Whole Genome Resequencing Reveals the Evolutionary History and Geographic Isolation of the eastern Asian Hickory (Carya)

Da Zhang  
Zhejiang A &F University

Rui-Feng Yang  
Zhejiang A &F University

Jia-Hui Chen  
Zhejiang A &F University

Lang-Bo Pan  
Zhejiang A &F University

Wei Duan  
Zhejiang A &F University

Qi-Xiang Zhang  
Zhejiang A &F University

Guo-Hua Xia  
Zhejiang A &F University

Youjun Huang  
hyj@zafu.edu.cn  
Zhejiang A &F University

Research Article

Keywords: evolutionary history, disjunctive distribution, endangered, resequencing, Carya, hickory

Posted Date: September 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3350839/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License

Additional Declarations: No competing interests reported.
Abstract

Background

The biogeographic characteristics of disjunctive distribution play a crucial role in plant geography and understanding mechanisms related to endangerment. Genome sequencing offers an opportunity to explore genetic relationships, population dispersal, and the floristic evolution of disjunctively distributed flora. In this study, we conducted comprehensive whole-genome deep resequencing of EA *Carya* to elucidate their evolutionary history and the factors contributing to disjunctive distribution.

Results

After conducting high-depth sequencing of five EA *Carya* species and filtering the data, we obtained a dataset comprising 3,869,439 high-quality SNPs. The phylogenetic analysis of the entire *Carya* revealed two monophyletic clades, separating EA *Carya* from NA *Carya*. Population genetic analysis indicated that *C. kweichowensis*, an endangered species, exhibited the lowest nucleotide diversity and the earliest divergence among the studied EA *Carya* species. Through selection elimination analysis, we identified a set of adaptive genes in *C. hunanensis* characterized by the highest nucleotide diversity and the widest distribution compared to other EA *Carya* species. These genes were primarily associated with mitochondrial function and alkaloid biosynthesis. Additionally, species distribution modeling predicted the suitable habitat for the five EA *Carya* species, revealing the potential distribution of the endangered *C. kweichowensis*. Importantly, our analysis indicated minimal spatial overlap in distribution regions among EA *Carya* species during different time periods.

Conclusions

The results of population historical dynamics combined with climate and geological changes support the hypothesis that EA *Carya* migrated from NA to Europe during the Paleogene or Eocene and continued their migration to Asia, subsequently moving eastward until reaching EA. The uneven regional distribution of EA *Carya* is believed to be a consequence of Quaternary climate fluctuations, mountain barriers hindering species dispersal, and the limited cold tolerance of these trees. EA *Carya* highlight the significant role of climate and geological changes in their regional distribution and migration routes in Asia. Furthermore, the discovery of potential habitats offers a promising avenue for the conservation of *C. kweichowensis*.

Background

The investigation of plant distribution has been ongoing since the 19th century [1,2], serving as a vital window into the study of ecology, plant interactions, potential trends in plant distribution, and endangered species [3–5]. The formation of plant distribution patterns reflects a complex interplay of various historical variables [6]. For instance, the consequences of changes in geological environment diversity are profound. While mountains promote species differentiation and provide refuges [7, 8], they can also serve
as barriers to species dispersal, reducing species diversification rates or even leading to extinctions [9, 10]. Plant distribution is profoundly influenced by climatic factors, with species' range expansions and contractions intricately linked to climate fluctuations [11, 12]. It's important to note that plant distribution is shaped not only by geological changes and climatic variations but also by genetic factors, evolutionary processes, and plant migrations [13, 14]. This complexity renders the delineation of plant distribution patterns one of the most intricate and challenging aspects of phytogeography [15]. In contemporary times, humanity confronts a pressing issue— the alarming loss of biodiversity. The delineation of plant distribution areas sheds light on the remarkable mechanisms through which plants adapt and persist in their ongoing battle with the environment. Moreover, it offers invaluable insights into future plant distribution trends and serves as a cornerstone for the conservation efforts aimed at safeguarding endangered species [16].

The disjunctive distribution of plants across East Asia and North America (EA-NA) has garnered considerable attention, emerging as a prominent focus in plant distribution research [17–19]. Within this context, the walnut family (Juglandaceae) stands out as a distinctive example of EA-NA disjunct species, drawing significant interest in previous investigations pertaining to its migratory history and lineage geography [20–22]. Here, we have identified an intriguing phenomenon within the Juglandaceae family, specifically among *Carya* - a group that exhibits not only intercontinental disjunctive distribution but also significant regional disjunctive distribution within EA [23]. In China, there are five *Carya* species: *C. hunanensis*, *C. kweichowensis*, *C. tonkinensis*, *C. dabieshanensis*, and *C. cathayensis*, each of which is distributed in different regions of southwestern and southeastern China, with minimal overlap between their respective habitats (Fig. 1a). They are collectively known as Asian hickory (*Sinocarya*) due to their shared characteristic of bearing naked winter buds. The EA environment typically fosters a higher rate of speciation and greater species diversity [22, 24]. However, in reality, the *Carya* species in ENA boast a larger number of species and higher species abundance than their EA *Carya* counterparts [25]. As of now, the underlying mechanisms contributing to the lower species diversity, reduced abundance, and regional disjunctive distribution in EA *Carya* remain a mystery. It is imperative to delve into the explicit dynamics of population history and examine the influence of environmental factors on the distribution of EA *Carya* especially in the context of the endangered *C. kweichowensis*.

Taking these concerns into account, we conducted a comprehensive census within the natural distribution areas of EA *Carya* in China, collecting samples directly from their native habitats. Our study aimed to assess the population structure of EA *Carya* and investigate genetic distinctions among different species. To achieve this, we conducted resequencing on representative samples and analyzed genetic variations within these five *Carya* species.

Furthermore, we sought to elucidate the population dynamics dating back to the Quaternary period. By integrating data on geological and climatic changes, we aimed to unravel the evolutionary history of EA *Carya* and shed light on the factors contributing to their regional disjunctive distribution. We believe that our findings will not only enhance our understanding of EA *Carya* but also provide valuable insights for phylogenetic research on other endangered plant species.
Results

Whole genome resequencing and phylogenetic construction of EA Carya

We performed comprehensive whole-genome resequencing with a high level of coverage on EA Carya sampled from representative natural habitats (Fig. 1a). This effort generated a total of 1.68 Tb of resequencing data. After stringent data filtering, each individual sample yielded an average of 22.4 Gb of clean data, with an average sequencing depth of 24.90×. The mean alignment rate reached 89.56%, and the average coverage attained 93.5% using the well-established *C. cathayensis* reference genome available on the NCBI website [26] (Additional file1: Table. S1), which we contributed to ourselves. These high alignment rates and sequencing depths significantly bolstered the reliability of our results. Based on the outcomes of the alignment, we identified and meticulously annotated a total of 3,869,439 high-quality single nucleotide polymorphisms (SNPs) through rigorous filtering. Notably, 39.36% of these SNPs were localized within intergenic regions, while 12.67% were situated within coding regions (Additional file2: Table. S2).

Plant populations with distinct geographical distributions often exhibit significant genetic variations. *Carya* demonstrates a clear disjunct distribution pattern between EA and NA. With the aim of exploring potential notable genetic disparities between the EA and NA *Carya* species, we incorporated 12 *Carya* species from NA into the construction of our phylogenetic tree. Simultaneously, *J. regia* from the *Juglans* of the Juglandaceae family was chosen as an outgroup specie.

Constructing a maximum likelihood tree using high-quality SNP data clearly segregates *Carya* samples from NA and EA, suggesting notable disparities between NA and EA *Carya* species. Within the phylogenetic tree, the *C. kweichowensis* population emerges as an independent branch. Compared to other EA *Carya* populations, the genetic divergence between the *C. cathayensis* and *C. dabieshanensis* populations is relatively minor. (Fig. 1b) Moreover, they exhibit conspicuous morphological similarities. Remarkably, previous studies erroneously classified them as the same species, aligning with the topology of the evolutionary tree we constructed. The five populations of EA *Carya* are visibly distinct within the tree. Their habitats exhibit minimal overlap and are largely segregated by mountain ranges, potentially serving as one of the reasons for the noticeable disparities. Additionally, the relationships among the five EA *Carya* species were observed to be inversely correlated with their geographical distribution distances.

We conducted additional principal component analysis (PCA) to delve into the variability observed among EA *Carya* species. The initial three principal components accounted for 54.88%, 28.60%, and 3.66% of the variance, respectively (Fig. 1c). Principal Component 1 (PC1) distinctly positioned *Carya. kweichowensis* apart from the other species, indicating its greater genetic divergence from the rest. This alignment was in line with our previously obtained topology, an observation that was further substantiated by model-based analyses of population admixture (K = 3, Fig. 1d). Principal Component 2 (PC2) effectively separated *C. kweichowensis, C. hunanensis,* and *C. tonkinensis,* while *C. cathayensis* and *C. dabieshanensis* exhibited closer proximity to each other. This segregation closely mirrored our established topology and was consistent with the findings from population ADMIXTURE analysis (K = 5).
The majority of the EA *Carya* species within Principal Component 3 (PC3) formed a closely clustered group. Intriguingly, three samples of *C. hunanensis* diverged from the main *C. hunanensis* cluster and exhibited a closer association with the other four EA *Carya* species. This outcome echoed our phylogenetic tree results and led us to hypothesize that this deviation might be attributed to potential hybridization events involving certain *C. hunanensis* samples and other EA *Carya* species, further validating our speculation by population ADMIXTURE analysis (K = 5).

**Population differences in EA Carya**

We estimated the nucleotide diversity ($\theta_\pi$) among different species of EA *Carya*. The nucleotide diversity of *C hunanensis* ($\pi = 1.5627 \times 10^{-3}$) significantly exceeded that of other EA *Carya* (Fig. 2a). Additionally, the population size of *C. hunanensis* exhibited the broadest distribution and was accompanied by the largest number of individuals [25]. Notably, *C. kweichowensis* is classified as critically endangered (CR) on the Red List of Higher Plants of China, owing to habitat fragmentation and an extremely low population size primarily resulting from human activities [27]. Being an endangered plant, it exhibited the lowest nucleotide diversity value ($\pi = 2.764 \times 10^{-4}$). The reduction in population size could potentially elevate the likelihood of inbreeding, consequently leading to a decline in adaptive potential [28]. Therefore, prioritizing the conservation of *C. kweichowensis* should primarily involve strategies to mitigate inbreeding and thereby alleviate the risk of extinction. Additionally, in the central and eastern regions of China, *C. cathayensis* is widely cultivated as a significant economic nut-bearing crop. Consequently, the nucleotide diversity of *C. cathayensis* remains at a diminished level, primarily attributed to human intervention and selective pressures during artificial breeding and selection processes.

Moreover, we assessed fixation indices ($F_{ST}$) to elucidate the extent of differentiation among populations. The findings revealed the highest $F_{ST}$ value (0.813) between *C. kweichowensis* and *C. cathayensis*, whereas the lowest $F_{ST}$ value (0.294) was observed between *C. dabieshanensis* and *C. cathayensis* (Fig. 2b). These outcomes were in line with geographical distribution patterns. As previously mentioned, the status of *C. cathayensis* and *C. dabieshanensis* being potentially the same species was acknowledged [23], further corroborating the depicted phylogenetic framework (Fig. 1D). Notably, all interspecific Fst values exceeded 0.25, indicating pronounced dissimilarities between species. This differentiation could be attributed to the significant impact of mountainous terrain, acting as barriers that isolate species, as proposed by studies [10].

The linkage disequilibrium (LD) decay distance represents the physical distance at which the correlation coefficient ($r^2$) between genetic variants decreases to half of its maximum value. In our study, we observed significant variation in LD decay distances among different Carya species. Specifically, LD decay distances were found to be 0.053 kb for *C. cathayensis*, 3.9 kb for *C. hunanensis*, 31.3 kb for *C. kweichowensis*, 39.3 kb for *C. tonkinensis*, and 77.9 kb for *C. dabieshanensis* (Fig. 2b). These values reflect variations in the extent of genetic linkage between pairs of variants across different species. Notably, *C. dabieshanensis* exhibits a remarkably limited distribution range, thriving on soils derived from granite, and experiences pronounced environmental selection pressures. Conversely, the habitat diversity
of C. cathayensis and C. hunanensis likely contributes to their relatively lower LD decay distances, as suggested by previous studies [29].

**Population history**

Addressing the endangerment of C. kweichowensis requires an examination of when its population began to decline, essentially exploring the origins of the bottleneck. We employed the pairwise sequentially Markovian coalescent (PSMC) approach to reconstruct the population history of EA Carya, obtaining a timeline spanning from 1.2 million years ago (mya) to 10 thousand years ago (kya) (Fig. 3a). We identified two distinct demographic fluctuations, attributed to the temperature rebound at the end of the Dagu glacial period (the second glacial period in China during the Quaternary, approximately 1.2–1.05 mya). This temperature increase resulted in an augmented effective population size (Ne) for each population, with C. hunanensis showing the most notable increase, peaking during the last interglacial (LIG, around 130 kya). Subsequently, all EA Carya populations experienced a decline in size around the Dali glacial period (approximately 110 – 100 kya) and underwent a significant bottleneck during the last glacial maximum (LGM, approximately 21 kya). These observations highlight the close correlation between population size and climatic fluctuations in EA Carya.

We employed Treemix to construct maximum likelihood trees of the EA Carya populations and to trace gene migration patterns using SNP data. Additionally, we utilized PAMLmcmctree to estimate the divergence times for each EA Carya. C. kweichowensis was initially differentiated around the middle Miocene (Fig. 3b), with evidence of gene flow occurring between it and C. tonkinensis (Fig. 3c). The divergence of C. tonkinensis and C. hunanensis took place during the late Miocene, accompanied by significant gene exchange between the two species. This phenomenon correlates with the more intricate population stratification observed in C. tonkinensis and C. hunanensis. The divergence of C. dabieshanensis and C. cathayensis occurred approximately 2.9 million years ago (mya), aligning with findings from previous studies [26]. Given the geographic locations of these species—with C. kweichowensis and C. tonkinensis in southwest China, C. hunanensis in central China, and C. dabieshanensis and C. cathayensis in eastern China—the migration route of EA Carya is evident from southwest to east in China. This observation is consistent with prior research [30], and the alignment between differentiation time and geographical distribution underscores the reliability of our findings.

**Environmental adaptation of C. hunanensis**

C. hunanensis possesses an extensive distribution range and exhibits the highest nucleotide diversity (\(\pi = 15.627 \times 10^{-4}\)), coupled with a substantial population size that has endured throughout its population history. This longevity of population size may be attributed to adaptive evolution, alongside environmental influences. In contrast, C. kweichowensis has attained a critically endangered (CR) status. By elucidating the mechanisms underlying environmental adaptation in C. hunanensis, we aim to offer novel insights for the subsequent conservation efforts of C. kweichowensis. To achieve this, we conducted a comprehensive analysis involving \(F_{ST}\) and \(\pi\)-based selective sweeps utilizing genomic data.
from both *C. hunanensis* and *C. kweichowensis*. Our findings revealed the presence of 47 candidate genes in *C. hunanensis* (Fig. 4a, Additional file3: Table. S3). Subsequently, we carried out a thorough Gene Ontology (GO) enrichment analysis and KEGG analysis on these identified genes, revealing a predominant association with mitochondrial function and alkaloid biosynthesis pathways (Fig. 4b, Fig. 4c). Notably, specific genes linked to mitochondrial processes, such as *AtNDB2* [31], and *LcAPX* [32], have been demonstrated to enhance plant resilience to environmental stress. Additionally, nicotinamide alkaloids have been recognized to augment plant tolerance to certain heavy metals, as exemplified by studies on plant responses [33].

The results of the KEGG enrichment analysis highlighted the significant involvement of alkaloid synthesis pathways, including terpenoids and isoquinolines. Comparable outcomes were observed for *C. hunanensis* when compared to other EA *Carya* species (Additional file4: Fig. S1). Terpenoids, which serve as precursors for essential compounds such as chlorophylls, carotenoids, and phytohormones, play a pivotal role in safeguarding plants against phytophagous herbivores and pathogens [34]. This potentially accounts for the heightened adaptability of *C. hunanensis* in comparison to other EA *Carya* species.

**Predicted EA Carya distribution in different periods**

To unravel the mechanisms underlying the disjunctive distribution of EA *Carya* species and identify potential habitats for the endangered *C. kweichowensis*, we employed Maxent to predict the species' distribution patterns in the present, Last Glacial Maximum (LGM), and Last Interglacial (LIG) periods. Notably, the present predictions align with the distribution characteristics of each EA *Carya* species. However, the potential habitat for *C. kweichowensis* was more expansive than initially anticipated, predominantly concentrated in southwestern China—an area of paramount importance for conservation efforts within the EA *Carya* (Fig. 5). Establishing favorable habitats is pivotal to prevent the decline of endangered plant populations [3] and prioritizing these high-suitability regions is crucial for augmenting the population of *C. kweichowensis*. In contrast to the present, suitable habitats during the LGM were notably reduced, especially for *C. tonkinensis* and *C. hunanensis*. During the LGM and LIG periods, the habitat range of *C. dabieshanensis* was significantly restricted, potentially due to robust environmental selection acting upon this species. Importantly, minimal overlap was observed among the most suitable habitats for the five species across the three periods. The climatic oscillations throughout the Quaternary era, combined with geographical barriers imposed by mountain ranges, likely contributed to the segregation of primary habitats [10].

**Discussion**

To explore the divergence history of EA *Carya* and the reasons behind their disjunct distribution in China, we conducted high-depth sequencing of five EA *Carya* species. This marks the first instance where a comprehensive phylogenetic relationship within the entire genus has been constructed, leveraging a substantial dataset of reliable SNPs and complemented with resequencing data from NA *Carya*. As a typical example of a disjunctly distributed species between EA and NA, the long-standing geographical
isolation has resulted in a distinct separation between NA *Carya* and EA *Carya* in terms of their phylogenetic relationships, consistent with previous research findings.

Furthermore, the migration routes of *Carya* are another focal point of investigation. Generally, NA is considered the cradle of the *Carya*. Based on fossil distribution and estimated divergence times, it is believed that *Carya* might have entered Europe from North America during the Paleocene or Eocene via the North Atlantic land bridge, subsequently spreading into Asia. Numerous geological factors in Eurasia have facilitated inter-regional biological exchange. For example, the closure of the Turgai Seaway after the Oligocene created more opportunities for plant migration between Europe and Asia [35.36]. Additionally, areas along the Himalayas and the central region of Tibet, with their warm climate habitats, served as biological exchange corridors within Eurasia during the Oligocene to early Neogene periods [37]. Examples of species that utilized these corridors for migration include European holly oaks (Quercus section Ilex Loudon) [35], *Elaeagnaceae* [38], and *Mahonia* (Berberidaceae) [39] Likewise, this could also serve as a key factor in the migration of *Carya* from Europe to EA. The differentiation between *Carya kweichowensis*, which split earliest in the southwestern direction of China, and *C. cathayensis* with the latest differentiation time is highly pronounced ($F_{ST} = 0.813$). Furthermore, based on the differentiation timeline, the order of differentiation among the five EA *Carya* species is from west to east. This provides additional support for the hypothesis that *Carya* migration in Asia likely occurred from west to east.

Disjunctive distribution of plants and the diverse flora across different regions are commonly observed phenomena. Typically, the underlying causes of this phenomenon are attributed to the effects of ongoing climate fluctuations [40.41]. The demographic history of EA *Carya* reveals that the major factor influencing effective population sizes (Ne) in these species was the glacial Quaternary period. Particularly noteworthy is the pronounced impact of the Dali sub-glacial period (approximately 110 – 10 kya) on the effective population sizes (Ne) of EA *Carya* (Fig. 3a). Harsh cold climates often act as barriers to the spread of plants. Fortunately, certain plants have evolved mechanisms for increased cold tolerance and adaptability to challenging environmental conditions [42.43].

It is evident that EA *Carya*, characterized by their exposed terminal buds, have not developed sufficient adaptations to cope with cold climates. This limitation significantly constrains their ability to expand their range. Throughout the Quaternary period, climatic fluctuations led many plant species to retreat to lower latitudes [44]. However, it's important to note that EA did not experience extensive glaciation. Furthermore, the topographical diversity within EA provided refuge habitats for certain plant species [24.45].

As previously mentioned, mountains can serve as both pathways for plant dispersal and formidable barriers that may even lead to species extinction [7]. With rising temperatures, the ability of species to radiate into their surrounding regions may be constrained, particularly in the presence of high mountains. All EA *Carya* species are primarily located in mountainous regions and their environs, a distribution pattern that aligns with the projections for the Last Glacial Maximum (LGM) as predicted by species distribution modeling (SDM) (Fig. 5B). The limited cold tolerance exhibited by EA *Carya* has confined their survival to the mountain refuges during periods of harsh cold. However, as the climate gradually warms,
these same mountains have become obstacles to the broader expansion of EA *Carya* species. Additionally, attaining an optimal ecological niche in lowland areas has proven to be more challenging than in mountainous regions during periods of rapid climatic fluctuations [46]. The frequent climatic shifts of the Quaternary era, coupled with the insufficient cold tolerance of EA *Carya* in EA, have impeded their ability to migrate from these refuges, resulting in the regional disjunctive distribution observed among EA *Carya* species.

The endangerment of *C. kweichowensis* raises a more serious concern. As a vital component of biodiversity, endangered plants require urgent conservation efforts due to their limited competitiveness and weak dispersal abilities [3]. Endangered species typically possess narrow ranges and heightened susceptibility to both human activities and climatic changes. Therefore, the identification of suitable habitats for endangered plants is paramount to prevent their extinction [47]. The potential habitat suitability areas for *C. kweichowensis* in the present period (1970–1990) appear to have been underestimated. These regions, characterized by high-quality habitats, hold significant potential for preservation through the establishment of nature reserves or through deliberate cultivation efforts aimed at sustaining *C. kweichowensis* populations.

**Conclusion**

In summary, our comprehensive high-depth sequencing of five EA *Carya* species in the Asian region strongly supports the differentiation of EA *Carya* and NA *Carya* as two distinct subgenera within the *Carya*. Furthermore, when considering the population dynamics in conjunction with climate and geological changes, it lends support to the hypothesis that hickories migrated from NA to Europe during the Paleocene or Eocene, then continued their migration into Asia, subsequently moving eastward until reaching East Asia. The isolated and scattered habitats of EA *Carya* species are believed to be a result of Quaternary climatic fluctuations, mountain ranges acting as barriers to species dispersal, and the limited cold tolerance of EA *Carya*. Predictions of suitable habitat areas for the narrowly distributed *C. kweichowensis* and the discovery of selectively advantageous genes in *C. hunanensis* offer valuable insights for the conservation of plant diversity. Our research underscores the significant roles of climate and geological changes in the regional disjunctive distribution and migration routes of *Carya*. Additionally, it highlights that the exploration of potential habitats can provide additional opportunities for the conservation of *C. kweichowensis*.

**Materials and methods**

**Sample Collection**

We collected accessions in natural distribution areas of 5 EA *Carya* species (Additional file5: Table. S4), including *C. hunanensis, C. kweichowensis, C. tonkinensis, C. dabieshanensis, C. cathayensis*, covering the distribution of *Carya* in China. The distance between samples in each species should be at least 2 km, in order to avoid too close the genetic relationship between the samples. Approximately 10–11 fresh
leaflets were collected in sealed bags, dried with silica gel, and stored finally at -80°C. We selected 75 representative accessions for subsequent the whole genome resequencing, including 21, 9, 8, 20, 17 accessions in *C. hunanensis*, *C. kweichowensis*, *C. tonkinensis*, *C. dabieshanensis* and *C. cathayensis*, respectively.

**Genome resequencing and quality control**

The genomic DNA was extracted with a total amount of 1.5µg per sample and used as input material for the DNA sample preparations (Additional file6: Table. S5). Sequencing libraries were generated using Truseq Nano DNA HT Sample preparation Kit (Illumina USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. These libraries were sequenced using the Illumina Hiseq X Ten platform. 350bp Pair-end reads were generated. Then, the raw reads (fastq format) were subjected to a series of quality control procedures to remove the low-quality reads (reads with ≥ 10% unidentified nucleotides (N); > 10 nt aligned to the adaptor, allowing ≤ 10% mismatches; > 50% bases having phred quality < 5; and putative PCR duplicates generated in the library construction process were removed). After strict filtering of sequencing data, high-quality clean data was obtained for subsequent analysis.

**Reads mapping and SNP calling**

The high quality paired-end reads were mapped to the *C. cathayensis* reference genome (ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100571/Cca.genome.fa.gz) using BWA (0.7.8) [48] with the command ‘mem -t 4 -k 32 –M’. In order to reduce mismatch generated by PCR amplification before sequencing, duplicated reads were removed by the help of SAMtools(1.17) [49] After alignment, we performed SNP calling on a population scale using a Bayesian approach as implemented in the package SAMtools. We then calculated genotype likelihoods from reads for each individual at each genomic location, and the allele frequencies in the sample with a Bayesian approach. The ‘mpileup’ command was used to identify SNPs with the parameters as ‘-q 1 -C 50 -D -m 2 -F 0.002 –u’. Then, to exclude SNP calling errors caused by incorrect mapping or InDels, only high quality SNPs (coverage depth ≥ 3, RMS mapping quality ≥ 20, maf ≥ 0.05, miss ≤ 0.1) were kept for subsequent analysis.

**Functional annotation of genetic variants**

SNP annotation was performed according to the *C. cathayensis* genome using the package ANNOVAR (Version: 2013.05.20) [50]. Based on the genome annotation, SNPs were categorized in exonic regions (overlapping with a coding exon), intronic regions (overlapping with an intron), splicing sites (within 2 bp of a splicing junction), upstream and downstream regions (within a 1 kb region upstream or downstream from the transcription start site), and intergenic regions. SNPs in coding exons were further grouped into synonymous SNPs or nonsynonymous SNPs. In addition, mutations causing stop gain and stop loss were also classified into this group.

**Phylogenetic tree and Population structure**
To elucidate the phylogenetic relationships from a whole-genome perspective, we initially employed Tassel (5.0) to sort and convert the SNP VCF file into phylip format. Subsequently, we utilized IQ-TREE (2.0.3) to construct a maximum likelihood tree based on the GTR model, followed by 1000 rounds of rapid bootstrap analysis. The population genetic structure was examined via an expectation maximization algorithm, as implemented in the program ADMIXTURE (1.23) [48]. We also conducted principal component analysis (PCA) to evaluate genetic structure using the software GCTA [51].

**Linkage disequilibrium and selective sweep**

The squared correlation coefficient ($r^2$) between pairwise SNPs was computed using the software Haploview. Parameters in the program were set as: ‘-n -dprime -minMAF 0.05’. The average $r^2$ value was calculated for pairwise markers in a 500-kb window and averaged across the whole genome. We calculated the genome-wide distribution of fixation index ($F_{ST}$) and the genetic diversity with a sliding-window approach (10-kb windows with 5-kb increments) using vcftools (0.1.16) [52]. We considered the windows with the top 5% $F_{ST}$ and log2($\theta$π ratio) simultaneously as candidate outliers under strong selective sweeps.

**Gene flow and PSMC analysis**

Using TreeMix [53] to infer phylogenetic differentiation and gene flow in 5 EA *Carya* species. Only SNPs with small allele frequencies and without any missing values were used to construct maximum likelihood trees and gene flow. The PSMC estimations were conducted using the default parameters tuned for human populations [54]. The utility fq2psmcfa (provided with the PSMC software) was used to convert this diploid consensus sequence to the required input format. The parameters were set as follows: ‘-N30 -t15 -r5’. Analysis of effective population size requires generation interval $g$ and nucleic acid mutation rate $\mu$. The generation interval is taken as the adult tree age of Juglans, 20 years; the nucleic acid mutation rate is taken from the closely related species Juglans, $\mu = 10.0 \times 10^{-8}$ [55].

**Species distribution modeling**

We used Maxent software (v3.4.1) to model the suitable areas of 5 EA *Carya* species in different periods [56]. The data of 19 bioclimatic variables were obtained from the World Climate Data website (http://www.worldclim.org/), the spatial resolution of LGM is 2.5 -arc minutes, and the spatial resolution of LIG and now (1970–1990) is 30-arc sec. Pearson correlation analysis was performed using SPSS 20.0 software. If $|r| > 0.8$ of the two environmental variables, the environmental variables with low contribution rate in the Maxent model prediction are eliminated. In Maxent, the test data is set to 25%, the training data set is 75%. Using ArcGIS 10.4 for reclassification function, the predicted suitability ($P$) was divided into: 1 ($P < 0.25$), 2 (0.25 < $P < 0.5$), 3 (0.5 < $P < 0.75$), 4 ( $P > 0.75$).

**Statement of permissions and licenses for plant or seed specimen collection**
We obtained the required permissions and licenses from the relevant authorities or institutions, including, to conduct the specimen collection activities. We followed all applicable local, national, and international regulations and guidelines governing the collection of plant or seed specimens.

Additionally, we confirm that all necessary ethical considerations and regulations regarding the collection, transportation, and use of these specimens were strictly adhered to. We took measures to ensure the sustainable and responsible collection of plant materials and adhered to best practices for biodiversity conservation.

In our study, the plant materials used were officially identified by Lang-Bo Pan.

Our plant samples have not been deposited in a public herbarium, and therefore, there is no herbarium accession number associated with them.

**Abbreviations**

SNP  
Single-nucleotide polymorphism  
NA  
North America  
EA  
East Asia  
KEGG  
Kyoto Encyclopedia of Genes and Genomes Enrichment Analysis  
GO  
Gene Ontology  
LD  
linkage disequilibrium  
$F_{ST}$  
F-statistics index  
SNP  
Single-nucleotide polymorphism

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.
Data Availability

The raw data of resequencing genome have been deposited in NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA860762. All data supporting this research result can be obtained in the paper.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was financially supported by the National Natural Science Foundation of China (31971672 and 32171815); the Key Research and Development Project of Zhejiang Province, China (2020C02005); the Zhejiang Key Research and Development project 2021C02037; the Yangtze River Delta Science and technology Innovation Joint project of Science and Technology Department of Anhui Province (202004G01020005); the Key Scientific and Technological Grant of Zhejiang for Breeding New Agricultural Varieties (2021C02066-12).

Author contributions


Acknowledgements

Not applicable.

References


Figures
Figure 1

Phylogenetic relationships and population structure of EA Carya.

(a) Geographical distribution of the EA Carya samples.

(b) The phylogenetic tree of Asian hickories based on genetic distance-based genome-wide SNP data.

(c) The principal component analysis of five EA Carya species.

(d) Model-based population assignment by ADMIXTURE analysis for K = 2–5.
Figure 2

Genetic differences among five EA *Carya* species.

(a) Nucleotide diversity of 5 EA *Carya* species and their $F_{ST}$ values. The value in each circle represents the nucleotide diversity of the species, while the values on the straight line represent the $F_{ST}$ value between the two populations.

(b) Decay of linkage disequilibrium of 5 EA *Carya* species.
Figure 3

Evolutionary history and gene flow of EA *Carya*.

(a) Demographic history of EA *Carya* inferred by PSMC.

(b) Estimation of divergence time for these 5 EA *Carya* species.

(c) Gene migration between different species, the arrow represents the direction of gene flow, the darker the color means the stronger the gene exchange.
Figure 4

Enrichment of candidate genes for environmental adaptation of *C. hunanensis*.

(a) Selected region of *C. hunanensis* based on FST & $\pi$.

(b) Gene ontology (GO) enrichment of the selected region in *C. hunanensis*.

(c) KEGG enrichment of selected region in *C. hunanensis*. The size of points represents the number of genes, and the q value represents the significance of the enrichment. Rich factor is the ratio of the annotated gene under this item to all genes in this item.
Figure 5

Species distribution modeling of EA *Carya*. (a,b,c,d,e) From top to bottom, the species distribution modeling of *C. cathayensis*, *C. dabieshanensis*, *C. hunanensis*, *C. kweichowensis* and *C. tonkinensis* in three periods (LIG, LGM, Present), respectively.

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

- Additionalfile1TableS1.xlsx
- Additionalfile2TableS2.xlsx
- Additionalfile3TableS3.xlsx
- Additionalfile4FigS1.jpg
- Additionalfile5TableS4.xlsx
- Additionalfile6TableS5.xlsx