

Supplementary Information for

**Ancient DNA Reveals the Population Interactions and A Neolithic
Patrilineal Community in Northern Yangtze Region**

Tingyu Yang,^{#,1,2} Jianing He,^{#,3} Chunmei Li,^{#,1,4} Chao Ning,^{#,3} Tianming Zhang,¹
Jincheng Wang,¹ Hai Zhang,³ Guanbo Wang,^{1,5} Xiaowen Jia,³ Chi Zhang,^{3,*} Shi
Yan,⁶ Xiannian Zhang,¹ Fan Wu,¹ Chaodong Wu,⁷ Xiaohong Wu,^{3,*} Yuhong
Pang,^{1,*} and Yanyi Huang^{1,2,5,8,*}

These authors contributed equally

* Correspondence to: yanyi@pku.edu.cn (Y. H.), pangyuhong@pku.edu.cn
(Y. P.), wuxh@pku.edu.cn (X. W.), csczc@pku.edu.cn (C. Z.)

This PDF file includes:

Supplementary Notes 1-3
Supplementary Figures 1-14
Supplementary References

Supplementary Note 1. Archaeological Background of Baligang ancient individuals

The middle and lower reaches of the Yellow River in northern China and the middle and lower reaches of the Yangtze River in southern China are the important founding areas of East Asian populations, which are also the centers of the origin and domestication of two major types of crop in the world, dryland millet and wetland rice, respectively. The development and spread of agriculture has driven rapid population expansion and the formation of cultures and complex societies.

The Baligang site^{1,2}, discovered in 1957 and excavated on a large scale from 1992 to 2010 in Dengzhou City, Henan Province, China (Fig. 1a-c), is located on the Nanyang basin, an area on a northern tributary of the middle Yangtze River, near a route of communication to the Yellow River. The earliest archaeological remains found here are 11 ash pits under the Yangshao culture layer, stratigraphically referred to as the Pre-Yangshao period (8,700-8,300 BP), also belonged to the Jiahu I Culture. Then after a hiatus, lots of materials belonging to the Yangshao culture period (6,300-5,000 BP) were successively excavated. These materials reveal pottery, house foundations, cemeteries, and storage pits (backfilled with ashy midden after being abandoned), which are related to the cultural of Yellow River. However, subsequent archaeological remains show an affiliation to the Yangtze River cultures, including Qujialing (5,000-4,500 BP) and Shijiahe culture (4,500-4,300 BP) periods, represented by houses, pits (probable storage pits) and urn burials. Above them are burial layers of the Late Longshan culture period (4,300-3,800 BP) and the Bronze age (mainly Western and Eastern Zhou, 3,000-2,500 BP), which once again point to northern cultural affinity.

1.1 Pre-Yangshao

The archeological remains excavated from Baligang site in pre-Yangshao period (8,700-8,300 BP) are very limited to 11 ash pits. No human skeleton remain is discovered, but the analysis of archaeobotanical remains provides evidence for domesticated rice (*Oryza sativa*) cultivation. The deficiency of dryland millet in this period, which is similar to the Jiahu site located in the upper reaches of Huai River, may imply the relations between them.

1.2 Yangshao Culture

The Yangshao Culture is a well-known Neolithic painted pottery culture that widely existed along the Yellow River basin in China³⁻⁵. It lasts more than two thousand years (7,000-5,000 BP), which has been considered to have profound impact on Chinese civilization and represents the typical evolution pattern of social structure in early China. It was first discovered in 1921 by Swedish geologist and archaeologist Johan Gunnar

Andersson at the site of Yangshao Village, Mianchi County, Henan Province, making the world realize for the first time that China has its own Neolithic culture.

Geographically, the Yangshao culture is centered on the Central Plains where the tributaries of the Yellow River and Huai River converge. It extends from the Great Wall and the Hetao Plain in the north, to the northwest of Hubei province in the south, Henan province in the east and Qinghai province in the west.

The dominating agricultural type of Yangshao culture is millet crop. Along with the rise of Yangshao culture, millet farming has rapidly developed and matured, which then widely spreaded to northern China. In Baligang site, although rice is overall the most important crop in all periods of Neolithic Age, the crop composition of the site has transformed to an assembly of rice (*Oryza sativa*) and millets (*Setaria italica* and *Panicum miliaceum*) since the Yangshao culture period. The proportion of foxtail millet (26%) and broomcorn millet (32%) together outnumbered rice (42%).

The typical pottery of Yangshao Culture is painted pottery, which developed to a great extent during this period⁶. The appearance and prosperity of painted pottery indicates that consciousness of pursuing aesthetic already emerged. The pottery of Baligang site in the period of Yangshao culture mainly belongs to red clay pottery or with some sand inside, a lot of painted pottery has been discovered.

The burials of the prosperous period of Yangshao culture display two types: the single burials and the large secondary joint burials. The single burials are typically located on the edge of villages and are divided into multiple graves, whereas the large secondary joint burials are generally in the center of a settlement within only one area of land. All the individuals found in secondary joint burial grave were exhumed from their primary burials and re-buried together. Interestingly, this kind of burial practice disappeared with the termination of the Yangshao culture^{7,8}.

In Baligang site, various burial types of the prosperous period of Yangshao culture have been discovered and displayed a transition over time. In the early period (about 6,000 BP), single burial (one individual in each grave) is dominant with extended supine mode. Later, along with cultural development and population growth, large secondary joint burials gradually prevailed at about 5,000 BP. Among them, grave No. 13 (M13) is the largest which is located between two rows of houses in the north and south, containing more than 90 human remains by column-and-row organization. The ages of individuals in the M13 grave span at least 200 years measured by random radiocarbon-dating, suggesting that some of the human bones in M13 were collected by a certain ethnic group in a period of time. This may represent a special social or community ceremonial practice.

Nine individuals belonging to Yangshao culture in Middle Neolithic were selected for aDNA analysis:

- BLG-M180: Male, 6166-5903 BP
- BLG-M184: Male, 5996-5759 BP
- BLG-M174: Male, 6170-5910 BP
- BLG-M261: Male, 5927-5750 BP
- BLG-M167: Male, 6104-5766 BP
- BLG-M179-1: Female, 5930-5747 BP
- BLG-M231: Male, 5998-5769 BP
- BLG-M76-2: Male, 5909-5726 BP
- BLG-M224-3: Male, 5585-5331 BP

Thirty-two individuals from the M13 grave belonging to Yangshao culture in Late Neolithic were selected for aDNA analysis, radiocarbon-dating to be approximate 5,000 BP:

- BLG-M13-20: Male
- BLG-M13-39A: Female
- BLG-M13-6: Female
- BLG-M13-67: Male
- BLG-M13-75: Male
- BLG-M13-76: Male
- BLG-M13-86: Female
- BLG-M13-69: Male
- BLG-M13-90: Male
- BLG-M13-4: Male
- BLG-M13-7: Male
- BLG-M13-8: Male
- BLG-M13-12: Male
- BLG-M13-14: Female
- BLG-M13-17: Female
- BLG-M13-25: Male
- BLG-M13-28: Female
- BLG-M13-30: Male
- BLG-M13-41B: Female
- BLG-M13-44: Male
- BLG-M13-45: Female
- BLG-M13-46: Male
- BLG-M13-51: Male
- BLG-M13-56: Male
- BLG-M13-64A: Female
- BLG-M13-68: Male
- BLG-M13-70: Female

- BLG-M13-77: Male
- BLG-M13-87: Male
- BLG-M13-88: Male
- BLG-M13-89: Male
- BLG-M13-TX4: Female

1.3 Qujialing Culture

Qujialing culture⁹⁻¹¹ is the late neolithic culture in the middle and lower reaches of the Yangtze River, which existed between 4,900-4,500 years ago. It was first discovered in Qujialing, Jingshan City, Hubei Province in the 1950s, and named after the site. Large-scale ancient cities began to be extensively built during this period, dozens of sites have been discovered in the middle reaches of the Yangtze River, representing a prehistoric settlement pattern characterized by many city ruins. Well-organized cities and religious ceremonies here also reveal stratified society, which might be the foundation for the construction of Chinese civilization.

The Qujialing culture is centered on the Jiangnan Plain, which geographically distributes from southwestern of Henan Province in the north, to Mengxi Sanyuan Palace, Hunan Province in the south, the Three Gorges in the west and Wuhan, Hubei Province in the east. The Qujialing culture also expands northward to the middle reaches of the Yellow River and profoundly affects Yangshao culture there.

During the period of Qujialing culture, a large number of tools and japonica rice chaffs have been unearthed, illustrating that agricultural production provided the main food for subsistence and rice occupied a dominant position. At this period in Baligang site, the proportion of rice increased to more than 85%, and foxtail millet decreased to 10%, broomcorn millet is even lower.

Muddy gray pottery was predominant in the early stage of Qujialing culture, followed by black and red pottery. In the late period of the Qujialing culture, brick-red colored sand pottery slightly increased. Throughout the Qujialing culture, pottery surfaces had been mostly plain polished, parts of the abdomen decorated with string patterns, and the underside of pots often pierced. A small number of painted pottery is engraved with shallow basket and carved patterns, while most painted pottery is engraved with black, gray and brown patterns on their surfaces. Geometric patterns with dots and lines are the common.

The burials of the Qujialing culture are mainly pit grave. Most of adults were singly buried in public cemeteries of the clan with extended supine mode.

Two individuals belonging to Qujialing culture in Late Neolithic were selected for aDNA analysis:

- BLG-H1284: Male, 4817-4529 BP

- BLG-H1416: Female , 4567-4417 BP

1.4 Shijiahe Culture

The Shijiahe culture is a late Neolithic culture centered on the middle Yangtze River region, which arises after the Qujialing culture in the same region^{11,12}. It exists between 4,500-4,200 years ago. The culture is named after the Shijiahe site complex in Tianmen, Hubei province, located on the Middle Yangtze Valley where more than 500 Shijiahe cultural sites have been discovered, including some famous sites like Dengjiawan, Xiaojiawuji, Qinglongquan, Qilihe, Xiangxiangdaijiping. The discovery of bronze pieces, jade ware and sacrificial relics indicates it has come to civilization.

Shijiahe Culture is centered on the Jiangnan Plain in the middle and lower reaches of the Han River which are the main tributary of the Yangtze River. Its distribution ranges from Huangmei area in the east, to Dongting Lake in the south, Han River (the upper reaches of the Danjiang River) in the west, and Tongbai Mountain and Dabie Mountain in the north.

Similar to the Qujialing culture period, rice was still predominately planted by people living on the Shijiahe culture sites, although millet was also used for food. In the Baligang site, the proportion of crops during this period is similar to that in Qujialing culture period.

Except for a small amount of handmade pottery, Shijiahe pottery is wheel pottery among which mud pottery is the most followed by sand pottery and a small amount of carbon pottery. The amount of gray pottery far exceeds that of black and red pottery. The surface of the utensils is mainly plain with common patterns of basket and string. Jar-like shape is most popular, especially the thick high neck jars and olive-shaped jars with a middle mouth. Meanwhile, many jade artifacts have been unearthed in the late phase of Shijiahe sites. The functions of the jade include sacrifice, decoration, tool and gift, etc.

The burials of the Shijiahe culture are mainly pit grave.

Four individuals belonging to Shijiahe culture in Late Neolithic were selected for aDNA analysis:

- BLG-M208: Female, 4403-4086 BP
- BLG-H182: Female
- BLG-H292: Female, 4227-3987 BP
- BLG-H197: Male, 4226-3984 BP

1.5 Longshan Culture

Generally, Longshan culture¹³ is a late Neolithic culture located in the middle and lower reaches of the Yellow River, named by the discovery of Longshan town in Zhangqiu,

Shandong province. It exists between 4,500-3,800 years ago. A large number of exquisitely polished black pottery has been unearthed, which is the main cultural heritage feature of this period.

Most of the Longshan cultural sites distribute in the Shandong Peninsula, and others have also been found in Shaanxi, Shanxi, Henan, Hebei, Jiangsu and Hubei provinces. The most prominent characteristic of the Longshan culture is the discovery of city sites.

The most important crop in the Longshan culture period is foxtail millet, whereas just a trace of broomcorn millet and some rice have been found. However, in the early period of Longshan culture, rice (80%) is predominately used by people living on the Baligang site then followed by foxtail millet (15%) and common millet (3%). In the late period, the proportion of rice decreases significantly and millet becomes more important. It's worth noting that wheat is first found in this period.

The pottery of the early Longshan culture in the Central Plains is mainly handmade gray, inheriting some elements from the Yangshao culture style. Most of the decorative patterns of pottery are basket, and some pottery is decorated with additional pile patterns even on the whole body used for strengthening. The pottery of the late Longshan culture is mainly gray pottery, the number of black pottery has increased compared to the Shijiahe culture period and red pottery only occupies a certain proportion. Most of them are still made by hand, but the technological innovation of wheel making has been further developed, and some pottery has been molded. Grain is decorated with rope grain. Basket grain is the most common, while a few square grain can still be seen.

The burials of the Longshan culture are mainly pit grave in rectangular shape which also include a small number of sarcophagus graves. Single burial is the main type, while secondary burial and joint burial are relatively less common. Adult urn coffins were also present during the late Longshan period.

Six individuals belong to Longshan culture in Late Neolithic were selected for aDNA analysis:

- BLG-M203: Male, 4090-3877 BP
- BLG-H616: Male, 4085-3876 BP
- BLG-H932: Female, 3966-3727 BP
- BLG-H1664: Female, 4083-3845 BP
- BLG-M209: Female, 3900-3719 BP
- BLG-H388: Female, 3966-3727 BP

1.5 Zhou Dynasty

The Zhou Dynasty is a historic period in Bronze age in China, existing between 3,046-2256 years ago. China's social system has experienced drastic changes during this period. Bronze casting sites are commonly found around the capitals of all countries. The bronze artifacts are mainly used as ritual vessels and musical instruments for sacrifice, feast or wedding and funeral ceremonies. Besides, iron tools are also widely used in handicrafts and agriculture.

In Baligang site, the proportion of millet is about 64% (54% for foxtail millet and 10% for broomcorn millet) and rice is 24%, wheat also increases to a certain extent in this period .

Six individuals belong to Eastern Zhou period in Bronze Age were selected for aDNA analysis:

- BLG-M198: Male, 2719-2434 BP
- BLG-M191: Female, 3059-2876 BP
- BLG-M194: Female, 3004-2855 BP
- BLG-M195: Male, 2995-2849 BP
- BLG-M204: Male, 2692-2348 BP

Supplementary Note 2. Second-degree kinship classification.

In order to identify the exact pedigree relationships between the 2nd degree pairs in BLG-M13, we derived a method utilizing multi-way IBD sharing based on the foundational principles of CREST¹⁴. Generally, different types of 2nd degree pairs—such as grandparent-grandchild (GP), avuncular (AV), or half siblings (HS)—exhibit distinct IBD sharing patterns with their mutual relatives. Quantitatively, as is described by Qiao et al., for a pair of 2nd degree relatives (x_1, x_2) and their mutual 2nd -5th degree relative y , two quantities R_1 and R_2 are defined as $R_i = \frac{\text{len}(\text{IBD}(x_1, x_2, y))}{\text{len}(\text{IBD}(x_i, y))}$ ($i \in \{1, 2\}$), where $\text{IBD}(s_1, s_2, \dots, s_n)$ denotes the set of IBD regions that all samples $s_1 \sim s_n$ share, and $\text{len}(\text{set})$ denotes the sum of genetic lengths (in Morgan) for a set of IBD segments. If y is related to both x_1 and x_2 solely through the most recent common ancestor(s) (MRCA(s)) of x_1 and x_2 , by counting the number of meioses separating individuals, one can easily calculate that the mean values of (R_1, R_2) are (0.5, 0.5), (0.25, 0.5) and (0.25, 1) when (x_1, x_2) are HS, AV and GP pairs, respectively (we take the convention that x_1 is genetically older than x_2 here and in the following text). Consequently, the 3 types of 2nd-degree pairs can be distinguished using (R_1, R_2) values, achieving high accuracy when combining information from multiple mutual relatives.

The original version of CREST operates on these principles, with the MRCA restriction of the mutual relative y easily achievable in modern genetic data. However, for extensive pedigrees spanning 5 or more generations, violations of this restriction can be common and challenging to avoid without sufficient information, severely impacting CREST's performance in such cases. To address this challenge, we adopted a straightforward approach by considering all possible violation cases and their occurrence probabilities. For instance, in the case of a GP pair (x_1, x_2) and an arbitrary mutual relative y , the theoretical mean values of (R_1, R_2) can be one of the followings, as is depicted in Supplementary Fig. 12A:

- I. (0.25, 1), for a y that fulfills CREST's MRCA assumption;
- II. (0.25, 0.5), for a y who is or is a direct descendant of a full sibling of x_2 's parent connecting x_1 and x_2 ;
- III. (0.5, 0.5), for a y who is or is a direct descendant of a half sibling of x_2 ;
- IV. (0.5, 0.25), for a y who is or is a direct descendant of a full sibling of x_2 ;
- V. (1, 0.25), for a y who is a direct descendant of x_2 .

Similar analysis can be done for AV and HS pairs, where these 5 cases of mean (R_1, R_2) values cover all possibilities. Building upon this, we performed simulations using PEDSIM to determine the occurrence rates of these cases in extensive pedigrees and the corresponding distributions of (R_1, R_2) . Pedigree generation

specifications are the same as is described in Supplementary Note 3.2, except for the introduction of multiple spouses to account for the occurrence of half siblings. In total, we simulated 100 random pedigrees, each spanning 8 to 10 generations and comprising 150 to 200 individuals, matching a likely case for the BLG-M13 population, as described in Supplementary Note 3.1. Every mutual 2nd to 5th degree relative of each pair of 2nd-degree relatives was classified into the aforementioned cases (a total of $3 \times 5 = 15$ cases), and corresponding (R_1, R_2) values were summarized (Supplementary Fig.12B).

Subsequently, kernel density estimation (KDE) models were generated for each case using Gaussian kernels and bandwidth chosen according to Scott's Rule (scipy.stats). A naïve Bayesian approach was then derived to classify a 2nd-degree pair given a set of (R_1, R_2) s calculated with their mutual relatives. For a specific (R_1, R_2) point, one can compute from the KDEs the probability it occurs under each case:

$$P((R_{1i}, R_{2i})|y \text{ type, pair type})$$

The probability it occurs under each 2nd-degree pair type can then be calculated as

$$P((R_{1i}, R_{2i})|\text{pair type}) = \sum_{y \text{ types}} P((R_{1i}, R_{2i})|y \text{ type, pair type})P(y \text{ type}|\text{pair type})$$

Where $P(y \text{ type}|\text{pair type})$ can be estimated from the occurrence rates of each case in the simulated dataset. Combining information from all (R_1, R_2) points assuming independence between points and using the Bayes' formula, one gets

$$\begin{aligned} P(\text{pair type} | (R_{11}, R_{21} \dots R_{1n}, R_{2n})) \\ &= \frac{P((R_{11}, R_{21} \dots R_{1n}, R_{2n})|\text{pair type})P(\text{pair type})}{\sum_{\text{pair types}} P((R_{11}, R_{21} \dots R_{1n}, R_{2n})|\text{pair type})P(\text{pair type})} \\ &\approx \frac{P(\text{pair type}) \prod_i P((R_{1i}, R_{2i})|\text{pair type})}{\sum_{\text{pair types}} P(\text{pair type})P(\text{pair type}) \prod_i P((R_{1i}, R_{2i})|\text{pair type})} \end{aligned}$$

Where the prior probabilities $P(\text{pair type})$ were set as uniform for all 3 types. The two possible directionalities within the GP or AV case were also assigned equal prior probabilities. Subsequently, the type and directionality of the 2nd-degree pair could be determined by accepting the case with the highest posterior probability.

Upon testing on 2nd degree pairs from 200 independently generated pedigrees spanning 8 to 10 generations and comprising 150 to 200 individuals, this simple method achieved precision rates of 77.8%, 70.9%, and 71.1% for GP, AV, and HS pairs, respectively. In the case of BLG-M13 data, it classified (BLG-M13-75, BLG-M13-7) as half siblings and (BLG-M13-90, BLG-M13-75) as well as (BLG-M13-90, BLG-M13-7) as avuncular, with BLG-M13-90 being genetically older. Given the current precision rates, it can be concluded that the core pedigree depicted in Supplementary Fig.12C represents the most probable scenario.

Supplementary Note 3. M13 kinship structure estimation.

3.1 Kinship structure estimation based on IBD sharing matrix simulation

According to IBD-sharing data inferred with ancIBD, sequenced individuals from BLG-M13 display extensive kinship connections. However, due to limited sample size, it is infeasible to directly determine lineages connecting these individuals, nor can conclusions be drawn on the kinship structure properties (e.g. Single or multiple lineages, total individual number of lineages) of the community they represent. Therefore, we turned to a probabilistic approach based on the overall IBD sharing pattern among sampled individuals. Generally, for a given target dataset like that of BLG-M13, we consider its constituent individuals as randomly sampled from a target population with kinship connections, the specific kinship structure of which we are interested in. (As a clarification, when we mention the "BLG-M13 population" later in the text, we also take this "sampling population" notion.) A basic intuition is that statistically, samples taken from populations with similar kinship structure should also exhibit analogous pairwise IBD sharing profiles, and vice versa. Based on this, we generated hypothetical populations with IBD sharing profiles by PEDSIM¹⁵ under various kinship structure assumptions (details on simulation specifications are discussed in the next section). From these populations, we drew samples and summarized pairwise total IBD sharing matrices. Subsequently, by observing those hypothetical populations whose samples show higher IBD sharing profile similarity with the target dataset, key kinship structure properties of the target population of interest can be approximated (Supplementary Fig.13A).

Specifically, for each simulated population, 50 rounds of independent sampling were performed. The sample number and sex ratio of each sample were set to match that of the target dataset. Subsequently, we computed the pairwise total length of valid IBD segments >20cM between sampled individuals, generating a IBD sharing matrix. To enhance comparability, male and female individuals were clustered separately using the UPGMA method according to pairwise total IBD length. The optimal leaf order was then applied to reorder the columns and rows of the IBD matrix (Supplementary Fig.13A, columns 2&3). The IBD matrix of the dataset of interest was filtered and reordered with the same procedure. Finally, we calculated the similarity index between this simulated matrix and the target matrix with a simple measurement $S(A, B) = 1 - \|A - B\| / (\|A\| + \|B\|)$, where $\|A\|$ represents the Forbenius norm of the matrix A . This process yielded a 2-dimensional similarity array, which was subsequently utilized for further analysis (Supplementary Fig.13A, columns 4&5).

The two main kinship structure properties we considered were the centrality C (defined as the population proportion taken up by the largest single lineage, as illustrated in in Supplementary Fig.13A) and the total individual number N . Estimation

of target population centrality mainly aims to discern whether it exhibits a dominant single lineage (resembling a "family tomb" scenario) or a combination of several largely independent lineages (akin to a "public burial" scenario). On another hand, estimation of the total individual number of the target population is primarily concerned with assessing the size of the community and the extent of community member inclusion in the burial.

We first developed and tested the centrality estimation procedure. In the simulation part, we used 6 total individual number intervals ranging from [50,100) to [400,500). For each interval, we independently simulated over 2,000 populations with a near-uniform centrality distribution from 0.2 to 1. Subsequently, sample extraction, matrix generation, and similarity calculation followed the previously described methodology. The simulated populations were then ranked based on the mean similarity indexes of their samples. The mean centrality of simulated populations with the top 10% mean similarities, $\bar{C}_{\text{top10\%}}$, is compared to that of simulated populations with the median 10% mean similarities, $\bar{C}_{\text{med10\%}}$. A high value of $\bar{C}_{\text{top10\%}}$ that significantly exceeds $\bar{C}_{\text{med10\%}}$ suggests a high centrality C of the target population, namely with a dominant single lineage. Specifically, we framed it as a classification problem: if $\bar{C}_{\text{top10\%}}$ exceeds a threshold t_1 and $\bar{C}_{\text{top10\%}} - \bar{C}_{\text{med10\%}}$ exceeds another threshold t_2 , we conclude that the target population comprises a dominant single lineage (positive result); otherwise, we conclude that the target population consists of multiple independent lineages (negative result).

To test the validity of this approach and find optimal t_1 and t_2 values, we independently simulated over 4,000 populations with total individual numbers ranging from 125 to 300 and centralities spanning from 1 to <0.2 as a test dataset. We then applied the procedure using IBD matrices generated with samples from these populations as target matrices. Taking $C > 0.8$ as the practical threshold for dominant single lineage, we summarized true positive rates (TPR) and false positive rates (FPR) of our method across different combinations of t_1 and t_2 . When t_1 varied from 0.85 to 0, with t_2 set to $0.8(t_1 - 0.5)$ (where the factor 0.8 is an insensitive parameter), the area under the ROC curve was approximately 0.73, indicating moderate classification performance (Supplementary Fig.13C). The optimal threshold values (t_1, t_2) were determined from the ROC curve to be (0.725, 0.18), maximizing the difference between TPR and FPR and yielding a precision rate of 67%, assuming equal prior probabilities for positive and negative samples.

Applying this procedure to BLG-M13 samples resulted in $\bar{C}_{\text{top10\%}} = 0.852$ and $\bar{C}_{\text{med10\%}} = 0.547$, significantly surpassing the predetermined thresholds. Notably, according to the previous test, setting $(t_1, t_2) = (0.85, 0.28)$ increased the method's precision rate to 78% (Supplementary Fig.13D). This value can be interpreted as a conservative estimate for the posterior probability that the BLG-M13 population has a

centrality $C > 0.85$, taking equal prior probabilities for $C > 0.85$ and $C \leq 0.85$. These results indicate that the BLG-M13 population likely comprises a dominant single lineage.

For simplicity, our population estimation procedure focused on populations comprised of dominant single lineages, a scenario likely applicable to BLG-M13, as discussed above. We independently simulated over 4,000 populations with $C = 1$ for each of 6 total individual number intervals from [50,100) to [400,500). Sample extraction, matrix generation, and similarity calculation were performed for each simulated population. Within each individual number interval, we computed the mean of the top 100 similarity values, as well as the mean similarity value of simulated populations exhibiting the top 50 mean similarity values. Ideally, either value would show a single-peak pattern over the individual number intervals, with the peak position indicating the most probable total individual number range for the target population. For BLG-M13 data, the former values displayed a single peak at (300,400), while the latter exhibited a single peak at (150,200) (Supplementary Fig.13E).

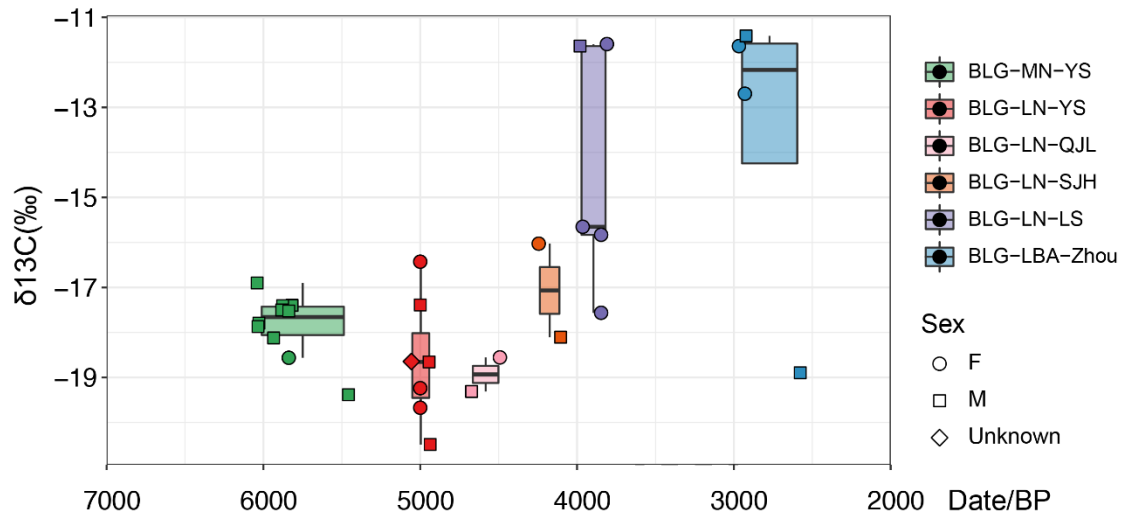
Regarding the inconsistency, rather than directly inferring the most likely individual number range from the peak positions, we estimated the posterior probability of different population ranges based on these specific results. We independently simulated over 1,500 populations with $C = 1$ and $N = 100 \sim 500$ as reference, ran the procedure using samples from them as proxies of real data, and summarized the results. The bounds on N were chosen based on archaeological estimations¹⁶. Both the distribution of matrix similarity peak position for top 100 samples, N_{ps} , and that for top 50 populations, N_{pp} , exhibited a notable shift towards higher values as N increased, as expected (Supplementary Fig.13F). The posterior probability distribution of N given test peak positions N_{ps} and N_{pp} could be approximated by counting these results. Specifically, for the BLG-M13 population, considering $N_{ps} \in [300,500)$ and $N_{pp} \in [150,300)$, and assuming a uniform prior distribution, it can be inferred that N exceeds 150 and 200 with posterior probabilities of 95.8% and 87.9%, respectively (Supplementary Fig.13G).

3.2 Pedigree simulation specifications

IBD information of populations with different kinship structures were simulated by PEDSIM¹⁵, which takes a custom lineage definition file and simulate recombination events across the lineage using sex-specific genetic maps and considering crossover interference. We use a custom script to generate lineage definition files. Key parameters it takes include generations, mean number of children (λ), population range (N), and outsider rate (p). Among these, the population range and outsider rate are mainly adjusted.

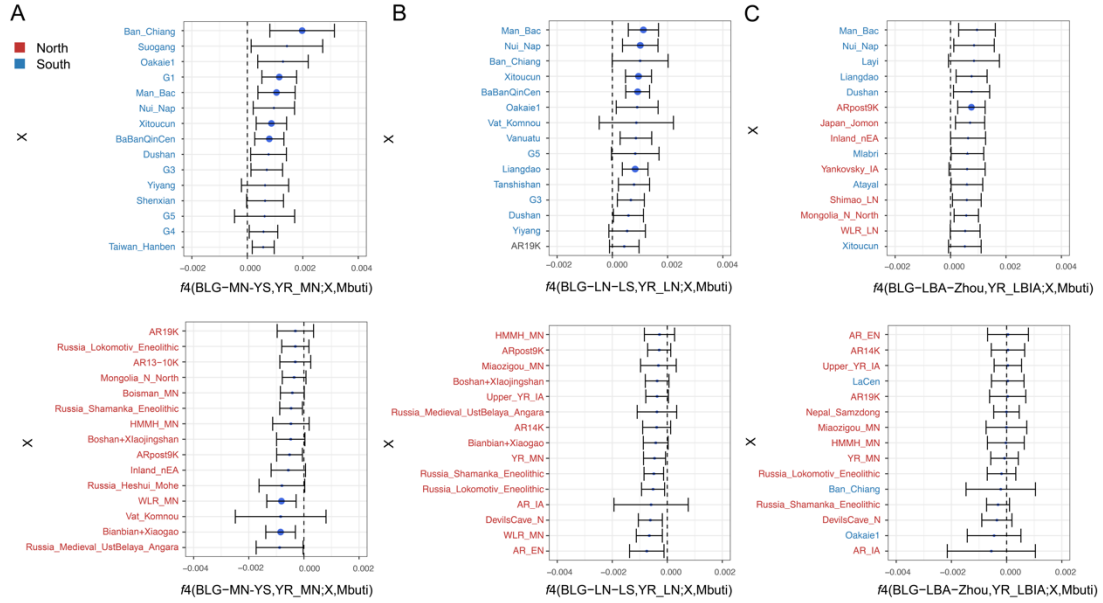
The outsider rate represents the ratio between the number of new founders and the number of offsprings from the previous generation in each generation. This parameter is directly adjusted to achieve different centralities. Under zero outsider rate, pedigrees with a single pair of founder individuals were generated. In each generation, each individual leaves a random number of children following an exponential distribution of $\lambda = 2$. Under a finite outsider rate p , for each generation, the total number of offsprings N_o from the previous generation is first determined from the exponential distribution of children number, after which the number of new founders is determined, following a normal distribution with a mean and variation $N_o p$. Additionally, in this case, the exponential factor of children number is set to $\lambda' = 2/(1 + p)$ to compensate for the contribution of new founders to the population. Under either setting, a small proportion (10%) of individuals are set as with random sex and leaving no offspring.

The population range is controlled by filtering in the actual simulation process. Under a certain population range setting, for all randomly generated populations, only those whose total individual numbers fall into the desired interval $(N_1, N_2]$ are used. Specifically, in the abovementioned centrality estimation part, we used population intervals $[50, 100)$, $[100, 150)$, $[150, 200)$, $[200, 300)$, $[300, 400)$ and $[400, 500)$, together with outsider rates 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7. For each pair of combination, 4,800 patrilineal populations with 9 or 10 generations (2,400 each) were generated. To avoid bias in further analysis, 2,000 populations with a near-uniform centrality distribution from 0.2 to 1 were sampled from each population interval as the final dataset. The test set is generated similarly, also with a near-uniform centrality distribution. In the population estimation part, for each population interval, 4,800 patrilineal populations with 9 or 10 generations (2,400 each) and zero outsider rate were generated.

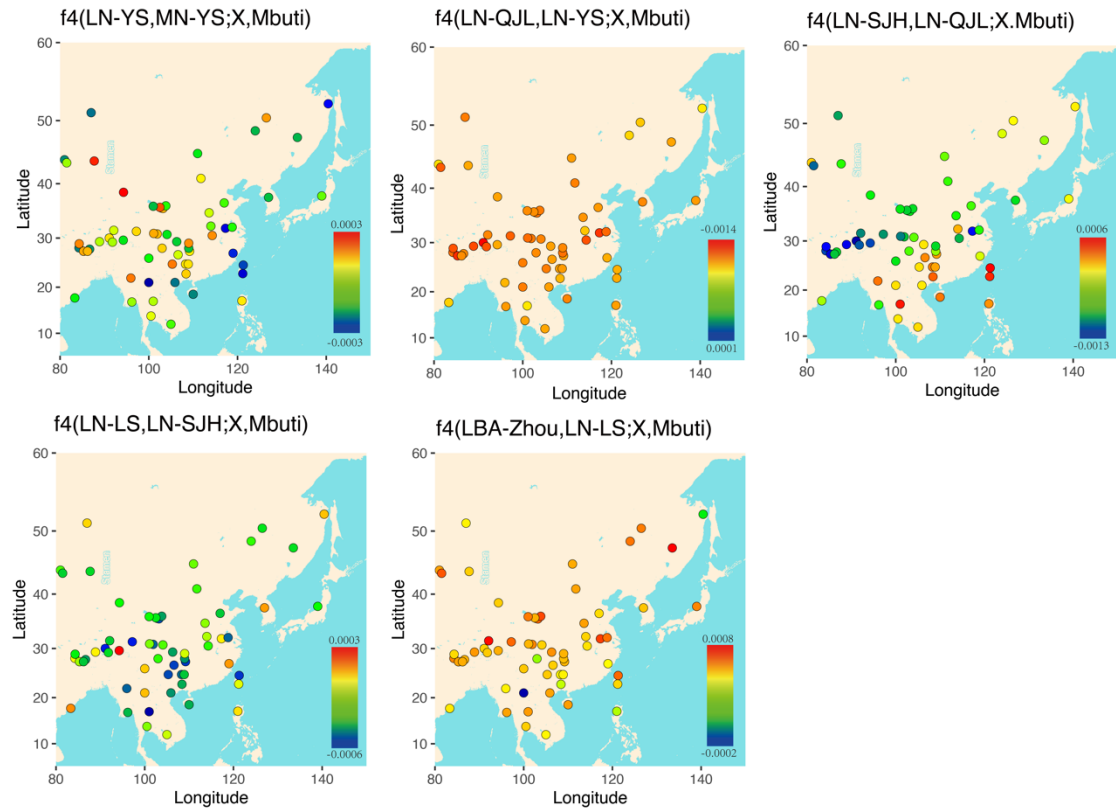


Supplementary Fig. 1. Stable isotopes analysis. The scatter plot shows the calendar ages of all period Baligang samples against their $\delta^{13}\text{C}$ values based on bone collagen stable isotope analysis, with individual sexes distinguished by different symbols. Each boxplot represents the distribution of points for each period, which show the median, first and third quartiles (lower and upper hinges), and whiskers extending to 1.5X interquartile range from the hinges.

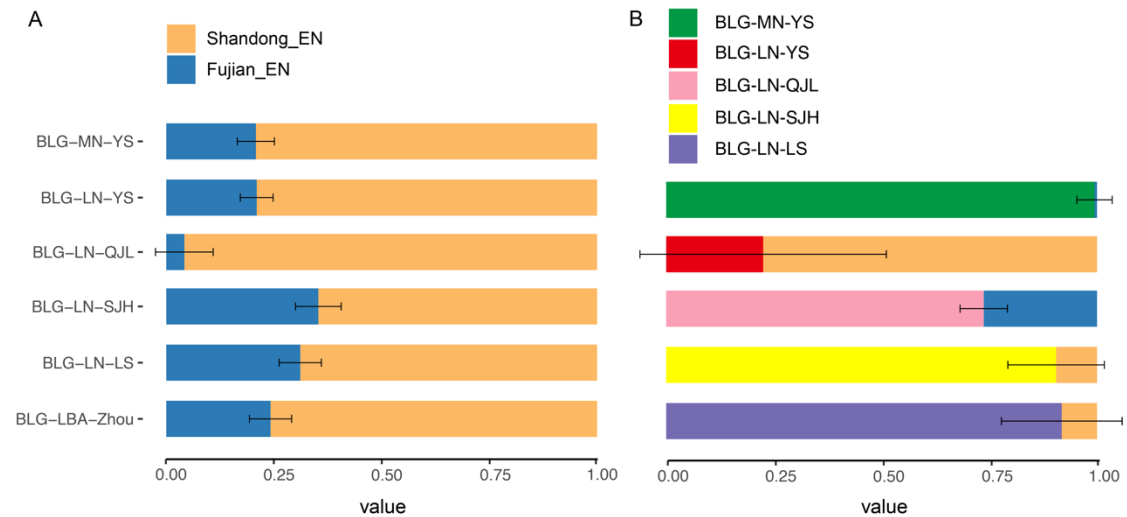
statistics across various ancient East Asians from diverse regions. Horizontal bars represent the point estimate ± 2 s.e.m, with significant values ($|Z| > 3$) indicated by larger symbols.



Supplementary Fig. 4. The genetic difference between Baligang and the Yellow River Basin populations from the same culture. For each period, we present 15 most positive (upper panels) and 15 most negative (lower panels) f_4 statistics across various ancient East Asians from diverse regions, briefly dividing them into north (red labels) and south (blue labels) East Asia. Horizontal bars represent the point estimate ± 2 s.e.m, significant values ($|Z| > 3$) are represented by large symbols. (A) $f_4(\text{BLG-MN-YS,YR_MN};X,Mbuti)$ from Yangshao cultural populations in Middle Neolithic period. (B) $f_4(\text{BLG-LN-LS,YR_LN};X,Mbuti)$ from Longshan cultural populations in Late Neolithic period. (C) $f_4(\text{BLG-LBA-Zhou-YS,YR_LBIA};X,Mbuti)$ from populations in Late Bronze Age period.

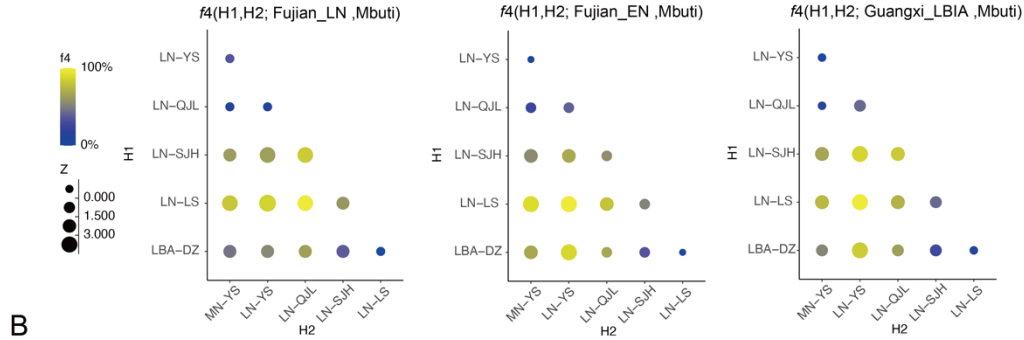


Supplementary Fig. 6. The different genetic affinities between each adjacent time of Baligang population to present-day East Asian and Southeast Asian populations. This figure is the supplement for Fig. 2d, red indicates more affinity to later period, whereas blue indicates greater affinity to early period, so the color gradient from red to blue presents the increased gene flow in each time interval. Source data are provided as a Source Data file.

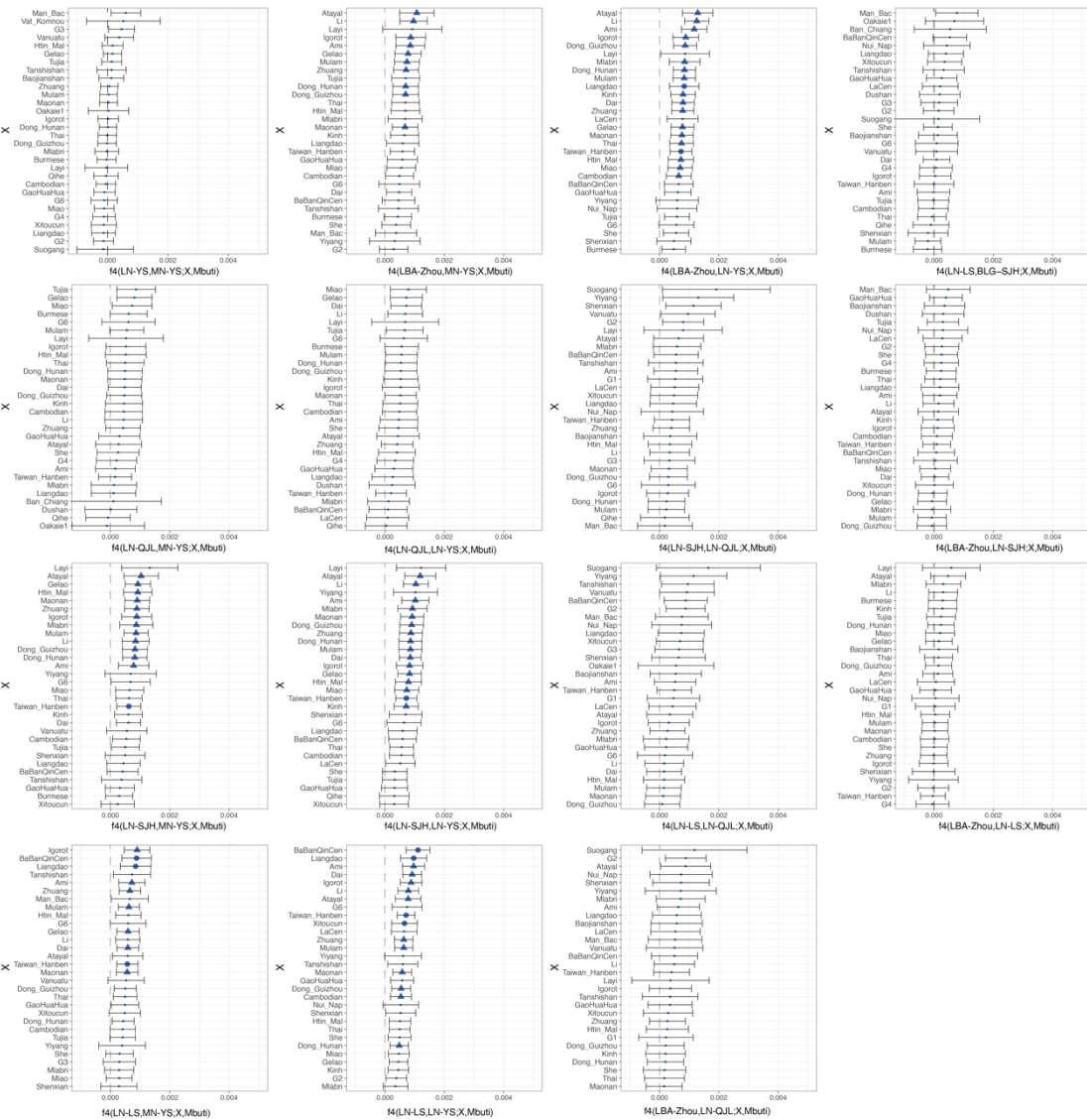


Supplementary Fig. 7. Admixture modeling of Baligang populations related to Fig. 3a. (A) QpAdm modeling of Baligang populations from diverse periods, suggesting an admixture between northern East Asian (nEA, represented by Shandong_EN) and southern East Asian (sEA, represented by Fujian_EN) genetic elements. Error bars indicate the standard errors of estimated ancestry proportions determined by 5-cM block jackknifing. (B) The corresponding proximal model connecting the earlier and later Baligang populations, and the increase in genetic components in each period corresponds to the distal model on the left.

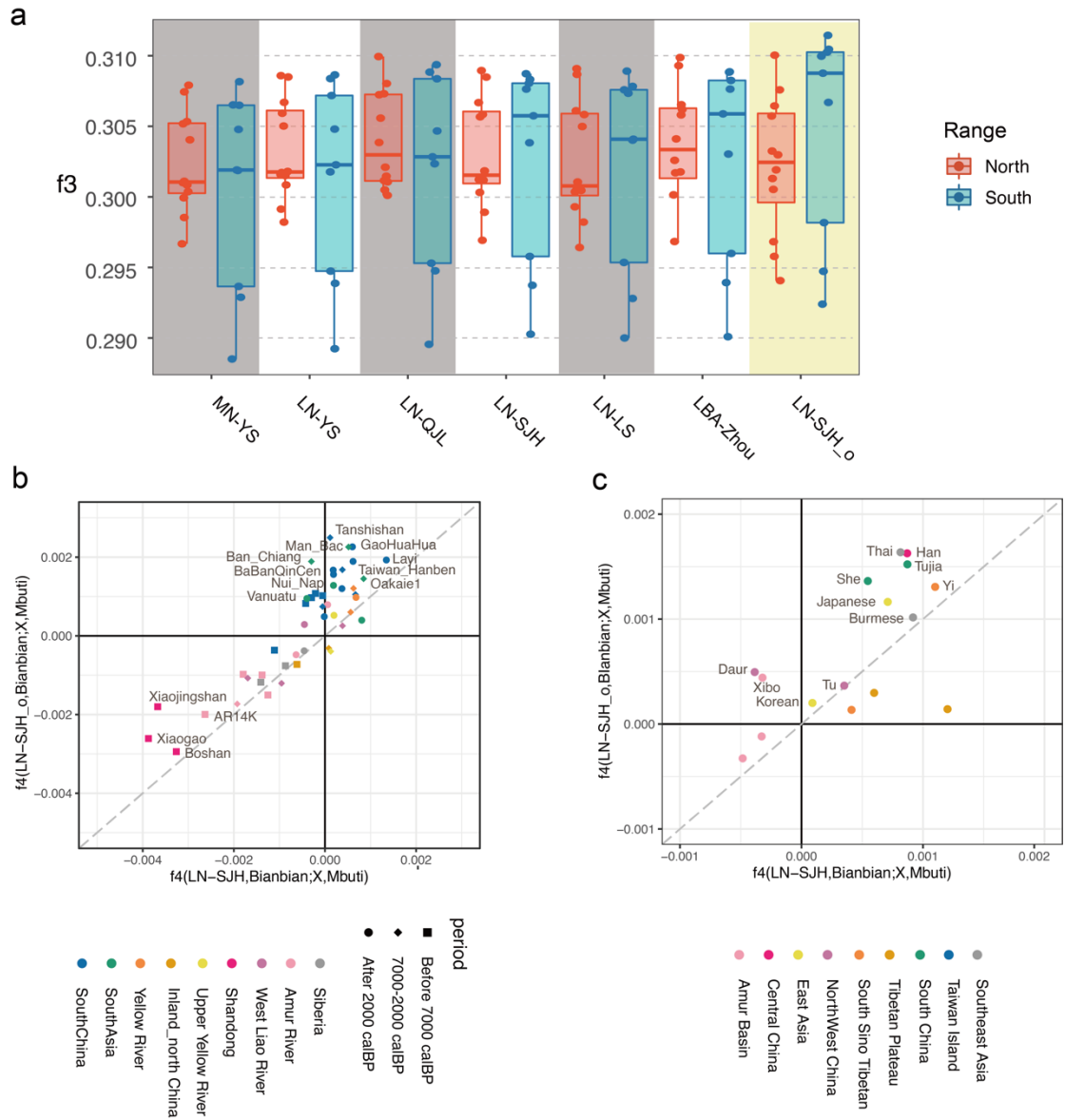
A



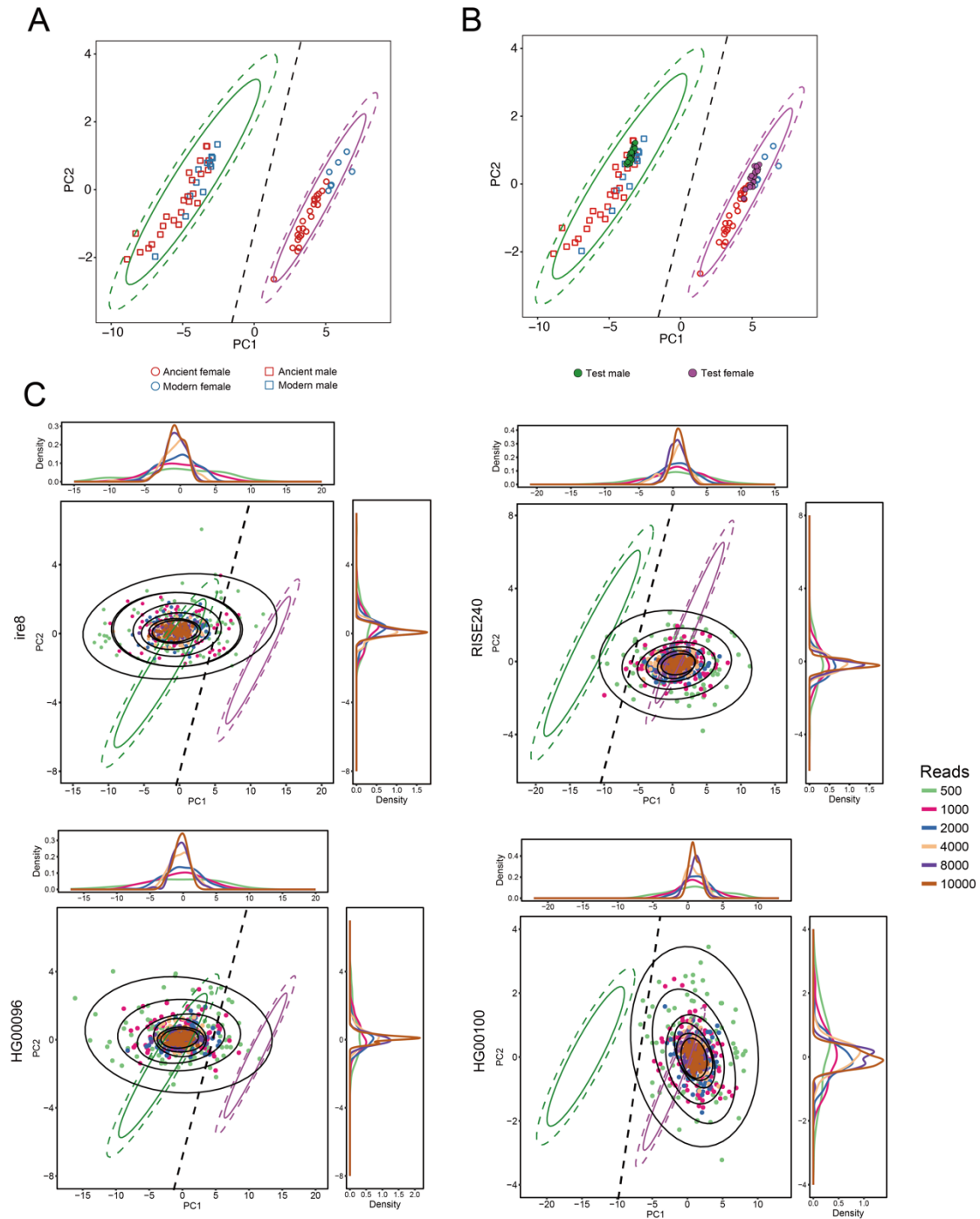
B



Supplementary Fig. 8. F4 statistics show the varied genetic affinity of southern populations to Baligang populations with series of time intervals. (A) Comparison of genetic affiliation to the ancient (Fujian_EN/ Fujian_LN/Guangxi_LBIA) southern East Asian populations by the f_4 statistics using the form: $f_4(H1:\text{later period}, H2:\text{earlier period}; \text{southern East Asian populations}, \text{Mbuti})$. (B) Supplement for Fig. 3b. Significant values are represented by large symbols, and ancient and modern populations are distinguished by different shapes.

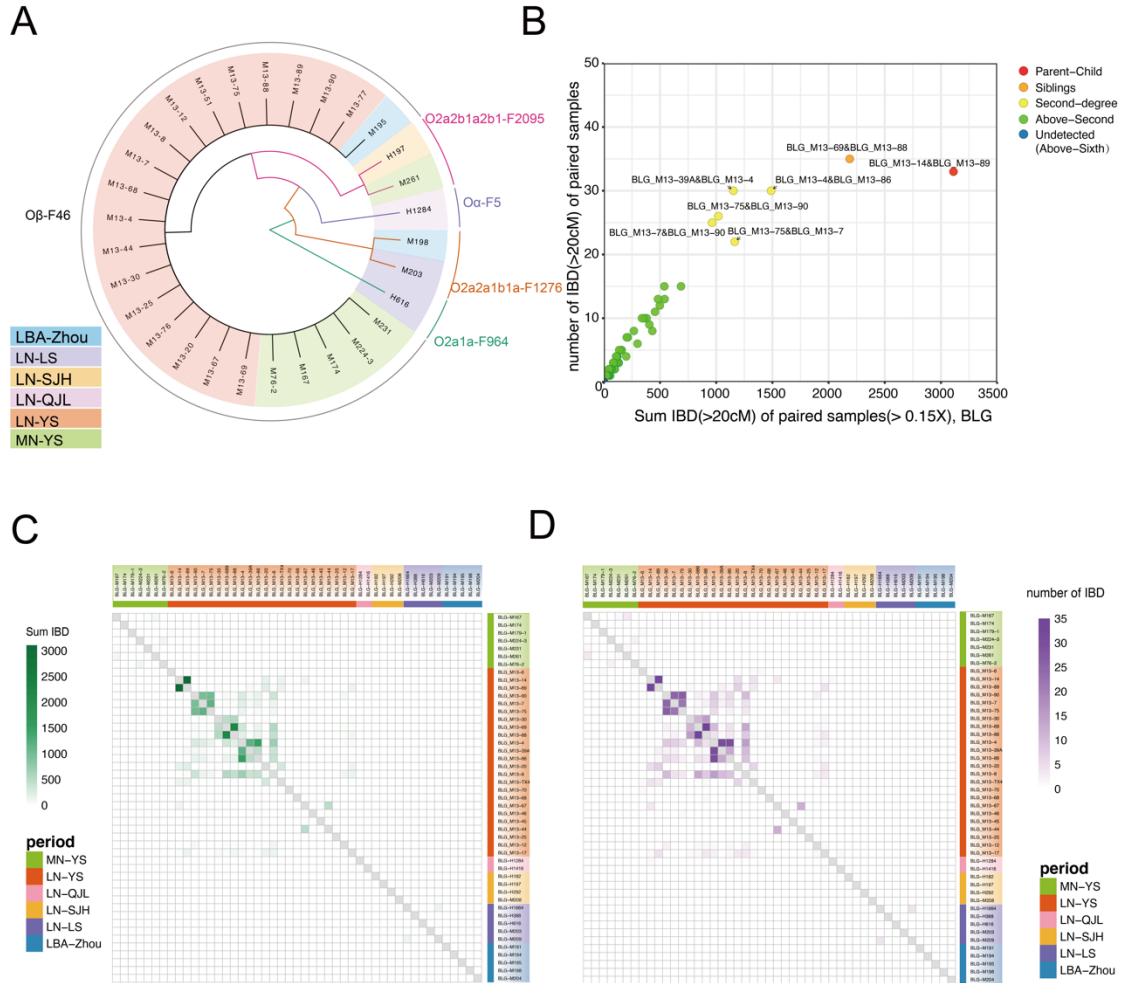


Supplementary Fig. 9. Comparison of genetic affiliation between the SJH outlier and contemporaneous population. (A) The different genetic affinities between each period of Baligang population to present-day East (n=11) and Southeast Asian(n=8) populations using outgroup f_3 -statistics with the form $f_3(\text{BLG, Test; Mbuti})$. Box plots show the median, first and third quartiles (lower and upper hinges), and whiskers extending to 1.5X interquartile range from the hinges. This figure is the supplement for Fig. 3e. (B-C) We compare $f_4(\text{BLG-SJH}_o, \text{Bianbian; X, Mbuti})$ with $f_4(\text{BLG-LN-SJH, Bianbian; X, Mbuti})$ (B/D) to determine the extra southern genetic component of outliers in the second quadrant. (B) use diverse ancient East Asian populations and (C) use present-day East Asian populations as test groups.

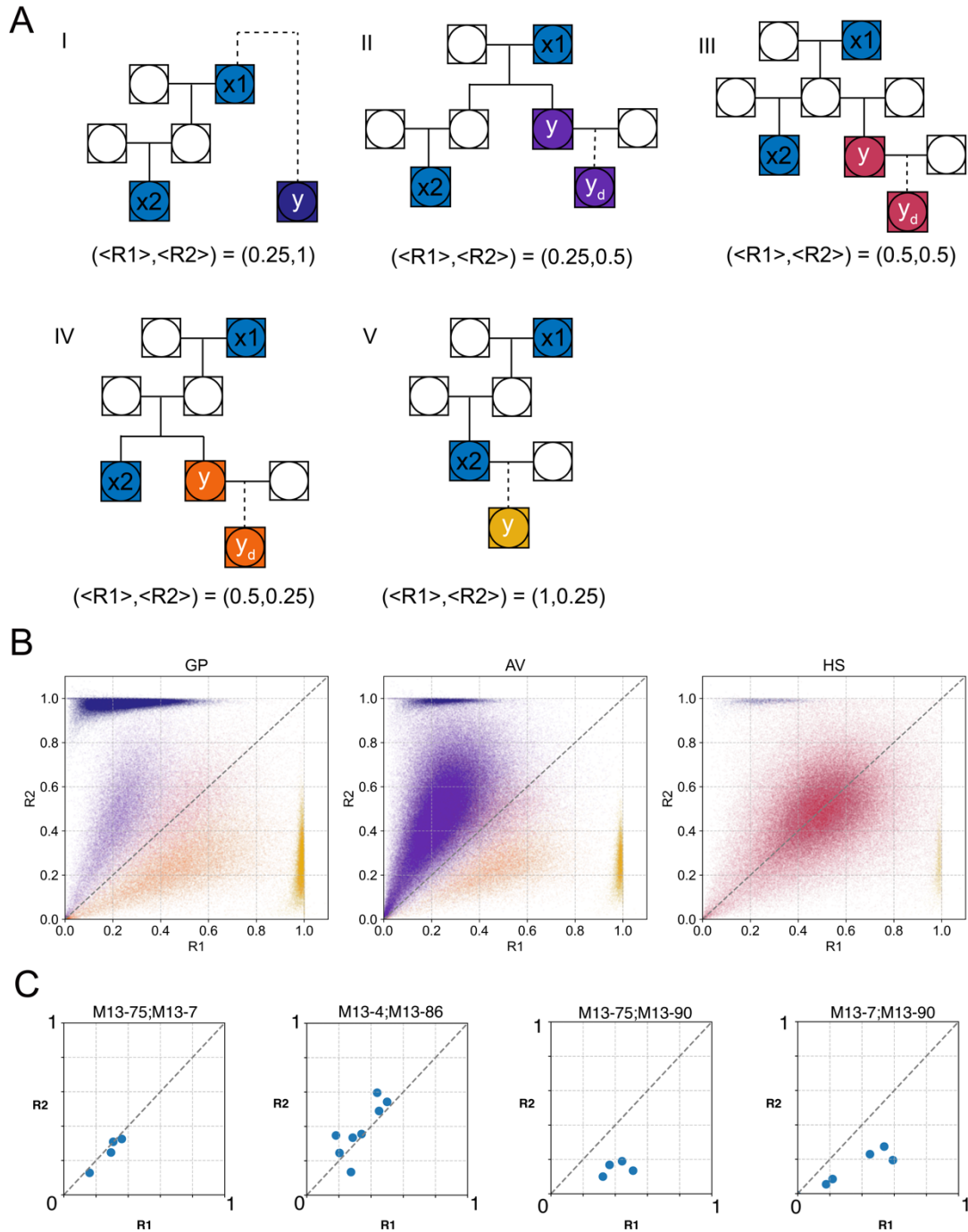


Supplementary Fig. 10. Sex determination through PCA-based classification. (A) A highly effective SVM classifier is trained using 67 individuals with known sex based on PCA. The green ellipse rings up the group of individuals representing males, while the purple ellipse rings up that for female with 95% CI (solid ring) and 99% CI (dashed ring) respectively. The oblique dashed line between two ellipses is the boundary for classification of two sexes. Source data are provided as a Source Data file. (B) Validation of the PCA-based approach for sex assignment using present-day human specimens with known sex. Shotgun sequencing reads from 22-male and 23-female test blood DNA are aligned and counted using regular workflow. All the ratios of normalized reads aligned to X chromosome to that aligned to all autosomes are calculated then log2 transformed, and PCA is performed. The location of dots representing

each test individual is plotted relative to ellipse region to verify the accuracy of sex determination. Source data are provided as a Source Data file. (C) Downsampling analysis of four individuals with known sex. Between 10,000 and 500 reads from one of 67 training individuals are randomly sampled 100 times. Dot plot expresses a significantly more dispersed distribution around the ellipse region along with the decrease of sampled reads. Black circles represent the region including 95% of dots corresponding to each downsampled reads. A more disperse distribution around the ellipse regions along with the decrease of sampled reads are replicated independent of sex and age of DNA sample.

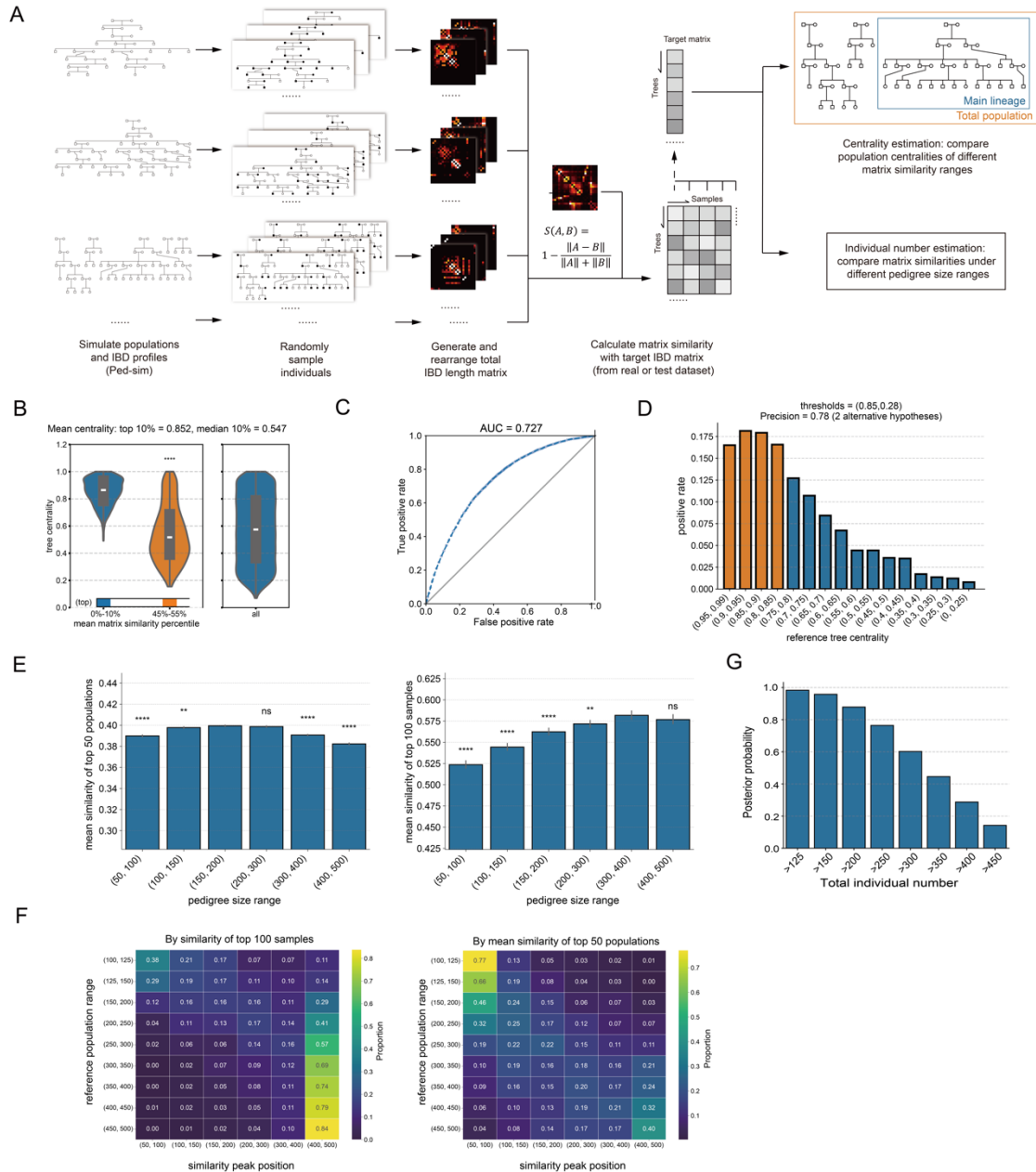


Supplementary Fig. 11. Kinship analysis of Baligang individuals. (A) Phylogenetic tree based on Y haplotypes of BLG males using pathPhynde by integrating Y-DNA data into present-day human phylogenies. (B) Pairs of individuals in grave M13 ($>0.15X$ coverage, $n=24$) are plotted by degrees of relatedness with different colors according to shared long blocks of IBD($>20cM$), relationships above the second degree are marked with text. (C,D) The heatmap shows pairwise IBD (Identity-by-descent) sharing between all Baligang individuals with more than 0.15X coverage ($n = 47$). (C) represent the sum length of shared long blocks IBD, and (D) represent the number of IBD segments. Deep-colored clusters reveal the links between individuals in grave M13 (BLG-LN-YS).



Supplementary Fig. 12. Classification of 2nd degree pairs. (A) Schematic diagram depicting cases of mutual relatives of a grandparent-grandson pair with different mean ($R1$, $R2$) values. Dashed lines represent direct descendent relationships. In cases II, III and IV, the notation y indicates the common ancestor of this family of mutual relatives, and y_d indicates descendant individuals. Both y and y_d belongs to the corresponding mutual relative case. (B) Scatter plot illustrating the distributions of ($R1$, $R2$) for the mutual relatives of the three types of 2nd degree pairs in simulated extensive pedigrees. Different colors represent different cases of mutual relatives classified according to the theoretical mean value of ($R1$, $R2$). Source data are provided as a Source Data file. (C) Scatter plot of CREST $R1$ and $R2$ values for second-degree pairs in BLG-M13. Each point represents a mutual relative of the second-degree pair of interest.

Inferred from the distribution patterns, pairs (M13-75, M13-7) and (M13-4, M13-86) present the highest likelihood of half-sibling, whereas pairs (M13-75, M13-90) and (M13-7, M13-90) are likely avuncular. Source data are provided as a Source Data file.

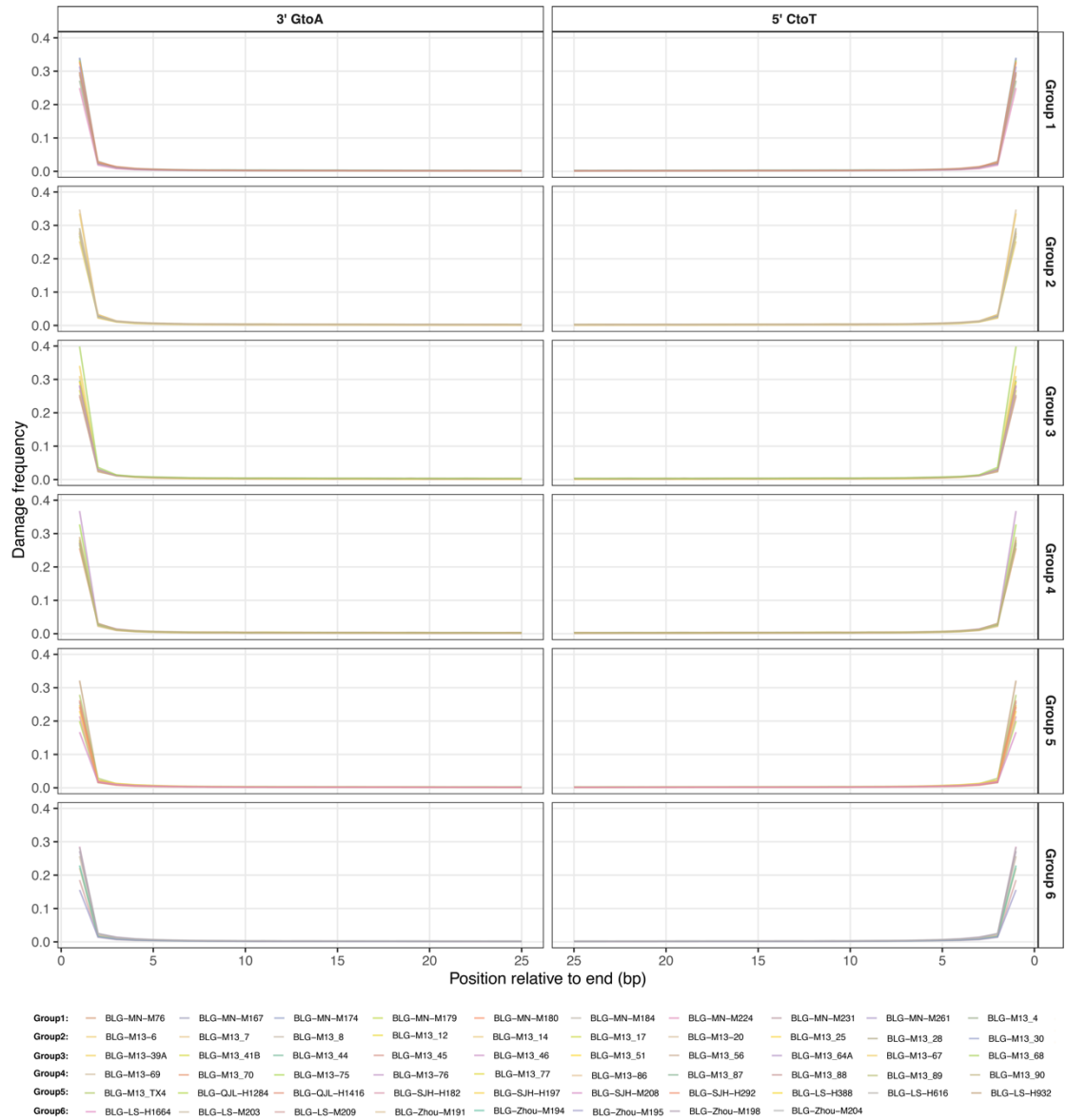


Supplementary Fig. 13. IBD-length-matrix-based population kinship structure estimation.

(A) Schematic workflow of the simulation and estimation process. Briefly, we used PEDSIM to generate virtual populations with IBD profiles under a variety of population kinship structure assumptions, from which individuals were sampled and their pairwise total IBD length matrices were generated. By inspecting the relationship between IBD matrix similarity with M13 data and population kinship structure property for simulated populations, the likelihood of different kinship structure assumptions on M13 were evaluated. See supplementary note 3 for details.

(B) Centrality estimation of the M13 population. We define centrality as the proportion of individuals in the largest single lineage (the main lineage) to the total individual number (see Supplementary Fig.12A). Simulated populations with the top 10% mean IBD matrix similarity to the M13 matrix show a high centrality ($n=1018$), exceeding those with the median 10% mean matrix similarity ($n=1018$), indicating that the M13 population is likely a near-single lineage. ****: $p < 0.0001$, from 2-sided Mann–Whitney U test ($u = 118639.5$, $p = 2.47e-199$). Source data are

provided as a Source Data file. (C) ROC curve of the centrality classification method at a true threshold of $C > 0.85$, tested on over 4,000 artificial populations. See supplementary notes 3 for details. Source data are provided as a Source Data file. (D) Positive rates of the centrality classification method for artificial populations with different centralities, using strict discrimination thresholds (0.85, 0.28). Orange bars indicate “true positives”, blue bars for “false positives”. Source data are provided as a Source Data file. (E) Total individual number estimation. Simulated populations with centrality=1 were divided into 6 total individual number intervals. In each interval, the mean matrix similarity with M13 of either top 100 samples or top 50 populations shows a single-peak pattern depending on the total population number. ****: $p < 0.0001$; *: $p < 0.05$; ns: $p > 0.05$, from 2-tailed t-test (left panel: $n=50$ for all groups; (50, 100) vs. (150,200): $p = 3.28e-31$; (100, 150) vs. (200,300): $p = 0.00158$; (200, 300) vs. (150,200): $p = 0.135$; (300, 400) vs. (150,200): $p = 8.00e-36$; (400, 500) vs. (150,200): $p = 1.19e-51$. right panel: $n=100$ for all groups; (50, 100) vs. (300,400): $p = 2.45e-41$; (100, 150) vs. (300,400): $p = 2.66e-24$; (150,200) vs. (300,400): $p = 1.18e-08$; (200, 300) vs. (300,400): $p = 0.00258$; (400, 500) vs. (300,400): $p = 0.160$). Source data are provided as a Source Data file. (F) Validation of the population estimation method on artificial data. The same procedure as in (E) were performed using IBD matrices from independently simulated populations as reference. The heatmap shows the distribution of mean matrix similarity peak position with different reference individual numbers. Then given the peak positions in (E) (G) Posterior probabilities of different total individual number ranges for the M13 population. See supplementary note 3 for details. Source data are provided as a Source Data file.



Supplementary Fig. 14. Terminal damage patterns of each individual, categorized into six groups for clarity.

References

- 1 Weisskopf, A., Deng, Z. H., Qin, L. & Fuller, D. Q. The interplay of millets and rice in Neolithic central China: Integrating phytoliths into the archaeobotany of Baligang. *Archaeol Res Asia* **4**, 36-45 (2015).
- 2 Deng, Z. H. *et al.* From Early Domesticated Rice of the Middle Yangtze Basin to Millet, Rice and Wheat Agriculture: Archaeobotanical Macro-Remains from Baligang, Nanyang Basin, Central China (6700-500 BC). *Plos One* **10** (2015).
- 3 Wang, C. *et al.* The development of Yangshao agriculture and its interaction with social dynamics in the middle Yellow River region, China. *Holocene* **29**, 173-180 (2019).
- 4 Peterson, C. E. & Shelach, G. Jiangzhai: Social and economic organization of a Middle Neolithic Chinese village. *J Anthropol Archaeol* **31**, 265-301 (2012).
- 5 Li, Y. P., Zhang, J. N., Zhang, X. H. & Zhao, H. T. Agriculture, the Environment, and Social Complexity From the Early to Late Yangshao Periods (5000-3000 BC): Insights From Macro-Botanical Remains in North-Central China. *Front Earth Sc-Switz* **9** (2021).
- 6 Zhang, K. The Spread and Integration of Painted Pottery Art along the Silk Road. (2021).
- 7 Lee, Y. K. & Zhu, N. Social integration of religion and ritual in prehistoric China. *Antiquity* **76**, 715-723 (2002).
- 8 Gao, Q. & Lee, Y. K. A Biological Perspective on Yangshao Kinship. *J Anthropol Archaeol* **12**, 266-298 (1993).
- 9 Li, Y. Y. *et al.* Palaeoecological records of environmental change and cultural development from the Liangzhu and Qujialing archaeological sites in the middle and lower reaches of the Yangtze River. *Quatern Int* **227**, 29-37 (2010).
- 10 Huan, X. J. *et al.* The Emergence of Rice and Millet Farming in the Zang-Yi Corridor of Southwest China Dates Back to 5000 Years Ago. *Front Earth Sc-Switz* **10** (2022).
- 11 Song, Z., Chen, J. & Zhao, Y. Rice domestication and the Yangtze River civilization. *Biodiversity Science (Chinese Series)* **26**, 346 (2018).
- 12 Chi, Z. The Qujialing–Shijiahe culture in the Middle Yangzi river valley. *A companion to Chinese archaeology*, 510-534 (2013).
- 13 Wang, L. & Cui, Y. F. The Spatiotemporal Pattern of Cultural Evolution Response to Agricultural Development and Climate Change From Yangshao Culture to Bronze Age in the Yellow River Basin and Surrounding Regions, North China. *Front Earth Sc-Switz* **9** (2021).
- 14 Qiao, Y., Sannerud, J. G., Basu-Roy, S., Hayward, C. & Williams, A. L. Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. *Am J Hum Genet* **108**, 68-83 (2021).
- 15 Marchini, J. *et al.* Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLOS Genetics* **15** (2019).
- 16 Zhang, C. *et al.* Study on the burial practice of tomb M13 of the Yangshao culture at Baligang site in Dengzhou City. *Chinese Archaeology* **20**, 132-138 (2020).