*) associated with the rice diet are showed in the local populations^{4,45}. In the current study, *ADH1B*, *ADH1A* and *ADH7* also showed a significant selection signal in the CPHC. However, the mutation in these genes was mainly as ancestral type in the ancient DNA samples from the Central Plains (Table 1). One possible explanation for this is that the *ADH* mutation in the modern populations of the Central Plains is the result of more recent mixing of the populations. The alcohol metabolism mutations in the Yangtze River basin (*e.g.* the rs1229984_T allele appeared around 7000 years ago) differed from the glycolipid metabolism mutations in the Yellow River basin (*e.g.* 6000 years ago, Table 1), suggesting that there are differences in crop preference and timing of domestication between northern and southern East Asian populations. This suggests that different populations in northern and southern China may have undergone different adaptive processes due to differences in crop domestication during the history of agricultural development.

The SVs found in this study will greatly increase the resources of structural variants in East Asia, suggesting that the SVs may be widespread in East Asia, implying the presence of these SVs may have contributed to the adaptive evolution of the phenotype in East Asian populations. Interestingly, we also found some high-frequency SVs associated with human glycolipid metabolism in the CPHC, and these were the major alleles (*AF* > 0.5) in East Asia. For example, *INS_27791* (*FGF21*) and *INS_03097* (*SLC19A2*) showed high frequencies and may be associated with adipose tissue production and glucose secretion^{60,61}. It can be assumed that the CPHC had a higher capacity for glycolipid metabolism, which may also have been adapted to the diet brought about by agricultural civilization. Some other East Asian-specific structural variants, such as *DEL_21699* (*TCFL5*) and *DEL_38406* (*FAM102A*) may be associated with esophageal squamous carcinoma and primary angle-closure glaucoma, which are highly prevalent diseases in East Asia^{57,58}. More interestingly, the diseases resulting from these two SVs may also be related to glycolipid metabolic adaptations. For example, it is well known that most wheat-based high-starch (hot food) diets in the Central Plains have higher temperatures. *TCFL5* and *FAM102A* can promote human M2 macrophage polarization and osteoclast differentiation, which increases glycolipid metabolism and reduces the energy burden of hot foods^{68,69}. However, prolonged exposure to hot food can damage the epithelial cells of the esophageal lining and cause esophageal squamous carcinoma⁵⁷. Meanwhile, hot foods induce an increase in oxidative stress, which promotes the development of primary angle-closure glaucoma^{70–72}.

In conclusion, we comprehensively characterized the genomic variants of CPHC, demonstrating the CPHC has a unique genetic component and higher genetic diversity compared with other East Asian populations. The

CPHC genome contains variants related to human glycolipid metabolism that may be the result of adaptation to agricultural civilization, providing important clues for further studies on the adaptive evolution of diet in agricultural populations.

Methods

Samples collection and whole-genome sequencing

We collected blood samples from 492 Han Chinese individuals from 16 cities in Henan Province, China. All volunteers were mainly within three generations of Han population, consented to blood collection and signed informed consent forms. Genomic DNA was extracted from blood using the Serum/Plasma Circulating DNA Kit (Tiangen, Beijing, China) according to the manufacturer's instructions. The DNA samples are then subjected to library construction and quality control according to standard Illumina library preparation protocols. Finally, all DNA libraries were sequenced using the HiSeq 2500 sequencing platform, and whole genome sequencing data with an average depth of $25.54 \times$ were obtained. Written informed consent was obtained from each individual before their inclusion in the study. All protocols of this study were approved by the Institutional Review Board of Zhengzhou University (ZZUIRB2009-002 and ZZUIRB2022-14) and the Institutional Review Board of The Second Affiliated Hospital of Zhengzhou University (2020013).

Variants calling, SV genotyping and annotation

Variants calling was followed the standard Genome Analysis Toolkit (GATK version 4.2.6) pipeline⁷³. We download the human reference genome assembly 38 from the resource bundle on the GATK website (<ftp://gsapubftpanonymous@ftp.broadinstitute.org/bundle/>). In brief, FastQC version 0.11.3 was used to check the quality of the sequencing data and to remove adapter sequences and low-quality data. Next, the sequencing reads were aligned to the human reference genome assembly 38 using the BWA-MEM version 0.7.15⁷⁴, and the results were converted into BAM files by SAMtools version 1.9⁷⁵. We further sorted and de-duplicated the BAM files using SAMtools version 1.9 and performed base quality score recalibration (BQSR and ApplyBQSR) on the BAM files to improve the reliability and accuracy of variants calling. The variants were then called for each sample's BAM file with GATK HaplotypeCaller to generate gVCF files and merged using GenotypeGVCFs to generate a population VCF file containing all 492 individuals. Finally, the merged VCF file was recalibrated for variant quality scores (VQSR and ApplyVQSR), and all SNPs and INDELs from the training set were recalibrated with 99% true sensitivity to generate the final variant dataset.

For SVs genotyping, we used the Paragraph, an accurate genotype that models SVs using sequence graphs and annotations¹¹. The CPHC and 54 world populations in HGP were included in our study. The set of high-confidence SVs we selected for genotyping was 107,590 SVs (insertions and deletions) based on fully phased genome assembly data from 35 individuals and used RepeatMasker to annotate the SVs in CPHC^{13,76}. Finally, we used Samplot and IGV to manually check the breakpoints of these structural variants in the CPHC^{77,78}.

Before annotating the variants, we conducted the following steps to filter low-quality SNPs. First, the SNPs with a genotype quality of less than 10 in more than 50% of the 492 CPHC individuals were removed. Subsequently, we calculated Hardy-Weinberg equilibrium "*p* value" and removed variants with *p* < 10^{-6} using PLINK version 1.09⁷⁹. Finally, we removed all multi-allelic variants using BCFtools version 1.12⁸⁰. Approximately 5.63 million SNPs and 540,000 INDELs were removed after these filtering steps. The final remaining SNPs and all SVs were annotated with genes, variant types and other information using the variant effect prediction tool¹².

Principal component analysis (PCA)

For the PCA analysis of SNPs, we downloaded the SNPs dataset from the 1000 Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/) and conducted Hardy-Weinberg equilibrium (HWE).

Subsequently, we removed variants with $p < 10^{-6}$, site missing rate > 0.05 , and $MAF < 0.05$. Finally, we extracted the SNPs that overlapped between the CPHC and the 1KGP and merged them into a single VCF file. For the merged VCF file, we used PLINK version 1.09 --indep-pairwise 50 10 0.5 to exclude linked variants ($R^2 > 0.5$), retaining approximately 0.17 million variants for PCA. However, for the PCA analysis of SVs, due to the insufficient number, we only removed variants with $p < 10^{-4}$, site missing rate > 0.2 , and individual absence rate > 0.5 . Next, we combined the SVs that were present in both the CPHC and the Human Genome Diversity Project (HGDP) into a single VCF file. Finally, the PCA of SNPs and SVs in the Central Han Chinese were conducted with the EIGENSOFT tool⁸¹ and visualized using the ggplot2 package in R.

Admixture analysis and LD decay

The kinship index between individuals was calculated using the King tool⁸², and the data of 20 individuals with kinship < 0.0442 were excluded. After performing individual filtering of the SNPs dataset, we used PLINK version 1.09 --indep-pairwise 50 10 0.5 to exclude highly linked variants ($R^2 > 0.5$), and about 0.13 million SNPs were included. Finally, we conducted admixture analysis in the CPHC as well as the CDX, KHV, CHS, CHB, JPT, CEU, and YRI populations from the 1KGP from $K = 2-7$ with ADMIXTURE tool, repeating the analysis 10 times for each K value under different seeds¹⁷. The K-value is a subpopulation parameter assumed by the Admixture tool for population structure analyses. We also calculated the R^2 within a maximum distance of 300 kb for the above populations using PopLDdecay tool to assess the extent of linkage disequilibrium⁸³.

Effective population size

To explore the timing of divergence between the CPHC and other populations, we analyzed effective population size of CPHC, CDX, KHV, CHS, CHB, JPT, CEU and YRI in the 1KGP. We phased the VCF file of each population using SHAPEIT version 4.2.2⁸⁴. And then the effective population size was analyzed by using MSMC2 with a generation time of 30 years while keeping other parameters at their default values¹⁹.

Detection of positive selection in the CPHC

We used VCFtools version 0.1.1379 to compute the whole-genome F_{ST} with a sliding window of 50 kb between the CPHC and the Cambodian population (JPZ)^{16,85}. The dataset of the Cambodian population was obtained from previous studies²³ and merged with the variants in the CPHC. We obtained a total of 5400 genomic regions, from which we identified the top 5% regions with the highest F_{ST} value as potential adaptation signals. The YRI and CEU populations in the 1KGP were selected as outgroups and 13,035 SNPs with $F_{ST} > 0.1$ and higher allele frequencies in the CPHC were selected as candidate adaptation signals. We also use the phased VCF file of the CPHC and JPZ to calculate XP-EHH value and the top 5% (XP-EHH > 1.697) were treated as the selected regions²⁴.

To investigate specific SVs in East Asian populations, we compared them with non-East Asian populations (Africans, Europeans, Americans, Middle Eastern, Central South Asians, and Oceanian populations) from HGDP. The Han Chinese population from the HGDP was also compared with the CPHC to ensure consistency across sequencing platforms. Finally, we chose SVs with frequencies greater than 0.2 in the CPHC and less than 0.02 in other non-East Asian populations of the HGDP as East Asian-specific variants. The SVs with frequency difference of more than 0.4 between the CPHC and the non-East Asian populations of the HGDP were selected as the high-frequency variants in East Asia.

Ancient DNA analysis

We analyzed the 21 typical mutation related to dietary metabolism using 14 ancient DNA samples from the Central Plains region (Lingbao city, Xiaowu site; Xingyang county, Wanggou village; Zhengzhou, Dahecun; Yuzhou city, Wadian site; Huaiyang, Pingliangtai site; Luohe city, Haojiatai site; Jiazuo

city, Niecun site) (Supplementary Fig. 6). The genotype calling was performed according to previous published protocols^{51,86}.

Statistics and reproducibility

Software used for the data analysis of this study is as follows: bcftools (<https://samtools.github.io/bcftools/bcftools.html>), FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), BWA, SAMtools (<https://github.com/samtools/>), GATK, VCFtools (<http://vcftools.sourceforge.net/>), PLINK, Repeatmasker (<https://www.repeatmasker.org/>), VEP, Samplot (<https://github.com/ryanlayer/samplot>), IGV, R, EIGENSOFT, ADMIXTURE, PopLDdecay (<https://github.com/BGI-shenzhen/PopLDdecay>), TreeMix (<https://github.com/carolindahms/TreeMix>), MSMC2. The databases used in the current study are described in the Methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequence data reported in this paper have been deposited in the Genome Sequence Archive (GSA)⁸⁷ in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA006704 (<https://ngdc.cncb.ac.cn/gsa-human/>). The BAM data are available under restricted access for privacy protection and access can be obtained by application on the website.

Received: 5 September 2024; Accepted: 17 February 2025;

Published online: 22 February 2025

References

- Jin, L. & Su, B. Natives or immigrants: modern human origin in East Asia. *Nat. Rev. Genet.* **1**, 126–133 (2000).
- Su, B. et al. Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last Ice Age. *Am. J. Hum. Genet.* **65**, 1718–1724 (1999).
- Luca, F., Perry, G. H. & Di Rienzo, A. Evolutionary adaptations to dietary changes. *Annu. Rev. Nutr.* **30**, 291–314 (2010).
- Cong, P. K. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).
- Cao, Y. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
- Zhang, P. et al. NyuWa Genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.* **37**, 110017 (2021).
- Tao, L. et al. Ancient genomes reveal millet farming-related demic diffusion from the Yellow River into southwest China. *Curr. Biol.* **33**, 4995–5002 e4997 (2023).
- Feng, J., Hu, P., Zhang, F. & Sun, J. HCHs and DDTs in Yellow River of Henan section—a typical agricultural area in China: levels, distributions and risks. *Environ. Geochem. Health* **38**, 1241–1253 (2016).
- Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
- Altshuler, D. M. et al. A global reference for human genetic variation. *Nature* **526**, 68–6 (2015).
- Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
- McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17** (2016). <https://doi.org/10.1186/s13059-016-0974-4>
- Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinform.* **25** (2009). <https://doi.org/10.1002/0471250953.bi0410s25>

14. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *Plos Genet.* **8**, e1002967 (2012).
15. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
16. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
17. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
18. Lu, Y. et al. Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapid LD decay in tropical than temperate germplasm in maize. *PLoS One* **6**, e24861 (2011).
19. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
20. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
21. Zhao, Z. J. New archaeobotanic data for the study of the origins of agriculture in China. *Curr. Anthropol.* **52**, S295–S306 (2011).
22. Lu, H. et al. Earliest domestication of common millet (*Panicum miliaceum*) in East Asia extended to 10,000 years ago. *Proc. Natl. Acad. Sci. USA* **106**, 7367–7372 (2009).
23. Zhang, X. et al. The distinct morphological phenotypes of Southeast Asian aborigines are shaped by novel mechanisms for adaptation to tropical rainforests. *Natl. Sci. Rev.* **9**, nwab072 (2022).
24. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
25. Chu, K. Y., Mellet, N., Thai, L. M., Meikle, P. J. & Biden, T. J. Short-term inhibition of autophagy benefits pancreatic beta-cells by augmenting ether lipids and peroxisomal function, and by countering depletion of n-3 polyunsaturated fatty acids after fat-feeding. *Mol. Metab.* **40**, 101023 (2020).
26. Harwood, J. L. Polyunsaturated fatty acids: conversion to lipid mediators, roles in inflammatory diseases and dietary sources. *Int. J. Mol. Sci.* **24**, 8838 (2023).
27. Vosa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
28. Zhu, Z. et al. E3 ubiquitin ligase Siah1 aggravates NAFLD through Scp2 ubiquitination. *Int. Immunopharmacol.* **124**, 110897 (2023).
29. Galora, S. et al. Association of rs1466535 LRP1 but not rs3019885 SLC30A8 and rs6674171 TDRD10 gene polymorphisms with abdominal aortic aneurysm in Italian patients. *J. Vasc. Surg.* **61**, 787–792 (2015).
30. Schumacher, J. A. et al. SH2 domain protein E and ABL signaling regulate blood vessel size. *PLoS Genet.* **20**, e1010851 (2024).
31. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
32. Luo, H. et al. Recent positive selection signatures reveal phenotypic evolution in the Han Chinese population. *Sci. Bull.* **68**, 2391–2404 (2023).
33. Gurunluoglu, K. et al. Global gene expression profiling in congenital diaphragmatic hernia (CDH) patients. *Funct. Integr. Genomics* **22**, 359–369 (2022).
34. Caliskan, Y. et al. LIMS1 risk genotype and T cell-mediated rejection in kidney transplant recipients. *Nephrol. Dial. Transplant.* **36**, 2120–2129 (2021).
35. Jiang, J., Wang, Y. E., Palazzo, A. F. & Shen, Q. Roles of nucleoporin RanBP2/Nup358 in acute necrotizing encephalopathy type 1 (ANE1) and viral infection. *Int. J. Mol. Sci.* **23**, 3548 (2022).
36. Navone, F. et al. Expression of KIF3C kinesin during neural development and in vitro neuronal differentiation. *J. Neurochem.* **77**, 741–753 (2001).
37. Emery, A. C., Eiden, M. V., Mustafa, T. & Eiden, L. E. Rapgef2 connects GPCR-mediated cAMP signals to ERK activation in neuronal and endocrine cells. *Sci. Signal* **6**, ra51 (2013).
38. Kurogi, K. et al. Human Cytosolic Sulphotransferase SULT1C3: genomic analysis and functional characterization of splice variant SULT1C3a and SULT1C3d. *J. Biochem.* **162**, 403–414 (2017).
39. Kamberov, Y. G. et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702 (2013).
40. Kothapalli, K. S. et al. Positive selection on a regulatory insertion-deletion polymorphism in FADS2 influences apparent endogenous synthesis of arachidonic acid. *Mol. Biol. Evol.* **33**, 1726–1739 (2016).
41. Coltell, O. et al. Genome-wide association study for serum omega-3 and omega-6 polyunsaturated fatty acids: exploratory analysis of the sex-specific effects and dietary modulation in mediterranean subjects with metabolic syndrome. *Nutrients* **12**, 310 (2020).
42. Koletzko, B. et al. FADS1 and FADS2 polymorphisms modulate fatty acid metabolism and dietary impact on health. *Annu. Rev. Nutr.* **39**, 21–44 (2019).
43. Fraga, L. N., Milenkovic, D., Lajolo, F. M. & Hassimotto, N. M. A. Association between single nucleotide polymorphisms of SULT1A1, SULT1C4, ABCC2 and phase II flavanone metabolites excretion after orange juice intake. *Nutrients* **14** <https://doi.org/10.3390/nu14183770> (2022).
44. Izawa, T. et al. ASXL2 regulates glucose, lipid, and skeletal homeostasis. *Cell Rep.* **11**, 1625–1637 (2015).
45. Peng, Y. et al. The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evol. Biol.* **10**, 15 (2010).
46. Yang, Z. et al. A genetic mechanism for convergent skin lightening during recent human evolution. *Mol. Biol. Evol.* **33**, 1177–1187 (2016).
47. Pu, Y. et al. Weakened tanning ability is an important mechanism for evolutionary skin lightening in East Asians. *J. Genet. Genomics* <https://doi.org/10.1016/j.jgg.2024.03.001> (2024).
48. Yang, Z. et al. Darwinian positive selection on the pleiotropic effects of KITLG explain skin pigmentation and winter temperature adaptation in Eurasians. *Mol. Biol. Evol.* **35**, 2272–2283 (2018).
49. Liu, J. M., Bitsue, H. K. & Yang, Z. H. Skin colour: A window into human phenotypic evolution and environmental adaptation. *Mol. Ecol.* <https://doi.org/10.1111/mec.17369> (2024).
50. Ma, X. & Xu, S. Archaic introgression contributed to the pre-agriculture adaptation of vitamin B1 metabolism in East Asia. *iScience* **25**, 105614 (2022).
51. Ning, C. et al. Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat. Commun.* **11** 2700 (2020).
52. Ma, H. et al. Genomic substructure in Yellow River Basin farmers during the transitional Yangshao-Longshan period. *J. Archaeol. Sci.* <https://doi.org/10.1016/j.jas.2025.106178> (2025).
53. Zhang, J., Ling, X., Fang, C. & Ma, J. Identification and validation of an eight-lncRNA signature that predicts prognosis in patients with esophageal squamous cell carcinoma. *Cell Mol. Biol. Lett.* **27**, 39 (2022).
54. Lin, C. et al. circRNA TCFL5 promote esophageal cancer progression by modulating M2 macrophage polarization via the miR-543-FMNL2 axis. *J. Oncol.* **2022**, 5075615 (2022).
55. Wang, L. D. et al. Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. *Nat. Genet.* **42**, 759–763 (2010).
56. Dawsey, S. M. et al. Squamous dysplasia and early esophageal cancer in the Linxian region of China: distinctive endoscopic lesions. *Gastroenterology* **105**, 1333–1340 (1993).
57. Abnet, C. C., Arnold, M. & Wei, W. Q. Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* **154**, 360–373 (2018).

58. Tham, Y. C. et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040. *Ophthalmology* **121**, 2081–2090 (2014).
59. Shi, H. et al. In-depth analysis of eight susceptibility loci of primary angle closure glaucoma in Han Chinese. *Exp. Eye Res.* **202**, 108350 (2021).
60. Lin, Z. et al. Adiponectin mediates the metabolic effects of FGF21 on glucose homeostasis and insulin sensitivity in mice. *Cell Metab.* **17**, 779–789 (2013).
61. Sun, C. et al. Recovered insulin production after thiamine administration in permanent neonatal diabetes mellitus with a novel solute carrier family 19 member 2 (SLC19A2) mutation. *J. Diab.* **10**, 50–58 (2018).
62. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
63. Medina-Munoz, S. G. et al. Demographic modeling of admixed Latin American populations from whole genomes. *Am. J. Hum. Genet.* **110**, 1804–1816 (2023).
64. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *Plos Genet.* **5**, e1000695 (2009).
65. Browning, S. R. et al. Ancestry-specific recent effective population size in the Americas. *Plos Genet.* **14**, e1007385 (2018).
66. Huang, X., Huang, S., Han, B. & Li, J. The integrated genomics of crop domestication and breeding. *Cell* **185**, 2828–2839 (2022).
67. Zhang, J. P. et al. Rice's trajectory from wild to domesticated in East Asia. *Science* **384**, 901–906 (2024).
68. Indo, Y. et al. Metabolic regulation of osteoclast differentiation and function. *J. Bone Min. Res.* **28**, 2392–2399 (2013).
69. Lin, S. et al. Targeting parvalbumin promotes M2 macrophage polarization and energy expenditure in mice. *Nat. Commun.* **13**, 3301 (2022).
70. Zhong, L. et al. High starch in diet leads to disruption of hepatic glycogen metabolism and liver fibrosis in largemouth bass (*Micropterus salmoides*), which is mediated by the PI3K/Akt signaling pathway. *Front. Physiol.* **13**, 880513 (2022).
71. Musiol, S. et al. Dietary digestible carbohydrates are associated with higher prevalence of asthma in humans and with aggravated lung allergic inflammation in mice. *Allergy* **78**, 1218–1233 (2023).
72. Li, S. et al. Relationship between oxidative stress biomarkers and visual field progression in patients with primary angle closure glaucoma. *Oxid. Med. Cell Longev.* **2020**, 2701539 (2020).
73. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
74. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
75. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
76. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, 48–4 (2021).
77. Belyeu, J. R. et al. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol.* **22**, 161 (2021).
78. Robinson, J. T., Thorvaldsdottir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34 (2017).
79. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
80. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
81. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
82. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
83. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M. & Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
84. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
85. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
86. Zhu, K. et al. Protocol for a comprehensive pipeline to study ancient human genomes. *STAR Protoc.* **5**, 102985 (2024).
87. Chen, T. et al. The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteom. Bioinforma.* **19**, 578–583 (2021).

Acknowledgements

We are grateful to all voluntary donors in this study. This work was supported by the National Natural Science Foundation of China (32370664 and 32070579 to Z.Y.), the Natural Science Foundation of Henan (222300420067 to Z.Y.), the Program for Science & Technology Innovation Talents in Universities of Henan Province (25HASTIT049 to Z.Y.), the Collaborative Innovation Project of Zhengzhou (Zhengzhou University) (20XTZX05014 to J.G. and 18XTZX12004 to W.T.) and Innovation and Entrepreneurship Training Program (2024cxcy091 to Z.Y.). This study is supported by the National Supercomputing Center in Zhengzhou.

Author contributions

Xiaoyang Qiao: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—Original draft, Writing—Review & Editing. Jianxiang Shi: Methodology, Software, Validation, Formal analysis, Investigation, Data curation. Hongen Xu: Methodology, Software, Validation, Formal analysis, Investigation, Data curation. Kai Liu: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Resources. Youwei Pu: Methodology, Software, Validation, Formal analysis, Writing—Original draft, Writing—Review & Editing. Xia Xue, Wangshan Zheng and Yongbo Guo: Methodology, Resources. Hao Ma: Software, Validation, Resources. Chuan-Chao Wang: Software, Validation, Resources. Habbom K. Bitsue: Validation, Resources. Xiaoyu Xu, Shanshan Wang, Jingru Zhao, Xiangqian Guo, Xinyue Hou, Xinwei Wang, Lei Peng, Zan Qiu and Bing Su: Resources. Wenxue Tang: Conceptualization, Methodology, Supervision, Funding acquisition. Yaoxi He: Conceptualization, Methodology, Supervision. Jiancheng Guo: Conceptualization, Methodology, Supervision, Funding acquisition. Zhaohui Yang: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing—Original draft, Writing—Review & Editing.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-07760-2>.

Correspondence and requests for materials should be addressed to Wenxue Tang, Yaoxi He, Jiancheng Guo or Zhaohui Yang.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Mengtan Xing and Aylin Bircan. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025