Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Genetic history of esophageal cancer group in southwestern China revealed by Y-chromosome STRs and genomic evolutionary connection analysis

Lihua Jia [a,1], Mengge Wang [b,c,**,1], Shuhan Duan [a,c,1], Jianghua Chen [a], Mei Zhao [a], Simeng Ji [d], Bingbing Lv [d], Xiucheng Jiang [a,c], Guanglin He [b,c,*], Junbao Yang [a,d,***]

[a] *Institute of Basic Medicine and Forensic Medicine, North Sichuan Medical College and Center for Genetics and Prenatal diagnosis, Affiliated Hospital of Northern Sichuan Medical College, Nanchong, Sichuan, 637007, China*
[b] *Center for Archaeological Science, Sichuan University, Chengdu, 610000, China*
[c] *Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China*
[d] *School of Laboratory Medicine, North Sichuan Medical College, Nanchong, Sichuan, 637000, China*

ARTICLE INFO

Keywords:
Genetic diversity
Population origin
Disease susceptibility
Han Chinese
Esophageal cancer

ABSTRACT

Genetic and environmental factors play crucial roles in the development of esophageal cancer (EC) and contribute uniquely or cooperatively to human cancer susceptibility. Sichuan is located in the interior of southwestern China, and the northern part of Sichuan is one of the regions with a high occurrence of EC. However, the factors influencing the high incidence rate of EC in the Sichuan Han Chinese population and its corresponding genetic background and origins are still poorly understood. Here, we utilized genome-wide single nucleotide polymorphisms (SNPs) and Y-chromosome short tandem repeats (Y-STRs) to characterize the genetic structure, connection, and origin of cancer groups and general populations. We generated Y-STR-based haplotype data from 214 Sichuan individuals, including the Han Chinese EC population and a control group of Han Chinese individuals. Our results, obtained from Y-STR-based population statistical methods (analysis of molecular variance (AMOVA), principal component analysis (PCA), and phylogenetic analysis), demonstrated that there was a genetic substructure difference between the EC population in the high-incidence area of northern Sichuan Province and the control population. Additionally, there was a strong genetic relationship between the EC population in the northern Sichuan high-incidence area and those at high risk in both the Fujian and Chaoshan areas. In addition, we obtained high-density SNP data from saliva samples of 60 healthy Han Chinese individuals from three high-prevalence areas of EC in China: Sichuan Nanchong, Fujian Quanzhou, and Henan Xinxiang. As inferred from the allele frequency of SNPs and sharing patterns of haplotype segments, the evolutionary history and admixture events suggested that the Han population from Nanchong in northern Sichuan Province shared a close genetic relationship with the Han populations from Xinxiang in Henan Province and Quanzhou in Fujian Province, both of which are regions with a high prevalence of EC. Our study illuminated the genetic profile and

* Corresponding author. Center for Archaeological Science, Sichuan University, Chengdu, 610000, China.
** Corresponding author. Center for Archaeological Science, Sichuan University, Chengdu, 610000, China.
*** Corresponding author. Institute of Basic Medicine and Forensic Medicine, North Sichuan Medical College and Center for Genetics and Prenatal diagnosis, Affiliated Hospital of Northern Sichuan Medical College, Nanchong, Sichuan, 637007, China.
E-mail addresses: Menggewang2021@163.com (M. Wang), guanglinhescu@163.com (G. He), yjb3589@vip.163.com (J. Yang).

[1] Lihua Jia, Mengge Wang, and Shuhan Duan contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e29867
Received 8 December 2023; Received in revised form 9 April 2024; Accepted 16 April 2024
Available online 26 April 2024

connection of the Northern Sichuan Han population and enriched the genomic resources and features of the Han Chinese populations in China, especially for the Y-STR genetic data of the Han Chinese EC population. Populations living in different regions with high incidences of EC may share similar genetic backgrounds, which offers new insights for further exploring the genetic mechanisms underlying EC.

1. Introduction

Molecular genetic markers of Y-chromosome short tandem repeats (Y-STRs) and single nucleotide polymorphisms (SNPs) provide critical clues for research on the kinship, patrilineal biogeographic ancestry, and genetic structure of different populations [1,2]. Y-STRs are located on the Y chromosome nonrecombination region and have a patrilineal inheritance mode [3]. Compared to the other parts of the genome, a distinctive feature of the Y chromosome is its high degree of geographic differentiation [4]. Rosser ZH et al. reported a highly significant partial correlation between genetics and geography by studying Y-chromosomal diversity in Europe [5]. SNPs are mostly double allele biallelic with highly abundant genomes and a low mutation rate, making them stable in the evolutionary process [6]. Previous studies have shown a close correlation between the geography and the genetic structure of Han Chinese individuals and that southern and northern populations are essentially distinct communities [7,8]. In recent years, SNPs and Y-STRs have become important tools for studying population disease susceptibility and adaptation, providing insights into the genetic architecture of various diseases. The results of Liu et al.'s analysis of Y-SNPs and Y-STRs both indicated that patients with esophageal cancer (EC) and gastric cardia cancer in the Chaoshan region have a closer genetic affinity with patients from the Taihang Mountains, and the

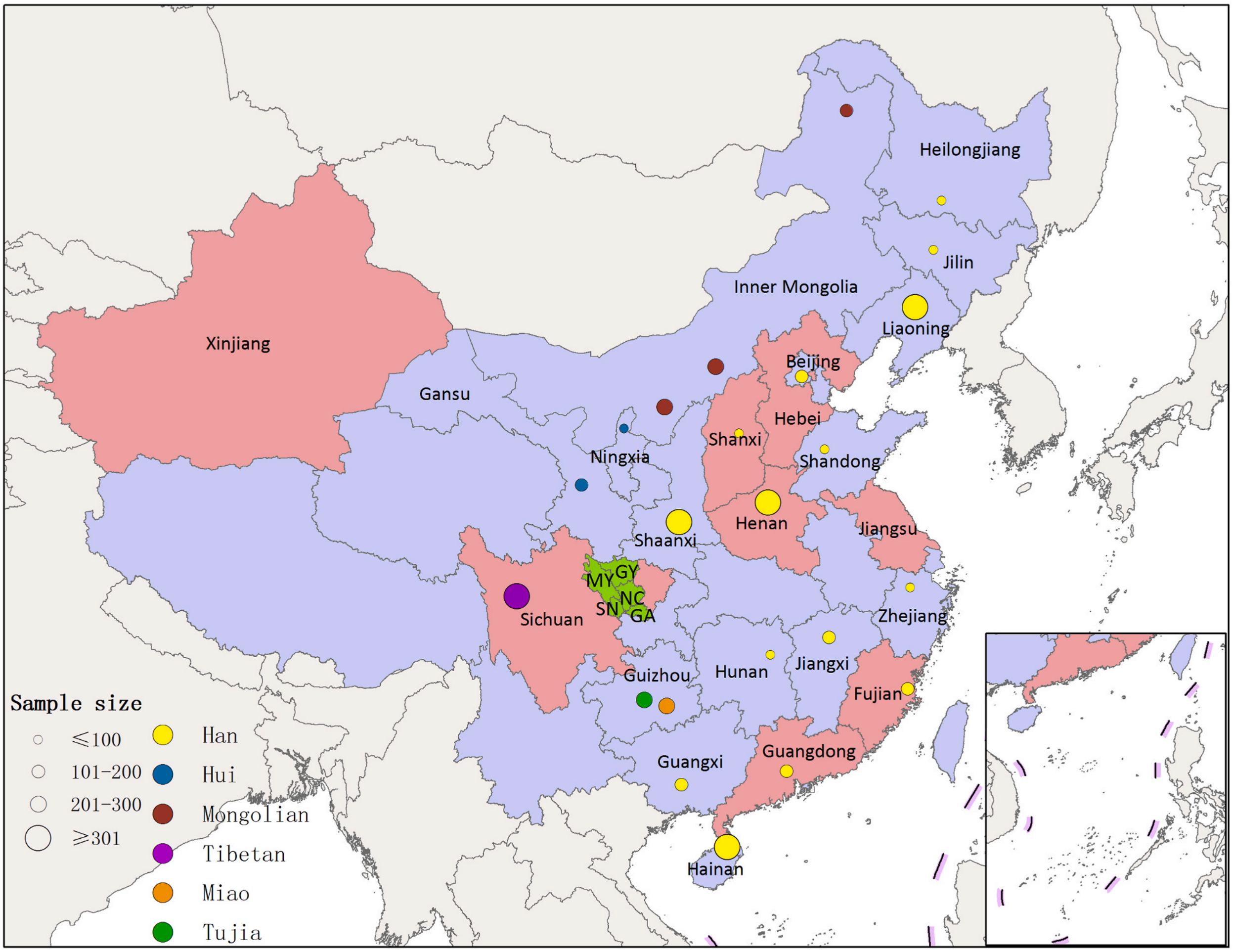


Fig. 1. The geographical distribution of high esophageal cancer incidence locations and 23 Chinese reference populations. The green shading denotes the study populations in northern Sichuan Province, while the red shading denotes the high-incidence area of EC in China. GY: Guangyuan; MY: Mianyang; NC: Nanchong; SN: Suining; GA: Guangan. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

relationship between patients and high-risk populations was even tighter [9]. He et al. revealed the unique natural selection signatures and biological adaptations of Hmong-Mien speakers through SNP-based allele and haplotype sharing and identified genetic loci linked to immune dysfunction, alcohol, and coronary heart disease in Hmong-Mien speakers in Southwest China's Guizhou region [10]. Combining SNPs and Y-STR analysis with the study of the genetic background of populations specific to certain diseases has revealed the genetic patterns of these groups, emphasizing the importance of considering genetic diversity in disease research.

EC is known for its uneven distribution across different regions, with a significantly greater incidence within specific geographic boundaries than in other areas. The high incidence of EC globally is concentrated in two main high-risk belts: one belt begins in north-central China and runs through Central Asia to northern Iran, and the other extends from eastern Africa to southern Africa [11,12]. The most significant areas in China with a high EC incidence include the Taihang Mountains, Northern Sichuan, Fujian, Guangdong, Northern Jiangsu, and Xinjiang (Fig. 1). Previous studies have investigated many etiologic factors involved in the development of EC, but none have convincingly linked these factors to its distinct geographic and population distribution characteristics [13–15]. Ku et al. analyzed germline variants in 515 patients with gastric, esophageal, and gastroesophageal junction cancers, and the results showed that approximately 10.6 % of EC patients carried germline variants associated with cancer susceptibility [16]. A multicenter, large cohort study reported that primary and secondary prevention methods for EC in high-risk areas have little effect on changing the genetic predisposition of the population [17]. The significance of genetic factors in the development of this disease is receiving increasing attention [18–20].

Migrants can provide vital insights into the causes of cancer, particularly in determining the extent to which genetic and environmental factors contribute to disease development. Sichuan Province, located in the interior of southwestern China, is the fifth most populous province in the country. Influenced by historical factors, the population of this province is primarily composed of immigrants from various parts of China. The genetic background of the Sichuan population has always been a topic of interest from various perspectives. Research utilizing Y-STRs has revealed that there is a closer genetic connection between the Han population of Chengdu and Guizhou [21], while the Han population in Luzhou shares a more intimate genetic relationship with those in Guangdong and Hunan [22]. These migratory activities have profoundly shaped the genetic structure and cultural characteristics of Sichuan's population, leading to diversity in dietary habits, cultural practices, and languages in the region. However, the northern part of Sichuan is a well-known high-incidence area for EC in China [23], and there is a significant difference in the incidence rates between high- and low-prevalence areas [24]. While population genetic studies in China often focus on minority ethnic groups [25–28], there is still a relative lack of research on the influencing factors, genetic background, and origins behind the high prevalence of EC among the Han population in the Sichuan region.

This research article aimed to study the genetic structure, linkage, and origin of the Han Chinese EC population in Sichuan Province and compare them with those of the general Han Chinese control population. To our knowledge, this is the first study to combine Y-STR molecular genetic markers and high-density SNP microarray assays to analyze the genetic background of populations from different geographic locations within high-incidence areas of EC. Finally, we obtained 29 Y-STR haplotype data from 214 individuals in the North Sichuan high-incidence EC region, including 117 Han Chinese EC patients and 97 Han Chinese control individuals. We also collected saliva samples from 60 unrelated healthy individuals from three high-prevalence areas of EC, Nanchong in northern Sichuan, Quanzhou in Fujian, and Xinxiang in Henan, for high-density SNP microarray analysis and comprehensive population genetic analysis. All of the data are reported for the first time.

2. Materials and methods

2.1. Subject recruitment and ethics approval

We recruited participants for Y-STR haplotype analysis from two groups: a Han Chinese population with EC and a Han Chinese control group, both of which were from the high-incidence EC region in northern Sichuan. The EC patients were those who presented at the Affiliated Hospital of North Sichuan Medical College between June 2022 and May 2023. The control group consisted of patients with nonneoplastic diseases who also presented at the same hospital during the same period and who resided in the same geographical area as the EC patients. All EC patients were confirmed through professional pathological tissue analysis. All patients with nonneoplastic diseases were also screened for tumor diseases through laboratory tests and imaging examinations. Ultimately, 117 Han Chinese EC patients and 97 Han Chinese control individuals were recruited (Supplementary Table 1). There was no statistically significant difference in the geographical distribution of the two populations ($P > 0.05$). Peripheral blood was collected from both groups with EDTA (Shanghai, BD) anticoagulant.

DNA was extracted from 117 EC patients using a Tiangen Blood Genomic DNA Extraction Kit (centrifugal column type; Tiangen, China). The extracted DNA was stored at −20 °C until use. The remaining 97 control group samples were processed differently: their blood was directly applied onto FTA classic cards and stored at room temperature. Disks with a diameter of 1.2 mm were prepared by punching holes in FTA blood spot cards, which then served as the amplification template for the 29 Y-STR multiplex amplifications, thus eliminating the need for a separate DNA extraction step. The Medical Ethics Committees of North Sichuan Medical College (NSMC [2022] 08) approved this study protocol.

In addition, we collected 60 saliva samples from unrelated healthy individuals, including 16 Han individuals in Nanchong, Northern Sichuan, 14 Han individuals residing in Henan Xinxiang, and 30 Han individuals from Fujian Quanzhou, from three high-incidence areas of EC for a high-density SNP microarray assay. All the data were reported for the first time. We extracted human genomic DNA from saliva using the PureLink Genomic DNA Kit (Thermo Fisher Scientific) and stored it at −20 °C for future use. The Medical Ethics Committees of West China Hospital of Sichuan University (2023-306) approved this study protocol.

The participants in this study had nonconsanguineous marriages and included their parents and grandparents, who were indigenous people who had lived in the sampling location for at least three generations. All individuals signed written informed consent before genetic analysis following the recommendations of the Helsinki Declaration [29].

2.2. Polymerase chain reaction (PCR), genotyping and quality control

The Y-STR haplotypes of 214 individuals from the Affiliated Hospital of North Sichuan Medical College were assayed using the Microreader™ 29Y Direct ID System Amplification Kit (Microread Genetics Incorporation, China) for multiplex Y-STR analysis, which included the following 29 Y-STR loci: DYS393, DYS570, DYS19, DYS392, DYS549, Y GATA H4, DYS460, DYS458, DYS481, DYS635, DYS448, DYS533, DYS456, DYS389I, DYS390, DYS389II, DYS438, DYS576, DYS391, DYS439, DYS437, DYS385a/b, DYS643, DYF387S1a/b, DYS627, DYS449, and DYS518. PCR was performed in a 10 μl reaction volume, which included 4 μl of PCR mix (Microreader ^{TM}2.5 × Master Mix I), 2 μl of primer mix (Microreader™ 29Y-D 5 × Primer Mix), 1 μl of template DNA (0.5 ng/L) or a 1.2 mm diameter disc on a blood filter paper card, and a volume adjusted to 10 μl with ddH_2O. According to the manufacturer's instructions, PCR amplification on a GeneAmp PCR System 9700 (ABI) was performed as follows: one cycle at 95 °C for 5 min; 28 cycles at 94 °C for 20 s, 59 °C for 90 s and 72 °C for 60 s; and one cycle at 60 °C for 45 min. The amplified products were detected by capillary electrophoresis on an ABI-3500 Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA, US). The genetic data were analyzed using GeneMapper ID-X (Thermo Fisher Scientific, Waltham, MA, US). Control DNA (M308) and ddH_2O were used as positive and negative controls, respectively.

Sixty saliva samples from unrelated healthy individuals from three high-prevalence areas of EC in China were genotyped via the Affymetrix WeGene V1 Array, and approximately 465K SNPs were obtained. PLINK v.1.9 [35] was first used for quality control to obtain the original data, and the parameters –geno 0.05 and –mind 0.05 were used to filter the missing SNPs. Second, KING2 [30] was used to calculate kinship coefficients and remove related samples within third generations before merging the data. Next, to explore the general genetic affinity and patterns at the East Asian scale, we combined our previously reported populations [31–35] that were also genotyped via Affymetrix chip and other publicly modern and ancient people from the Allen Ancient DNA Resource (AADR) with our target groups, forming a dataset for autosomal analysis including 119114 SNPs, which is also called the low-density human origin (HO) dataset. This dataset contained 1212 individuals from 123 populations.

2.3. Y-STR-based population genetics analysis and statistical analysis

2.3.1. Population genetics analysis based on 24 common Y-STR gene loci

To reconstruct the phylogenetic relationships among our two study populations and 23 reference populations along administrative divisions and ethnic regions in mainland China and increase the discriminative power, we analyzed 24 common Y-STR gene loci shared by 25 populations, including DYS19, DYF387S1b, DYS389I, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS449, DYS456, DYS458, DYS460, DYS481, DYS518, DYS533, DYS570, DYS576, DYS627, DYS635 and YGATAH4, for analysis of molecular variance (AMOVA), principal component analysis (PCA) and phylogenetic tree construction.

2.3.2. Population genetics analysis based on six common Y-STR loci

To delve deeper into the genetic background correlation between the high-risk EC populations across various geographical areas, we merged the present Y-STR haplotype data from the population in the northern Sichuan high-incidence area with previously published Y-STR data from high-risk populations in three high-prevalence areas for EC and performed population genetics analyses based on six common Y-STR gene loci. The six shared Y-STR loci included DYS389I, DYS389b, DYS390, DYS391, DYS392 and DYS393. The high-risk populations from areas at high risk of EC included Henan (n = 48, consisting of 29 local residents from Linxian and 19 from Xinxiang), Fujian (n = 74, with 37 local residents from Jinjiang and 37 from Putian) and Chaoshan (n = 119, consisting of 89 local residents from Nanao, Chaoyang, Chenghai and Chaozhou and 30 EC patients from Chaoyang, Jieyang, Shantou and surrounding areas) [36]. Moreover, Y-STR data from other regional groups historically or geographically associated with these high-prevalence areas were included in this study. Populations with sample sizes less than 100 were excluded.

2.3.3. Statistical analysis

Statistical analysis was performed with SPSS (IBM SPSS Statistics 25), and the Bonferroni correction was applied to compensate for type I errors in multiple comparisons. Allele frequencies, haplotype frequencies, genetic diversity (GD), Nei genetic distances, and analysis of molecular variance (AMOVA) between populations were calculated by the GenAlEx 6.5 program [37,38]. To estimate the population's genetic structure, we computed Φ_{PT} via AMOVA, a genetic differentiation measure similar to F_{st} for binary or haploid data [39,40]. The significance of the analysis was tested using 999 random permutations [41,42]. All haplotypes with null alleles and triplicated or duplicated alleles were removed, and DYS389I was subtracted from DYS389II for the multicopy locus to obtain DYS389b. Phylogenetic relationships were reconstructed based on the R_{st} genetic matrix using Molecular Evolutionary Genetics Analysis 11 (MEGA 11.0) software [43]. Principal component analysis (PCA) based on the allele frequencies of each Y-STR locus was performed using the online tool Majorbio Cloud Platform (https://cloud.majorbio.com/page/tools/).

2.4. SNP-based population genetic modeling

2.4.1. Principal component analysis (PCA)

PCA was performed using Smart software in the EIGENSOFT (v.6.1.4) [44] package (default and additional parameters: numoutlieriter: 0 and lsqproject: YES.) Strongly linked SNPs were pruned using PLINK (v.1.9) (parameters: -indep-pairwise 200 25 0.4) [45]. On this basis, we performed PCA on 1212 individuals from 123 ancient and modern populations. Moreover, modern populations were regarded as the background; ancient individuals from the Yellow River Basin (YRB), Amur River_WestLiao River (AMR_WLR), Siberia, Nepal, and South China (Guangxi/Fujian/Taiwan Island) were projected onto two-dimensional plots [46].

2.4.2. Model-based ADMIXTURE analysis

We adopted the model-based maximum likelihood clustering algorithm in unsupervised mode and used the abovementioned low-density HO datasets. We used ADMIXTURE (1.3.0) [47] and ran the analysis 100 times based on the default parameters from K = 2 to K = 20 in bootstrap sequences with different random seeds. In addition, we calculated the cross-validation errors and selected six predefined ancestral sources because they are optimal K values.

2.4.3. Painting chromosomes and fineSTRUCTURE analysis

To evaluate the genetic differences or similarities between our target populations and other modern populations in the context of East Asia, we eliminated ancient or unrelated populations and preserved 68 modern populations. Furthermore, to dissect population stratifications, we first used SHAPEIT [48] (Segmented Haplotype Estimation & Imputation Tool, v2) software to haplotype phase the genome-wide data from 68 modern populations based on the low-density HO datasets and with the default parameters (–burn 10 –prune 10 –main 30). Then, we employed ChromoPainter [49] to paint and compute the shared haplotype of 883 individuals and obtain the dendrogram. Moreover, we used R packages with fineSTRUCTURE [50] to explore population structure and phylogenetic relationships based on the above fine-scale information.

2.4.4. Pairwise F_{st} genetic distances

We used PLINK (v.1.9) [45,51] and our in-house scripts to calculate pairwise fixation indices (F_{st}) [52], which reflect genetic relationships by measuring pairwise genetic distance.

2.4.5. TreeMix analyses

Moreover, to further infer phylogenetic relationships and evaluate potential gene flow events, we constructed a phylogenetic tree using TreeMix (v.1.13) [53] and generated migration events varying from 0 to 7 with the default parameters. Finally, we found the optimal model when two migration edges were fitted in the model.

3. Results

3.1. Population genetic structure analysis based on Y-STRs

3.1.1. Y-STR haplotype diversity in Northern Sichuan populations

This research used two approaches for Y-STR PCR amplification: the extracted DNA template amplification method and FTA direct amplification. In the Microreader™ 29Y Fluorescence Detection System, all 29 Y-STR loci were amplified efficiently, and better genotyping results were obtained. A total of 214 individuals from the Han Chinese EC and nonneoplastic Han Chinese control populations were successfully genotyped. The allele frequencies and gene diversity of the 29 Y-STRs are shown in Supplementary Tables 2–5. Of these, a total of 116 haplotypes were detected from 117 Han Chinese EC samples, with one being shared and 115 being unique. In the 97 Han Chinese control population samples, a total of 96 haplotypes were detected, with one being shared and 95 being unique. Additionally, two different triplications (35,36, 37, and 35, 37, 38) were observed at the DYF387S1 marker, and one triplication (32,33,34) was observed at the DYS449 marker. The GD values for the Han Chinese EC population ranged from 0.406 (DYS438) to 0.881 (DYS385a/b), and those for the Han Chinese control population ranged from 0.381 (DYS437) to 0.893 (DYS439). Among the 29 Y-STRs, 233 and 225 alleles were detected in the EC and control groups, respectively, and the cumulative individual identifiability was >0.999999. Excluding the multicopy locus, we compared allele frequencies at the remaining loci between the Han Chinese EC population and the Han Chinese control population. Our analyses revealed a statistically significant difference in the allele frequency distribution at the DYS576 locus ($P = 0.047$). Multiple comparisons revealed that the frequency distributions of allele 18 at the DYS576 locus in the Han Chinese EC population group and the Han Chinese control population group were 26.96 % and 43.75 %, respectively ($P = 0.0098$), and the frequency distributions of allele 19 were 31.30 % and 15.63 %, respectively ($P = 0.0089$) (Supplementary Table 6).

3.1.2. Phylogenetic relationship reconstruction along mainland Chinese administrative and ethnic divisions

The 23 reference populations were distributed around the high incidence of EC in China (Fig. 1), including Han_Beijing (n = 114) [54], Han_Fujian (n = 117) [54], Han_Guangdong (n = 112) [54], Han_Guangxi (n = 104) [54], Han_Shandong (n = 85) [54], Han_Shanxi (n = 95) [54], Han_Zhejiang (n = 94) [54], Han_Jiangxi (n = 108) [54], Han_Heilongjiang (n = 89) [54], Han_Hunan (n = 98) [54], Han_Henan (n = 1411) [54,55], Han_Jilin (n = 88) [56], Han_Liaoning (n = 566) [56], Han_Hainan (n = 473) [57], Han_Shaanxi (n = 430) [58], Miao_Guizhou (n = 206) [59], Tujia_Guizhou (n = 246) [59], Hui_Gansu (n = 146) [60], Hui_Ningxia (n

= 83) [60], Tibetan_Sichuan (n = 499) [61], Mongolian_Hohhot (n = 226) [62], Mongolian_Hulunbuir (n = 223) [62] and Mongolian_Ordos (n = 200) [62]. We calculated Φ_{PT} and corresponding *P* values using AMOVA to provide a valid measure of population genetic differentiation. A smaller Φ_{PT} value indicates a closer genetic relationship. The Sichuan Tibetan and Fujian Han populations showed the greatest differences in genetic differentiation ($\Phi_{PT} = 0.114$, $P < 0.001$). The largest difference in genetic differentiation among the Han Chinese populations was between Han_Jiangxi and Han_Hainan ($\Phi_{PT} = 0.031$, $P < 0.001$), and the largest difference in genetic differentiation among the minority populations was between Miao_Guizhou and Tibetan_Sichuan ($\Phi_{PT} = 0.091$, $P < 0.001$). Our two studied populations showed the closest genetic relatedness to the Han populations from northern China, such as Jilin, Liaoning and Henan ($P > 0.05$) (Supplementary Table 7).

We calculated Nei's genetic distances among 25 Chinese populations. The phylogenetic tree was constructed by the unweighted pair group method with arithmetic average (UPGMA) method to explore genetic differentiation (Fig. 2A). Here, two major branches were observed in the phylogenetic tree: the Sichuan-Tibetans were clustered separately, with the rest of the population clustered and showing further subclusters. The subcluster showed that ethnic minorities and Han Chinese were clustered together separately. The results indicated that the genetic affinities of the Han groups in China were generally similar, while those of the Han and minority groups were more distant. Moreover, the southern and northern Han Chinese populations clustered separately, which was consistent with the previously observed genetic phenomenon of a genetic gradient between the southern Han and northern Han Chinese populations [63–65]. Notably, the two northern Sichuan populations were more closely related to the Han Chinese population in northern

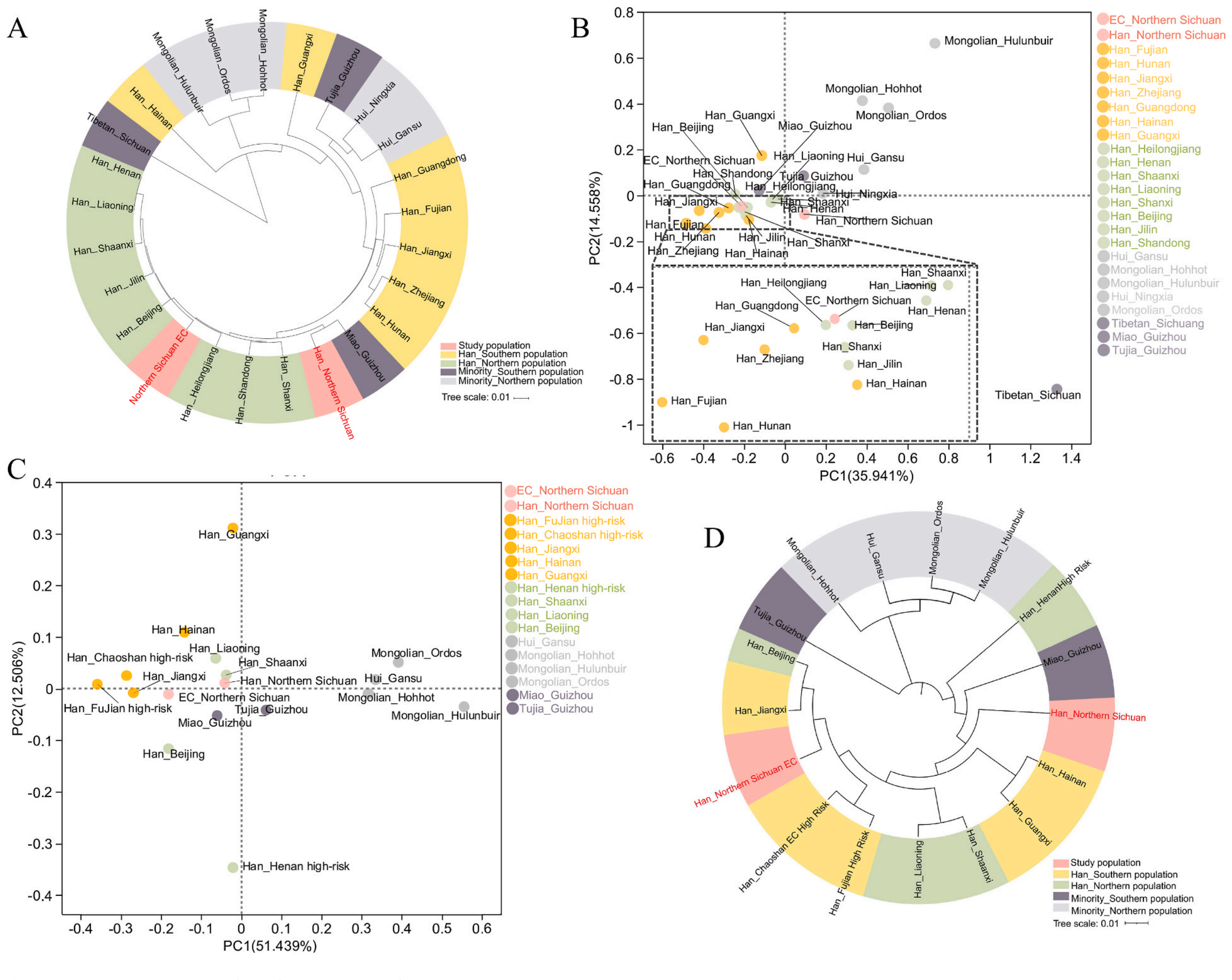


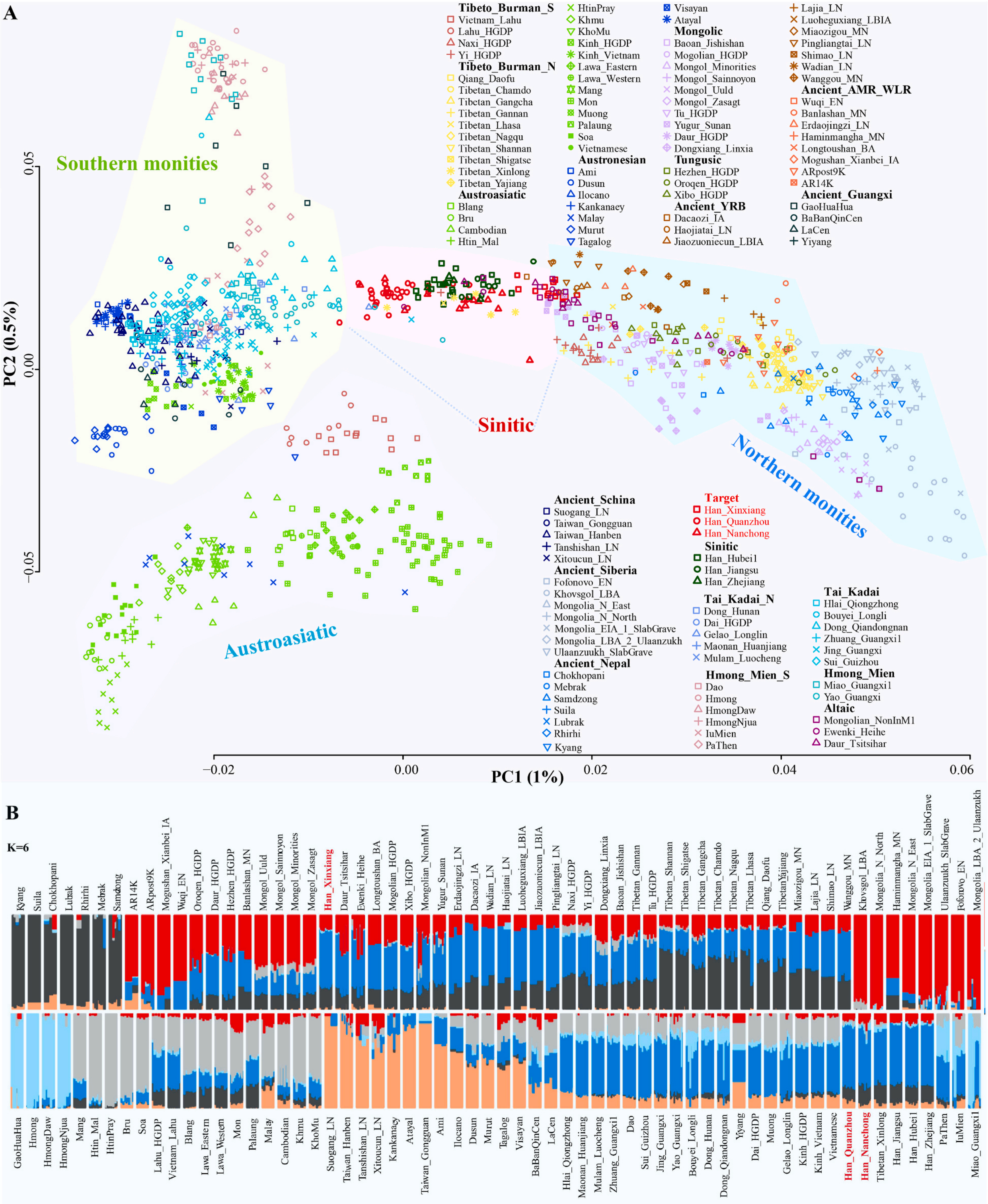
Fig. 2. Genetic structure inferred from Y-STR data.

(A) The phylogenetic tree reconstructed based on 24 shared Y-STR loci using the UPGMA method demonstrated the phylogenetic relationships between populations in the high-incidence area of EC in northern Sichuan Province and 23 reference populations in China.

(B) PCA analyses among populations in the high-incidence area of EC in northern Sichuan Province and 23 reference populations based on 24 common Y-STR loci in China.

(C) PCA based on six common Y-STR loci showing the relationships among the high-risk populations for esophageal cancer and other reference populations.

(D) The neighbor-joining phylogenetic tree constructed from the R_{st} genetic distance matrix among the EC high-risk and reference populations based on the six common Y-STR loci.



(caption on next page)

Fig. 3. Genomic affinity of three Han Chinese populations in different regions of China with high incidence rates of esophageal cancer inferred from genome-wide SNP data.
(A): Principal component analysis (PCA) of 1212 individuals from 123 populations revealed close genetic relationships among the Han_Nanchong, Han_Xinxiang, and Han_Quanzhou populations. Analysis was conducted based on the genetic variations of 68 modern populations and 52 ancient populations, from which ancient people were projected. All included individuals were color-coded by geographical division.
(B): Model-based ADMIXTURE results displaying the ancestry composition of 68 modern populations with the least error (K = 6). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

China than to the southern population. PCA based on allele frequencies at each locus showed that the northern Sichuan Han Chinese EC population clustered together with the majority Chinese northern Han Chinese population, and the Sichuan Han Chinese control population and other ethnic minority populations appeared more scattered (Fig. 2B). The results suggested that the Han EC population from the high-incidence area in northern Sichuan Province had a closer genetic relationship with the Han populations from northern China.

3.1.3. Geographically different high-risk EC group locations may have similar genetic backgrounds

A total of 15 reference populations were included for population genetic structure analysis, including Han_Beijing (n = 114) [54], Han_Guangxi (n = 104) [54], Han_Jiangxi (n = 108) [54], Han_Liaoning (n = 556) [56], Han_Hainan (n = 473) [57], Han_Shaanxi (n = 430) [58], Miao_Guizhou (n = 206) [59], Tujia_Guizhou (n = 246) [59], Hui_Gansu (n = 146) [60], Mongolian_Hohhot (n = 226) [62], Mongolian_Hulunbuir (n = 223) [67], and Mongolian_Ordos (n = 200) [67]. Our analysis showed that the frequency of the DYS393*12 allele was 0.44 in the EC high-risk population in Henan, while the rest of the EC population and the EC high-risk population possessed the high-frequency alleles DYS389I*12 and DYS393*12 (frequencies ≥0.5). (Supplementary Table 8). We performed a PCA that utilizes dimensionality reduction to reveal the structure hidden behind complex data. The first two principal components accounted for 63.95 % of the total variance (Fig. 2C). We found that in the northern Sichuan high-incidence area, the Han Chinese EC population was closer to the high-risk EC populations from Chaoshan and Fujian but farther from the high-risk Henan population. We constructed a neighbor-joining (N-J) tree based on Nei's distance (Fig. 2D). The phylogenetic tree showed two branches: the northern minorities such as the Hulunbeier Mongols and the Gansu Hui grouped into one main branch, while the remaining populations formed another large branch. In this phylogenetic tree, the northern Han Chinese and northern Sichuan Han Chinese EC populations were clustered together and then clustered with the high-risk EC populations in Chaoshan and Fujian, which was similar to the PCA results. As a result of the present study, the high-risk groups of EC in Chaoshan, Fujian, and northern Sichuan were closely genetically related. These findings suggested that high-risk groups of EC in different geographic regions may share similar genetic backgrounds.

3.2. SNP-based population genetic modeling analysis

3.2.1. Analysis of genetic relationships and ancestral composition in three regions with a high incidence of EC in China

Sichuan has undergone various waves of migration throughout its history, leading to a diverse population with a wide range of ethnic and linguistic backgrounds. These include Sino-Tibetan (ST), Hmong-Mien (HM), and Tai-Kadai (TK) speakers. The PCA conducted at the East Asian scale revealed four genetically distinct clusters (Fig. 3A): (1) the Southern Minority Cluster, comprising populations that speak Austroasiatic (AA), Tai-Kadai (TK), and Hmong-Mien (HM) languages; (2) Northern minorities, grouped by their Mongolic, Tibeto-Burman_N, and Altaic languages; (3) the Han Chinese people Cluster, composed of Sino-Tibetan speaking populations; and (4) the remaining groups comprise the genetic clusters that include the South Asian island speaking populations. The three target groups in the high-EC-incidence areas of Nanchong in northern Sichuan, Quanzhou in Fujian, and Xinxiang in Henan were located in an intermediate position between the southern minority clusters and the northern minority clusters and closely clustered together, suggesting that they are genetically close. Among these, the Henan Xinxiang population partially overlaps with the Mongolian, Ewenki, and Daur populations of the Altaic language family and is genetically closer to the Pingliangtai_LN population of the Ancient_YRB.

To analyze the proportions of mixed ancestral components among the three Han Chinese populations from high-prevalence regions of EC, we performed a model-based ADMIXTURE analysis on 123 populations. When K = 6, we observed the following six East Asian ancestries: Ancient_Schina (orange), Ancient_Nepal (black), Sinitic (dark blue), Hmong_Mien_S (light blue), Austroasiatic (gray), Ancient_Siberia (red), (Fig. 3B). The dark blue ancestry component widely existed in the Han Chinese populations and their neighbors. Moreover, similar ancestral progenitor compositions were observed in the samples from the Han Chinese population in three high-risk areas [Nanchong (Sinitic 59.4 %, Ancient_Nepal 20.1 %, Ancient_Siberia 7.5 %), Quanzhou (Sinitic 65.8 %, Ancient_Nepal 14.3 %, Ancient_Schina 9.9 %), Xinxiang (Sinitic 60.3 %, Ancient_Nepal 22.4 %, Ancient_Siberia 13.3 %)]. The results showed that the Han ethnicity is primarily based on the ancient Han people, while also incorporating a significant amount of bloodlines and culture from various other ethnic groups. The complex genetic history and migration of the Han population have had a significant influence on the current genetic makeup of the Han gene pool in China. From our research findings, it is evident that the widespread presence of Han components (represented in dark blue) reflects the ubiquity of Han cultural and genetic traits. The existence of ancient Siberian elements (in red) and ancient Nepalese elements (in black) indicates the impact of ancient migrations and gene flows on the genetic structure of the Han population [66]. Our ADMIXTURE analysis supported the similarity in the genetic background of Han populations from high-incidence EC areas, providing further evidence to uncover the genetic and historical connections between these groups.

3.2.2. Populations in three geographically distinct regions with a high prevalence of EC have close genetic affinity and a common origin

To further elucidate genetic similarities or differences among populations in three geographically distinct regions with high rates of EC and to explore their fine-scale population substructures, we further explored genetic differentiation and substructure using haplotype-based fineSTRUCTURE and mixed-model reconstruction based on the allele spectra of 883 individuals (Fig. 4A). The haplotype-based clustering patterns were consistent with the model-based ADMIXTURE ancestry composition results: most individuals in the three Han studied populations had the greatest degree of primary simulated ancestry. Next, we analyzed the F_{st} genetic differentiation indices for our three studied populations and modern East Asian populations (Supplementary Table 9). A lower F_{st} score indicates greater genetic similarity between populations. Our findings revealed that the three studied populations have close genetic affinities. The genetic distance between Han_Nangchong and Han_Xingxiang (0.000730204) was smaller than that between Han_Quanzhou (0.00116002). We created a heatmap based on the values that display the kinship relationships among these 68 modern East Asian populations (Fig. 4B). ADMIXTURE mixed analyses, fineSTRUCTURE, and F-statistical analyses revealed strong genetic affinities between our study populations in the three regions with a high prevalence of EC.

We also reconstructed a maximum likelihood-based TreeMix to describe the migration and integration among 68 modern populations through gene flow analyses (Fig. 4C). The results showed that their genetic branching is consistent with their linguistic categorization and geographic divisions. Our study populations from the three areas with high EC prevalence clustered with Altaic-speaking Mongolians in the northern region and Tibetan-Burmese-speaking populations. Specifically, when two gene flow events

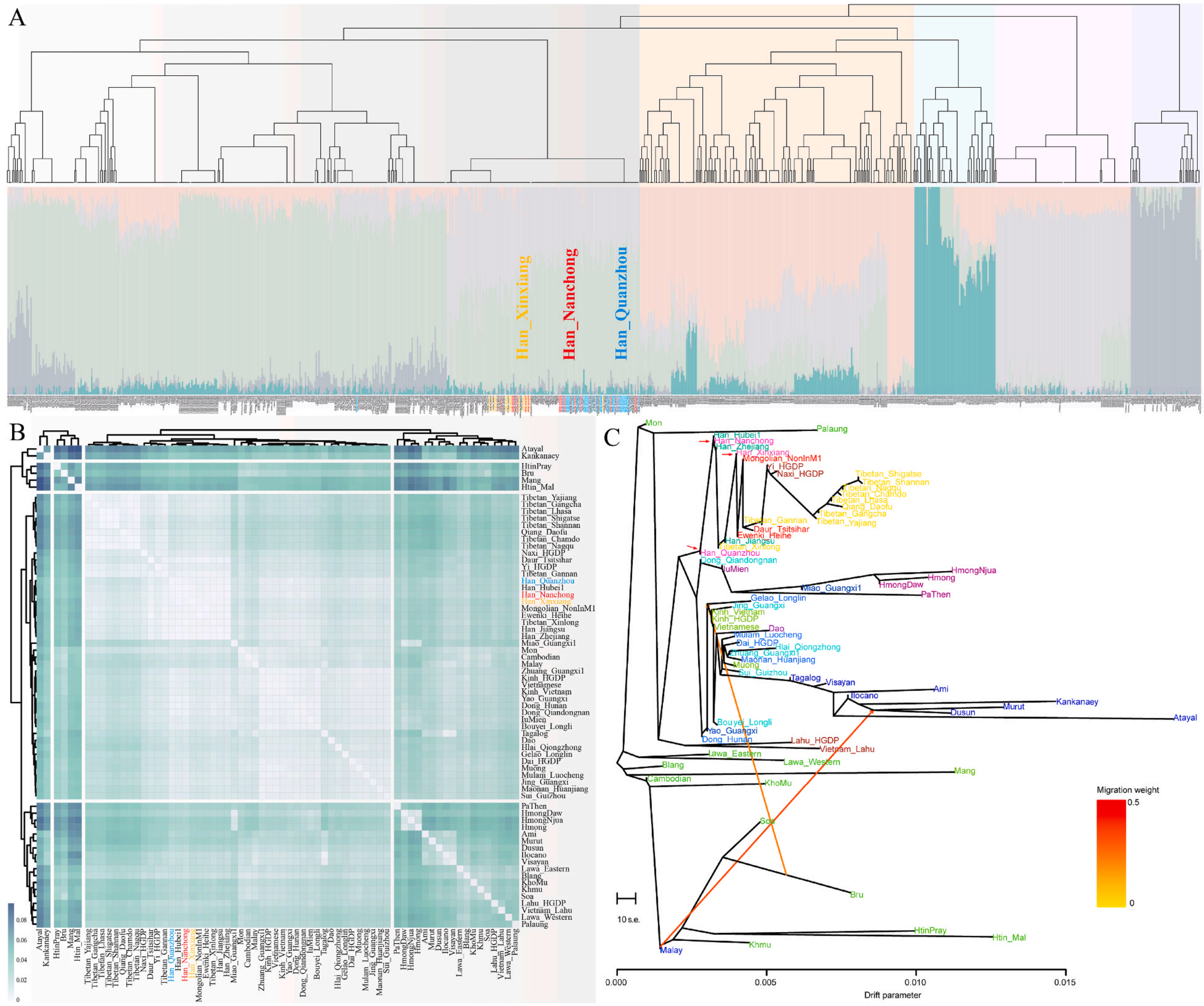


Fig. 4. Fine-scale population relationships of 68 modern populations and gene flow events.
(A) FineSTRUCTURE analysis based on the allele spectra of 883 individuals from 68 modern populations.
(B) A heatmap based on F_{st} genetic distance revealed the kinship relationships among 68 modern East Asian populations.
(C) Migration and integration among 68 modern populations and gene flow events (migration edges = 2). There was a close genetic affinity among the Han_Nanchong, Han_Xinxiang, and Han_Quanzhou populations.

occurred, the Nanchong Han population clustered first with the Quanzhou Han population and then with the Xinxiang Han population. The clustering patterns were consistent with the results of the PCA, ADMIXTURE, and F_{st} genetic differentiation analyses described above. The Han populations from Nanchong and Quanzhou diverged from their northern ancestral roots and migrated to southern regions of China.

4. Discussion

The development and progression of tumors is a complex and multistage process. Environmental factors are likely to interact with genetic susceptibility factors, burdening susceptible groups and increasing the likelihood of developing tumors [67]. Migrants, who move between regions at different risk levels for disease risk, provide a unique opportunity to study the impact of environmental factors and genetic predispositions on the risks that trigger the development of particular cancers [68,69]. A study suggested that the high incidence of EC in central-western Brazil may be due to substantial internal migration from southern Brazil [68]. Huang et al. reported that the elevated incidence of EC in coastal areas is likely due to migration from high-risk EC areas in the Central Plains [70], and this finding was also confirmed by Li et al. through the study of mitochondrial DNA markers [71]. These studies provide further insight into the critical role of the genetic background of a population in the development of EC in high-risk areas. Moreover, further research into the genetic background of EC in these populations is needed.

Sichuan is situated in southwestern China and is surrounded by mountains, which historically made transportation difficult and created natural geographic isolation. Due to wars and disasters, Sichuan has experienced extreme population loss [72]. To restore economic and social order in Sichuan, the government has implemented a series of policies to encourage migrants from different parts of the country to move into Sichuan and rebuild their homes. The most recent major population migration in Sichuan occurred between the late Ming and early Qing dynasties. The government organized and encouraged large-scale immigration on a national level during 1694–1851, which resulted in a significant increase in the population from less than 90,000 to over 48 million [73,74]. The continuous mixing of populations and geographic isolation have formed the specific genetic structure and history of the Sichuan population and complicated the gene pool of Han Chinese in Sichuan. It is worth noting that in Sichuan Province, only the northern region has a higher incidence of EC. In large-scale migration movements, immigrants tend to settle and live with people from their origin. Given that the genetic background of the population is likely to play an essential role in the predisposition to EC, whether the geographic clustering of EC cases in this region is related to the genetic background of the population deserves close attention and further study [12]. Fortunately, the large and diverse immigrant population in Sichuan provides a natural opportunity to investigate the genetic background of populations in regions with a high prevalence of EC.

Previous studies have shown that increasing the number of Y-STR loci can improve discrimination and facilitate phylogenetic analysis [75–77]. Building on this foundation, we employed 29 Y-STR loci to perform a comprehensive genetic analysis of the Han population in the high-incidence EC region of northern Sichuan Province, which marked the first collection of Y-STR haplotype data specifically for Han EC patients in this area. For people living in areas with a high incidence of EC, although they are in the same environment, the incidence of individual tumors varies greatly, suggesting a strong relationship between the development of tumors and the genetic background of different individuals. In this study, we used a case-control strategy and found that there was a significant difference in the allele frequency distribution of the DYS576 locus between Han EC patients and the local control population in northern Sichuan. Furthermore, we examined the genetic relationships between the Han population from the EC high-incidence area in Sichuan Province and other reference populations across China using 24 shared Y-STR loci. Our findings suggested a close genetic affinity between the populations from the EC high-incidence area in northern Sichuan and Han populations from northern China, such as Jilin, Liaoning and Henan. Additionally, to gain a deeper understanding of the genetic backgrounds of high-risk EC populations across different geographical areas, we investigated the genetic background associations between the Northern Sichuan Han Chinese EC population and those populations at risk in three regions with a high prevalence of EC in China based on six shared Y-STR loci. The results showed that all these Han Chinese high-risk EC populations had high-frequency alleles of DYS389I*12 and DYS393*12 and shared a similar genetic background. This finding is consistent with the studies by Li et al. [71] and Huang et al. [36], which identified the same genetic background among the high-risk populations for EC in Chaoshan, Fujian, and Henan. Compared to autosomal variation, the Y chromosome exhibits a smaller effective population size and displays more pronounced geographical stratification [78]. To avoid potential biases that might arise from relying solely on Y-STR markers, we added autosomal SNP microarray analysis to more accurately determine the genetic structure of the population. To this end, we collected saliva samples from healthy individuals in high-incidence areas of the EC, including Nanchong, Quanzhou, and Xinxiang, and performed high-density SNP analysis on the autosomes and genetic modeling. Through these analyses, we meticulously traced the genetic origins of the Han population in the high-incidence area of northern Sichuan Province and found that they have a high degree of genetic similarity with Han populations in North China, as well as close genetic ties and similar ancestral components with populations from Quanzhou and Xinxiang, which are also regions with a high incidence of EC.

It is worth mentioning that previous studies have indicated that out of the six regions in China identified as high-risk areas for EC, five are located in the central or northern regions [79]. The molecular genetic data obtained from patrilineal Y chromosome and autosomal high-density SNP microarray analyses in this study further supported the notion that the population in the high-incidence area of EC in northern Sichuan may have originated from the northern regions. Additionally, this study revealed that populations from different EC high-incidence areas have a closer genetic affinity and a similar ancestral composition. The high incidence of EC in northern Sichuan may be related to the historical high-risk migrant populations from northern China. These findings not only provide genetic evidence for the origin and migration of the population in northern Sichuan but also offer a partial explanation for the geographical clustering of EC incidence in China from a genetic perspective.

However, this study has several limitations. First, current research on Y-STRs in China has focused predominantly on ethnic minority groups, while studies on the Han population, particularly groups linked to EC diseases, are comparatively limited. Our study did not include Y-STR analysis data from a larger Han population from the southwestern region of China or additional data from known high-risk EC populations, which may limit our comprehensive understanding of the genetic background of EC and the depth of analysis. These data are crucial for revealing the genetic characteristics of EC in depth, and future research should aim to fill this gap to more accurately assess the contribution of population genetic backgrounds to the risk of developing EC. Second, although we made considerable efforts to rule out any potential influences from the genetic background of the population in other areas, we were still unable to rule out the possibility of population stratification [80,81]. Our results may deserve validation in a larger cohort. Therefore, follow-up studies of the genetic background of the Han Chinese EC population should integrate larger clinical cohorts, including multicenter Han Chinese EC populations, larger haplotype databases, diverse genetic markers, and different detection methods. Nonetheless, our study also provides Y-STR profiling data for underrepresented populations, which is essential for enhancing the population genetics data of northern Sichuan Province and will contribute to exploring the genetic mechanism of EC.

5. Conclusions

In summary, our research findings indicated that there was a strong genetic association between the high-incidence populations of EC in the northern Sichuan region and those in the Fujian and Chaoshan areas. The Han population in Nanchong, a high-incidence area of EC in northern Sichuan, was closely related to the Han populations in Henan Xinxiang and Fujian Quanzhou, both of which were also regions with high incidences of EC. The results suggested that the genetic background of populations plays a significant role in the pathogenesis of EC in high-incidence areas. These research outcomes were crucial for a deeper understanding of the group genetic characteristics of EC among the Han population and could facilitate further exploration of the molecular genetic mechanisms of EC.

Ethics declarations

The study protocol was approved by the Medical Ethics Committees of West China Hospital of Sichuan University (2023-306) and North Sichuan Medical College (NSMC [2022] 08).

Funding

This work was supported by grants from the Excellent Teacher Research Fund of North Sichuan Medical College (MS2021003), the Nanchong Social Science Research "Fourteenth Five-Year" Plan Project Fund (NC23C102), and the National Natural Science Foundation of China (82202078).

Data availability statement

The allele frequency data derived from human samples have been deposited in the National Omics Data Encyclopedia (NODE, http://www.biosino.org/node). Reference genotype data for ancient and modern individuals were collected from the Allen Ancient DNA Resource (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data, version 54.1).

The access and use of the data complied with the regulations of the People's Republic of China on the administration of human genetic resources. Requests for access to data can be directed to Guanglin He (Guanglinhescu@163.com).

CRediT authorship contribution statement

Lihua Jia: Writing – original draft, Visualization, Data curation. **Mengge Wang:** Writing – review & editing, Methodology. **Shuhan Duan:** Writing – original draft, Visualization. **Jianghua Chen:** Validation, Data curation. **Mei Zhao:** Methodology, Data curation. **Simeng Ji:** Methodology, Data curation. **Bingbing Lv:** Investigation, Data curation. **Xiucheng Jiang:** Investigation. **Guanglin He:** Writing – review & editing, Funding acquisition, Conceptualization. **Junbao Yang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mengge Wang reports statistical analysis and writing assistance were provided by Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China. Guanglin He reports financial support was provided by Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

AA	Austroasiatic
AADR	Allen Ancient DNA Resource
AMOVA	Analysis of molecular variance
AMR_WLR	Amur River_WestLiao River
EC	Esophageal cancer
F_{st}	Fixation index
GD	Genetic diversity
HM	Hmong-Mien
HO	Human Origin
MEGA	Molecular Evolutionary Genetics Analysis
PCA	Principal component analysis
PCR	Polymerase chain reactions
SNPs	Single nucleotide polymorphisms
ST	Sino-Tibetan
TK	Tai-Kadai
YRB	Yellow River Basin
Y-STRs	Y-chromosome short tandem repeats

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29867.

References

[1] G. He, J. Liu, M. Wang, X. Zou, T. Ming, S. Zhu, H.-Y. Yeh, C. Wang, Z. Wang, Y. Hou, Massively parallel sequencing of 165 ancestry-informative SNPs and forensic biogeographical ancestry inference in three southern Chinese Sinitic/Tai-Kadai populations, Forensic Sci. Int.: Genetics 52 (2021).

[2] F. Song, M. Song, H. Luo, M. Xie, X. Wang, H. Dai, Y. Hou, Paternal genetic structure of Kyrgyz ethnic group in China revealed by high-resolution Y-chromosome STRs and SNPs, Electrophoresis 42 (19) (2021) 1892–1899.

[3] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (6) (2004) 435–445.

[4] M.T. Seielstad, E. Minch, L.L. Cavalli-Sforza, Genetic evidence for a higher female migration rate in humans, Nat. Genet. 20 (3) (1998) 278–280.

[5] Z.H. Rosser, T. Zerjal, M.E. Hurles, M. Adojaan, D. Alavantic, A. Amorim, W. Amos, M. Armenteros, E. Arroyo, G. Barbujani, et al., Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language, Am. J. Hum. Genet. 67 (6) (2000) 1526–1543.

[6] S.A. C, Accessing genetic variation: genotyping single nucleotide polymorphisms, Nat. Rev. Genet. 2 (12) (2001) 930–942.

[7] J.Y. Chu, W. Huang, S.Q. Kuang, J.M. Wang, J.J. Xu, Z.T. Chu, Z.Q. Yang, K.Q. Lin, P. Li, M. Wu, et al., Genetic relationship of populations in China, Proc. Natl. Acad. Sci. U. S. A. 95 (20) (1998) 11763–11768.

[8] B. Su, J. Xiao, P. Underhill, R. Deka, W. Zhang, J. Akey, W. Huang, D. Shen, D. Lu, J. Luo, 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 (6) (1999) 1718–1724.

[9] S. Liu, B. Huang, H. Huang, X. Li, G. Chen, G. Zhang, W. Lin, D. Guo, J. Wang, Z. Yu, et al., Patrilineal background of esophageal cancer and gastric cardia cancer patients in a Chaoshan high-risk area in China, PLoS One 8 (12) (2013) e81670.

[10] G. He, J. Wang, L. Yang, S. Duan, Q. Sun, Y. Li, J. Wu, W. Wu, Z. Wang, Y. Liu, et al., Genome-wide allele and haplotype-sharing patterns suggested one unique Hmong-Mein-related lineage and biological adaptation history in Southwest China, Hum. Genom. 17 (1) (2023) 3.

[11] E.C. Smyth, J. Lagergren, R.C. Fitzgerald, F. Lordick, M.A. Shah, P. Lagergren, D. Cunningham, Oesophageal cancer, Nat. Rev. Dis. Prim. 3 (1) (2017).

[12] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 68 (6) (2018) 394–424.

[13] C. Yu, H. Tang, Y. Guo, Z. Bian, L. Yang, Y. Chen, A. Tang, X. Zhou, X. Yang, J. Chen, et al., Hot tea consumption and its interactions with alcohol and tobacco use on the risk for esophageal cancer: a population-based cohort study, Ann. Intern. Med. 168 (7) (2018) 489–497.

[14] A. Yokoyama, T. Omori, T. Yokoyama, Alcohol and aldehyde dehydrogenase polymorphisms and a new strategy for prevention and screening for cancer in the upper aerodigestive tract in East Asians, Keio J. Med. 59 (4) (2010) 115–130.

[15] J. Andrici, G.D. Eslick, Hot food and beverage consumption and the risk of esophageal cancer: a meta-analysis, Am. J. Prev. Med. 49 (6) (2015) 952–960.

[16] G.Y. Ku, Y. Kemel, S.B. Maron, J.F. Chou, V. Ravichandran, Z. Shameer, A. Maio, E.S. Won, D.P. Kelsen, D.H. Ilson, et al., Prevalence of germline alterations on targeted tumor-normal sequencing of esophagogastric cancer, JAMA Netw. Open 4 (7) (2021) e2114753.

[17] J. Zhou, K. Sun, S. Wang, R. Chen, M. Li, J. Gu, Z. Fan, G. Zhuang, W. Wei, Associations between cancer family history and esophageal cancer and precancerous lesions in high-risk areas of China, Chinese Med J 135 (7) (2022) 813–819.

[18] C.L. Carter, N. Hu, M. Wu, P.Z. Lin, C. Murigande, G.E. Bonney, Segregation analysis of esophageal cancer in 221 high-risk Chinese families, J. Natl. Cancer Inst. 84 (10) (1992) 771–776.

[19] W. Garavello, E. Negri, R. Talamini, F. Levi, P. Zambon, L. Dal Maso, C. Bosetti, S. Franceschi, C. La Vecchia, Family history of cancer, its combination with smoking and drinking, and risk of squamous cell carcinoma of the esophagus, Cancer Epidemiol. Biomarkers Prev. 14 (6) (2005) 1390–1393.

[20] W. Zhang, J.E. Bailey-Wilson, W. Li, X. Wang, C. Zhang, X. Mao, Z. Liu, C. Zhou, M. Wu, Segregation analysis of esophageal cancer in a moderately high-incidence area of northern China, Am. J. Hum. Genet. 67 (1) (2000) 110–119.

[21] H. Wang, J. Mao, Y. Xia, X. Bai, W. Zhu, D. Peng, W. Liang, Genetic polymorphisms of 17 Y-chromosomal STRs in the Chengdu han population of China, Int. J. Leg. Med. 131 (4) (2017) 967–968.

[22] J. Fu, B. Song, J. Qian, T. He, H. Chen, J. Cheng, J. Fu, Genetic polymorphism analysis of 24 Y-STRs in a han Chinese population in Luzhou, southwest China, Genes 14 (10) (2023).

[23] Y. Lin, Y. Totsuka, B. Shan, C. Wang, W. Wei, Y. Qiao, S. Kikuchi, M. Inoue, H. Tanaka, Y. He, Esophageal cancer in high-risk areas of China: research progress and challenges, Ann. Epidemiol. 27 (3) (2017) 215–221.

[24] C.C. Abnet, M. Arnold, W.Q. Wei, Epidemiology of esophageal squamous cell carcinoma, Gastroenterology 154 (2) (2018) 360–373.

[25] X. Bin, R. Wang, Y. Huang, R. Wei, K. Zhu, X. Yang, H. Ma, G. He, J. Guo, J. Zhao, et al., Genomic insight into the population structure and admixture history of tai-kadai-speaking sui people in Southwest China, Front. Genet. 12 (2021) 735084.
[26] B. Ma, J. Chen, X. Yang, J. Bai, S. Ouyang, X. Mo, W. Chen, C.C. Wang, X. Hai, The genetic structure and east-west population admixture in northwest China inferred from genome-wide array genotyping, Front. Genet. 12 (2021) 795570.
[27] M. Wang, W. Du, R. Tang, Y. Liu, X. Zou, D. Yuan, Z. Wang, J. Liu, J. Guo, X. Yang, et al., Genomic history and forensic characteristics of Sherpa highlanders on the Tibetan Plateau inferred from high-resolution InDel panel and genome-wide SNPs, Forensic Sci Int Genet 56 (2022) 102633.
[28] H. Zhang, L. Yun, Y. Li, J. Zhang, J. Wu, J. Yan, Y. Hou, Haplotype of 12 Y-STR loci of the PowerPlex Y-system in Sichuan Han ethnic group in west China, Forensic Sci. Int. 175 (2–3) (2008) 244–249.
[29] W.M. Association, World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects, Bull. World Health Organ. 79 (4) (2001) 373.
[30] N.A. Tinker, D.E. Mather, Kin - software for computing kinship coefficients, J. Hered. 84 (3) (1993) 238, 238.
[31] X. Bin, R. Wang, Y. Huang, R. Wei, K. Zhu, X. Yang, H. Ma, G. He, J. Guo, J. Zhao, et al., Genomic insight into the population structure and admixture history of tai-kadai-speaking sui people in Southwest China, Front. Genet. 12 (1718) (2021) 735084.
[32] X. Huang, Z.-Y. Xia, X. Bin, G. He, J. Guo, A. Adnan, L. Yin, Y. Huang, J. Zhao, Y. Yang, Genomic insights into the demographic history of the Southern Chinese, Frontiers in Ecology and Evolution 10 (2022) 853391.
[33] G.L. He, M.G. Wang, X. Zou, H.Y. Yeh, C.H. Liu, C. Liu, G. Chen, C.C. Wang, Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity, J. Systemat. Evol. 61 (1) (2022) 230–250.
[34] G. He, Z. Wang, J. Guo, M. Wang, X. Zou, R. Tang, J. Liu, H. Zhang, Y. Li, R. Hu, et al., Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping, Eur. J. Hum. Genet. 28 (8) (2020) 1111–1123.
[35] M.G. Wang, G.L. He, X. Zou, P.Y. Chen, Z. Wang, R.K. Tang, X.M. Yang, J. Chen, M.Q. Yang, Y.X. Li, et al., Reconstructing the genetic admixture history of Tai-Kadai and Sinitic people: insights from genome-wide SNP data from South China, J. Systemat. Evol. 61 (1) (2022) 157–178.
[36] H. Huang, Y-Chromosome Evidence: Genetic Structure Analysis Among Populations from Southern Littoral and North-central Esophageal Cancer High-Risk Areas in China Shantou University，PhD Dissertation, 2011.
[37] R.O.D. Peakall, P.E. Smouse, Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research, Mol. Ecol. Notes 6 (1) (2006) 288–295.
[38] R. Peakall, P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update, Bioinformatics 28 (19) (2012) 2537–2539.
[39] R. Peakall, P.E. Smouse, D.R. Huff, Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss Buchloë dactyloides, Mol. Ecol. 4 (2) (1995) 135–148.
[40] D.R. Huff Rp, P. E, Smouse RAPD variation within and among natural populations of outcrossing buffalograss [Buchlo dactyloides (Nutt.) Engelm.], Theor. Appl. Genet. 86 (1993) 927–934.
[41] P.E.S.J.C. Long, Matrix correlation analysis in anthropology and genetics, Yearbk. Phys. Anthropol. 35 (1992) 187–213.
[42] E. Peter, Smouse JCLaRRS: multiple regression and correlation extensions of the mantel test of matrix correspondenc, Syst. Zool. 35 (1986) 627–632.
[43] K. Tamura, G. Stecher, S. Kumar, F.U. Battistuzzi, MEGA11: molecular evolutionary genetics analysis version 11, Mol. Biol. Evol. 38 (7) (2021) 3022–3027.
[44] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history, Genetics 192 (3) (2012) 1065–1093.
[45] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (1) (2015) s13742, 13015-10047-13748.
[46] H. McColl, F. Racimo, L. Vinner, F. Demeter, T. Gakuhari, J.V. Moreno-Mayar, G. van Driem, U. Gram Wilken, A. Seguin-Orlando, C. de la Fuente Castro, et al., The prehistoric peopling of Southeast Asia, Science 361 (6397) (2018) 88–92.
[47] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19 (9) (2009) 1655–1664.
[48] O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes, Nat. Methods 9 (2) (2012) 179–181.
[49] G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, S. Myers, A genetic atlas of human admixture history, Science (New York, NY) 343 (6172) (2014) 747–751.
[50] D.J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data, PLoS Genet. 8 (1) (2012) e1002453.
[51] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (3) (2007) 559–575.
[52] B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution 38 (6) (1984) 1358–1370.
[53] J.K. Pickrell, J.K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data, PLoS Genet. 8 (11) (2012) e1002967.
[54] M. Lang, H. Liu, F. Song, X. Qiao, Y. Ye, H. Ren, J. Li, J. Huang, M. Xie, S. Chen, et al., Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population, Forensic Sci Int Genet 42 (2019) e13–e20.
[55] R. Bai, Y. Liu, J. Zhang, M. Shi, H. Dong, S. Ma, R.F. Bai, M. Shi, Analysis of 27 Y-chromosomal STR haplotypes in a Han population of Henan province, Central China, Int. J. Leg. Med. 130 (5) (2016) 1191–1194.
[56] W. Song, S. Zhou, W. Yu, Y. Fan, X. Liang, Genetic analysis of 42 Y-STR loci in Han and Manchu populations from the three northeastern provinces in China, BMC Genom. 24 (1) (2023) 578.
[57] H. Fan, X. Zhang, X. Wang, Z. Ren, W. Li, R. Long, A. Liang, J. Chen, T. Song, Y. Qu, et al., Genetic analysis of 27 Y-STR loci in Han population from Hainan province, southernmost China, Forensic Sci Int Genet 33 (2018) e9–e10.
[58] L. Li, L. Yao, X. He, H. Gong, Y. Deng, M. Luan, G. He, F. Jia, P. Chen, Haplotype diversity and phylogenetic characteristics for guanzhong han population from northwest China via 38 Y-STRs using Yfiler™ platinum amplification system, Molecular Genetics & Genomic Medicine 8 (5) (2020).
[59] L. Luo, L. Yao, S. Chai, H. Zhang, M. Li, J. Yu, X. Hu, C. Li, Y. Bian, P. Chen, Forensic characteristics and population construction of two major minorities from southwest China revealed by a novel 37 Y-STR loci system, R. Soc. Open Sci. 8 (7) (2021).
[60] M. Xie, F. Song, J. Li, M. Lang, H. Luo, Z. Wang, J. Wu, C. Li, C. Tian, W. Wang, et al., Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs, Forensic Sci. Int.: Genetics 41 (2019) 11–18.
[61] F. Song, M. Xie, B. Xie, S. Wang, M. Liao, H. Luo, Genetic diversity and phylogenetic analysis of 29 Y-STR loci in the Tibetan population from Sichuan Province, Southwest China, Int. J. Leg. Med. 134 (2) (2020) 513–516.
[62] M. Wang, G. He, X. Zou, J. Liu, Z. Ye, T. Ming, W. Du, Z. Wang, Y. Hou, Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels, Forensic Sci. Int.: Genetics 54 (2021).
[63] Bo Wen, Hui Li, Daru Lu, Xiufeng Song, Feng Zhang, Yungang He, Feng Li, Yang Gao, Xianyun Mao, Liang Zhang, et al., Genetic evidence supports demic diffusion of Han culture, Nature 431 (7006) (2004) 302–305.
[64] J.Y. Chu, W. Huang, S.Q. Kuang, J.M. Wang, J.J. Xu, Z.T. Chu, Z.Q. Yang, K.Q. Lin, P. Li, M. Wu, et al., Genetic relationship of populations in China, Proc. Natl. Acad. Sci. U. S. A. 95 (20) (1998) 11763–11768.
[65] J. Chen, H. Zheng, J.-X. Bei, L. Sun, W-h Jia, T. Li, F. Zhang, M. Seielstad, Y.-X. Zeng, X. Zhang, et al., Genetic structure of the han Chinese population revealed by genome-wide SNP variation, Am. J. Hum. Genet. 85 (6) (2009) 775–785.
[66] M.A. Yang, X. Fan, B. Sun, C. Chen, J. Lang, Y.C. Ko, C.H. Tsang, H. Chiu, T. Wang, Q. Bao, et al., Ancient DNA indicates human population shifts and admixture in northern and southern China, Science (New York, NY) 369 (6501) (2020) 282–288.
[67] P.A. Cowin, M. Anglesio, D. Etemadmoghadam, D.D. Bowtell, Profiling the cancer genome, Annu. Rev. Genom. Hum. Genet. 11 (2010) 133–159.
[68] D.R. e Silva, M.P. Curado, J.C. de Oliveira, High incidence of esophageal cancer in central-western Brazil: a migrant effect? Eur. J. Cancer Prev. : the official journal of the European Cancer Prevention Organisation (ECP) 22 (3) (2013) 235–243.
[69] A.J. Swerdlow, M.G. Marmot, A.E. Grulich, J. Head, Cancer mortality in Indian and British ethnic immigrants from the Indian subcontinent to England and Wales, Br. J. Cancer 72 (5) (1995) 1312–1319.

[70] H. Huang, M. Su, X. Li, H. Li, D. Tian, Y. Gao, Y. Guo, Y-chromosome evidence for common ancestry of three Chinese populations with a high risk of esophageal cancer, PLoS One 5 (6) (2010) e11118.

[71] X.Y. Li, M. Su, H.H. Huang, H. Li, D.P. Tian, Y.X. Gao, mtDNA evidence: genetic background associated with related populations at high risk for esophageal cancer between Chaoshan and Taihang Mountain areas in China, Genomics 90 (4) (2007) 474–481.

[72] Y. Sun, M. Wang, Q. Sun, Y. Liu, S. Duan, Z. Wang, Y. Zhou, J. Zhong, Y. Huang, X. Huang, et al., Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people, J. Genet. Genom. (2023). S1673-8527(23)00210-2.

[73] L. Shiping, Population History of Sichuan, 1987.

[74] G. Jianxiong, C. Shuji, W. Songdi, A Brief History of Chinese Immigration, 1993.

[75] M. Shi, R. Bai, X. Yu, J. Lv, B. Hu, Haplotype diversity of 22 Y-chromosomal STRs in a southeast China population sample (Chaoshan area), Forensic Sci. Int.: Genetics 3 (2) (2009) e45–e47.

[76] M. Wang, Z. Wang, Y. Zhang, G. He, J. Liu, Y. Hou, Forensic characteristics and phylogenetic analysis of two Han populations from the southern coastal regions of China using 27 Y-STR loci, Forensic Sci. Int.: Genetics 31 (2017) e17–e23.

[77] Y. Zhou, C. Shao, L. Li, Y. Zhang, B. Liu, Q. Yang, Q. Tang, S. Li, J. Xie, Genetic analysis of 29 Y-STR loci in the Chinese Han population from Shanghai, Forensic Sci. Int.: Genetics 32 (2018) e1–e4.

[78] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (8) (2003) 598–612.

[79] L. Ke, Mortality and incidence trends from esophagus cancer in selected geographic areas of China circa 1970-90, Int. J. Cancer 102 (3) (2002) 271–274.

[80] M.L. Freedman, D. Reich, K.L. Penney, G.J. McDonald, A.A. Mignault, N. Patterson, S.B. Gabriel, E.J. Topol, J.W. Smoller, C.N. Pato, et al., Assessing the impact of population stratification on genetic association studies, Nat. Genet. 36 (4) (2004) 388–393.

[81] J. Marchini, L.R. Cardon, M.S. Phillips, P. Donnelly, The effects of human population structure on large genetic association studies, Nat. Genet. 36 (5) (2004) 512–517.