frontiers | Research Topics

Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation

Edited by
Shengqian Xia, Yue Yaojing, Nikica Šprem, Jianhai Chen and Yongjie Wu

Published in
Frontiers in Genetics
Frontiers in Ecology and Evolution

ISSN 1664-8714
ISBN 978-2-83251-416-0
DOI 10.3389/978-2-83251-416-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation

Topic editors

Shengqian Xia — The University of Chicago, United States
Yue Yaojing — International Livestock Research Institute (Ethiopia), Ethiopia
Nikica Šprem — University of Zagreb, Croatia
Jianhai Chen — The University of Chicago, United States
Yongjie Wu — Sichuan University, China

Citation

Xia, S., Yaojing, Y., Šprem, N., Chen, J., Wu, Y., eds. (2023). *Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-416-0

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 10 January 2023
DOI 10.3389/fgene.2022.1104836

OPEN ACCESS

EDITED AND REVIEWED BY
Luis Diambra,
National University of La Plata,
Argentina

*CORRESPONDENCE
Jianhai Chen,
jianhaichen@uchicago.edu

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 22 November 2022
ACCEPTED 07 December 2022
PUBLISHED 10 January 2023

CITATION
Chen J, Šprem N, Wu Y and Xia S (2023), Editorial: Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation. *Front. Genet.* 13:1104836. doi: 10.3389/fgene.2022.1104836

Editorial: Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation

Jianhai Chen[1]*, Nikica Šprem[2], Yongjie Wu[3] and Shengqian Xia[4]

[1]Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, China, [2]Department of Fisheries, Apiculture, Wildlife Management and Special Zoology, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia, [3]Key Laboratory of Bio-resources and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan, China, [4]Department of Ecology and Evolution, The University of Chicago, Chicago, IL, United States

KEYWORDS

evolutionary young genes, genetic novelty, crop improvement, evolutionary biology, population genetics, biomarker, DNA sequencing

Editorial on the Research Topic
Application in evolutionary novelties and diversities: Medicine, agriculture, and conservation

Since the early days of DNA sequencing using the Sanger dideoxy synthesis (Sanger et al., 1977) and Maxam–Gilbert chemical cleavage (Maxam and Gilbert, 1977) methods, multiple improved platforms have been developed and commercialized to produce higher throughput and longer DNA sequences. Example of these advances include the widespread exploration of high-throughput Next-Generation Sequencing (Slatko et al., 2018) (NGS) and long-reads oriented Third Generation Sequencing (Bleidorn, 2016) (TGS). The ever-growing amount of multi-omics data has subsequently innovated the research paradigm for biological questions, especially the population genetics which emphasized the classical evolutionary questions on genetic novelties and biodiversity.

As an old science, evolutionary biology occupies a central position in the biological sciences. Most evolutionary models and theories were developed early with the establishment of evolutionary biology and population genetics. Until recently, however, questions about the genetic basis of evolutionary novelty and biodiversity could be answered directly through in-depth analysis of sequences and molecular biology technologies. Our knowledge of biodiversity and novelty is growing rapidly and enormously in species of academic, social, and economic importance. In light of this critical bridging role of sequencing data and biotechnology between questions and applications, we have launched a Frontiers Research Topic: "*Application in Evolutionary Novelties and Diversities: Medicine, Agriculture, and Conservation, to yield more insightful intellectual chemicals.*"

Broadly speaking, the Research Topic aims to promote practical applications of genetic markers in medicine, agriculture, and conservation and exploration with evolutionary novelties and/or diversity. The novel genetic elements of a species have long attracted special interest. Among these markers, evolutionary young genes are known to play important roles in phenotypic novelties with medical importance (Suzuki et al., 2018), adaptive evolution (Deng et al., 2010; Long et al., 2013), agricultural potential applications (Xia et al., 2016; Chen et al., 2019; Zhang et al., 2020), genomic evolution of pest species (Miller et al., 2022), *etc.* In this Research Topic, young genes were identified and analyzed in a plant species, Olea europaea subsp. Cuspidate, a widespread close relative of olive. Notably, a burst (19.5%) of gene transposition events was detected in the common ancestor of olive subspecies, suggesting the importance of the emergence of new genes in the evolution of olive species.

In addition to evolutionarily young genes, other biomarkers, such as microsatellites, are also extensively explored for conservation genetics. In this Research Topic, the nuclear genetic diversity and structure of wild populations of Anastrepha ludens were analyzed with microsatellite markers. Gálvez-Reyes et al. examined nine microsatellite loci and answered interesting questions about population diversity, structure, gene flow, and population size of the Mexican fruit fly, *Anastrepha ludens*, an important pest that causes widespread damage to a range of fruit crops in Mexico. In addition, Sun et al. conducted a landscape genetic study and linked genetic differentiation to bioclimatic factors for the parasitoid *Stethynium empoasca*. Guan et al. produced a genome assembly for *Luehdorfia taibai,* an endangered butterfly endemic to the Qinling Moutains in China with extremely small populations, which showcases the whole-genome as a powerful and complete biomarker to study conservation biology. Ke et al. conducted temporal sampling and network analysis on *Plutella xylostella* and uncovered rapid population change and dynamic migration patterns in overwintering regions of this cosmopolitan pest. Guo et al. conducted a mendelian randomization study based on single-nucleotide polymorphisms (SNPs) to analyze the causal relationship between vitamin D and uterine fibroids, demonstrating the applications of genomic variants in addressing questions of cause and effect with medical considerations. Mamat et al. performed population genetics for *Tolai hares*, *Lepus tolai*, in Xinjiang, China, using genome-wide SNPs from SLAF-seq and mitochondrial markers. Mu et al. performed transcriptome analysis to understand the mechanisms underlying the adaptation of plateau pika, *Ochotona curzoniae*, on the Qinghai-Tibet Plateau. Tian and Ma proposed to use natural hybrids in horticulture to accelerate the breeding of woody ornamental plant. Although the specific questions are different, these studies demonstrate the power of genetic, genomic, or transcriptomic data at species or population level in address questions involving biodiversity and evolution.

In summary, this Research Topic has gained updated knowledge input from multidiscipline scientists who studied genetic markers considering the importance of novelty and diversity. Future fine-scale studies and multi-omics data from more diverse species and geographical populations would enrich our understanding of the evolutionary significance of genetic novelty and biodiversity on our planet.

Author contributions

JC wrote the manuscript and all authors took part in improving the manuscript. YW designed the conceptual framework and helped the arrangement of background knowledge. SX supplied ideas on plant breeding and improved manuscript. NŠ provided critical revision and insightful improvement on the manuscript.

Funding

This study was supported by the fifth batch of technological innovation research projects in Chengdu (2021-YF05-01331-SN), the Postdoctoral Research and Development Fund of West China Hospital of Sichuan University (2020HXBH087), the Short-Term Expert Fund of West China Hospital (139190032).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Syst. Biodivers.* 14, 1–8.

Chen, J., Mortola, E., Du, X., Zhao, S., and Liu, X. (2019). Excess of retrogene traffic in pig X chromosome. *Genetica* 147, 23–32. doi:10.1007/s10709-018-0048-5

Deng, C., Cheng, C.-H. C., Ye, H., He, X., and Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21593–21598. doi:10.1073/pnas.1007883107

Long, M., Vankuren, N. W., Chen, S., and Vibranovski, M. D. (2013). New gene evolution: Little did we know. *Annu. Rev. Genet.* 47, 307–333. doi:10.1146/annurev-genet-111212-133301

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–564. doi:10.1073/pnas.74.2.560

Miller, D., Chen, J., Liang, J., Betrán, E., Long, M., and Sharakhov, I. V. (2022). Retrogene duplication and expression patterns shaped by the evolution of sex chromosomes in malaria mosquitoes. *Genes (Basel)* 13, 968. doi:10.3390/genes13060968

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467. doi:10.1073/pnas.74.12.5463

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.* 122, e59. doi:10.1002/cpmb.59

Suzuki, I. K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., et al. (2018). Human-specific NOTCH2NL genes expand cortical neurogenesis through delta/notch regulation. *Cell* 173, 1370–1384. e1316. doi:10.1016/j.cell.2018.03.067

Xia, S., Wang, Z., Zhang, H., Hu, K., Zhang, Z., Qin, M., et al. (2016). Altered transcription and neofunctionalization of duplicated genes rescue the harmful effects of a chimeric gene in Brassica napus. *Plant Cell* 28, 2060–2078. doi:10.1105/tpc.16.00281

Zhang, Z., Fan, Y., Xiong, J., Guo, X., Hu, K., Wang, Z., et al. (2020). Two young genes reshape a novel interaction network in Brassica napus. *New Phytol.* 225, 530–545. doi:10.1111/nph.16113

frontiers | Frontiers in Ecology and Evolution

TYPE Brief Research Report
PUBLISHED 04 August 2022
DOI 10.3389/fevo.2022.955246

OPEN ACCESS

EDITED BY
Yongjie Wu,
Sichuan University, China

REVIEWED BY
Arong Luo,
Institute of Zoology (CAS), China
Xiangqun Yuan,
Northwest A&F University, China

*CORRESPONDENCE
Huateng Huang
huanghuateng@snnu.edu.cn
Sheng-Quan Xu
xushengquan@snnu.edu.cn

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 28 May 2022
ACCEPTED 19 July 2022
PUBLISHED 04 August 2022

CITATION
Guan D-L, Zhao L, Li Y, Xing L-X, Huang H and Xu S-Q (2022) Genome assembly of *Luehdorfia taibai*, an endangered butterfly endemic to Qinling Moutains in China with extremely small populations. *Front. Ecol. Evol.* 10:955246. doi: 10.3389/fevo.2022.955246

Genome assembly of *Luehdorfia taibai*, an endangered butterfly endemic to Qinling Moutains in China with extremely small populations

De-Long Guan[1], Lu Zhao[1], Yufei Li[2], Lian-Xi Xing[3], Huateng Huang[1]* and Sheng-Quan Xu[1]*

[1]College of Life Sciences, Shaanxi Normal University, Xi'an, China, [2]School of Public Health, Xi'an Jiaotong University, Xi'an, China, [3]College of Life Sciences, Northwest University, Xi'an, China

Conservation genomic resources over the past decade has drastically improved, since genomes can be used to predict diverse parameters vital to conservation management. *Luehdorfia taibai* is an endemic butterfly only found in restricted aeras in middle-west China and is critically endangered. It was classfied as a vunerlable (VN) species in the "China species red list." Here we generated 34.38 Gb of raw DNA sequencing reads and obtained a high-qualified draft genome assembly of *L. taibai*. The final genome is ~683.3 Mb, with contig N50 size of 10.19 Mb. Further, 98.6% of single-copy orthologous genes have been recovered by BUSCO. An estimated 42.34% of the genome of *L. taibai* consists of repetitive elements. Combined with gene prediction and transcriptome sequencing, genome annotation produced 15,968 protein-coding genes. Additionally, a nearly 1:1 orthology ratio of syntenic blocks between *L. taibai* and its closest genome *Luehdorfia chinensis* suggested that the genome structures have not changed much after speciation. The genome of *L. taibai* have not undergone a whole genome duplication event. Population dynamics analyses indicates that *L. taibai* has an extremely low heterozygosity of 0.057%, and its population size has declined dramatically over the past 10 thousand years. Our study describes a draft genome assembly of the *L. taibai*, the first implication of this species. We consider the globally overexploited of the host plants is not the main reason to threaten *L. taibai*. The genome will provide advice for the conservation to the economically important *Luehdorfia* lineage and this specific species.

KEYWORDS

China species red list, genome assembly, genome annotation, conservation management, *Luehdorfia taibai*

Abbreviations: BUSCO, Benchmarking Universal Single-Copy Orthologs; GO, gene ontology; LINE, long interspersed nuclear elements; WGD, whole genome duplication; Ka, non-synonymous substitution rate; Ks, synonymous substitution rate; PSMC, pairwise sequentially Markovian coalescent.

Introduction

Conservation genomics has rapidly developed in popularity over the past decade as genomic data has become increasingly valuable for answering conservation questions (Hohenlohe et al., 2021). By integrating high-quality reference genomes, scientists can gather detailed information about a species, such as its effective population size, genetic drift, and gene flow (Wright et al., 2020), providing essential benchmarks in assessing the protection status of species (Wu et al., 2020). However, the insects were less concerned than vertebrates, on which conservation management often focused (Podsiadlowski et al., 2021). With their high esthetical attractiveness, butterflies could serve as flagship species for conservation projects. Developing genome assemblies of endangered butterfly species could provide a more detailed understanding of their evolutionary history and contributes to their conservation.

The genus *Luehdorfia*, which belongs to the tribe Zerynthiini, is one of the rarest genera of butterflies in East Asia (Dong et al., 2016). The genus comprises only four species, two endemic to China (*Luehdorfia chinensis* and *L. taibai*); (Liu et al., 2013; Xing et al., 2014). *L. taibai* (NCBI txid: 367834) is a relatively recently established species [recognized in the 1994 (Chou, 1994)]. It has a restricted distribution range in the alpines of the Qinling Mountains in China (Chou, 1999; Guo et al., 2014). This species was categorized as vulnerable (VN) by the "China species red list." A field survey in 2010 recorded less than 250 extant mature individuals on the south slopes of the Qinling Mountains (Xing et al., 2014; Dong et al., 2016). Further survey work in three consecutive summers from 2011 to 2013 recovered about 100∼200 larvas from six counties in Qinling Mountains per year and observed less than three mature individuals during the eclosion seasons per day (Guo, 2013). Such a small population size and low eclosion rate suggest this species faces a high threat of extinction and should be re-evaluated as "Endangered" (Dong et al., 2014; Guo et al., 2014; Fang et al., 2019). *L. taibai* is also classified as a species of "Beneficial or Have Important Economic and Scientific Research Value" by the National Forestry and Grassland Administration. In light of this threatened status, an effective conservation strategy is urgently needed (Dong et al., 2014; Guo et al., 2014). Here, we present a high-quality genome assembly of *L. taibai*, which sheds light on its demographic history and can serve as a critical resource for future population genomics research and conservation efforts.

Materials and methods

Sample collection and sequencing

Two *L. taibai* larvae and one adult individual were collected from Huxian County, Shaanxi province, China, in May 2019. All samples were immediately transferred into liquid nitrogen and stored for DNA/RNA extractions. We used one larva for DNA sequencing and reserved the other larva and the adult individual for transcriptome sequencing.

High-quality genomic DNA was extracted from the selected larva using the Qiagen DNAeasy Tissue kit. One library for nanopore sequencing was constructed with 50 μg DNA following the standard protocols. A total of ∼34.38 Gb of raw reads were produced, with a read N50 value of 33,504 bp. Another 5 μg DNA was used for short reads sequencing. One Illumina library was constructed according to the standard protocol and sequenced on the Illumina HiSeq X-ten platform (Nair et al., 2018), generating a total of 18.32 Gb raw data of 150 bp paired-end reads. Total RNA was isolated using the Qiagen RNeasy tissue kit for the other larva and the adult sample. After reverse transcription, two cDNA libraries were sequenced using the same Illumina platform. A total of 7.17 Gb of paired-end reads were generated. All sequencing was performed by Beijing Biomarker Biotechnology Co. Ltd (Beijing biomarker biotechnology co, LTD, Beijing, China).

Genome assembly and quality assessment

Long reads generated by nanopore sequencing were cleaned first. *De novo* genome assembly was carried out using the Nextdenovo v2.5.0 software with the read length cut-off and seed length cut-off value set to 12 and 20 Kb, respectively, Guiglielmoni et al. (2021). The raw assembly was polished using the Nexpolish v1.4.0 software (Hu et al., 2020) with Illumina short reads for three rounds. Then, the haplotigs were removed using PurgeHaplotigs (Roach et al., 2018) with default parameters. Finally, we assess the integrity of the genome assembly using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.3 (Simao et al., 2015) and Quast v5.2.0 (Gurevich et al., 2013) software.

Identification of repetitive elements, protein-coding gene prediction, and genome-guided transcriptome assembly

Repeatmasker v4.0.7 (Tarailo-Graovac and Chen, 2009) with the insect library of the Repbase and Repeatmodeler v1.0.8 (Flynn et al., 2020) were used to identify and mask repetitive elements on the genome assembly. Protein-coding genes were predicted with the masked genome assembly using a combination of *ab initio* and homology-based prediction methods. The transcriptomic data was initially used in PASA v2.5.1 (Haas et al., 2003) to obtain the top

500 gene models, which were then applied in Augustus v3.3.1 (Burge and Karlin, 1997) for *ab initio* prediction. For the homology-based prediction, annotated protein sequences of four closely related species (*Papilio machaon, P. bianor, Kallima inachus,* and *Parnassius apollo*) were downloaded from the Genbank and imported into GeneWise (Birney et al., 2004). Finally, we merged all predictions to produce a non-redundant raw gene set in Evidence Modeler (Haas et al., 2008). Functional annotation of the gene set was conducted by querying the protein sequences against the InterProscan database (Jones et al., 2014) with a customized searching script. The final gene set only retains the annotated genes. Based on this final gene set, genome-guided transcriptome assemblies were carried out using HiSat2 v2.1.3 (Kim et al., 2015) and Stringtie v1.3.5 (Pertea et al., 2015) with default parameters. Differentially expressed genes between two samples were identified using edgeR (R package; Robinson et al., 2010).

Phylogenomic and comparative genomic analysis

Five closely related butterfly reference genomes—*Luehdorfia chinensis*, *Papilio machaon*, *Papilio bianor*, *Parnassius apollo*, and *Parnassius orleans* were downloaded from NCBI for identifying orthologous genes and gene families using Orthofinder v2.3.8 (Emms and Kelly, 2019). The first three species belong to the subfamily Parnassiidae, and the other two species belong to Papilioninae, another subfamily in Papilionidae (He et al., 2022). All single-copy orthologous genes shared across all six genomes were selected and aligned in MAFFT v7.4 (Katoh and Standley, 2013). With default settings, we used RaxML v8.2.12 (Stamatakis, 2014) to build a maximum-likelihood (ML) phylogeny with the concatenated sequences. Based on the ML tree topology, divergence times and nucleotide substitution rates were estimated using R8S v1.81 (Sanderson, 2003). The lowest chi-2 cross-validation score was used to select the best method in the calculation. From the website www.time-tree.org, we selected two calibration points: the divergence between *Papilio machaon* and *P. bianor* [20.3 Mya; (Condamine et al., 2012)] and that between *Parnassius apollo* and *P. orleans* [13.4 Mya; (Condamine et al., 2018)].

To examine patterns of genome evolution, we applied MCscanX (Wang et al., 2012) with default parameters to infer collinear syntenic blocks (defined as having at least five collinear genes, blast *e*-value set as 1e-10) within *L. taibai* and between *L. taibai* and its sister species, *L. chinensis*. The expansion and contraction of gene families were examined in CAFÉ v4.0 (Abramova et al., 2021) using the ultrametric tree derived in R8S.

Inference of demographic history

We inferred the *L. taibai* population size history using the Pairwise Sequentially Markovian Coalescent model [PSMC v0.6.5; Nadachowska-Brzyska et al. (2016)]. Illumina pair-end reads were mapped onto the genome assembly using bwa v0.7.17 mem (Li and Durbin, 2010). Genome consensus sequences based on the read alignments were generated with the mpileup utility in samtools v1.9 (Li et al., 2009) and the vcfutils.pl script from the psmc package. Then, we estimated the effective population size (Ne) using psmc with the "-p" option set to "$28 \times 2 + 3 + 5$" as in a previous butterfly study (Yang et al., 2020). The result was scaled assuming a generation time of 1 year and a mutation rate of 3.59e-09 per site per generation—the rate estimated from our R8S analyses.

Results and discussions

De novo genome assembly

The genome of *L. taibai* was assembled into 232 contigs and had a size of 683.3 Mb. The *de novo* genome assembly is of high quality with N50 of 10.19 Mb, L50 of 24, and the most extended contig length of 26.79 Mb (**Table 1**). Over 99.9% of the nanopore reads can be mapped back to the assembly—most long reads were incorporated. Blasting (BLASTN) assembled contigs against the database of known sequencing adaptors did not find any potential matches. In addition, Quast analysis showed a high mapping rate to *L. chinensis* genomes (98.17% contigs can be aligned). BUSCO analysis revealed that 98.6% of the 5,286 expected Lepidoptera single-copy orthologous genes are complete on the assembled genome. Only 0.2% of duplicated BUSCOs indicate the absence of haplotigs. Mapping all Illumina short reads back to the genome shows one peak in the coverage depth distribution, confirming a neglectable proportion of haplotigs (**Supplementary Figure 1**). Furthermore, most of the short reads from transcriptome sequencing can be mapped to the genome assembly as well—the mapping rates of RNA-seq

TABLE 1 Summary statistics of the genome assembly of *L. taibai*.

Characteristics	Statistic values
Genome size	683,338,409 bp
Contig number	232
Longest contig length	26,798,033 bp
Shortest contig length	36,976 bp
Rate of GC	37.94%
Contig N50	10,191,599 bp
Contig L50	24
Contig N90	1,849,070
Contig L90	83

data of the adult and larval samples were 95.26 and 95.42%, respectively. Hence, our *de novo* genome assembly showed no obvious assembly error and is primarily complete regarding functional elements. Compared to the reported genome of *L. chinensis* (N50 of 2.39 Mb, and with 1,362 scaffolds), our genome assembly of *L. taibai* has higher connectivity and integrity. It could be a better reference for future studies on the genome evolution of the genus *Luehdorfia*.

Identification of repetitive elements and gene finding

In total, the repeat sequences comprised 42.34% (289.33 Mb) of the *L. taibai* genome. Interspersed Repeats occupied the most—282.34 Mb. Among them, the retroelements and long interspersed nuclear elements are the most abundant subtypes, accounting for 4.43% (30.29 Mb) and 3.80% (26.01 Mb) of the assembly. For non-coding RNAs, we identified 107 ribosomal RNA and 3918 transfer RNA sequences. A summary of the repeat annotations is provided in **Supplementary Table 1**.

After masking repetitive sequences, we identified 15,968 protein-coding genes on the assembled genome. The mean lengths of the gene, coding DNA sequence (CDS), and intron were 12,582.95, 2,195.44, and 10,377.87 bp, respectively. The average number of CDS and exons per gene were 6.19 and 6.52, respectively. Gene functions are annotated based on protein domain conservation using Interproscan, which determines motifs and domains by querying protein sequences against 21 public databases, including the Pfam, PANTHER, Gene3D, and CDD. The InterProscan iprterm database annotated the most number of genes—83.69% (13,364) of the gene set, followed by PANTHER and Pfam, which were 78.85% (12,592) and 76.54% (12,222), respectively. In addition, 9,273 genes were annotated with gene ontology (GO) terms (see **Supplementary Table 2** for a summary of the gene functional annotations).

The larval and adult samples expressed high proportions of genes—85.04% (13,580) and 83.52% (13,338) of the gene set, respectively. 12,600 genes were shared between

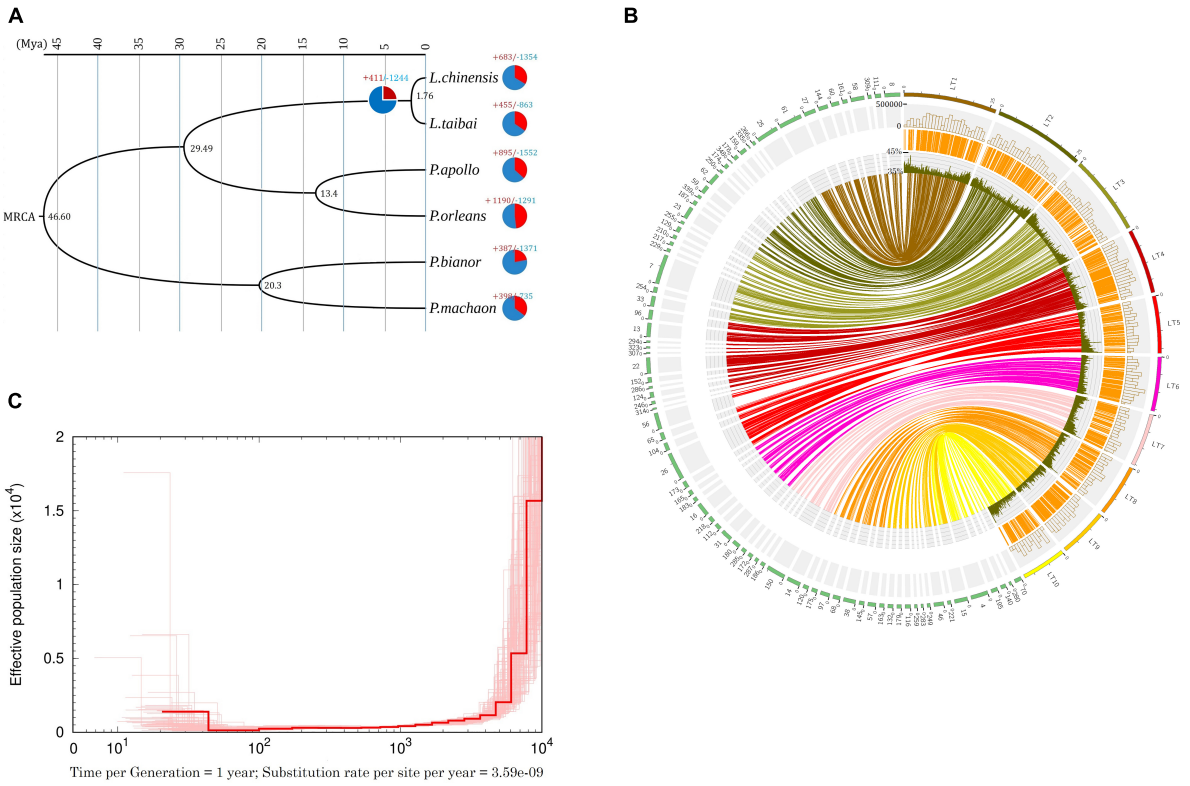


FIGURE 1
(A) Time-calibrated phylogeny of the six butterfly species based on single-copy orthologous genes. Numbers on the pie charts indicate the numbers of gene families that experienced expansion (red) or contractions (blue); the number beside each node denotes the estimated divergence time (million years ago). **(B)** Circos plot of syntenic blocks between *L. taibai* (longer colorful blocks on the right) and *L. chinensis* (shorter green blocks on the left) genome. Only the longest ten contigs in *L. taibai* are shown. The internal lines link the collinear gene pairs, and the outer circles (from inside outward) represent GC content (sliding window of 1 Mb), expressed gene location, and SNP density (sliding window of 1 Mb). The outermost numbers and strings represent the names of the contigs for *L. taibai* and the scaffold number for *L. chinensis*. **(C)** Estimated historical effective population size (Ne) of *L. taibai* with bootstrap results (thin lines).

samples, and 2,251 genes were differentially expressed (FDR adjusted *P*-value < 0.01). Enrichment analysis revealed several significantly enriched GO terms (see **Supplementary Table 3** for a Table showing the results) among these differentially expressed genes. In particular, the most significantly enriched BP (Biological Pathway) terms include reproduction (GO:0000003), reproductive process (GO:0022414), and developmental process (GO:0032502), concordant with the different developmental stages of the two samples.

Phylogenetic analysis

Across the six butterfly genomes, OrthoFinder identified a total of 14,924 orthologous and/or paralogous groups of genes. Among them, 5,923 are single-copy orthologous genes shared by the six genomes. The concatenated alignment comprises 20,534,736 amino-acid sites. The derived ML tree is well-supported such that all bootstrapping values are 100% (**Figure 1A**). The time-calibrated tree shows that *L. taibai* and *L. chinensis* are sister species with a divergence time estimated at ∼1.76 Mya (**Figure 1A**). The clade of these two species clusters with the apollo butterflies (Genus *Parnassius*; divergence time around 29.49 Mya) and then with the swallowtail butterflies (Genus *Papilio*; divergence time around 46.60 Mya).

Comparative genomics for *L. taibai*

Regarding gene families, *L. taibai* has 46 unique (254 genes), 455 expanded and 863 contracted gene families (**Figure 1A**). Functional enrichment analysis of the 211 significantly expanded (adjusted *P*-value < 0.01) gene families reveal several significantly enriched MF (Molecular Function) terms, including the heterocyclic compound binding (GO:1901363) and structural constituent of chromatin (GO:0030527). The most significantly enriched BP terms are the multi-organism process (GO:0051704) and metabolic process (GO:0008152; **Supplementary Table 4**). On the branch leading to *L. taibai* and *L. chinensis*, there are 411 predicted expanded and 1,244 contracted gene families. Several BP terms are enriched among the 63 significantly expanded gene families, including metabolic process (GO:0008152) and cellular process (GO:0009987).

Genome collinearity analyses inferred 10,260 collinear gene pairs from 369 syntenic blocks between *L. taibai* and *L. chinensis*. Each block's average number of genes reached an astonishing value of 27.81. The overall genomic gene collinearity between *L. taibai* and *L. chinensis* revealed a nearly 1:1 orthology ratio, indicating similar genomic structures in these two species (**Figure 1B**).

Genetic diversity and demographic history of *L. taibai*

The assembled *L. taibai*'s genome has 395,579 heterozygous sites, corresponding to a heterozygosity rate of 0.057%. This heterozygosity is extremely low compared to other species in the Papiliondae family. For example, *Papilio bianor*, a common species with effective population size (Ne) size over 10 million, has a heterozygosity of 1.81%. This low heterozygosity rate is comparable to the Gaint Panda (0.049%), a famous animal with worldwide conservation interest (Westbury et al., 2018).

The PSMC analysis indicates that around 10 thousand years ago, the effective population size (Ne) of *L. taibai* experienced a rapid decline and then stayed at a deficient level ever since (**Figure 1C**). This pattern is contrary to common ideas about why *L. taibai* became an endangered species. This species currently only oviposits on *Saruma henryi* Oliv., a species used in traditional Chinese herbal medicine. Over the past decades, this host species has undergone excessive exploration, which is considered the leading cause of the population decline in *L. taibai* (Zhou et al., 2010). If the over-exploitation of host plants is the primary reason, we would only observe a sharp population size decline in recent history, just as in the walrus (Shafer et al., 2015), and whales (Morin et al., 2021). The rapid decline of *L. taibai* occurred about 7000∼10,000 years ago long predated any anthropogenic activities on their host plants. Also, a diet experiment on *L. taibai* showed that under starvation, these butterflies will alternate and expand host plants (Guo, 2013). That is, a shortage of *S. henryi* might not necessarily cause the butterfly population to collapse.

Nevertheless, geological analyses of the local climate history of the Qinling Mountains showed that the local temperature significantly raised in the early Holocene (Fang and Hou, 2011; Li et al., 2015). The time coincided with the population decline we observed in *L. taibai* (**Figure 1C**). *L. taibai* lives at mid-high altitudes (above 1,500 m ASL). It is likely that this butterfly has adapted to a cold environment for most of their life cycle. The rising temperature could severely impede the growth and productivity of the butterfly, leading to population decline. If the main reason for *L. taibai*'s low effective population size was climate change, the effect of it should be the focus of future conservation and population management.

Conclusion

Here, we have assembled and annotated the genome of *Luehdorfia taibai* using a combination of Nanopore long-read

and Illumina short-read sequencing. This is the first such effort for this species and the genus *Luehdorfia*. The extremely low heterozygosity of *L. taibai* and its demographic history suggest that this species should be a priority for conservation management, and conservation efforts should focus on the impact of climate change.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GenBank under BioProject accession numbers: PRJNA615396 and PRJNA615348.

Author contributions

D-LG and LZ performed the bioinformatics analyses and wrote the manuscript. YL and L-XX collected and identified *L. taibai* samples for this research. HH and S-QX conceived the study. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Fundamental Research Funds for the Central Universities (GK201903063, GK202105003, and TD2020041Y). This work was partly supported by the National Natural Science Foundation of China (No. 31872273).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.955246/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Depth distribution of Illumina short reads mapped to the genome assembly of *L. taibai*.

SUPPLEMENTARY TABLE 1
A summary of the repeat annotations in *L. taibai*.

SUPPLEMENTARY TABLE 2
Functional annotation results from the Interproscan databases.

SUPPLEMENTARY TABLE 3
GO enrichment for differently expressed genes in larval and adult samples of *L. taibai*.

SUPPLEMENTARY TABLE 4
GO enrichment for gene families significantly expanded in *L. taibai*.

References

Abramova, A., Osińska, A., Kunche, H., Burman, E., and Bengtsson-Palme, J. (2021). CAFE: A software suite for analysis of paired-sample transposon insertion sequencing data. *Bioinformatics* 37, 121–122. doi: 10.1093/bioinformatics/btaa1086

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Chou, I. (1994). *Monographia rhopalocerorum sinensium (Monograph of Chinese butterflies) Henan*, Vol. 1. Setúbal: Scientific and Technological Publishing House, 408.

Chou, I. (1999). *Monographia rhopalocerorum sinensium*, Revised Edn. Zhengzhou: Henan Scientific and Technological Publishing House.

Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. H., and Sanmartín, I. (2018). Testing the role of the red queen and court jester as drivers of the macroevolution of apollo butterflies. *Syst. Biol.* 67, 940–964. doi: 10.1093/sysbio/syy009

Condamine, F. L., Sperling, F. A., Wahlberg, N., Rasplus, J. Y., and Kergoat, G. J. (2012). What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. *Ecol. Lett.* 15, 267–277. doi: 10.1111/j.1461-0248.2011.01737.x

Dong, S., Jiang, G., and Hong, F. (2014). Advances in conservation biology of the rare and threatened butterfly genus *Luehdorfia*(Lepidoptera: Papilionidae). *Chin. J. Appl. Environ. Biol.* 20, 1139–1144.

Dong, Y., Zhu, L. X., Wang, C. B., Zhang, M., and Ding, P. P. (2016). The complete mitochondrial genome of *Luehdorfia chinensis* Leech (Lepidoptera: Papilionidae) from China. *Mitochondrial DNA B Resour.* 1, 198–199. doi: 10.1080/23802359.2016.1155084

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238. doi: 10.1186/s13059-019-1832-y

Fang, L. J., Zhang, Y. L., Gao, K., Ding, C. P., and Zhang, Y. J. (2019). Butterfly communities along the Heihe River Basin in Shaanxi Province, a biodiversity conservation priority area in China. *J. Insect Conserv.* 23, 873–883. doi: 10.1007/s10841-019-00184-4

Fang, X., and Hou, G. (2011). Synthetically reconstructed holocene temperature change in China. *Sci. Geogr. Sin.* 31, 385–393.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K., and Flot, J. F. (2021). Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* 22:303. doi: 10.1186/s12859-021-04118-3

Guo, Z., Gao, K., Li, X., and Zhang, Y. (2014). Study on the bionomics and habitat of *Luehdorfia taibai* (Lepidoptera : Papilionidae). *Acta Ecol. Sin.* 34, 6943–6953.

Guo, Z.-Y. (2013). *The conservation biology of the endangered butterfly Luehdorfia taibai*. Xianyang: Northwest A&F University.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

He, J. W., Zhang, R., Yang, J., Chang, Z., Zhu, L. X., Lu, S. H., et al. (2022). High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. *Zool. Res.* 43, 367–379. doi: 10.24272/j.issn.2095-8137.2021.303

Hohenlohe, P. A., Funk, W. C., and Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Mol. Ecol.* 30, 62–82. doi: 10.1111/mec.15720

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255. doi: 10.1093/bioinformatics/btz891

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Li, F., Hou, G., Chongyi, E., and Jiang, Y. (2015). Integrated reconstruction of the holocene temperature series of Qinghai-Tibet plateau. *Arid Zone Res.* 32, 716–725.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, G., Jiang, G. F., Pang, H. C., and Hong, F. (2013). The mitochondrial genome of the Chinese special butterfly *Luehdorfia chinensis* Leech (Lepidoptera: Papilionidae). *Mitochondrial DNA* 24, 211–213. doi: 10.3109/19401736.2012.748043

Morin, P. A., Archer, F. I., Avila, C. D., Balacco, J. R., Bukhman, Y. V., Chow, W., et al. (2021). Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol. Ecol. Resour.* 21, 1008–1020. doi: 10.1111/1755-0998.13284

Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. J. M. E. (2016). PSMC-analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol. Ecol.* 25, 1058–1072. doi: 10.1111/mec.13540

Nair, S. S., Luu, P. L., Qu, W., Maddugoda, M., Huschtscha, L., Reddel, R., et al. (2018). Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. *Epigenetics Chromatin* 11:24. doi: 10.1186/s13072-018-0194-0

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Podsiadlowski, L., Tunström, K., Espeland, M., and Wheat, C. W. (2021). The genome assembly and annotation of the Apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. *Genome Biol. Evol.* 13:evab122. doi: 10.1093/gbe/evab122

Roach, M. J., Schmidt, S. A., and Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460. doi: 10.1186/s12859-018-2485-7

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Shafer, A. B., Gattepaille, L. M., Stewart, R. E., and Wolf, J. B. (2015). Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: In silico evaluation of power, biases and proof of concept in Atlantic walrus. *Mol. Ecol.* 24, 328–345. doi: 10.1111/mec.13034

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4:Unit 4.10. doi: 10.1002/0471250953.bi0410s25

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Westbury, M. V., Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., et al. (2018). Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Mol. Biol. Evol.* 35, 1225–1237. doi: 10.1093/molbev/msy037

Wright, B. R., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., and Grueber, C. E. (2020). A demonstration of conservation genomics for threatened species management. *Mol. Ecol. Resour.* 20, 1526–1541. doi: 10.1111/1755-0998.13211

Wu, M. Y., Low, G. W., Forcina, G., van Grouw, H., Lee, B. P. Y., Oh, R. R. Y., et al. (2020). Historic and modern genomes unveil a domestic introgression gradient in a wild red junglefowl population. *Evol. Appl.* 13, 2300–2315. doi: 10.1111/eva.13023

Xing, L. X., Li, P. F., Wu, J., Wang, K., and You, P. (2014). The complete mitochondrial genome of the endangered butterfly *Luehdorfia taibai* Chou (Lepidoptera: Papilionidae). *Mitochondrial DNA* 25, 122–123. doi: 10.3109/19401736.2013.800506

Yang, J., Wan, W., Xie, M., Mao, J., Dong, Z., Lu, S., et al. (2020). Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*. *Mol. Ecol. Resour.* 20, 1080–1092. doi: 10.1111/1755-0998.13185

Zhou, T. H., Qian, Z. Q., Shan, L., Guo, Z. G., Huang, Z. H., Liu, Z. L., et al. (2010). Genetic diversity of the endangered Chinese endemic herb *Saruma henryi* Oliv. (Aristolochiaceae) and its implications for conservation. *Popul. Ecol.* 52, 223–231.

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 15 August 2022
DOI 10.3389/fevo.2022.948640

OPEN ACCESS

EDITED BY
Nikica Šprem,
University of Zagreb, Croatia

REVIEWED BY
Raul Ruiz,
United States Department of Agriculture (USDA), United States
Eugenia Zarza,
El Colegio de la Frontera Sur, Mexico

*CORRESPONDENCE
Nancy Gálvez-Reyes
nancygalvez@ecologia.unam.mx
Daniel Piñero
pinero@unam.mx

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 20 May 2022
ACCEPTED 18 July 2022
PUBLISHED 15 August 2022

CITATION
Gálvez-Reyes N, Salvador-Figueroa M, Santini NS, Mastretta-Yanes A, Núñez-Farfán J and Piñero D (2022) Nuclear genetic diversity and structure of *Anastrepha ludens* wild populations evidenced by microsatellite markers. *Front. Ecol. Evol.* 10:948640. doi: 10.3389/fevo.2022.948640

Nuclear genetic diversity and structure of *Anastrepha ludens* wild populations evidenced by microsatellite markers

Nancy Gálvez-Reyes[1,2]*, Miguel Salvador-Figueroa[3], Nadia S. Santini[4,5], Alicia Mastretta-Yanes[5,6], Juan Núñez-Farfán[1] and Daniel Piñero[1]*

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2]Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Instituto de Biociencias, Universidad Autónoma de Chiapas, Tapachula, Mexico, [4]Departamento de Ecología Funcional, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [5]Consejo Nacional de Ciencia y Tecnología, Mexico City, Mexico, [6]Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Mexico City, Mexico

The Mexican fruit fly, *Anastrepha ludens*, is an important pest that causes widespread damage to a number of fruit crops in Mexico. The sterile insect technique (SIT) is commonly used for its control. However, the existence of natural barriers can give rise to a population structure in neutral loci and possibly behavioral or adaptive traits that interfere with SIT. For this reason, it is important to understand the genetic diversity and structure of *A. ludens* populations and to better understand the evolutionary ecology and population processes in view of possible expansions and possible host shifts due to climate change. We genotyped nine nuclear DNA (nDNA) microsatellite loci among fruit fly populations collected from five biogeographic areas within Mexico, namely, the Mexican Plateau, the Northeastern Coastal Plain, the Pacific Coast, the Gulf Coast of Mexico, and the Soconusco, and a laboratory strain. The nuclear genetic diversity was moderate (from *He* = 0.34 to *He* = 0.39) within the wild mexfly population. We found that populations were clustered in three genetic groups (K = 3). The diversity and the genetic structure of *A. ludens* are determined by environmental and geological conditions, as well as local conditions like anthropogenic perturbation, which would produce population expansion and the existence of possible predators that would affect the population density. Gene flow showed recent migration among populations. The laboratory strain showed fewer diversity than the wild samples. Large values of current and ancestral population size suggest high resistance to climatic changes, probably due to biological attributes, such as its polyphagous, multivoltine, and high dispersal characteristics. In particular, ecosystem fragmentation and perturbation as well as the existence of new plant hosts would probably increase the abundance of flies.

KEYWORDS

variability, structure, migration, biogeographic zone, nuclear, microsatellites

Introduction

The patterns of genetic variation in wild populations are the consequence of changes over time and space determined by the combined effect of evolutionary forces such as genetic drift, gene flow, and selection (Nielsen and Slatkin, 2013). Understanding these patterns is particularly important for planning management strategies to control invasive and pathogenic species, such as insect pests (Krafsur, 2005).

Fruit flies of the genus *Anastrepha* (Diptera: Tephritidae) are the most harmful pest of commercial fruit crops and comprise more than 300 currently recognized species distributed in the tropical American continent (Norrbom et al., 1999; Norrbom and Korytkowski, 2012, 2009). In Mexico, 37 species of economic importance have been described, including *Anastrepha obliqua*, *Anastrepha serpentina*, *Anastrepha striata, Anastrepha fraterculus*, and *Anastrepha ludens* (Hernandez-Ortiz and Aluja, 1993; Aluja, 1994; Hernandez-Ortiz, 2007). These species cause losses of up to $710 million per year (Reyes et al., 2000; Enkerlin, 2005).

In this study, we focus on *A. ludens* (Loew), known as the Mexican fruit fly (mexfly), which is widely distributed in Mexico and Central America (Hernandez-Ortiz and Aluja, 1993; Foote, 1994). Female flies harness fruits as oviposition substrate and develop larvae causing damage to up to 60 varieties of fruits, such as citrus, mangoes, peach, guava, sapodilla fruits, and hot peppers (Hernández-Ortíz, 1992; Hernandez-Ortiz and Aluja, 1993; Aluja, 1994; Thomas, 2004).

The sterile insect technique (SIT) is a method currently used to control *A. ludens* populations. The SIT involves the production of sterile male flies whose sterility is activated by gamma radiation. These flies are then released into infested areas. Sterile males mate with wild females, who then generate infertile eggs (Klassen and Curtis, 2005; Vreysen, 2005; Pérez-Staples et al., 2021). The sterilization of males of different species of flies does not affect the biology of the insect (Walder and Calkins, 1993; Rull and Barreda-Landa, 2007; Mahmoud, 2010; Panduranga et al., 2022). In the field, it is a method that allows reducing the natural population affecting the population dynamics. However, concerns about SIT effectiveness remain. For example, a sterile laboratory fly colony established in Mexico resulted in the loss of fertility among native flies (Orozco-Dávila et al., 2007). Similarly, the colony establishment must include specimens with broad genetic diversity (Parker et al., 2021). Also, the "degradation" of laboratory stocks occurs *via* adaptation by selection and inbreeding, with a loss of field competitiveness (Krafsur and Ouma, 2021). Predating pressure is absent in colonies, and sexual selection that operates in nature may be relaxed during laboratory adaptation (Krafsur and Ouma, 2021). Inbreeding causes a loss of genetic variation that could, in principle, reduce competitiveness, and a substantial loss of heterozygosity may occur if a release strain is formed from too few founding insects, followed by a prolonged "bottleneck" in colony size (Krafsur and Ouma, 2021). This could also be related to the loss of competitiveness of mexfly males in the field due to the artificial selection of mass-reared flies. The SIT premise supposes that there is a high genetic similarity between wild and laboratory strains that allow mating. However, sterile males usually have lower mating success with wild females compared with wild males (Pérez-Staples et al., 2013). After many decades of mass reproduction in captivity, the success of this approach has decreased because of mating incompatibility, low competitiveness in reproductive behavior (sexual selection and courtship patterns), and shorter longevity of the sterile versus the wild males (Rull et al., 2005; Rull and Barreda-Landa, 2007). Also, population isolation, such as mass rearing, leads to the appearance of homozygous individuals and inbreeding depression, which in turn causes inefficient mating (Alberti et al., 2002; Rull and Barreda-Landa, 2007). The loss of genetic variation in mass-reared insects may also reflect the loss of the wild genotypes and the loss of their natural vigor under laboratory conditions (Zygouridis et al., 2014). Furthermore, management strategies probably affect the SIT control method in a negative way by diminishing genetic variation (Dupuis et al., 2019; Ruiz-Arce et al., 2019). This impact could be more important for a species with large population sizes and this is distributed in a wide geographic area (e.g., Shi et al., 2005).

Molecular analyses based on population genetics studies on *A. ludens* using mitochondrial sequence polymorphisms (Ruiz-Arce et al., 2015), biochemical markers such as isoenzymes and molecular dominant amplified fragment length polymorphism (AFLP) markers (Malavasi and Morgante, 1982; Molina-Nery et al., 2014; Pecina-Quintero et al., 2020, 2009; Ruiz-Montoya et al., 2020), and single-nucleotide polymorphisms (SNPs) throughout the genome (Dupuis et al., 2019) have contributed to the understanding of the origin of mexfly populations and indicate mixed results stemming from the extent of the geographic range analyzed and in a lesser degree from the genetic markers (dominant–codominant) employed to survey populations. Particularly, at a small scale, AFLP markers detect very low (Nuevo León, Tamaulipas; Pecina-Quintero et al., 2009) or high genetic structure (Veracruz, Nuevo León, and Tamaulipas; Pecina-Quintero et al., 2020). When using biochemical markers such as allozymes and surveying seven populations in the different States of Mexico, with different climates and vegetation types, low genetic structure for *A. ludens* was detected (Molina-Nery et al., 2014). In contrast, populations of this species in Chiapas, Mexico, show high genetic variation and low genetic structure between localities and moderate structure when populations are grouped by host plant used (Ruiz-Montoya et al., 2020). The latter results are in agreement with the findings of Ruiz-Arce et al. (2015), they performed the most comprehensive phylogeographic study on the genetic structure of *A. ludens* to date, with two mitochondrial regions (COI and ND6) covering 67 populations along the whole range of the species and collections from the northern and southern

Mexican territory to Central America. Ruiz-Arce et al. (2015), found a low genetic structure between the two major groups (northŰsouth) divided by the Tehuantepec Isthmus, accounting ca. 6% of genetic variance, and significant genetic variance within groups (ca. 25%). Furthermore, this study supports a Southern Mexico/Central American origin for *A. ludens* (Ruiz-Arce et al., 2015). However, SNP analyses of the same populations did not detect differences in genetic diversity between populations of Central America and others, but they did detect a higher level of genetic structuring (F_{ST} = 0.09). With these markers, three main genetic clusters were identified, namely, west Mexico, east Mexico/Texas, and Isthmian Central America; high divergence in the studied strains of *A. ludens* and an explicit biogeographic analysis are suggested to identify the ancestral range of *A. ludens* (Dupuis et al., 2019). However, further molecular analyses to determine the diversity and the structure based on nuclear microsatellites may be used to better understand the dynamics of mexfly populations.

During the past decade, nuclear microsatellite tools developed for *Anastrepha* species have increased exponentially (Boykin et al., 2010; Islam et al., 2011; Lanzavecchia et al., 2014; Manni et al., 2015; Ruiz-Arce et al., 2019). Microsatellites are polymorphic DNAs that comprise sequences of repeated nucleotides; usually, the repeated units are composed of two to six nucleotides (motifs) (Hancock, 1999). The number of repeated units in a specific locus may differ, thus constituting alleles. Microsatellites are codominant, a relevant property that makes these markers suitable for population genetics studies, such as, genetic structure, gene flow, and genetic relationships between populations. Microsatellites are distributed along the whole genome of organisms, and although it seems that they are not strictly randomly distributed, they are present in coding and noncoding sequences. Theoretically, microsatellites are expected to have lower occurrence in coding regions due to purifying selection (see Carneiro-Vieira et al., 2016 and references therein). Although microsatellites have higher rates of mutation than other genetic markers, variations between loci and alleles within a locus have been documented (Jin et al., 1996; Carneiro-Vieira et al., 2016). Short-length microsatellite repeats appear to have lower mutation rates (Schug et al., 1997). The high polymorphism level of microsatellites makes them more suitable than other genetic markers (Carneiro-Vieira et al., 2016). Relative reproducibility in different laboratories and high-throughput genotyping (Barker, 2002; Selkoe and Toonen, 2006) have proven to be very useful to study populations. Specifically, microsatellite markers have been designed for *A. obliqua*, and microsatellite amplification in *A. ludens* DNA has been successfully tested (Islam et al., 2011).

The correlation of genetic variation with the geographic distribution of *A. ludens* populations has not been explored in detail. In this study, we analyzed *A. ludens* populations from five biogeographic areas within Mexico, namely, the Mexican Plateau, the Northeastern Coastal Plain, the Pacific Coast, the Gulf Coast of Mexico, and Soconusco, as well as a laboratory strain (25 years old = 152 generations, used for developing SIT). We used nine selected microsatellite markers to achieve the following three goals, (i) to investigate the genetic structure of the *A. ludens* population from different biogeographic areas in Mexico; (ii) to analyze if wild *A. ludens* populations from Mexico are genetically different from the laboratory strain; and (iii) to propose recommendations for pest management. We hypothesized that the population structure among *A. ludens* populations would be influenced by climate and the biogeographic characteristics of its distributional area. Additionally, we expect lower diversity in laboratory strain samples than in wild population samples.

Materials and methods

Sampling and genomic DNA extraction

We collected *A. ludens* adults from five biogeographic provinces that included the Mexican Plateau (MP), the Northeastern Coastal Plain (NCP), the Pacific Coast (PC), the Gulf Coast of Mexico (GC), and the Soconusco (SOC). These areas have been defined using the proposal by Rzedowski (1978), and it is based on plant species composition in different localities (see **Supplementary Table 1**). Additionally, a laboratory strain (LAB) of ∼25 years equivalent to 152 generations, currently in production and being used by the mass rearing facility (Orozco-Dávila et al., 2007, 2017), was used as a reference group to analyze the levels of genetic variation and inbreeding. The specimens were provided by the Department of Colonization and Breeding of Fruit Flies, the Development Methods of the Moscafrut Program (Metapa de Domínguez, Chiapas), and the Instituto de Ecología-INECOL (Xalapa, Veracruz, México). Flies were sorted for each locality (10 insects of each sex, 20 from each locality). In total, 120 specimens were stored at −20°C until processed for further analyses. Deoxyribonucleic acid (DNA) extractions and purifications were performed with a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, United States) following the instructions of the manufacturer.

Genotyping microsatellite markers

We amplified DNA using nine fluorescently labeled nuclear microsatellite primers (GenBank access key: JF292994, JF292995, JF292996, JF293003, JF292998, JF293004, JF292999, JF293001, and JF293006). Amplification of microsatellite loci was carried out following the protocol described by Islam et al. (2011). PCR products of approximately 150–240 bp were confirmed on 1% agarose gel dyed with 0.1% ethidium bromide. The samples were processed at the Caver Biotechnology, University of Illinois (Champaign, United States). We determined the size of each peak by using Peak Scanner software 1.0 (Applied Biosystems).

Estimation of genetic diversity

We calculated the number of alleles (*Ae*), the observed (*Ho*) and expected (*He*) heterozygosity, Hardy-Weinberg equilibrium (HWE), allelic richness (A), and effective population size (*Ne*) with Arlequin 3.5.2.2 (Excoffier and Lischer, 2010). Also, we assessed the inbreeding coefficient (F_{IS}) and the statistics *F* using GenePop 4.6 (Rousset, 2008).

Structure and gene flow of populations

We estimated the degree of genetic differentiation from R_{ST} and F_{ST} values using Arlequin 3.5.2.2. The genetic grouping was performed using Samova 1.0, and the analysis of molecular variance (AMOVA) was performed using Arlequin 3.5.2.2. The genetic structure for collections was examined using the ancestry admixture model runs with 100,000 burn-in, 1,000,000 MCMC, and 25 iterations for each value *K* (1−8) with STRUCTURE 2.3.4 (Pritchard et al., 2000). We estimated the most probable *K* value, the highest value of ΔK was calculated by implementing the Evanno method (Evanno et al., 2005) using Structure Harvester v0.6.94 (Earl and vonHoldt, 2012), the structure runs were summarized using CLUMPP (Jakobsson and Rosenberg, 2007), and the results were plotted in Distruct v1.1 and CLUMPP v1.1.2 (Rosenberg, 2004). The dendrogram was calculated with the genetic distance using the Neighbor-Joining with Population v1.2.30 (Langella, 1999). The tree was exported, visualized, and edited using figtree-1.4.3 (Rambaut, 2012). For the isolation by distance (IBD), we calculated population pairwise F_{ST} values (Slatkin's Distance) with 9,999 permutations (Arlequin 3.5.2.2), and we tested IBD using a Mantel test with 9,999 permutations by testing for a correlation between genetic distance and geographic distance (Arlequin 3.5.2.2). We detected the genetic barriers using the Monmonier algorithm based on linearized F_{ST} with Barrier 2.2 (Slatkin, 1995; Manni et al., 2004) and evaluated the gene flow between genetic groups with Migrate-N 3.6 (Beerli, 2006).

Results

Genetic diversity of populations

The genetic diversity exhibited 40 alleles from wild fly collections, two alleles for the laboratory strain, nine private alleles from the wild fly, and eleven private alleles among all collections (**Supplementary Table 2**). The HWE test for wild and laboratory flies showed homozygous excess at the nine microsatellite loci, but we only found significant deviations in six of them (**Supplementary Table 3**). The highest allelic richness was observed for Soconusco (*A* = 3.22), and the lowest allelic richness was observed for the laboratory strain (*A* = 2.11). The highest genetic diversity of Nei was found in the Gulf Coast of Mexico (*He* = 0.39) and Soconusco (*He* = 0.38). The lowest values were found in the population of the Mexican Plateau (*He* = 0.35) and the laboratory strain (*He* = 0.34). The positive values and significant inbreeding coefficients (F_{IS} $p < 0.05$) showed heterozygosity deficiency for all loci (**Table 1**).

Structure and gene flow of populations

The comparison of pairwise R_{ST} and F_{ST} values displayed genetic differentiation among all populations with microsatellites (**Table 2**), principally, among SOC and PC (R_{ST} = 0.191; F_{ST} = 0.065) and LAB among SOC (R_{ST} = 0.154; F_{ST} = 0.110). The analysis of molecular

TABLE 1 Genetic diversity of nine nDNA microsatellite loci.

Biogeographic regions	nDNA microsatellite loci						
	Ae	**A**	**I**	***Ho***	***He***	**F_{IS}**	***P***
Mexican Plateau	1.79	3.00	0.62	0.14	0.35	0.59*	88.9
SE	0.26	0.47	0.16	0.04	0.09	0.12	
Northeastern Coastal Plain	1.93	2.56	0.61	0.31	0.36	0.15*	66.7
SE	0.35	0.56	0.18	0.09	0.09	0.09	
Pacific Coast	1.99	2.89	0.64	0.26	0.36	0.30*	88.9
SE	0.42	0.69	0.19	0.06	0.09	0.11	
Gulf Coast of Mexico	2.05	3.0	0.68	0.20	0.39	0.49*	77.8
SE	0.39	0.58	0.19	0.07	0.09	0.12	
Soconusco	1.91	3.22	0.66	0.25	0.38	0.35*	100
SE	0.29	0.68	0.17	0.06	0.09	0.08	
laboratory strain	1.67	2.11	0.51	0.21	0.34	0.38*	77.8
SE	0.19	0.31	0.13	0.08	0.08	0.16	
Global	1.8	2.79	0.62	0.23	0.36	0.26	83.3
SE	0.13	0.23	0.07	0.03	0.04	0.05	4.8

We used 20 samples per population. Values are means and standard errors (SE).
*indicates significant differences, $p < 0.05$.
Ae: No. of effective alleles, A: allelic richness, I: Shannon index, *Ho*: observed heterozygosity, *He*: expected heterozygosity, F_{IS}: inbreeding coefficient, P: polymorphism percentage.
Ne values scaled to θ mutation rate in black.

TABLE 2 Pairwise comparison R_{ST} and F_{ST} in nDNA microsatellites loci.

R_{ST} \ F_{ST}	MP	NCP	PC	GC	SOC	LAB
Mexican Plateau (MP)	–	0.030	0.009	0.017	0.047*	0.046*
Northeastern Coastal Plain (NCP)	0.104*	–	0.048*	0.001	0.010	0.044*
Pacific Coast (PC)	0.013	0.145*	–	0.028	0.065*	0.049*
Gulf Coast of Mexico (GC)	0.033	0.007	0.066	–	0.026	0.053*
Soconusco (SOC)	0.156*	0.005	0.191*	0.038	–	0.110*
laboratory Strain (LAB)	0.008	0.102*	0.0005	0.035	0.154*	–

*indicates significant level differences, $p < 0.05$. Ne values scaled to θ mutation rate in black.

variance presented Φ_{ST} = 0.115 in nDNA microsatellite loci among wild populations and the laboratory strain (**Table 3**). According to the genetic structure analysis, the best K value was three genetic populations, K = 3 showed the highest value of ΔK, and the second most relevant K value is 4, suggesting ancestry on one fraction of mixed alleles, probably genetic migrants, except Soconusco and the laboratory strain (**Figure 1A** and **Supplementary Figure 1**). Moreover, the dendrogram supported a pattern of genetic clustering with three groups (**Figure 1C**). The Mantel test did not show a significant relationship between geographic and genetic distances with microsatellites $< r = 0.275$, $p > 0.05 >$ (**Supplementary Figure 2**).

We identified three main genetic barriers contributing to the population structure: "a," "b" (bootstrap support = 100), and "c" (bootstrap support = 90). Genetic barrier "a" separates populations to the east of the country (Northeastern Coastal Plain and Gulf Coast). Genetic barrier "b" likely is located at the Isthmus of Tehuantepec, separating Soconusco from the rest of the populations. Genetic barrier "c" separates the Mexican Plateau from the Pacific Coast (**Figure 1B**). For genetic flow, our result about the most probable migration model was based on the Neighbor-Joining dendrogram (**Table 4**) calculated with Migrate-N 3.6 [AICmin-AICi = 0, p(Bezier) = 1]. The migration was observed only among neighboring populations (Soconusco↔Gulf Coast of Mexico↔Northeastern Coastal Plain, and Mexican Plateau↔Pacific Coast), with migration being higher from south to north, and it was not found between the two groups (**Figure 1B** and **Table 4**).

Discussion

By studying the genetic variation of nDNA microsatellites in *A. ludens*, the data reveal moderate genetic diversity and population structure (K = 3) within species, including wild populations and laboratory strains. The genetic diversity and population structure may be influenced by climate with a close association in the species' distribution (Santos et al., 2020) and the biogeographic characteristics of the regions of this study, because we identified genetic barriers delineating populations, showing high bootstrap support. Those genetic barriers are delineated according to physiographic features from the Sierra Madre Oriental and the Sierra Madre Occidental, the Sierra Madre del Sur, and the Isthmus of Tehuantepec. Thus, migration was observed only among neighboring populations.

Genetic variability

Moderate Nei's genetic diversity and low allelic richness can be observed on nDNA among populations of *A. ludens*. Ruiz-Arce et al. (2015) reported low genetic diversity of *A. ludens* (hd = 0.58, π = 0.00184) with 68 haplotypes (COI+ND6). The low allelic richness and moderate Nei's genetic diversity in wild populations were reported in the same species (Malavasi and Morgante, 1982; Pecina-Quintero et al., 2009; Molina-Nery et al., 2014). Also, relative low diversity for the laboratory strain has been previously reported in nuclear markers (Isozymes: A = 1.55 and 1.95) within this species (Malavasi and Morgante, 1982; Molina-Nery et al., 2014). A moderate genetic diversity within our samples, when compared with other *Anastrepha* species that have been studied with nuclear microsatellites (*A. fraterculus*, Lanzavecchia et al., 2014; Parreño et al., 2014; *A. obliqua*, Ruiz-Arce et al., 2019; *A. suspensa*, Boykin et al., 2010), is likely due to its geographic distribution and effective population size, which are associated with host availability and climate conditions as rain and environmental heterogeneity have a strong influence on the distribution of the species (Aluja, 1994; Celedonio-Hurtado et al., 1995; Hernandez-Ortiz, 2007). Also, the strong wind dispersal capability (up to 135 km) of *A. ludens*, or human-assisted spread through the movement of infested fruits and pupae on the ground, probably has influenced the moderate genetic diversity (Christenson and Foote, 1960; Aluja, 1994; European Food Safety Authority [EFSA] et al., 2019, 2021). Natural populations of *A. ludens* are discontinuously distributed, and gene flow varies among these populations depending on historical, geographical, and environmental factors and locality conditions such as humidity, host variability, competition for space, nutrition or mating, parasitism, and competition for resources with another species (Krafsur and Ouma, 2021; Parker et al., 2021). The geographic range of many pest species is very large, and locally adapted populations probably may exist. There may be local selection regimes that cause one or more populations to differ biologically in ways that

TABLE 3 Analysis of molecular variance and genetic differentiation index in nDNA microsatellite loci among wild populations and the laboratory strain.

Source of variation	Degrees of freedom	Sum of squares	Estimated variance	Percent variation	F-statistics	*P*
Among groups	1	1419	11.042Va	11.3	Φ_{SC} = 0.002	0.001
Among populations	4	277	0.198Vb	0.2	Φ_{ST} = 0.115	0.003
Within populations	234	20204	86.344Vc	88.5	Φ_{CT} = 0.113	0.103
Total	239	22000	97.6	100		

No. of groups maximized and genetic differentiation between populations by SAMOVA in nDNA microsatellite loci (K_{max} = 2, Φ_{ST} = 0.115) showed 2 groups: (1) NCP, GC, and SOC and (2) MP, PC, and LAB.

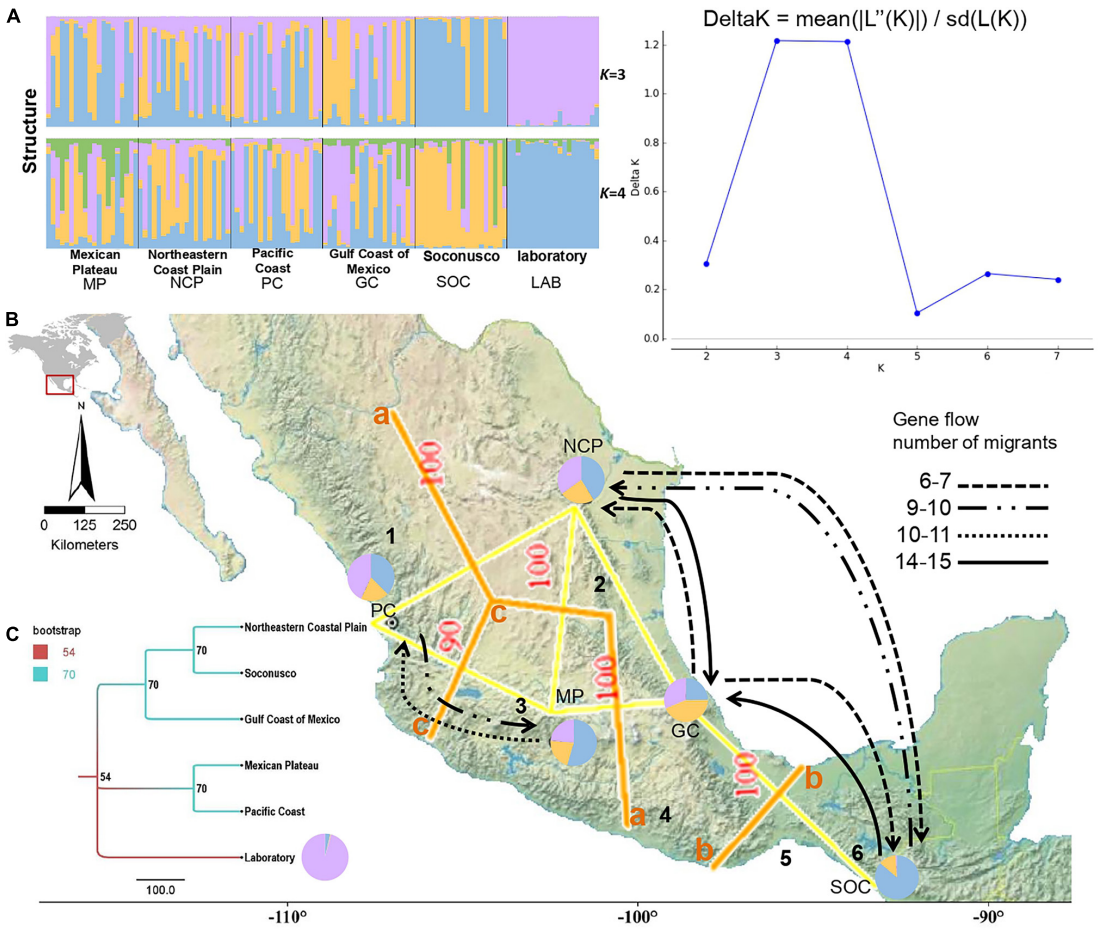


FIGURE 1
(A) Genetic structure of mexfly based on the nDNA microsatellite loci estimated using STRUCTURE and visualized with Distruct. The bar plot and pie chart indicating the blue, yellow, and purple colors of graphics indicate three genetic groups with *Admixture* model: K = 3, 25 iterations of each K (1–8) and delta K. **(B)** Map of five collected localities in Mexico showing the three main genetic barriers with nDNA and the orange line (a, b, and c) found by the Monmonier algorithm; thickness and the number on the side of the barriers indicate the percentage of bootstrap support. The map shows the migration of mexfly populations, and the arrows represent the effective number of migrants in a generation. The main physiographic features and/or regions' biogeographical boundaries for fruit fly are delimited with lines: (1) SMOcc: Sierra Madre Occidental, (2) SMO: Sierra Madre Oriental, (3) TMVB: Trans Mexican Volcanic Belt, (4) SMS: Sierra Madre del Sur, (5) IT: Isthmus of Tehuantepec, (6) SCh: Sierra Madre de Chiapas. **(C)** Genetic distance of mexfly populations obtained by *Neighbor-Joining* of nDNA.

could make the SIT less effective due to mating barriers. Such barriers might be ecological, temporal, or behavioral (Krafsur and Ouma, 2021; Parker et al., 2021).

The low allelic richness reveals a decrease in genetic diversity that can be explained by a lower effective population size when compared with other *Anastrepha* species where nuclear microsatellites have also been used (*A. fraterculus*, Lanzavecchia et al., 2014; Parreño et al., 2014; *A. obliqua*, Ruiz-Arce et al., 2019; *A. suspensa*, Boykin et al., 2010), probably due to adaptive response to environmental and demographic changes, and that indicates that *A. ludens* is at the limit of the tropical distribution. Also, allelic richness is sensitive to the effects of short and severe bottlenecks that affect the distribution of allele frequencies (Franks et al., 2011; Nielsen and Slatkin, 2013). Thus, temporal allelic richness in mexfly is likely due to effective population size variation during fruiting, related to host availability, and due to their polyphagous and multivoltine behavior (Hernandez-Ortiz and Aluja, 1993; Thomas, 2003; Aluja and Mangan, 2008).

Structure, genetic barriers, and migration of populations

We found moderate differentiation, three genetic groups within *A. ludens* populations were inferred by STRUCTURE, and no pattern of isolation by distance was detected. This is similar to other studies that have found moderate differentiation among groups F_{ST} = 0.302 (Ruiz-Arce et al., 2015), F_{ST} = 0.058 (Pecina-Quintero et al., 2009), F_{ST} = 0.105 (Molina-Nery et al., 2014), and F_{ST} = 0.047 (Dupuis et al., 2019) within populations

TABLE 4 Migration's model and no. of migrants between populations.

Migration models	Bezier LH	Harmonic media HL	K	AIC	(AICmin-AICi)	*P* (Bezier)
Among groups	−340882.47	29.52	25	681814.94	445602.22	0
Barriers	−149018.66	−502.75	11	298059.32	61846.6	0
Neighbor-joining	−118093.36	−727.56	13	236212.72	0	1

	Destiny				
Origin	**MP**	**NCP**	**PC**	**GC**	**SOC**
Mexican Plateau	**Θ = 0.90238**	0	*M = 10.74*	0	0
Northeastern Coastal Plain	0	**Θ = 0.98350**	0	*M = 14.07*	*M = 6.84*
Pacific Coast	*M = 9.76*	0	**Θ = 0.95519**	0	0
Gulf Coast of Mexico	0	*M = 5.87*	0	**Θ = 1.07390**	*M = 7.24*
Soconusco	0	*M = 9.93*	0	*M = 13.88*	**Θ = 1.07398**

The parameters used were a model of Brownian mutation for microsatellites, an analysis strategy with Bayesian inference, heating iterations of 10,000, and chain 1.0×10^7 iteration samples every 1,000. Four chains sampled temperatures T1 = 1.0, T2 = 1.5, T3 = 3.0, and T4 = 1.0×10^6. Migration model most probably based on Akaike information criterion (AIC) using Bezier's estimators was third with $K = 13$ and probability of model $\Delta AIC = 0$.
AIC, Akaike information criterion. No. of migrants significant (Nm > 1) based on *Neighbor-Joining*. Ne values scaled to θ mutation rate in black and M mutation rate in cursive.
Ne values scaled to θ mutation rate are in bold.

from Mexico. Specifically, the Soconusco population presented a high differentiation with respect to others; this is probably because the Soconusco population is isolated by the mountains of the Sierra Madre del Sur and by the environmental barrier due to trade winds from the Isthmus of Tehuantepec (Lugo-Hubp, 1990). Also, this supports the notion that the origin of this fly may have been southern Mexico/Central America (Ruiz-Arce et al., 2015).

Our results, based on population structure, demonstrated that the laboratory strain remains genetically differentiated and has lower genetic diversity than wild populations, probably due to inbreeding associated with mass rearing. Similar results were found by Dupuis et al. (2019) that included rearing strains of *A. ludens* to provide context to the structure of wild populations, and all rearing strains were quite divergent from the wild populations. Despite the genetic differentiation, the laboratory strain has been refreshed with wild populations of Mexico, principally Chiapas (Orozco-Dávila et al., 2007). However, the original laboratory strain is a mixture of a strain similar to the initial one from Mission, Texas, United States, whose specimens were collected in Nuevo León and Tamaulipas (México), and wild flies collected from different regions in Mexico for renewing laboratory strain (Orozco-Dávila et al., 2007). Until now, the original strain is still used for mass rearing (Orozco-Dávila et al., 2007, 2017). Also, mass rearing involves breeding about 175 million pupas weekly, intended for the liberation of these flies (Domínguez et al., 2010). Although rearing lines are to be highly bottlenecked and inbreeding causes a loss of genetic variation that could reduce the competitiveness of males (Krafsur and Ouma, 2021), these lines can still be expected to continue to evolve over time and may share some artificially selected traits due to common artificial rearing conditions (Dupuis et al., 2019). A progressive loss of diversity in laboratory colonies, coupled with a progressive decay in physiological quality control indices (e.g., egg hatch, larval development time, pupa size, emergence, flight ability, pheromone production, vision, longevity, and mating compatibility) (FAO/IAEA/USDA, 2019; Parker et al., 2021), could provide evidence of degradation and possibly predict a decline in field competitiveness (Krafsur and Ouma, 2021).

Nevertheless, the genetic structure found in wild populations due to genetic barriers is similar to a metapopulation dynamic (Hanski and Gilpin, 1991; Hanski, 1998; Wegier et al., 2011), that is, a group of spatially separated populations of a species that interact at some level. Each population functions in relative independence from other populations and may eventually become extinct, but immigrants may recolonize another population that is declining or that has become extinct, so the metapopulation persists as long as the colonization rate is equal to the extinction rate (Hanski and Gilpin, 1991; Hanski, 1998; Hanski and Gaggiotti, 2004). The population structure corresponds to the biogeographic provinces in Mexico (Rzedowski, 1978). For example, the main genetic barrier (**Figure 1**, letter b) matches the existence of physical barriers from the Sierra Madre Oriental and the Sierra Madre Occidental. A second genetic barrier, which is isolating the Soconusco population, is likely related to the Sierra Madre del Sur and the Isthmus of Tehuantepec, the inter-oceanic pass of low altitude in Mexico where the winds range from 18 to 90 km/h (**Figure 1**, letter b). Similar to previous population genetic studies, our population structuring is high and matches with the biogeographic zones, for example, Dupuis et al. (2019) found that the west Mexico cluster is bounded using SNPs by the southwestern and western extents of the Sierra Madre

Occidental, while Ruiz-Arce et al. (2015) found the support for genetic structure between the northern and southern parts of *A. ludens*' range, corresponding to the Isthmus of Tehuantepec, using two mitochondrial genes (Ruiz-Arce et al., 2015).

The migration model indicated that natural barriers strongly influence dispersion among populations. However, between neighbor populations, bidirectional genetic flow was significant. This suggests that migration is an important factor in the variability of genetic interchange and in reducing the genetic structure. The more the gene flow, the less the population structure. Therefore, migration under a stepping-stone model, where nearby populations have higher bidirectional gene flow, seems to better explain the distribution of genetic diversity rather than IBD alone. In addition, mexfly has been a species with the adaptation capacity to diverse climates. The gene flow and the genetic structure could be influenced by the high dispersion that occurred due to considerable distance (>30 km) by wind, climate conditions, altitude, and dissemination of larvae for fruits' commerce activity, allowing mexfly to find the availability of hosts and the refuge for reproduction (Alberti et al., 2002; Shi et al., 2005; Aluja et al., 2009).

Implications of genetic diversity for sterile insect technique management of *Anastrepha ludens*

At present, old concerns about SIT effectiveness are still applicable (Krafsur, 2005; Pérez-Staples et al., 2021). The results of the current study of mexfly are relevant to understanding the reduced effectiveness of SIT programs. This study found that the genetic diversity of the laboratory strain is lower than that from the wild populations of *A. ludens*; also, the population structure estimates for the laboratory strain revealed less admixture than wild populations, probably due to massive breeding. This factor may contribute negatively to the SIT efficacy. Furthermore, local landscape, natural barriers, and metapopulation structure may contribute to the moderate genetic diversity, gene flow leading to genetic panmixia, moderate differentiation, and genetic structure found in wild mexfly species.

The population size of *A. ludens* has been kept high, and there have been small changes in allele frequencies caused by the effect of genetic drift and bottleneck. Regarding our results, considering that the diversity and Ne of the strain are lower than the wild populations, probably due to the bottleneck effect and inbreeding, they could contribute to reduced effectiveness of SIT (Krafsur and Ouma, 2021), so one proposal would be to control the pupae stage where *A. ludens* are less likely to disperse. Also, genetic data such as inbreeding coefficients and genetic distances should be estimated continuously as part of the quality control program for target populations and for laboratory colonies and their source populations (Krafsur and Ouma, 2021). According to our STRUCTURE results, genetic data showed considerable admixture in wild populations, and it could be by natural spread. *A. ludens* is considered a strong flier (Centre for Agriculture and Bioscience International [CABI], 2019), which flies as far as 135 km (Christenson and Foote, 1960) using the wind for displacement (Aluja, 1994). The maximum natural spreading distance that *A. ludens* is expected to cover in 1 year is approximately 9.4 km (with a 95% uncertainly range of 1–34 km). However, the introduction of the pest into areas could occur by human-assisted spread through the movement of infested fruits, such as imports of fruit commodities or fruits in passenger luggage and pupae in soil or another growing medium with host plants (European Food Safety Authority [EFSA] et al., 2019, 2021). As such, we suggest establishing population management units; the first control unit would comprise Soconusco–Gulf Coast of Mexico–Northeastern Coastal Plain, and the second would integrate the Mexican Plateau and the Pacific Coast. In particular, the states of Veracruz (in the Gulf Coast of Mexico) and Tamaulipas (in Northeastern Coastal Plain) grow and harvest citrus. Similarly, Sinaloa and Nayarit (in Pacific Coast) and Chiapas (in Soconusco) grow and harvest mangoes (SAGARPA, 2022). Where each management unit could be colonized by sterile specimens derived from the same unit. This would also require developing strains compatible with each local population, by utilizing genomics differentiation in *A. ludens* populations and associations with different host plants and predicting the geographic and environmental ranges and relative abundances of invasive species using ecological niche modeling to give policies on a scientific basis (Santos et al., 2020; Aguirre-Ramirez et al., 2021; Gutierrez et al., 2021).

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, and the input files of the microsatellite loci from this project have been deposited in the Dryad repository, available at: https://doi.org/10.5061/dryad.xwdbrv1gw; further inquiries can be directed to the corresponding authors.

Ethics statement

The animal study was reviewed and approved by the Instituto de Ecología and posgrado en Ciencias Biologicas, UNAM.

Author contributions

DP, MS-F, and NG-R conceived and designed the study. NG-R performed the laboratory work, processed the microsatellites data, and performed the analyses.

DP, NS, AM-Y, MS-F, and JN-F supervised the analyses and contributed to the discussion. NG-R and DP led the manuscript writing with contributions from all authors. All authors approved the final version of the manuscript.

Funding

Funding was provided by the Instituto de Ecología, UNAM to DP and NS.

Acknowledgments

We thank Larissa Guillén-Conde and Emilio Hernández-Ortiz for providing the samples. We sincerely thank reviewers of the manuscript Marco Suárez-Atilano and Julio Cesar García-Zebadúa for the useful comments to the manuscript. This manuscript represents fulfillment of the graduate program of maestría en Ciencias Biologicas of the Universidad Nacional Autónoma de México (UNAM) to NG-R. We are grateful for the scholarship of the Consejo Nacional de Ciencia y Tecnología, México (CONACyT: 412884).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.948640/full#supplementary-material

References

Aguirre-Ramirez, E., Velasco-Cuervo, S., and Toro-Perea, N. (2021). Genomic traces of the fruit fly anastrepha obliqua associated with its polyphagous nature. *Insects* 12:1116. doi: 10.3390/insects12121116

Alberti, A. C., Rodriguero, M. S., Cendra, P. G., Saidman, B. O., and Vilardi, J. C. (2002). Evidence indicating that argentine populations of *Anastrepha fraterculus* (Diptera: Tephritidae) belong to a single biological species. *Ann. Entomol. Soc. Am.* 95, 505–512.

Aluja, M. (1994). Bionomics and management of anastrepha. *Annu. Rev. Entomol.* 39, 155–178. doi: 10.1146/annurev.en.39.010194.001103

Aluja, M., and Mangan, R. L. (2008). Fruit fly (Diptera: Tephritidae) host status determination: critical conceptual, methodological, and regulatory considerations. *Annu. Rev. Entomol.* 53, 473–502. doi: 10.1146/annurev.ento.53.103106.093350

Aluja, M., Rull, J., Pérez-Staples, D., Díaz-Fleischer, F., and Sivinski, J. (2009). Random mating among *Anastrepha ludens* (Diptera: Tephritidae) adults of geographically distant and ecologically distinct populations in Mexico. *Bull. Entomol. Res.* 99, 207–214. doi: 10.1017/S0007485308006299

Barker, G. C. (2002). Microsatellite DNA: a tool for population genetic analysis. *Trans. R. Soc. Trop. Med. Hyg.* 96, S21–S24. doi: 10.1016/S0035-9203(02)90047-7

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22, 341–345. doi: 10.1093/bioinformatics/bti803

Boykin, L. M., Shatters, R. G., Hall, D. G., Dean, D., and Beerli, P. (2010). Genetic variation of *Anastrepha suspensa* (Diptera: Tephritidae) in Florida and the caribbean using microsatellite DNA markers. *J. Econ. Entomol.* 103, 2214–2222. doi: 10.1603/EC10128

Carneiro-Vieira, M. L., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol.* 39, 312–328.

Celedonio-Hurtado, H., Aluja, M., and Liedo, P. (1995). Adult population fluctuations of anastrepha species (Diptera: Tephritidae) in tropical orchard habitats of chiapas, Mexico. *Environ. Entomol.* 24, 861–869. doi: 10.1093/ee/24.4.861

Centre for Agriculture and Bioscience International [CABI] (2019). *Invasive Species Compendium. Datasheet Anastrepha ludens (Mexican fruit fly) 19/11/1919*. Wallingford: CAB International.

Christenson, L. D., and Foote, R. H. (1960). Biology of fruit flies. *Annu. Rev. Entomol.* 5, 171–192.

Domínguez, J., Artiaga, T., Solís, E., and Hernandez, E. (2010). "Métodos de colonización y cría masiva," in *Moscas de la Fruta: Fundamentos y Procedimientos Para su Manejo*, eds P. Montoya, J. Toledo, and E. Hernández (México D.F: S y G Editores).

Dupuis, J. R., Ruiz-Arce, R., Barr, N. B., Thomas, D. B., and Geib, S. M. (2019). Range-wide population genomics of the Mexican fruit fly: toward development of pathway analysis tools. *Evol. Appl.* 12, 1641–1660. doi: 10.1111/eva.12824

Earl, D. A., and vonHoldt, B. M. (2012). Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Enkerlin, W. R. (2005). "Impact of fruit fly control programmes using the sterile insect technique," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Dordrecht: Springer Netherlands), 651–676. doi: 10.1007/1-4020-4051-2_25

European Food Safety Authority [EFSA], Baker, R., Gilioli, G., Behring, C., Candiani, D., Gogin, A., et al. (2019). *Anastrepha Ludens – Pest Report to Support Ranking of EU Candidate Priority Pests by the EFSA Working Group on EU Priority Pests*. Parma: EFSA.

European Food Safety Authority [EFSA], Schenk, M., Mertens, J., Delbianco, A., and Graziosi I Vos, S. (2021). *Pest Survey Card on Anastrepha ludens*. Parma: EFSA.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620.

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

FAO/IAEA/USDA (2019). *Product Quality Control for Sterile Mass-Reared and Released Tephritid Fruit Flies, Version 7.0*. Vienna: International Atomic Energy Agency.

Foote, B. A. (1994). Handbook of the fruit flies (Diptera: Tephritidae) of America North of Mexico. *Ann. Entomol. Soc. Am.* 87, 400–401. doi: 10.1093/aesa/87.3.400

Franks, S. J., Pratt, P. D., and Tsutsui, N. D. (2011). The genetic consequences of a demographic bottleneck in an introduced biological control insect. *Conserv. Genet.* 12, 201–211. doi: 10.1007/s10592-010-0133-5

Gutierrez, A. P., Ponti, L., Neteler, M., Suckling, D. M., and Cure, J. R. (2021). Invasive potential of tropical fruit flies in temperate regions under climate change. *Commun. Biol.* 4:1141. doi: 10.1038/s42003-021-02599-9

Hancock, J. M. (1999). "Microsatellites and other simple sequences: genomic context and mutational mechanisms," in *Microsatellites. Evolution and Applications*, eds D. B. Goldstein and C. Schlötterer (Oxford: Oxford University Press).

Hanski, I. (1998). Metapopulation dynamics. *Nature* 396, 41–49. doi: 10.1038/23876

Hanski, I., and Gaggiotti, O. (2004). "1 - metapopulation biology: past, present, and future," in *Ecology, Genetics and Evolution of Metapopulations*, eds I. Hanski and O. E. Gaggiotti (Burlington: Academic Press).

Hanski, I., and Gilpin, M. (1991). Metapopulation dynamics: brief history and conceptual domain. *Biol. J. Linn. Soc.* 42, 3–16. doi: 10.1111/j.1095-8312.1991.tb00548.x

Hernández-Ortíz, V. (1992). *El Género Anastrepha Schiner en México (Diptera: Tephritidae), Taxonomía, Distribución y sus Plantas Huéspedes*. Xalapa: Sociedad Mexicana de Entomología.

Hernandez-Ortiz, V. (2007). "Diversidad y biogeografía del género anastrepha en Mexico," in *Moscas de La Fruta En Latinoamérica (Diptera: Tephritidae): Diversidad, Biología y Manejo*, ed. V. Hernández-Ortiz (México: Distrito Federal).

Hernandez-Ortiz, V., and Aluja, M. (1993). Listado de especies del genero neotropical anastrepha (Diptera: Tephritidae) con notas sobre su distribución y plantas hospederas. *Folia Entomol. Mex.* 88, 89–105.

Islam, M. S., Ruiz-Arce, R., and McPheron, B. A. (2011). Microsatellite markers for the West Indian fruit fly (Anastrepha obliqua) and cross species amplification in related pest species. *Conserv. Genet. Resour.* 3, 549–551. doi: 10.1007/s12686-011-9401-z

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jin, L., Macaubas, C., Hallmayer, J., Kimura, A., and Mignot, E. (1996). Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc. Natl. Acad. Sci. U S A.* 93, 15285–15288.

Klassen, W., and Curtis, C. F. (2005). "History of the sterile insect technique," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Dordrecht: Springer Netherlands).

Krafsur, E. S. (2005). "Role of population genetics in the sterile insect technique," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Dordrecht: Springer Netherlands).

Krafsur, E. S., and Ouma, J. O. (2021). "Role of population genetics in the sterile insect technique," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Boca Raton, FL : CRC Press).

Langella, O. (1999). *Populations 1.2.30 [WWW Document]*. Available online at: https://bioinformatics.org/populations/ (accessed 30 March, 2022).

Lanzavecchia, S. B., Juri, M., Bonomi, A., Gomulski, L., Scannapieco, A. C., Segura, D. F., et al. (2014). Microsatellite markers from the "South American fruit fly" *Anastrepha fraterculus*: a valuable tool for population genetic analysis and SIT applications. *BMC Genet.* 15(Suppl. 2):S13. doi: 10.1186/1471-2156-15-S2-S13

Lugo-Hubp, J. (1990). El relieve de la República Mexicana. *Revista Mexicana Ciencias Geol.* 9, 82–111.

Mahmoud, M. F. (2010). Effect of gamma radiation on the sterility and quality of male peach fruit fly, *Bactrocera zonata* (Saunders)(Diptera: Tephritidae). *Egyptian J. Biol. Pest Control* 20, 71–77.

Malavasi, A., and Morgante, J. S. (1982). Genetic variation in natural populations of anastrepha (Diptera: Tephritidae). *Rev. Bras Genet.* 5, 263–278.

Manni, F. E., Guérard, E., and Heyer, E. (2004). Geographic patterns of (Genetic, Morphologic, Linguistic) variation: how barriers can be detected by using monmonier's algorithm. *Hum. Biol.* 76, 173–190.

Manni, M., Lima, K. M., Guglielmino, C. R., Lanzavecchia, S. B., Juri, M., Vera, T., et al. (2015). Relevant genetic differentiation among brazilian populations of Anastrepha fraterculus (*Diptera, tephritidae*). *ZooKeys* 2015, 157–173. doi: 10.3897/zookeys.540.6713

Molina-Nery, M. C., Ruiz-Montoya, L., Zepeda-Cisneros, C. S., and Liedo, P. (2014). Genetic structure of populations of *Anastrepha ludens* (Diptera: Tephritidae) in Mexico. *Fla. Entomol.* 97, 1648–1661.

Nielsen, R., and Slatkin, M. (2013). *An Introduction to Population Genetics*. Sunderland MA: Sinauer Assoc.

Norrbom, A. L., and Korytkowski, C. A. (2009). A revision of the *Anastrepha robusta* species group (Diptera: Tephritidae). *Zootaxa* 2182, 1–91. doi: 10.11646/zootaxa.2182.1.1

Norrbom, A. L., and Korytkowski, C. A. (2012). New species of *Anastrepha* (Diptera: Tephritidae), with a key for the species of the megacantha clade. *Zootaxa* 3478, 510–552. doi: 10.11646/zootaxa.3478.1.43

Norrbom, A. L., Zucchi, R. A., and Hernández-Ortiz, V. (1999). "Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based on morphology," in *Fruit Flies (Tephritidae)*, eds M. Aluja, and A. Norrbom (Boca Raton, FL: CRC Press), 317–360. doi: 10.1201/9781420074468

Orozco-Dávila, D., Hernández, R., Meza, S., and Domínguez, J. (2007). Sexual competitiveness and compatibility between mass-reared sterile flies and wild populations of *Anastrepha ludens* (Diptera: Tephritidae) from different regions in Mexico. *Fla. Entomol.* 90, 19–26.

Orozco-Dávila, D., Quintero, L., Hernández, E., Solís, E., Artiaga, T., Hernández, R., et al. (2017). Mass rearing and sterile insect releases for the control of *Anastrepha spp*. pests in mexico–a review. *Entomol. Exp. Appl.* 164, 176–187.

Panduranga, G. S., Sharma, K., Singh, B., and Sharma, R. K. (2022). Effect of gamma irradiation on quality parameters, sterility and mating competitiveness of melon fly. *Bactrocera cucurbitae* (Coquillett). *Int. J. Trop. Insect Sci.* 42, 875–883.

Parker, A. G., Vreysen, M. J. B., Bouyer, J., and Calkins, C. O. (2021). "Sterile insect quality control/assurance," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Boca Raton, FL: CRC Press), 399–440.

Parreño, M. A., Scannapieco, A. C., Remis, M. I., Juri, M., Vera, M. T., Segura, D. F., et al. (2014). Dynamics of genetic variability in *Anastrepha fraterculus* (Diptera: Tephritidae) during adaptation to laboratory rearing conditions. *BMC Genet.* 15:S14. doi: 10.1186/1471-2156-15-S2-S14

Pecina-Quintero, V., Jiménez-Becerril, M. F., Ruiz-Salazar, R., Núñez-Colín, C. A., Loera-Gallardo, J., Hernández-Delgado, S., et al. (2020). Variability and genetic structure of *Anastrepha ludens* Loew (Diptera: Tephritidae) populations from Mexico. *Int. J. Trop. Insect Sci.* 40, 657–665. doi: 10.1007/s42690-020-00117-8

Pecina-Quintero, V., López Arroyo, J. I., Loera Gallardo, J., Rull, J., Rosales Robles, E., Cortez Mondaca, E., et al. (2009). Genetic differences between *Anastrepha ludens* (Loew) populations stemming from a native and an exotic host in NE Mexico. *Agric. Téc. En México* 35, 323–331.

Pérez-Staples, D., Díaz-Fleischer, F., and Montoya, P. (2021). The sterile insect technique: success and perspectives in the neotropics. *Neotrop. Entomol.* 50, 172–185. doi: 10.1007/s13744-020-00817-3

Pérez-Staples, D., Shelly, T. E., and Yuval, B. (2013). Female mating failure and the failure of 'mating' in sterile insect programs. *Entomol. Exp. Appl.* 146, 66–78. doi: 10.1111/j.1570-7458.2012.01312.x

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Rambaut, A. (2012). *FigTree v.1.4.2*. Available online at: http://tree.bio.ed.ac.uk/software/figtree/

Reyes, F. J., Santiago, M. G., and Hernandez, M. P. (2000). "The Mexican fruit fly eradication programme," in *Proceedings of the area-wide control of fruit flies and other insect pests. International conference on area-wide control of insect pests, and the 5th international symposium on fruit flies of economic importance, 28 MayŰ5 June 1998*, ed. K. H. Tan (Pulau Pinang: Penerbit Universiti Sains Malaysia), 377–380.

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Not.* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for windows and linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Ruiz-Arce, R., Islam, M. S., Aluja, M., and McPheron, B. A. (2019). Genetic variation in *Anastrepha obliqua* (Diptera: Tephritidae) in a highly diverse tropical environment in the Mexican state of veracruz. *J. Econ. Entomol.* 112, 2952–2965. doi: 10.1093/jee/toz223

Ruiz-Arce, R., Owen, C. L., Thomas, D. B., Barr, N. B., and McPheron, B. A. (2015). Phylogeographic structure in *Anastrepha ludens* (Diptera: Tephritidae) populations inferred with mtDNA sequencing. *J. Econ. Entomol.* 108, 1324–1336. doi: 10.1093/jee/tov082

Ruiz-Montoya, L., Vallejo, R. V., Haymer, D., and Liedo, P. (2020). Genetic and ecological relationships of *Anastrepha ludens* (Diptera: Tephritidae) populations in Southern Mexico. *Insects* 11:E815. doi: 10.3390/insects11110815

Rull, J., and Barreda-Landa, A. (2007). Colonization of a hybrid strain to restore male *Anastrepha ludens* (Diptera: Tephritidae) mating competitiveness for sterile insect technique programs. *J. Econ. Entomol.* 100, 752–758. doi: 10.1603/0022-0493(2007)100[752:coahst]2.0.co;2

Rull, J., Brunel, O., and Mendez, M. E. (2005). Mass rearing history negatively affects mating success of male *Anastrepha ludens* (Diptera: Tephritidae) reared for sterile insect technique programs. *J. Econ. Entomol.* 98, 1510–1516. doi: 10.1093/jee/98.5.1510

Rzedowski, J. (1978). *Vegetación de México*. México, DF: Limusa.

SAGARPA (2022). *Secretaría de Agricultura y Desarrollo Rural, 2022*. Available online at: https://www.gob.mx/agricultura#335 (Accessed July 7, 2022).

Santos, R. P. D., Silva, J. G., and Miranda, E. A. (2020). The past and current potential distribution of the fruit fly *Anastrepha obliqua* (Diptera: Tephritidae) in South America. *Neotrop. Entomol.* 49, 284–291. doi: 10.1007/s13744-019-00741-1

Schug, M. D., Mackay, T. F., and Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nat. Genet.* 15, 99–102.

Selkoe, K. A., and Toonen, R. J. (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol. Lett.* 9, 615–629. doi: 10.1111/j.1461-0248.2006.00889.x

Shi, W., Kerdelhue, C., and Ye, H. (2005). Population genetics of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae), in Yunnan (China) based on mitochondrial dna sequences. *Environ. Entomol.* 34, 977–983. doi: 10.1603/0046-225X-34.4.977

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139, 457–462. doi: 10.1093/genetics/139.1.457

Thomas, D. B. (2003). Reproductive phenology of the Mexican fruit fly, *Anastrepha ludens* (Loew) (Diptera: Tephritidae) in the sierra madre oriental, Northern Mexico. *Neotrop. Entomol.* 32, 385–397. doi: 10.1590/S1519-566X2003000300002

Thomas, D. B. (2004). Hot peppers as a host for the Mexican fruit fly *Anastrepha ludens* (Diptera: Tephritidae). *Fla. Entomol.* 87, 603–608.

Vreysen, M. J. B. (2005). "Monitoring sterile and wild insects in area-wide integrated pest management programmes," in *Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*, eds V. A. Dyck, J. Hendrichs, and A. S. Robinson (Dordrecht: Springer Netherlands), 325–361. doi: 10.1007/1-4020-4051-2_12

Walder, J. M. M., and Calkins, C. O. (1993). Effects of gamma radiation on the sterility and behavioral quality of the Caribbean fruit fly, *Anastrepha suspensa* (Loew)(Diptera: Tephritidae). *Scientia Agrícola* 50, 157–165.

Wegier, A., Piñeyro-Nelson, A., Alarcón, J., Gálvez-Mariscal, A., Álvarez-Buylla, E. R., and Piñero, D. (2011). Recent long-distance transgene flow into wild populations conforms to historical patterns of gene flow in cotton (*Gossypium hirsutum*) at its centre of origin. *Mol. Ecol.* 20, 4182–4194. doi: 10.1111/j.1365-294X.2011.05258.x

Zygouridis, N. E., Argov, Y., Nemny-Lavy, E., Augustinos, A. A., Nestel, D., and Mathiopoulos, K. D. (2014). Genetic changes during laboratory domestication of an olive fly SIT strain. *J. Appl. Entomol.* 138, 423–432. doi: 10.1111/jen.12042

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 25 August 2022
doi: 10.3389/fgene.2022.868540

The *De Novo* Genome Assembly of *Olea europaea* subsp. *cuspidate*, a Widely Distributed Olive Close Relative

Tao Wu[†], *Ting Ma*[†], *Tian Xu, Li Pan, Yanli Zhang, Yongjie Li* and Delu Ning**

Institute of Economic Forest, Yunnan Academy of Forestry and Grassland, Kunming, China

The olive complex, comprising six subspecies, is a valuable plant for global trade, human health, and food safety. However, only one subspecies (*Olea europaea* subsp. *europaea*, OE) and its wild relative (*Olea europaea* subsp. *europaea* var. *sylvestris*, OS) have genomic references, hindering our understanding of the evolution of this species. Using a hybrid approach by incorporating Illumina, MGI, Nanopore, and Hi-C technologies, we obtained a 1.20-Gb genome assembly for the olive subspecies, *Olea europaea* subsp. *cuspidate* (OC), with contig and scaffold N50 values of 5.33 and 50.46 Mb, respectively. A total of 43,511 protein-coding genes were predicted from the genome. Interestingly, we observed a large region (37.5 Mb) of "gene-desert" also called "LTR-hotspot" on chromosome 17. The gene origination analyses revealed a substantial outburst (19.5%) of gene transposition events in the common ancestor of olive subspecies, suggesting the importance of olive speciation in shaping the new gene evolution of OC subspecies. The divergence time between OC and the last common ancestor of OE and OS was estimated to be 4.39 Mya (95% CI: 2.58–6.23 Mya). The pathways of positively selected genes of OC are related to the metabolism of cofactors and vitamins, indicating the potential medical and economic values of OC for further research and utilization. In summary, we constructed the *de novo* genome assembly and protein-coding gene pool for *Olea europaea* subsp. *cuspidate* (OC) in this study, which may facilitate breeding applications of improved olive varieties from this widely distributed olive close relative.

Keywords: *Olea europaea* subsp. *cuspidate*, olive subspecies, wild olive, genome, Hi-C

OPEN ACCESS

Edited by:
Zefeng Yang,
Yangzhou University, China

Reviewed by:
Guodong Rao,
Chinese Academy of Forestry, China
Zhiqiang Wu,
Agricultural Genomics Institute at Shenzhen (CAAS), China

****Correspondence:***
Yongjie Li
liyongjie107@126.com
Delu Ning
ningdelu@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *02 February 2022*
Accepted: *09 May 2022*
Published: *25 August 2022*

Citation:
Wu T, Ma T, Xu T, Pan L, Zhang Y, Li Y and Ning D (2022) The De Novo Genome Assembly of Olea europaea subsp. cuspidate, a Widely Distributed Olive Close Relative. Front. Genet. 13:868540. doi: 10.3389/fgene.2022.868540

INTRODUCTION

As "the queen of vegetable oils" and "a symbol of peace," *Olea europaea* subsp. *europaea* (OE) is one of the most widespread and socioeconomically important oil crops in the Mediterranean Basin. It is well-acknowledged that olive domestication is one of the most important events in human agricultural civilization. This event was initiated in the Near East around 4,000–6,000 years ago, and now olive trees have been planted in more than 40 countries due to its distinguished nutritional value (Kostelenos et al., 2017). Apart from its agricultural and economic importance, olive oil also has great medical importance due to its high-value health compounds, including monounsaturated free fatty acids, squalene, phytosterols, and phenols, which may exert favorable effects on inflammation, free radicals, gut microbiota, and carcinogenesis (Borzì et al., 2019).

Although genome-wide features of this species have been investigated extensively (Cruz et al., 2016; Rao et al., 2021), *O. europaea* is not a singular and isolated species. OE is one member of the olive compound species, a well-known species complex with a total of six subspecies members. This evolutionary complexity renders the subspecies nearly comparable in scientific relevance due to their close and complicated relationship. The six natural subspecies distribute over a wide range of the Old World and comprise OE (the cultivated olive), which was genetically domesticated from "wild olive" (*Olea europaea* subsp. *europaea* var. *sylvestris*, OS), and the other five wild relatives (subsp. *cuspidate*, OC; subsp. *laperrinei*; subsp. *maroccana*; subsp. *cerasiformis,* and subsp. *guanchica*) (Green, 2002; Sebastiani and Busconi, 2017). The five wild relatives of OE thrive in Africa, Asia, Europe, and the islands of the Indian Ocean (Mauritius and Madagascar) (Besnard et al., 2013).

Wild olive relatives also have multiple economically important and promising properties, including resistant and strong growth characteristics. Their intersubspecific hybrid with cultivated olive can amplify the genetic basis of the existing olive germplasm resources. OC has many common names, such as African olive, Indian olive, brown olive, and wild olive, probably due to its wide distribution in China, Iran, India, and at higher elevations in North, East, and South Africa (Green, 2002) and its close relationship with OE. Natural hybridization does exist between OE/OS and OC (Hannachi et al., 2008). Experimental hybrids between a domesticated olive variety and a wild relative of the same genus or subspecies were also reported in several studies (Besnard et al., 2001; Ma et al., 2014; Cáceres et al., 2015; Niu et al., 2020; Li et al., 2021). OC is frequently used as a graft rootstock for olive to provide vigor and possible resistance against olive fungal diseases. Grafting experiments in China showed that the survival rate of grafted seedlings was high, but the grafted plants were prone to "little feet," an appearance of a big top and a small bottom at the association interface because of the slow growth of OC (rootstock) and the rapid growth of OE (scion) (Shi et al., 1991).

In recent years, the Yunnan Academy of Forestry and Grassland has successfully bred a few olive varieties by crossbreeding of *O. europaea* subsp. *europaea* cv. Frantoio as the female parent and OC as the male parent. One of these hybrid varieties, named Yunza 3 or Jinyefoxilan, characterized by a lepidote trichome under the leaf blade, was registered as a new variety of horticultural plants in Yunnan Province, China (Ma et al., 2014). The fruit of Yunza 3 is oval, the average weight of a single fruit is 1.50 g, the pulp rate is 68.90%, and the oil content of the whole fresh fruit is 16.00%. Due to its strong adaptability and high vigor in southwest China (Yunnan province), this hybrid variety is extensively used as rootstock. The survival rate of grafted olive is high, without the "little feet" phenomenon (Ma et al., 2014). This breeding achievement strongly demonstrates the great potential of OC to improve the agroeconomic traits of olive.

To cultivate and improve new olive varieties based on intersubspecific crossing, it is necessary to further understand the genomic information of more wild subspecies. By November 2021, the whole genomes of three olive varieties from OE [cv. Leccino (Barghini et al., 2014), cv. Farga (Cruz et al., 2016), cv. Arbequina (Rao et al., 2021)], and a wild olive tree from OS (called oleaster, *Olea europaea* subsp. *europaea* var. *sylvestris*) (Unver et al., 2017) have been sequenced. The genome information of OC is not available, except for chloroplast genome data (Besnard et al., 2011). Despite the agricultural importance, there is still no high-quality genome reference for OC (subsp. *cuspidata*). There is no doubt that the reference genome has fundamental importance in aiding the target-gene sequencing and short-read mapping and in molecular breeding, population diversity, and genotype–phenotype association study. The lack of this basic data strongly hinders our understanding of genomic evolution, diversity, oil biosynthesis, and local adaptation of this important plant complex. Here, we studied the genome of subsp. *cuspidata* by incorporating Illumina, MGI, Nanopore, and Hi-C technology, which would provide insights on the adaptive evolution, molecular breeding, genomic novelty, and phylogenetic relationship of the olive complex.

MATERIALS AND METHODS

Sampling, Sequencing, *De Novo* Assembling, and Annotation

The taxonomy of the investigated OC sample was identified by Dr. Yong-Kang Sima, a professional taxonomist from the Yunnan Academy of Forestry and Grassland. This sample is now deposited in Kunming Arboretum, Yunnan province of China (voucher specimen Wu20056, N 25°9′13″, E 102°45′9″). The standard preparation procedures before sequencing, including DNA and RNA extraction and Hi-C library construction, were based on the requirements of specific sequencers. In total, five tissues, namely, leaves, roots, twigs, bark, and fruits were used for RNA-seq in Illumina platform. For DNA-seq, 65.68 Gb short-reads (300 bp PE) and 96.5 Gb Nanopore long-reads were obtained from the DNBSEQ-T7 and PromethION platform, respectively. The raw reads were filtered using the fastp preprocessor (Chen S. et al., 2018). To achieve chromosome-level assembly, we further generated 129.21 Gb data of the paired-end Hi-C reads (150 bp) from the DNBSEQ-T7 platform (MGI). We conducted the karyotyping of OC to determine the number of chromosomes using rooted cuttings, which have active meristems of mitosis suitable for detecting clear chromosomes. The root tips were treated with nitrous oxide to obtain sufficient cells at mitosis metaphase for staining with DAPI and telomere repetitive sequences (TTTAGGG) 6.

A genome survey was conducted using GenomeScope (Vurture et al., 2017) for heterozygosity and repeat content. The genome size was estimated with the mean values of gce 1.0.2 with *k-mer* 17, 19, and 21 (Liu et al., 2013). The basecalling output from the PromethION platform was treated using Guppy (Wick et al., 2019). Only the reads with mean quality scores >7 were retained and further corrected using NextDenovo software with parameters "reads_cutoff:2k, seed_cutoff:18k" (https://github.com/Nextomics/NextDenovo) (Wang et al., 2019). The assembling processes include the correction module using

NextCorrect and the assemble module using NextGraph, with default parameters. Subsequently, Nextpolish software was used to polish the genome with short-reads four times and long-reads three times (sgs_options = -max_depth 100) (Hu et al., 2020). The paired-end Hi-C reads were filtered by fastp to remove the adapter and low-quality reads (Phred Score >15, and 5 > number of Ns in the reads) (Chen S. et al., 2018). The obtained assembly was further corrected with 3d-DNA five times and manually tuned with Juicebox Assembly Tools v1.9.8 (Dudchenko et al., 2018). During scaffolding, to facilitate contig ordering and revise the misjoin, we mapped the OC draft genome to OE assembly using minimap2 with parameter "-xasm10" (Li, 2018) and visualized the major structure variations with dotPlotly (https://github.com/tpoorten/dotPlotly). Subsequently, the pseudo-structural variations caused by misassembling were manually corrected by examining the HI-C matrix with Juicebox following the official manual (https://aidenlab.org/assembly/manual_180322.pdf). The genome assessments were conducted by using LTR_retriever (Ou et al., 2018), mapping rate of short-read data by BWA (Li and Durbin, 2009), and N50 values with QUAST (Gurevich et al., 2013), with default parameters.

The RepeatMasker v2.0.3 was used for repeat annotation following the manual-recommended parameters (Tarailo-Graovac and Chen, 2009). To aid gene annotation, a total of ~25 Gb RNA-sequencing (RNA-seq) clean pair-ended reads from five tissues, namely, leaves, roots, twigs, barks, and fruits were generated using Illumina HiSeq platform. All libraries were *de novo* assembled separately and subsequently merged using the TransABySS v2.0.1 manual pipeline (Robertson et al., 2010). The protein-coding and non-coding gene structural annotation was conducted using the MAKER2 pipeline (Cantarel et al., 2008) by incorporating transcriptome mapping, *de novo* gene predictions, and homology predictions with OS proteins from the NCBI (GCF_002742605.1). The majorly used softwares from MAKER2 pipeline include blast + tools (Camacho et al., 2009), exonerate v2.2.0 (Keller et al., 2011), hmm-E and GeneMark-ES (Borodovsky and Lomsadze, 2011), and augustus (Stanke and Morgenstern, 2005).

The high-throughput sequencing data files are available at the GenBank database (https://www.ncbi.nlm.nih.gov/), with SRA accession numbers: SRR17299471 and SRR17299472. The associated BioProject and BioSample numbers are PRJNA785068 and SAMN23526758, respectively. The genome assembly of OC is available under NCBI accession number JAKWBP000000000.

Gene Family and Species Evolution

For species evolution, we organized "dataset A″ to address the questions related to phylogeny and divergence time. The dataset A covers three subspecies of olive (OC, OS, and OE), in addition to the other five species of eudicots without gene annotations. Five species, namely, *Jasminum sambac, Forsythia suspensa*, *Fraxinus pennsylvanica*, *Fraxinus excelsior*, and *Osmanthus fragrans* with reference genomes but without gene annotations were retrieved from the NCBI (**Supplementary Table S1A**). *Arabidopsis thaliana* was further added as an outgroup species. To facilitate species phylogeny analysis, we used a "proxy" approach based on dataset B, which involves 10 species with available gene annotations from the NCBI (**Supplementary Table S1B**). In detail, these species/subspecies include *Arabidopsis thaliana*, *Arachis hypogaea*, *Elaeis guineensis*, *Glycine max*, *Helianthus annuus*, *Juglans sigillata*, *Ricinus communis*, *Sesamum indicum*, *Olea europaea* subsp. *europaea* var. *sylvestris*, and *Olea europaea* subsp. *europae* cv "Arbequina" (**Supplementary Table S1B**).

In simple terms, the strategy of the "proxy" approach is that the "one-to-one" single-copy orthologous genes were identified from dataset B and then mapped to dataset A to re-analyze orthologous gene groups. In detail, based on the "one-to-one" orthologous genes obtained from dataset B with OrthoFinder v2.5.4 (Li et al., 2003), we locally annotated the corresponding homologous genes for dataset A using BRAKER2 with only homology prediction (Brůna et al., 2021). Then, these homologous genes were fed into OrthoFinder v2.5.4 again to obtain "one-to-one" orthologous single-copy genes for dataset A (**Supplementary Table S2**).

In detail, the orthologous genes, phylogeny, and divergence time were analyzed as follows. The OrthoFinder v2.5.4 with default parameters was used for gene family and orthologous gene identification (Li et al., 2003). Only the longest transcript was used for protein sequence comparison with BLAST tools (Altschul et al., 1997). We estimated the evolutionary topology with FastTree-2 (Price et al., 2010), an approximately maximum-likelihood (ML) method, using the combined sequences of "one-to-one" single-copy gene families, with bootstrap replicates set to 1,000. MCMCTREE in PAML v4.8a was used to estimate the divergence time of these species (Yang, 2007). The divergence calibration was based on the divergence time between "*Osmanthus fragrans*" and "*Olea europaea*" (7-45Mya) from the time-tree database (http://www.timetree.org). The sequence alignment and filtering were based on MAFFT v7.49 (Katoh and Standley, 2013) and Glocks (parameter: b5 = h) (Castresana, 2000).

For gene family evolution, we only analyzed dataset B (**Supplementary Table S3**), which has available gene annotations from the NCBI. The CAFE v4.2.1 (Computational Analysis of gene Family Evolution) package (De Bie et al., 2006) was used to analyze gene family expansion and contraction with a significant level of p-value < 0.01 across ancestral nodes, leading to olive species.

Fast Evolution and Positive Selection Analysis

Identifying genes under positive selection is a common way to detect genes with novel functions and molecular adaptation, which has been successfully applied in both plants and animals (Yang and Dos Reis, 2010; Zhang et al., 2011; Chen J. et al., 2018). In this study, the branch model and branch-site model in PAML packages (v4.8a) were used to detect fast evolution and positively selected genes, based on dataset B with available gene annotations. The branch model was analyzed by comparing the "free-ratio model" with the "one-ratio model" and choosing only the significant genes and those evolving fastest in OC. The subsequent genes were identified by

comparing Model A (assuming the focal branch under positive selection indicated by Ka/Ks > 1) with the null model (Ka/Ks of any site was ≤1). The statistical significance of the likelihood ratio test (LRT) was determined with "chi2" function in PAML. The positive selected sites were further determined using the Bayesian method (BEB, Bayes empirical method) with a probability value of over 0.95.

Whole-Genome Duplication and Transposed Gene Duplications

Whole-genome duplication (WGD) analysis was conducted by the 4DTv method (four-fold synonymous third-codon transversion) and Ka/Ks estimation in MCScanX with default parameters (Wang et al., 2012). The gene duplication event dating was determined using MCScanX-transposed (Wang et al., 2013). Specifically, the gene duplication types were categorized into tandem duplication, proximal duplication, segmental duplication, and transposed gene duplication. The oldest branch of the synteny block was used as a proxy for the gene ages of transposed genes. The retrogenes, or the RNA-based gene duplications, were identified using a method similar to that used by Betrán et al. (Betrán et al., 2002), with the BLASTP parameters including identity value >60%, length mapping coverage >80%, and an E-value < 0.000001.

Structural Variation Identification

We first tried the SyRI and the "assembly-to-assembly" approach for SV identification (Chen et al.; Goel et al., 2019). However, these approaches are better for references at the population level or with higher DNA identity. We further conducted the SV identification based on comparing OC long reads to OS and OE references with a dual-mode alignment strategy. In detail, the reads were mapped to a reference with two commonly used mappers, Minimap2 and NGMLR, which are integrated in a software named Vulcan (Fu et al., 2021). Minimap2 is a highly fast long-read mapper, implementing a time-efficient alignment approach involving a two-piece affine gap model and a faster chaining process (Li, 2018). NGMLR is designed to make use of a convex scoring matrix to better distinguish the read error from the SV signal (Sedlazeck et al., 2018). For SV calling, we utilized Sniffles (version 2.0.3) and filtered out imprecise and low reads supporting SVs (<3) (Sedlazeck et al., 2018).

RESULTS AND DISCUSSION

Genome Assembly of *Olea europaea* subsp. *Cuspidata*

Before performing *de novo* genome assembly, we estimated the genomic featuring parameters including genome size, heterozygosity, and repeat content to roughly assess the complexity of the *O. europaea* subsp. *cuspidata* genome with *k-mer* analysis (Chor et al., 2009; Liu et al., 2013; Vurture et al., 2017), which is the most frequently used method for genomic survey. Compared to the previously reported OE genome (Rao et al., 2021), OC has a higher level of heterozygosity (2.28% vs. 1.09%), a comparable level of repeat content (54.5% vs. 56.18%), and a slightly smaller genome size (1.2G vs. 1.3G).

In total, 65.68-Gb MGI DNA-seq short reads (300 bp PE, 54.7×), 129.21-Gb Hi-C paired-end reads (107.7×), and 96.5-Gb Nanopore long-reads (80.4×) were obtained following data filtering. The draft contigs were constructed with short reads and Nanopore long reads, followed by semi-automatic scaffolding with 3D-DNA (Dudchenko et al., 2018). After manually revising the orders and orientations of super-scaffolds with Hi-C interaction signals, we achieved an anchor rate of 87.95% to place the initial contigs to scaffolds. We observed a clear aggregation of 23 super-scaffolds, which are also OC chromosomes, with the lengths from 28.38 to 87.93 Mb (**Supplementary Table S4**). All other scaffolds or contigs are shorter than 0.8 Mb and have no clear signals of interaction with any chromosome (**Figure 1A**). We further validated the total chromosome number of 23 in OC by using the karyotyping of DAPI staining (**Supplementary Figure S1**) and telomere staining with repetitive sequences (TTTAGGG) 6 (**Figure 1B**).

The genome size of the final OC reference was estimated to be 1.20 Gb. The longest scaffold and contig are 87.93 and 17.29 Mb, respectively. The lengths of the contig and scaffold at 50% of total genome length (N50) are 5.33 and 50.46 Mb, respectively (**Table 1**), which are greater than those of the previously published wild olive tree oleaster (contig N50, 25.49 Kb; scaffold N50, 228.62 Kb) (Unver et al., 2017) and the cultivated *Olea europaea* genome (contig N50, 4.67 Mb; scaffold N50 42.60 Mb) (Rao et al., 2021). We mapped the original clean short and long reads to the novel *de novo* OC genome assembly which was used as a reference. The mapping rate of MGI short-read data against the OC reference is 99.91%, which is almost the same as that of Nanopore long reads (99.38%). In addition, the LAI score (12.95) indicates the high quality of the OC genome that has reached the reference level, based on previous assessment of multiple species (Ou et al., 2018).

Annotation of the *O. europaea* subsp. *Cuspidata De Novo* Genome Assembly

To evaluate the continuity of both assembly and protein-coding genes, we conducted BUSCO analysis to assess the completeness and redundancy of the OC assembly and proteins based on the fractions of conserved genes (Manni et al., 2021). BUSCO assessment revealed that 94.1% of 1,440 plant conserved genes are complete in OC assembly, similar to the level of the OE assembly reported previously (Rao et al., 2021) and much higher than that of OS. Similar patterns were found for assembly and protein completeness (**Table 2**), suggesting high level of integrity and completeness of the OC genome. The repeat annotation based on RepeatMasker revealed that repeats, including DNA elements, LINE, SINE, LTR, satellite, simple repeats, and unknown elements, account for 74.22% of genome sequences (**Table 3**), which is higher than the estimation based on the *k-mer* survey. The top three abundant repeat elements are LTR, DNA elements, and LINE, accounting for 62.76%, 11.03%, and 2.48%, respectively (**Table 3**).

To understand inter- and intra-assembly synteny, we conducted a whole-genome alignment between OC and OE

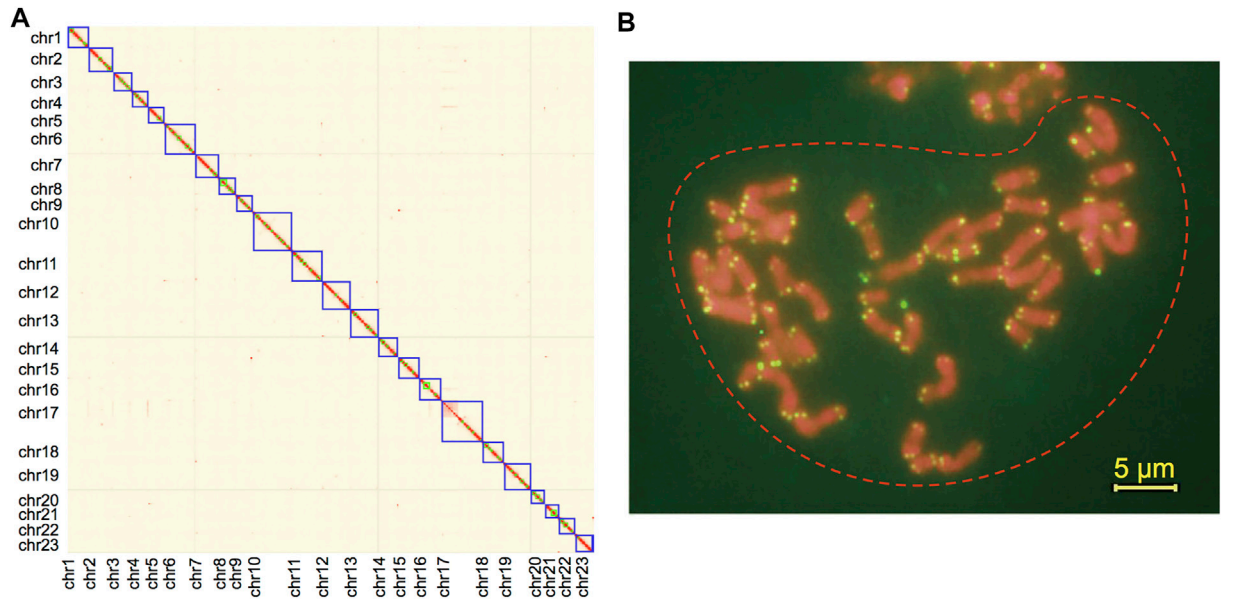


FIGURE 1 | (A) Intensity signal heatmap of the Hi-C chromosome for *Olea europaea* subsp. *cuspidata* (OC). **(B)** Karyotype by telomere staining with repetitive sequences (TTTAGGG)6. Note: the yellow bar indicates 5 μm.

TABLE 1 | Summary of the *de novo* genome assembly of OC and the comparison with two related species, *Olea europaea* subsp. *europaea* var. *sylvestris* (OS) and *Olea europaea* subsp. *europae* cv "Arbequina" (OE).

Assembly	OC	OS	OE
No. of sequences (≥ 50,000 bp)	187	2,104	849
No. Total length (≥ 50,000 bp)	1,183,913,677	985,700,118	1,098,745,707
No. of sequences	1,078	41,219	962
Largest sequence (bp)	87,931,667	46,026,434	68,066,766
Total length (bp)	1,196,933,720	1,141,142,775	1,102,969,454
GC (%)	35.36	35.4	34.33
N50 (bp)	50,460,234	12,567,911	42,601,851
N75 (bp)	41,133,639	174,775	35,395,138

TABLE 2 | BUSCO assessment of genome and gene continuity.

	Assembly proteins	Percentage (%)	Annotation proteins	Percentage (%)
Complete BUSCOs	1,356	94.1	1,393	96.7
Complete Single-Copy BUSCOs	1,036	71.9	997	69.2
Complete Duplicated BUSCOs	320	22.2	396	27.5
Fragmented BUSCOs	20	1.4	25	1.7
Missing BUSCOs	64	4.5	22	1.6
Total BUSCO groups searched	1,440	100	1,440	100

and between OC and OS based on Minimap2 (Li, 2018) (**Figure 2A**). We also conducted self-alignment using MCScanX software with collinear genes of OC (Wang et al., 2012) (**Figure 2B**). The cross-assembly comparison revealed that OC has the highest alignment identity rates to OS rather than OE, suggesting closer distance from OC to wild olive (OS) than from OC to domestic olive (OE) (**Figure 2A**). Unexpectedly, based on these two alignments, we found a "gene-desert" region on chromosome 17 of OC (0–37.5 Mb, **Figure 2C**). Only four genes, including phytochrome B–like gene, transposable element gene, arginine methyltransferase–interacting related gene, and zinc finger BED domain–containing related gene, are found within this region. BLASTP search against the database of RefSeq non-redundant proteins revealed that these genes are genetically nearest to OS, consistent with the overall pattern of the other chromosomes. Among the four genes, the

TABLE 3 | Annotation summary statistics for repeats of the OC reference genome.

Type	Repbase TEs (%)	TE proteins (%)	De novo (%)	Combined TEs (bp)	Combined TEs (%)
DNA elements	1.73	0.74	9.76	157,780,557	11.03
LINE	0.37	0.24	2.19	35,494,175	2.48
SINE	0	0	0.06	922,395	0.06
LTR	16.29	12.32	61.52	897,920,339	62.76
Satellite	0.16	0	0.29	6,373,488	0.45
Simple repeat	0	0	0.02	348,505	0.02
Unknown	0.01	0	3.77	32,645	3.79
Total	18.37	13.3	72.14	54,150,429	74.22

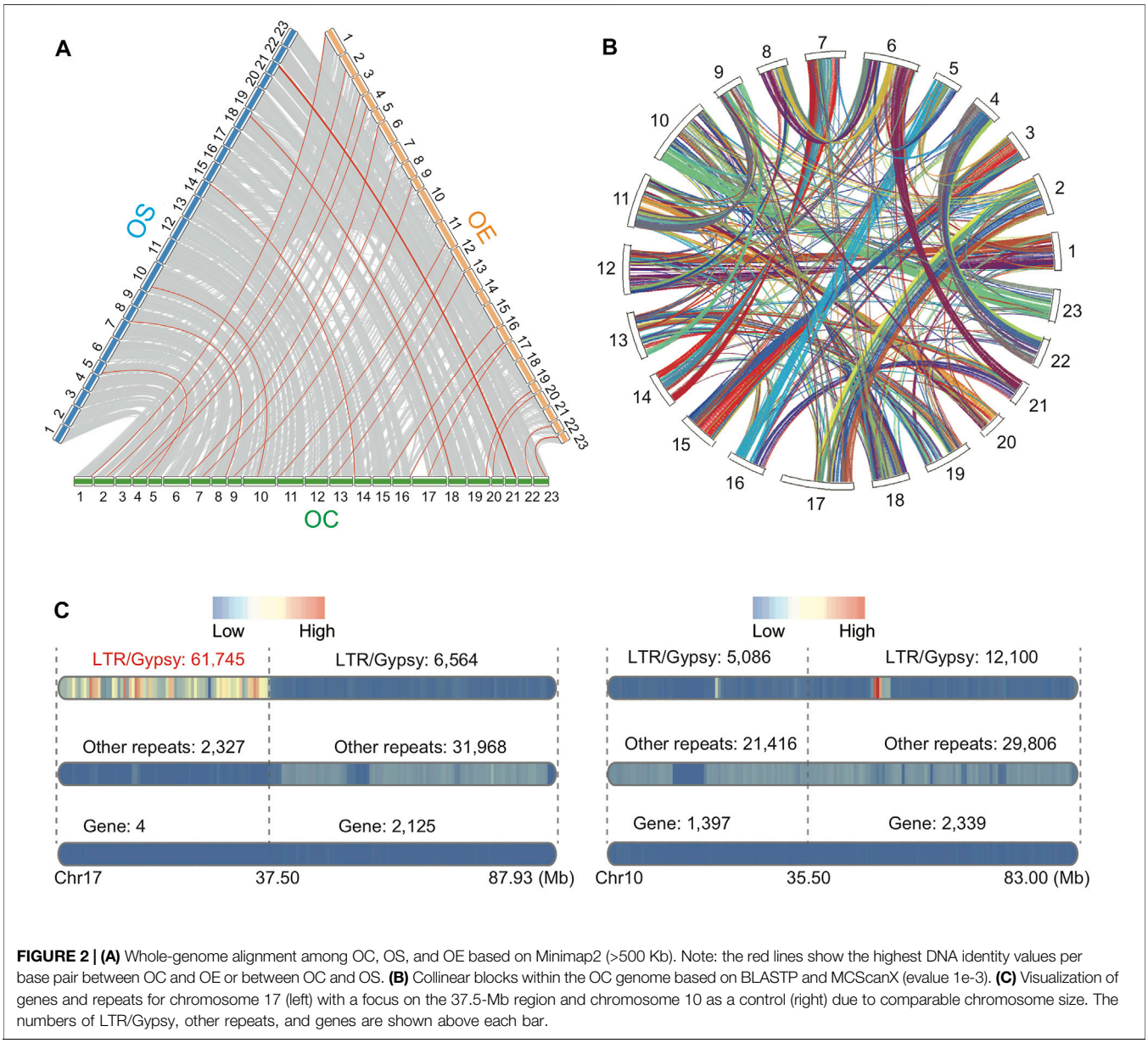


FIGURE 2 | (A) Whole-genome alignment among OC, OS, and OE based on Minimap2 (>500 Kb). Note: the red lines show the highest DNA identity values per base pair between OC and OE or between OC and OS. **(B)** Collinear blocks within the OC genome based on BLASTP and MCScanX (evalue 1e-3). **(C)** Visualization of genes and repeats for chromosome 17 (left) with a focus on the 37.5-Mb region and chromosome 10 as a control (right) due to comparable chromosome size. The numbers of LTR/Gypsy, other repeats, and genes are shown above each bar.

phytochrome B–like gene and the arginine methyltransferase–interacting related gene are particularly interesting due to their known roles in light-controlled chromatin compaction and methylation regulation (Tessadori et al., 2009; Cho et al., 2012; Zhang et al., 2019). In addition, we uncovered that this "gene-desert" region is also the "LTR-

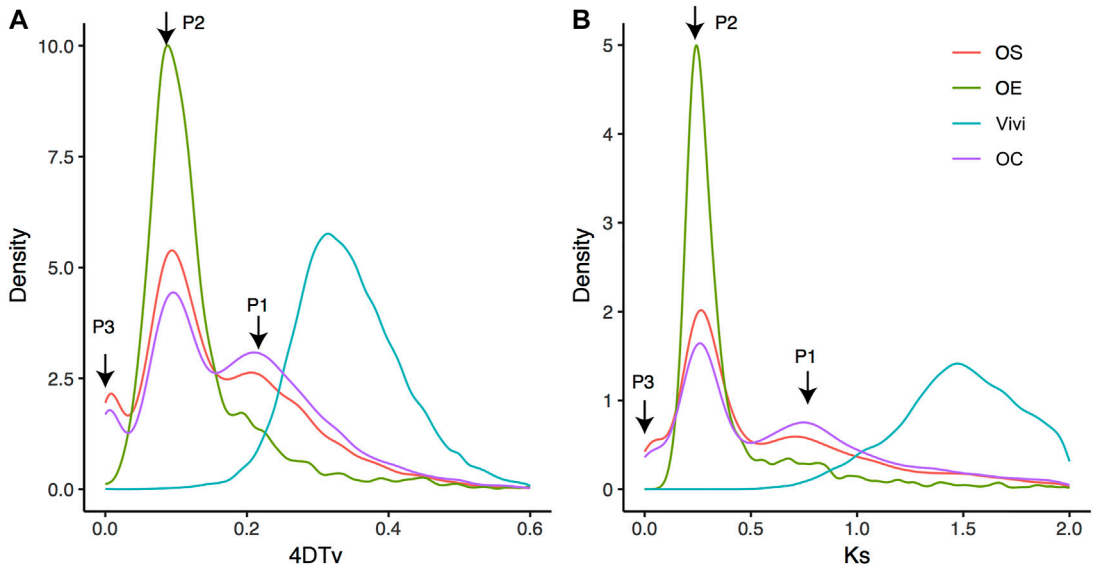


FIGURE 3 | Transversion of four-fold degenerate site (i.e., 4DTv) distribution **(A)** and Ks distribution **(B)**. Abscissa and ordinate represent 4DTv or Ks distance and the percentage of homologous gene pairs, respectively. Note: Vivi represented *Vitis vinifera*, which was used to show a more ancient WGD event. P1, P2, and P3 are three recent peaks. Species evolution in phylogeny and divergence time.

hotspot", with the highest density of retrotransposon LTR/Gypsy (61,745/63,462, 97.29%) among all types of repeats (**Figure 2C**). Interestingly, within chromosome 17, 90.39% of LTR/Gypsy repeats reside in the 37.5-Mb region (61,745 out of 68,309), suggesting significant local enrichment (χ^2 test, $p < 0.00001$). This region covers 2.98 to 16.40 times higher number of LTR/Gypsy (61,745) than other complete chromosomes, which range from 3,766 in chromosome 21 to 20,731 in chromosome 6. This finding may pave the way for future study on the olive region of "gene-desert" but "LTR-hotspot".

Whole-Genome Duplication

It has long been known that whole-genome duplication is one of the most important evolutionary forces driving phenotypic diversity during plant speciation. Previous reports have revealed that the OS genome contains WGD events that are specific to Oleaceae (Unver et al., 2017). Here, we identified collinear blocks at the intraspecies level for three annotated genomes (OC, OS, and OE). Then, based on paralogous genes within these collinear blocks, we analyzed the whole-genome evolution events using 4DTv (transversion of four-fold degenerate site) and Ks (synonymous substitution rate) values (**Figure 3**). Both 4DTv and Ks demonstrated two major peaks (P1 and P2) for OC, OS, and OE, supporting their status as a species complex. In addition, OE and OC have a third minor peak (P3). No observable P3 peak in OS is possible due to synteny loss caused by the fragmented nature of the OS current reference (scaffold N50 is only 228.62 Kb). Most likely, the peaks indicate three rounds of WGD events at the same time in the genome evolution of Oleaceae species.

To examine the possibility of whole-genome triplication (WGT) underlying the three peaks, we analyzed the depth of collinear genes within the three peaks. The depths were determined with the "dissect multiple alignment" function of MCScanX based on collinear blocks of OC self-alignment (Wang et al., 2012). If WGT causes the three peaks, most of the genes of the peaks would have a collinear depth of 2, corresponding to a total of three collinear blocks. Interestingly, different from the expectation of WGT, depth 1 (1107 in P1, 2165 in P2, and 40 in P3) is higher than depth 2 (987 in P1, 888 in P2, and 15 in P3) for all the three peaks, suggesting that three rounds of WGD may have a significant role in shaping OC genome evolution. We also uncovered a dominant proportion of OC genes (73.76%, 24015 genes), retained due to the WGD events or segmental duplications, than other types of duplicates (5331 transposed duplications, 1673 tandem duplications, and 1535 proximal duplications). This composition of paralogs is similar to the pattern previously reported in *Glycine max*, which was also attributable to the WGD event (Wang et al., 2012). Absolute time inference revealed that P1, P2, and P3 occur at 69.38–81.88 Mya, 34.69–40.94 Mya, and 4.34–5.12 Mya, respectively.

To understand the phylogeny of Oleaceae (OC, OS, and OE) in eudicots, we organized a dataset A covering other five related species, namely, *Jasminum sambac*, *Forsythia suspensa*, *Fraxinus pennsylvanica*, *Fraxinus excelsior*, and *Osmanthus fragrans*, with *Arabidopsis thaliana* as an outgroup species (**Supplementary Table S1B**). To address the issue of unavailability of public gene annotations for these species, we used a "proxy" method. We identified 1,463 single-copy orthologous groups based on dataset B of 11 species/subspecies with their publicly available annotations (**Supplementary Tables 1B, 3**). Then, orthologous genes from dataset B were mapped to dataset A and inferred orthologous genes for eudicot species with OrthoFinder (Emms and Kelly, 2019).

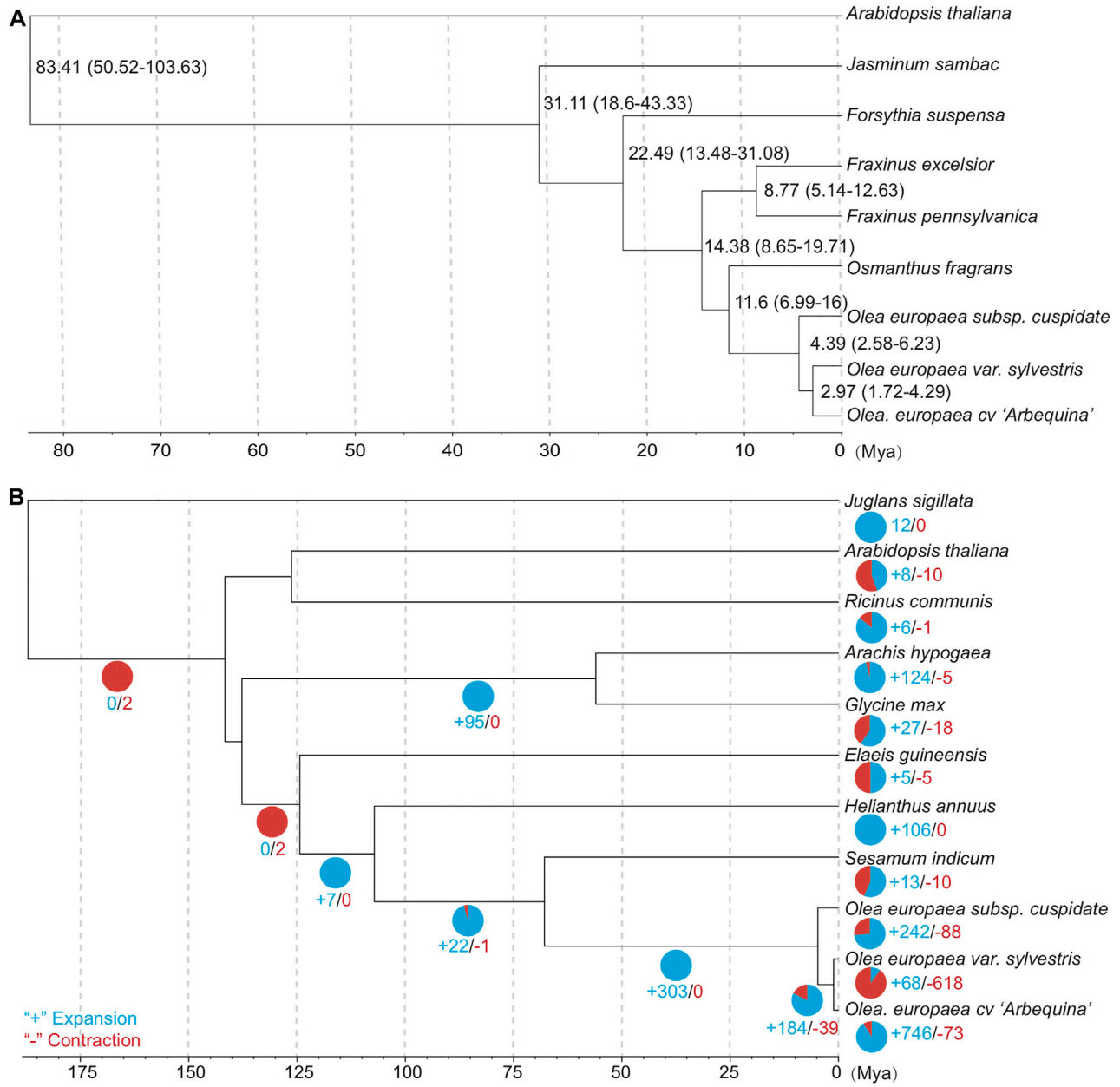


FIGURE 4 | (A) Time-tree of OC and its related subspecies/species (dataset A). The median and 95% confidential interval (95% CI) are shown near the nodes. **(B)** Contraction and expansion statistics of gene families based on dataset B due to their available gene annotations.

Finally, we obtained 1,247 groups of "one-to-one" single-copy orthologous genes to estimate the topology and divergence time of eudicots based on dataset A (**Supplementary Table S2**). The phylogeny and divergence time were estimated using the approximately maximum-likelihood method for each single-copy gene group (Whelan and Goldman, 2001; Yang, 2007). The closest relationship was found among the olive subspecies, consistent with our expectation about the recent evolution of the olive complex (**Figure 4A**). The divergence time between OC and the last common ancestor of OE and OS was estimated to be 4.39 Mya (95% CI: 2.58–6.23 Mya). Interestingly, this time range is roughly the same with P3 peak at 4.34–5.12 Mya, suggesting the contribution of the most recent WGD event on the divergence of olive subspecies.

Gene Family Evolution in Terms of Expansion and Contraction

For gene family evolution, we analyzed the expansion and contraction patterns based on 11 species/subspecies of dataset B due to their available gene annotations. The ultrametric tree was estimated with r8s to transform the species tree into a time

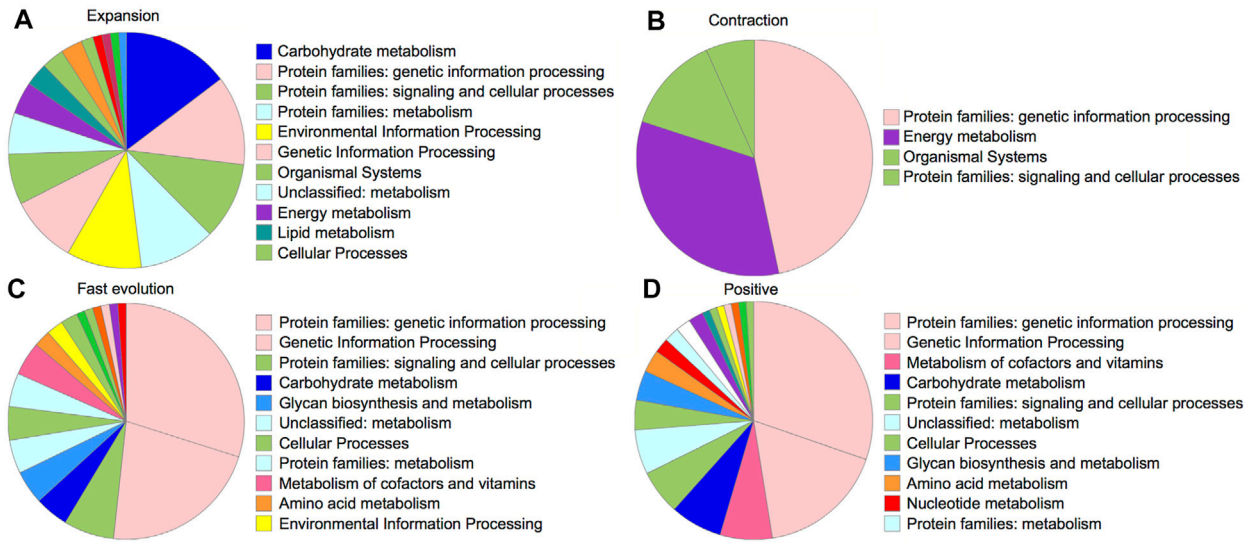


FIGURE 5 | KEGG pathways for 242 gene families under expansion **(A)**, 88 gene families under contraction **(B)**, 192 genes with signals of faster evolution in OC **(C)**, and 73 positively selected genes **(D)**.

tree (Sanderson, 2003). We identified 242 gene families that expanded and 88 gene families that contracted during OC genome evolution after OC speciation (**Figure 4B**). For the expanded gene families, the KEGG analyses (**Figure 5A**) based on BlastKOALA (Kanehisa et al., 2016) revealed that the enriched pathways include carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, and genetic information processing. The expanded genes include alcohol dehydrogenase, isocitrate dehydrogenase (NAD+), S-(hydroxymethyl) glutathione dehydrogenase, dihydropyrimidine dehydrogenase (NADP+), polyphenol oxidase, L-ascorbate oxidase, homocysteine methyltransferase, phospholipid: diacylglycerol acyltransferase, *etc* (**Supplementary Table S5**). The contracted gene families majorly involve genetic information processing and energy metabolism (**Figure 5B**, **Supplementary Table S6**). These results indicate that some gene families related to traits with potential economic value, such as lipid metabolism, are under gene expansion rather than contraction, which may need further study and exploration.

Gene Sequence Evolution Related to Selection

To identify genes under OC-specific positive selection, we conducted branch model and branch-site model tests using CODEML in PAML software (Yang, 2007). Among 1,463 "one-to-one" orthologous genes, 40.05% of genes (586) were detected to significantly deviate from the null model of neutral evolution *via* the branch model analysis by comparing the "free-ratio model" with the "one-ratio model" ($p < 0.05$, χ^2 test). The "free-ratio model" allows the Ka/Ks ratio to be flexibly modeled, thus providing a Ka/Ks ratio for each branch to compare. By ranking Ka/Ks ratios across species, we found 13.12% of genes (192) with the highest Ka/Ks in the OC genome, suggesting their OC-specific fast evolution. 45.30% of these genes can be mapped into KEGG biological processes, including genetic information processing, glycan biosynthesis and metabolism, carbohydrate metabolism, and lipid metabolism (**Supplementary Table S7**, **Figure 5C**). The pathway analysis revealed that these genes could be categorized into 84 pathways, with metabolism and biosynthesis of secondary metabolites as the top two pathways with the most abundant genes (18 and 7 genes, respectively). Among the 192 significantly faster evolution genes ($p < 0.05$), 125 genes have Ka/Ks ratios >1, suggesting that these genes are under positive Darwinian selection.

We further conducted the branch-site model analysis by focusing only on the OC branch to identify OC-specific positively selected genes. The branch-site model detected that 7.18% of orthologous genes (105) may be under significant positive selection during OC evolution, with only 17 being shared with the branch model result, suggesting the importance of using complementary methods during the positive selection analysis. There are 73 genes, out of 105 positively selected genes identified with the branch-site model, showing at least one site with a significant positive selection signal (probability >0.95) inferred with the Bayes Empirical Bayes (BEB) analysis. KEGG analysis revealed that the pathways of these positively selected genes are related to the processes involving genetic information processing and the metabolism of cofactors and vitamins (**Figure 5D**). Interestingly, consistent with the expectation of oil-related traits in OC, some positively selected genes are related to lipid metabolism processes, including glycerophospholipid metabolism, ether lipid metabolism, and sphingo-lipid metabolism (**Supplementary Table S7**). These results indicate the potential medical and economic values of OC for further research and utilization.

TABLE 4 | Summary of gene numbers and gene family numbers.

Species	No. of families	No. of genes	No. of genes in families	Unclustered genes<	Unique family
Arabidopsis thaliana	3,906	19,614	13,532	6,082	65
Arachis hypogaea	11,047	49,359	48,914	445	858
Elaeis guineensis	4,741	21,783	16,793	4,990	91
Glycine max	9,424	41,092	38,809	2,283	212
Helianthus annuus	5,822	31,783	27,110	4,673	134
Juglans sigillata	5,612	25,769	20,310	5,459	225
Ricinus communis	3,051	18,161	9,847	8,314	44
Sesamum indicum	4,585	22,010	15,187	6,823	69
Olea europaea subsp. cuspidate	8,988	43,511	39,008	4,503	364
Olea europaea subsp. *europaea* var. *sylvestris*	7,432	37,104	31,025	6,079	80
Olea europaea subsp. *europae* cv 'Arbequina'	9,386	48,032	43,344	4,688	585

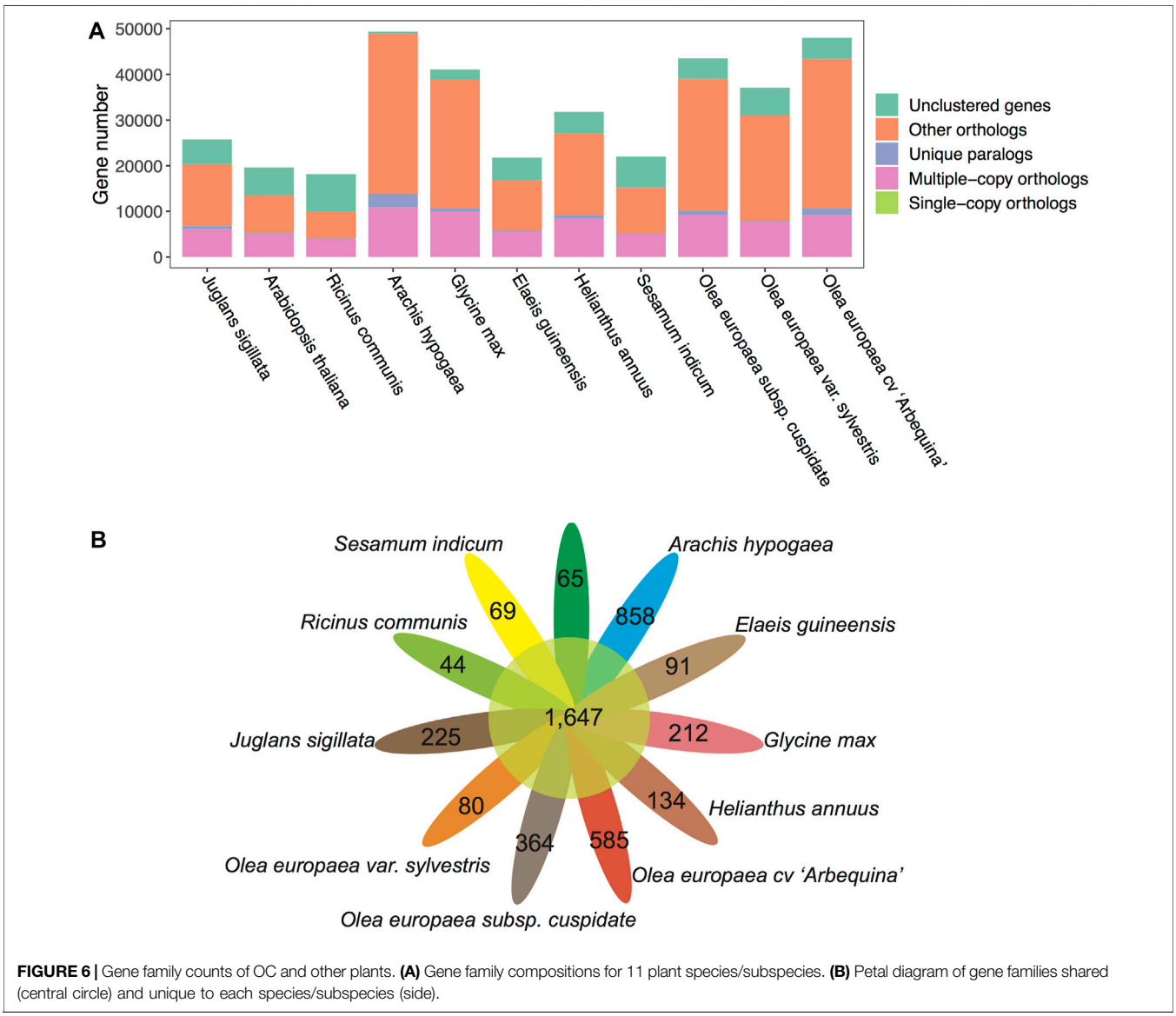


FIGURE 6 | Gene family counts of OC and other plants. **(A)** Gene family compositions for 11 plant species/subspecies. **(B)** Petal diagram of gene families shared (central circle) and unique to each species/subspecies (side).

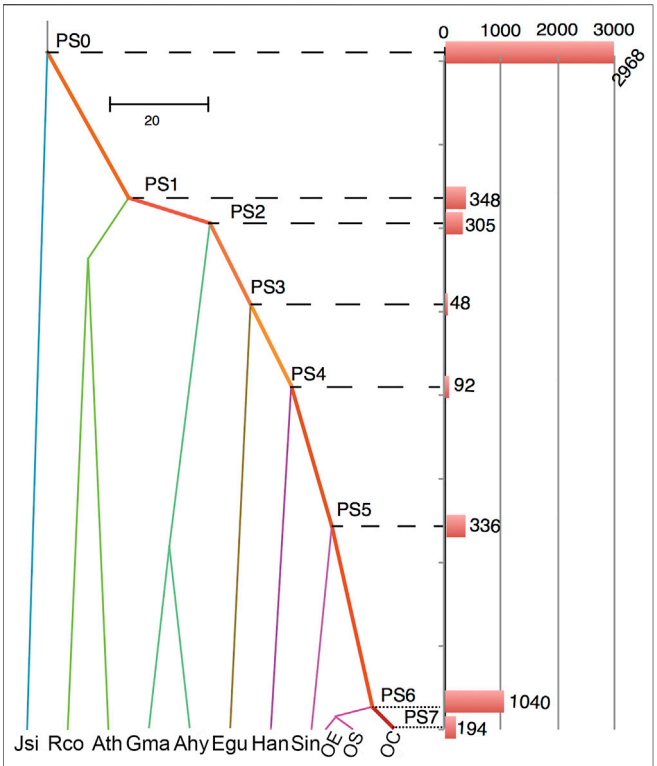


FIGURE 7 | Evolutionary dating of transposed duplicated genes. PS indicates phylogenetic stages of ancestral nodes. The bars show the number of transposed genes that originated at these stages. The abbreviations and their full names are Ahy (*Arachis hypogaea*), Ath (*Arabidopsis thaliana*), Egu (*Elaeis guineensis*), Gma (*Glycine max*), Han (*Helianthus annuus*), Jsi (*Juglans sigillata*), Rco (*Ricinus communis*), and Sin (*Sesamum indicum*).

To understand the evolution of gene families, we conducted a comparative genomics analysis by incorporating other well-annotated genomes. Based on the Markov Cluster Algorithm (MCL), a fast and scalable unsupervised cluster algorithm for graphs, we identified a total of 73,994 distinct gene families (BLASTP *E-value*≤1e-10) (**Table 4** and **Figure 6A**). Based on the constitution of shared or unique gene families, we found that OC has comparable numbers with OE in terms of gene family number, gene numbers within families, and unclustered gene numbers, strongly reflecting their much better gene annotation and assembly quality than OS. For a unique gene family in each species, OC is 0.62 times lower than OE (364 vs. 585) but 4.55 times higher than OS (364 vs. 80) (**Figure 6B**).

The Evolutionary Dating of Transposed Genes

New genes, including gene duplications, are known as one of the most important drivers of phenotypic innovations in species and populations (Chen et al., 2022a; Long and Langley, 1993; Long et al., 2013; Xia et al., 2016; Chen et al., 2019; Xia et al., 2021). To understand how new gene duplications have contributed to the evolution of OC, we categorized the genes into segmental duplication, tandem duplication, proximal duplication, and transposed duplication through synteny sharing or breaking of protein-coding genes. The transposed duplicated genes were further mapped into the phylogenetic tree that leads to our focal genome OC. Hence, we can understand how OC gradually disseminated duplicated genes into new chromosome context by DNA- or RNA-based transposition processes. RNA-based transposed genes (1111 genes), which are known as retroposed genes or retrogenes (Emerson et al., 2004; Chen et al., 2019), were found to account for 20.84% of all gene transpositions. Among eight evolutionary branches leading to OC, we found a substantial outburst (19.5%) of gene transposition events in PS6 (**Figure 7**), which is the common ancestor of olive subspecies, suggesting the importance of new gene evolution in shaping olive speciation. Interestingly, this outburst of new genes seems to occur simultaneously with the minor WGD event (P3) that happened at 4.34–5.12 Mya. A previous study in bamboos has revealed the connections between recent WGD and new gene origination in both time and function (Jin et al., 2021). Our study provides further evidence on the close relationship between transposition and WGD events, which is worthy of further investigation.

The Structural Variation Identification

Although the "assembly-to-assembly" approach has been successfully used to identify SVs in other species (Chen et al., 2022b; Goel et al., 2019), we failed to obtain results from this method, probably due to the known phenomenon of higher rearrangements in plants than in animals. We further identified structural variations (SVs) using Sniffles V2.0.3 (Sedlazeck et al., 2018) and a dual-alignment strategy implemented in Vulcan (Fu et al., 2021). Vulcan explores the advantages of two efficient mappers, Minimap2 (Li, 2018) and NGMLR (Sedlazeck et al., 2018), to improve the accuracy and efficiency of mapping. Here, after mapping OC long reads to OS and OE, we obtained four types of SVs, namely, deletions, duplications, insertions, and inversions (**Table 5**; **Supplementary Tables S8, S9**). We found that the number of three types of SVs (deletions, insertions, and inversions) between OC and OS is lower than that between OC and OE, suggesting a comparatively closer relationship between OC and OS. This finding is consistent with our synteny mapping result that the nucleotide identity is higher between OC and OS than between OC and OE (**Figure 2A**). It is well-established that SVs have higher functional impacts than SNPs (Alonge et al., 2020; Chen et al., 2022a). Thus, it is promising to identify the SVs associated with critical traits at the population level, based on larger sample size. Since reliable SV calling procedures require a high-quality genome reference, our study may pave the way for further studies

TABLE 5 | Number summary of SVs (>50bp) numbers between OC and other two references (OE and OS).

Reference	Deletions	Duplications	Insertions	Inversions
OS	41,283	67	34,866	100
OE	70,180	59	52,152	149

of population genomics, genomic selection, and functional genomics.

CONCLUSION

The olive complex includes both wild and domestic subspecies, distributed in a wide range of temperate regions globally. *Olea europaea* subsp. *cuspidata* (OC) is one of the closest wild relatives of the olive tree (*O. europaea* subsp. *europaea,* OE), the symbol of peace and prosperity. Despite its close relationship with OE and great value in crossbreeding, OC still has no high-quality genomic reference, hindering its application in breeding and performance improvement. In this study, we used the most cutting-edge technologies in genomic sequencing, including Nanopore long-reads, Hi-C, second-generation sequencing, and RNA-seq, to conduct *de novo* genome assembly for an OC sample. The reference quality of OC is comparable to that of OE in terms of parameters, including scaffold N50 (50.46 Mb) and completeness of protein-coding genes (96.7%). On chromosome 17, we uncovered a particularly large region of "gene-desert" and "LTR-hotspot," possibly associated with the two genes *in situ*, phytochrome B–like gene and arginine methyltransferase–interacting related gene, which are related to chromatin compaction and gene methylation. We uncovered the recent divergence of OC from wild and domestic olive trees at 4.39 Mya, consistent with the complicated diversification process of all olive subspecies. The reference of OC would promote its future use in both scientific research and breeding applications.

STATEMENT FOR MATERIAL COLLECTION

Leaves of a single plant of *Olea europaea* subsp. *cuspidate* from Kunming arboretum, Yunnan Province, China (N 25°9′13″, E 102°45′9″) were collected for genome sequencing. Five kinds of tissues, namely, leaves, roots, twigs, bark, and fruits from the same plant were collected for RNA-seq to aid gene annotation. A specimen identified by Dr. Yong-Kang Sima was deposited at the Herbarium of the Yunnan Academy of Forestry and Grassland, Kunming City, China, under the voucher number Wu20056.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

TW, YL, TM, and DN conceived and designed the project. TW, LP, and YL collected the plant materials. TW and YL performed all the data analyses under the supervision of DN. TX, LP, and YZ performed the karyotype examination. TW and TM were major contributors in writing the manuscript. All authors contributed to and approved the final manuscript.

FUNDING

This work was financially supported by the National Key Research and Development Project (2019YFD1001205) and the Yunnan Provincial Science and Technology Major Project (202102AE090012).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.868540/full#supplementary-material

REFERENCES

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell.*182 (1), 145–161.e123. doi:10.1016/j.cell.2020.05.021

Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Barghini, E., Natali, L., Cossu, R. M., Giordani, T., Pindo, M., Cattonaro, F., et al. (2014). The Peculiar Landscape of Repetitive Sequences in the Olive (*Olea Europaea* L.) Genome. *Genome Biol. Evol.* 6 (4), 776–791. doi:10.1093/gbe/evu058

Besnard, G., Baradat, P., Chevalier, D., Tagmount, A., and Bervillé, A. (2001). Genetic Differentiation in the Olive Complex (*Olea Europaea*) Revealed by RAPDs and RFLPs in the rRNA Genes. *Genet. Resour. Crop Evol.* 48 (2), 165–182. doi:10.1023/A:1011239308132

Besnard, G., Hernández, P., Khadari, B., Dorado, G., and Savolainen, V. (2011). Genomic Profiling of Plastid DNA Variation in the Mediterranean Olive Tree. *BMC Plant Biol.* 11 (1), 80–12. doi:10.1186/1471-2229-11-80

Besnard, G., Khadari, B., Navascués, M., Fernández-Mazuecos, M., El Bakkali, A., Arrigo, N., et al. (2013). The Complex History of the Olive Tree: from Late Quaternary Diversification of Mediterranean Lineages to Primary Domestication in the Northern Levant. *Proc. R. Soc. B* 280 (1756), 20122833. doi:10.1098/rspb.2012.2833

Betrán, E., Thornton, K., and Long, M. (2002). Retroposed New Genes Out of the X in Drosophila. *Genome Res.* 12 (12), 1854–1859. doi:10.1101/gr.6049

Borodovsky, M., and Lomsadze, A. (2011). Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES. *Curr. Protoc. Bioinforma.* 35 (1), bi0406s35. doi:10.1002/0471250953.bi0406s35

Borzì, A., Biondi, A., Basile, F., Luca, S., Vicari, E., and Vacante, M. (2019). Olive Oil Effects on Colorectal Cancer. *Nutrients.* 11 (1), 32. doi:10.3390/nu11010032

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-Ep+ and AUGUSTUS Supported by a Protein Database. *NAR Genomics and Bioinformatics* 3 (1), lqaa108. doi:10.1093/nargab/lqaa108

Cáceres, M. E., Ceccarelli, M., Pupilli, F., Sarri, V., and Mencuccini, M. (2015). Obtainment of Inter-subspecific Hybrids in Olive (*Olea Europaea* L.). *Euphytica.* 201 (2), 307–319. doi:10.1007/s10681-014-1224-z

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and Applications. *BMC Bioinforma.* 10 (1), 1–9. doi:10.1186/1471-2105-10-421

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an Easy-To-Use Annotation Pipeline Designed for Emerging Model Organism Genomes. *Genome Res.* 18 (1), 188–196. doi:10.1101/gr.6743907

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17 (4), 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Chen, J., Mortola, E., Du, X., Zhao, S., and Liu, X. (2019). Excess of Retrogene Traffic in Pig X Chromosome. *Genetica.* 147 (1), 23–32. doi:10.1007/s10709-018-0048-5

Chen, J., Ni, P., Li, X., Han, J., Jakovlić, I., Zhang, C., et al. (2018a). Population Size May Shape the Accumulation of Functional Mutations Following Domestication. *BMC Evol. Biol.* 18 (1), 4. doi:10.1186/s12862-018-1120-6

Chen, J., Zhang, P., Chen, H., Wang, X., He, X., Zhong, J., et al. (2022a). Whole-genome Sequencing Identifies Rare Missense Variants of WNT16 and ERVW-1 Causing the Systemic Lupus Erythematosus. *Genomics.* 114 (3), 110332. doi:10.1016/j.ygeno.2022.110332

Chen, J., Zhong, J., He, X., Li, X., Ni, P., Safner, T., et al. (2022b). The De Novo Assembly of a European Wild Boar Genome Revealed Unique Patterns of Chromosomal Structural Variations and Segmental Duplications. *Anim. Genet.* 53, 281–292. doi:10.1111/age.13181

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018b). Fastp: an Ultra-fast All-In-One FASTQ Preprocessor. *Bioinformatics.* 34 (17), i884–i890. doi:10.1093/bioinformatics/bty560

Cho, J.-N., Ryu, J.-Y., Jeong, Y.-M., Park, J., Song, J.-J., Amasino, R. M., et al. (2012). Control of Seed Germination by Light-Induced Histone Arginine Demethylation Activity. *Dev. Cell.* 22 (4), 736–748. doi:10.1016/j.devcel.2012.01.024

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic DNA K-Mer Spectra: Models and Modalities. *Genome Biol.* 10 (10), R108–R110. doi:10.1186/gb-2009-10-10-r108

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016). Genome Sequence of the Olive Tree, *Olea Europaea*. *GigaSci.* 5 (1), 29. doi:10.1186/s13742-016-0134-5

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a Computational Tool for the Study of Gene Family Evolution. *Bioinformatics.* 22 (10), 1269–1271. doi:10.1093/bioinformatics/btl097

Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., et al. (2018). The Juicebox Assembly Tools Module Facilitates De Novo Assembly of Mammalian Genomes with Chromosome-Length Scaffolds for under $1000. *bioRxiv.* doi:10.1101/254797

Emerson, J. J., Kaessmann, H., Betra´n, E., and Long, M. (2004). Extensive Gene Traffic on the Mammalian X Chromosome. *Science.* 303 (5657), 537–540. doi:10.1126/science.1090042

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* 20 (1), 238. doi:10.1186/s13059-019-1832-y

Fu, Y., Mahmoud, M., Muraliraman, V. V., Sedlazeck, F. J., and Treangen, T. J. (2021). Vulcan: Improved Long-Read Mapping and Structural Variant Calling via Dual-Mode Alignment. *GigaScience.* 10 (9). doi:10.1093/gigascience/giab063

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: Finding Genomic Rearrangements and Local Sequence Differences from Whole-Genome Assemblies. *Genome Biol.* 20 (1), 277. doi:10.1186/s13059-019-1911-0

Green, P. S. (2002). A Revision of Olea L. (Oleaceae). *Kew Bull.* 57 (1), 91–140. doi:10.2307/4110824

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics.* 29 (8), 1072–1075. doi:10.1093/bioinformatics/btt086%JBioinformatics

Hannachi, H., Sommerlatte, H., Breton, C., Msallem, M., El Gazzah, M., Ben El Hadj, S., et al. (2008). Oleaster (Var. *Sylvestris*) and Subsp. *Cuspidata* Are Suitable Genetic Resources for Improvement of the Olive (*Olea Europaea* Subsp. *Europaea* Var. *Europaea*). *Genet. Resour. Crop Evol.* 56 (3), 393–403. doi:10.1007/s10722-008-9374-2

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a Fast and Efficient Genome Polishing Tool for Long-Read Assembly. *Bioinformatics.* 36 (1), 2253–2255. doi:10.1093/bioinformatics/btz891

Jin, G., Ma, P.-F., Wu, X., Gu, L., Long, M., Zhang, C., et al. (2021). New Genes Interacted with Recent Whole-Genome Duplicates in the Fast Stem Growth of Bamboos. *Mol. Biol. Evol.* 38 (12), 5752–5768. doi:10.1093/molbev/msab288

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* 428 (4), 726–731. doi:10.1016/j.jmb.2015.11.006

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi:10.1093/molbev/mst010

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A Novel Hybrid Gene Prediction Method Employing Protein Multiple Sequence Alignments. *Bioinformatics.* 27 (6), 757–763. doi:10.1093/bioinformatics/btr010

Kostelenos, G., Kiritsakis, A., and Shahidi, F. (2017). *Olive Tree History and Evolution*. Oxford, UK: Wiley.

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics.* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H. (2018). Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics.* 34 (18), 3094–3100. doi:10.1093/bioinformatics/bty191

Li, J., Ji, X., Wang, Z., Zeng, Y., and Zhang, J. (2021). Morphological, Molecular and Genomic Characterization of Two Inter-subspecific Hybrids between Olive Cultivars and Olive Subspecies. *Horticulturae.* 7 (6), 138. doi:10.3390/horticulturae7060138

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13 (9), 2178–2189. doi:10.1101/gr.1224503

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. 2013. Estimation of Genomic Characteristics by Analyzing K-Mer Frequency in De Novo Genome Projects. arXiv. doi:10.48550/arXiv.1308.2012

Long, M., and Langley, C. H. (1993). Natural Selection and the Origin of Jingwei , a Chimeric Processed Functional Gene in Drosophila. *Science.* 260 (5104), 91–95. doi:10.1126/science.7682012

Long, M., VanKuren, N. W., Chen, S., and Vibranovski, M. D. (2013). New Gene Evolution: Little Did We Know. *Annu. Rev. Genet.* 47, 307–333. doi:10.1146/annurev-genet-111212-133301

Ma, T., Ning, D., and Yang, W. (2014). Breeding of a New Olive Cultivar 'Jinyefoxilan. *Zhongguo Guoshu (China Fruits).* (6), 3–4.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 38 (10), 4647–4654. doi:10.1093/molbev/msab199

Niu, E., Jiang, C., Wang, W., Zhang, Y., and Zhu, S. (2020). Chloroplast Genome Variation and Evolutionary Analysis of *Olea Europaea* L. *Genes.* 11 (8), 879. doi:10.3390/genes11080879

Ou, S., Chen, J., and Jiang, N. (2018). Assessing Genome Assembly Quality Using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46 (21), e126. doi:10.1093/nar/gky730

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PloS one.* 5 (3), e9490. doi:10.1371/journal.pone.0009490

Rao, G., Zhang, J., Liu, X., Lin, C., Xin, H., Xue, L., et al. (2021). De Novo assembly of a New *Olea Europaea* Genome Accession Using Nanopore Sequencing. *Hortic. Res.* 8 (1), 64. doi:10.1038/s41438-021-00498-y

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De Novo assembly and Analysis of RNA-Seq Data. *Nat. Methods.* 7 (11), 909–912. doi:10.1038/nmeth.1517

Sanderson, M. J. (2003). r8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock. *Bioinformatics.* 19 (2), 301–302. doi:10.1093/bioinformatics/19.2.301

Sebastiani, L., and Busconi, M. (2017). Recent Developments in Olive (*Olea Europaea* L.) Genetics and Genomics: Applications in Taxonomy, Varietal Identification, Traceability and Breeding. *Plant Cell. Rep.* 36 (9), 1345–1360. doi:10.1007/s00299-017-2145-9

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., et al. (2018). Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing. *Nat. Methods.* 15 (6), 461–468. doi:10.1038/s41592-018-0001-7

Shi, Z., Luo, F., Li, Y., Yang, F., Xie, K., and Yang, W. (1991). Study on the Rootstock (*Olea Ferruginea*) for Grafting Olive. *Acta Bot. Yunnanica.* 13 (1), 65–74.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a Web Server for Gene Prediction in Eukaryotes that Allows User-Defined Constraints. *Nucleic Acids Res.* 33 (Suppl. l_2), W465–W467. doi:10.1093/nar/gki458

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* Chapter 4 (1), bi0410s25. doi:10.1002/0471250953.bi0410s2510.1002/0471250953.bi0410s25

Tessadori, F., van Zanten, M., Pavlova, P., Clifton, R., Pontvianne, F., Snoek, L. B., et al. (2009). PHYTOCHROME B and HISTONE DEACETYLASE 6 Control Light-Induced Chromatin Compaction in *Arabidopsis thaliana*. *PLoS Genet.* 5 (9), e1000638. doi:10.1371/journal.pgen.1000638

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of Wild Olive and the Evolution of Oil Biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 114 (44), E9413–E9422. doi:10.1073/pnas.1708621114

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast Reference-free Genome Profiling from Short Reads. *Bioinformatics.* 33 (14), 2202–2204. doi:10.1093/bioinformatics/btx153

Wang, X., Xu, W., Wei, L., Zhu, C., He, C., Song, H., et al. (2019). Nanopore Sequencing and De Novo Assembly of a Black-Shelled Pacific Oyster (*Crassostrea gigas*) Genome. *Front. Genet.* 10, 1211. doi:10.3389/fgene.2019.01211

Wang, Y., Li, J., and Paterson, A. H. (2013). MCScanX-Transposed: Detecting Transposed Gene Duplications Based on Multiple Colinearity Scans. *Bioinformatics.* 29 (11), 1458–1460. doi:10.1093/bioinformatics/btt150

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity. *Nucleic Acids Res.* 40 (7), e49. doi:10.1093/nar/gkr1293

Whelan, S., and Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* 18 (5), 691–699. doi:10.1093/oxfordjournals.molbev.a003851

Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. *Genome Biol.* 20 (1), 129. doi:10.1186/s13059-019-1727-y

Xia, S., Ventura, I. M., Blaha, A., Sgromo, A., Han, S., Izaurralde, E., et al. (2021). Rapid Gene Evolution in an Ancient Post-transcriptional and Translational Regulatory System Compensates for Meiotic X Chromosomal Inactivation. *Mol. Biol. Evol.* 39 (1). doi:10.1093/molbev/msab296

Xia, S., Wang, Z., Zhang, H., Hu, K., Zhang, Z., Qin, M., et al. (2016). Altered Transcription and Neofunctionalization of Duplicated Genes Rescue the Harmful Effects of a Chimeric Gene in Brassica Napus. *Plant Cell.* 28 (9), 2060–2078. doi:10.1105/tpc.16.00281

Yang, Z., and Dos Reis, M. (2010). Statistical Properties of the Branch-Site Test of Positive Selection. *Mol. Biol. Evol.* 28 (3), 1217–1228. doi:10.1093/molbev/msq303

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24 (8), 1586–1591. doi:10.1093/molbev/msm088

Zhang, C., Wang, J., Xie, W., Zhou, G., Long, M., and Zhang, Q. (2011). Dynamic Programming Procedure for Searching Optimal Models to Estimate Substitution Rates Based on the Maximum-Likelihood Method. *Proc. Natl. Acad. Sci. U.S.A.* 108 (19), 7860–7865. doi:10.1073/pnas.1018621108

Zhang, J., Jing, L., Li, M., He, L., and Guo, Z. (2019). Regulation of Histone Arginine Methylation/demethylation by Methylase and Demethylase (Review). *Mol. Med. Rep.* 19 (5), 3963–3971. doi:10.3892/mmr.2019.10111

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 30 August 2022
DOI 10.3389/fgene.2022.986724

OPEN ACCESS

EDITED BY
Yongjie Wu,
Sichuan University, China

REVIEWED BY
Eswara Reddy,
Institute of Himalayan Bioresource Technology (CSIR), India
Bin Zhu,
China Agricultural University, China

*CORRESPONDENCE
Minsheng You,
msyou@fafu.edu.cn
Shijun You,
sjyou@fafu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 05 July 2022
ACCEPTED 08 August 2022
PUBLISHED 30 August 2022

CITATION
Ke F, Li J, Vasseur L, You M and You S (2022), Temporal sampling and network analysis reveal rapid population turnover and dynamic migration pattern in overwintering regions of a cosmopolitan pest.
Front. Genet. 13:986724.
doi: 10.3389/fgene.2022.986724

Temporal sampling and network analysis reveal rapid population turnover and dynamic migration pattern in overwintering regions of a cosmopolitan pest

Fushi Ke[1,2,3,4†], Jianyu Li[1,2,3,5†], Liette Vasseur[1,2,6], Minsheng You[1,2,3*] and Shijun You[1,2,3,7*]

[1]State Key Laboratory of Ecological Pest Control for Fujian-Taiwan Crops, Institute of Applied Ecology, Fujian Agriculture and Forestry University, Fuzhou, China, [2]Joint International Research Laboratory of Ecological Pest Control, Ministry of Education, Fuzhou, China, [3]Ministerial and Provincial Joint Innovation Centre for Safety Production of Cross-Strait Crops, Fujian Agriculture and Forestry University, Fuzhou, China, [4]School of Biological Sciences, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China, [5]Institute of Plant Protection, Fujian Academy of Agricultural Sciences, Fuzhou, China, [6]Department of Biological Sciences, Brock University, St. Catharines, ON, Canada, [7]BGI-Sanya, Sanya, China

Genetic makeup of insect pest is informative for source-sink dynamics, spreading of insecticide resistant genes, and effective management. However, collecting samples from field populations without considering temporal resolution and calculating parameters related to historical gene flow may not capture contemporary genetic pattern and metapopulation dynamics of highly dispersive pests. *Plutella xylostella* (*L.*), the most widely distributed Lepidopteran pest that developed resistance to almost all current insecticides, migrates heterogeneously across space and time. To investigate its real-time genetic pattern and dynamics, we executed four samplings over two consecutive years across Southern China and Southeast Asia, and constructed population network based on contemporary gene flow. Across 48 populations, genetic structure analysis identified two differentiated insect swarms, of which the one with higher genetic variation was replaced by the other over time. We further inferred gene flow by estimation of kinship relationship and constructed migration network in each sampling time. Interestingly, we found mean migration distance at around 1,000 km. Such distance might have contributed to the formation of step-stone migration and migration circuit over large geographical scale. Probing network clustering across sampling times, we found a dynamic *P. xylostella* metapopulation with more active migration in spring than in winter, and identified a consistent pattern that some regions are sources (e.g., Yunnan in China, Myanmar and Vietnam) while several others are sinks (e.g., Guangdong and Fujian in China) over 2 years. Rapid turnover of insect swarms and highly dynamic metapopulation highlight the importance of temporal sampling and network analysis in investigation of source-sink relationships and thus effective pest management of *P. xylostella*, and other highly dispersive insect pests.

KEYWORDS

temporal sampling, kinship analysis, population network, dynamic metapopulation, insect pest

Introduction

Insect pests are abundant in all agricultural regions of the world, and many are difficult to control due to their rapid development of insecticide resistance and active migration. For these pests, molecular evidence of genetic pattern and gene flow may provide important guidance in understanding the colonization history (Yang et al., 2012; Kirk et al., 2013; Cao et al., 2019; You et al., 2020), tracking insecticide resistance (Taylor et al., 2021), and investigating metapopulation dynamics (Dinsdale et al., 2012; Elameen et al., 2020; Perera et al., 2020). Much work has been done on the genetic structure and phylogeography of widely distributed insect pests, such as *Locusta migratoria* (Ma et al., 2012), *Frankliniella occidentalis* (Yang et al., 2012; Cao et al., 2019), *Nilaparvata lugens* (Stål) (Hu et al., 2019) and *Plutella xylostella* (You et al., 2020). However, little is known for contemporary dynamics of insect pests that migrate actively all year-round (Fu et al., 2014) and form migration circuit between source-sink populations (Gao et al., 2020). Analysis of real-time pattern of genetic variation and contemporary gene flow thus would help to clarify source-sink metapopulation dynamics and insecticide resistance development.

Unlike vertebrate migrants, migration circuits of insects can involve multiple generations and experience different levels of connectivity due to their shorter generation time and lower degree of control over movement (flight) directions (Gao et al., 2020). Most population genetic studies of highly dispersive species report a pattern of low genetic differentiation and high gene flow (Jiang et al., 2010; Wei et al., 2013; Elameen et al., 2020), which is not unexpected, but not informative. This is due to the frequently applied genetic parameter N_m refers to historical gene flow and thus inadequate for contemporary population dynamics. In addition, *Nm* calculated by F_{ST} need several assumptions that real populations are likely violating and do not provide accurate estimation of migration (Whitlock and Mccauley, 1999). To better understand the contemporary genetic pattern and migration dynamics, choosing the suitable spatial and temporal resolution as well as the right metrics thus is vital (Osborne and Woiwod, 2002).For insect species with partial migration (i.e., a population consisting of migrants and residents) (Menz et al., 2019), pedigree-based approaches can be useful in identifying contemporary migration events (Chen et al., 2021). Furthermore, quantifying contemporary gene flow between populations by counting kinship assignments (Wang, 2014) can lead to the construction of population networks and identification of source-sink relationships. Population network analysis that uses populations (sites) as nodes and genetic similarity (or gene flow) as edges (Dyer, 2015) could further identify key nodes that affect temporal connectivity of metapopulation.

Plutella xylostella (diamondback moth, DBM) has developed resistance to almost all insecticides, causing enormous damage to global cruciferous vegetable and oil crop industries (Furlong et al., 2013; Li et al., 2016). It has been recorded in over 140 countries and regions worldwide, including Svalbard Island near the north pole (Coulson et al., 2002) and subantarctic Marion Island (Chown and Crafford, 1987). DBM is one of the most successful moths according to Miyata' classification (Miyata, 1983). It is characterized by having high migration ability and with large-scale distribution of host plants. Migration and colonization of DBM are regionally and temporally heterogeneous. For example, DBM can overwinter in tropic and subtropic regions, but can only breed at particular time intervals at higher latitudes, where it must recolonize every year under suitable weather conditions (Chen et al., 2021; Ma et al., 2021). This forms a dynamic metapopulation (Ke et al., 2019) between overwintering habitats (e.g., lower latitude) and seasonally breeding regions (e.g., higher latitude). In northern Europe, warm currents from western Russia facilitate colonization of DBM at high-latitude area close to the north pole, while its presence in Finland is due to northward wind from Estonia (Tiilikkala et al., 1996). In England, winds from the mainland Europe bring DBM seasonally, with individuals of the early summer coming from East and Southeast Europe, and those arriving in late summer are from the eastern regions of Baltic Sea (Chapman et al., 2002). In Eastern Asia, mass migration of DBM is known to cross the Bohai Strait, with population density varying with the season and peaking in June and August (Fu et al., 2014). Monsoons (Ke et al., 2015) and typhoons (Kohno et al., 2004) also aid the migration of DBM across the open seas, resulting in their colonization of islands far from the continent. Dosdall et al. (2001) suggest that continuously strong winds in North America bring DBM from Southern United States and Mexico into Canada. Overall, spatially and temporally varied migration of DBM is influenced by external factors that formed the "weather window," including wind strength and direction (Coulson et al., 2002).

East Asia is one of the most important regions for mass insect migration. Every year, a large number of insects crossing the sea (e.g., Bohai gulf and South China Sea) are recorded by radar (Guo et al., 2020; Zhou et al., 2021). Among them, *P. xylostella* is an important agricultural pest that migrate temporally and regionally (Fu et al., 2014; Wang

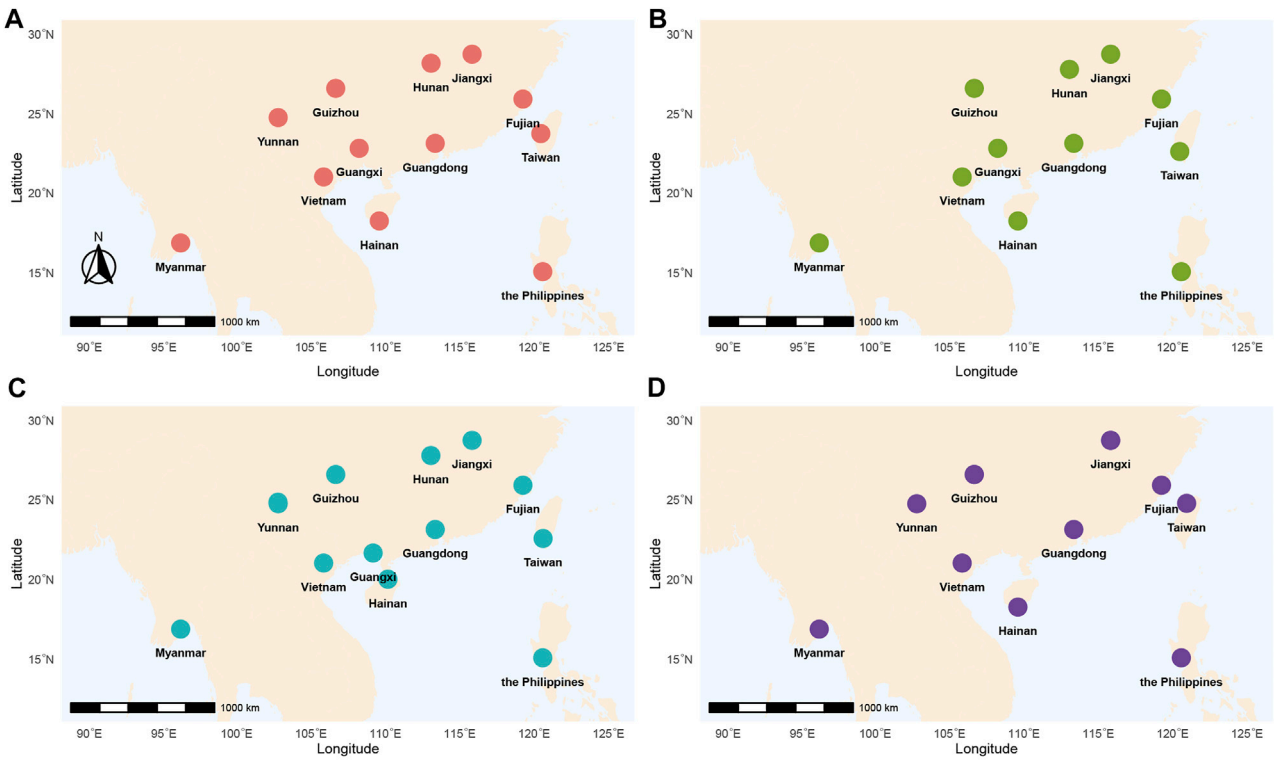


FIGURE 1
Geographical locations of *P. xylostella* sampling sites across Southern China and Southeast Asia. Different colors represented time1 (from April to June 2016) **(A)**, time 2 (from October 2016 to January 2017) **(B)**, time 3 (from April to June 2017) **(C)**, time 4 (from November 2017 to January 2018) **(D)**, respectively. Refer to Supplementary Table S1 for detailed information of collecting samples.

et al., 2022). Widely distributed host plants (e.g., cultivated cruciferous vegetables) and suitable weather make East Asia an ideal place for DBM's reproduction and migration throughout the year (Fu et al., 2014; Li et al., 2016). In this study, we hypothesized that temporal variation of DBM population dynamics has contributed to the differentiation of contemporary genetic pattern, source-sink relationships, and regional population clustering. To test this hypothesis, we conducted longitudinal sampling of *P. xylostella* populations in Southern China and Southeastern Asia during two consecutive years and investigated contemporary genetic makeup of these populations. We first probed temporal variation of genetic pattern in overwintering regions of DBM populations. Further, we applied kinship-based inference of gene flow to investigate contemporary metapopulation dynamics. Finally, by constructing population network of temporally sampled populations in each sampling time interval, we identified key populations that were important in regional dynamics. Our study represented the first documentation that employed temporal sampling and network analysis based on contemporary gene flow.

Materials and methods

Sampling design

We implemented a longitudinal sampling of *P. xylostella* populations in four Asian countries including China, Myanmar, the Philippines, and Vietnam (Figure 1; Table 1). Specifically, one site from each of the nine provinces in China (Fujian, Guangdong, Guangxi, Guizhou, Hainan, Hunan, Jiangxi, Taiwan, Yunnan), and three additional sites from Myanmar, the Philippines, and Vietnam were selected and sampled twice a year (first sampling: April, May and June; second sampling: December and January) from April 2016 to January 2018. Note that three sites failed to be sampled at one specific sampling time (Yunnan in time 2, Hunan in time 4 and Guangxi in time 4, Figure 1; Table 1). In each site, we collected individuals at a distance of at least 3 m (to avoid sampling too many individuals with close kinship) and stored them in 95% alcohol. The samples were further stored at −80°C before being used for DNA extraction. Without any congeneric species of DBM reported or identified in this region (You et al., 2020), we verified the species identity of all sampled individuals

TABLE 1 Sampling information of *P. xylostella* populations across Southern China and Southeast Asia.

Sampling time code	Population	Sampling site	Latitude	Longitude	Sample size	Sampling date
time1	FJ1	Fujian	25.94392	119.25313	32	2016/5/12
	FL1	the Philippines	15.10136	120.58555	32	2016/6/21
	GD1	Guangdong	23.16017	113.35634	31	2016/6/16
	GX1	Guangxi	22.85021	108.2448	32	2016/6/3
	GZ1	Guizhou	26.61723	106.66834	32	2016/6/28
	HN1	Hainan	18.29119	109.59779	32	2016/4/28
	JX1	Jiangxi	28.76363	115.83489	32	2016/4/24
	MY1	Myanmar	16.90561	96.24384	32	2016/7/9
	TW1	Taiwan	23.77326	120.45462	32	2016/6/10
	VN1	Vietnam	21.04872	105.85288	29	2016/7/7
	XX1	Hunan	28.1974	113.07853	25	2016/7/7
	YN1	Yunan	24.78023	102.79866	32	2016/6/25
time2	FJ2	Fujian	25.94392	119.25313	20	2016/12/6
	FL2	the Philippines	15.10136	120.58555	32	2017/1/16
	GD2	Guangdong	23.16017	113.35634	32	2017/1/9
	GX2	Guangxi	22.85021	108.2448	32	2017/12/18
	GZ2	Guizhou	26.61723	106.66834	32	2016/12/23
	HN2	Hainan	18.29119	109.59779	32	2017/1/6
	JX2	Jiangxi	28.76363	115.83489	29	2016/10/25
	MY2	Myanmar	16.90561	96.24384	32	2016/12/22
	TW2	Taiwan	22.64479	120.4785	29	2016/10/22
	VN2	Vietnam	21.04872	105.85288	32	2016/12/18
	XX2	Hunan	27.81154	113.06472	32	2016/12/20
time3	FJ3	Fujian	25.94392	119.25313	32	2017/4/27
	FL3	the Philippines	15.10136	120.58555	32	2017/5/6
	GD3	Guangdong	23.16017	113.35634	32	2017/5/23
	GX3	Guangxi	21.68073	109.17835	32	2017/4/29
	GZ3	Guizhou	26.61723	106.66834	32	2017/6/6
	HN3	Hainan	20.03826	110.17471	5	2017/5/15
	JX3	Jiangxi	28.76363	115.83489	32	2017/5/17
	MY3	Myanmar	16.90561	96.24384	32	2017/6/4
	TW3	Taiwan	22.59524	120.60754	32	2017/5/24
	VN3	Vietnam	21.04872	105.85288	21	2017/6/1
	XX3	Hunan	27.81154	113.06472	32	2017/5/25
	YN3	Yunnan	24.78023	102.79866	32	2017/6/5
	YC3	Yunan	24.87667	102.78666	14	2017/5/22
time4	FJ4	Fujian	25.94392	119.25313	32	2017/11/7
	FL4	the Philippines	15.10136	120.58555	32	2017/11/20
	GD4	Guangdong	23.16017	113.35634	32	2017/11/25
	GB4	Guizhou	26.61831	106.66114	32	2017/12/2
	GL4	Guizhou	26.61723	106.66834	32	2017/12/1
	HN4	Hainan	18.29119	109.59779	32	2018/1/26
	JX4	Jiangxi	28.76363	115.83489	32	2017/12/15
	JX5	Jiangxi	28.76363	115.83489	6	2018/1/6
	MY4	Myanmar	16.90561	96.24384	32	2017/11/28
	TW4	Taiwan	24.80418	120.94269	32	2018/1/26
	VN4	Vietnam	21.04872	105.85288	32	2017/11/26
	YN4	Yunan	24.78023	102.79866	32	2017/11/29

using morphological characteristics following You & Wei (You, 2007) instead of sequencing mitochondrial *COI* of every individual (You et al., 2020).

DNA extraction and microsatellite genotyping

We extracted total genome DNA individually by using Gentra Puregene Blood Kit (QIAGEN) following the instruction manual. The extracted DNA of each individual was tested based on OD 260/280 ratio measured by UV-1600 Spectrophotometer. All the individuals were with high DNA concentration and quality, and suitable for microsatellite (i.e., simple sequence repeats, SSR) genotyping.

A total of 15 published SSR markers were selected for genotyping (Esselink et al., 2006; Ke et al., 2015). The PCR program with three primers based on Schuelke' method (Schuelke, 2000) was employed. A more detailed description of the PCR reaction system can be found in our previous study (Ke et al., 2015). Briefly, the total volume of each PCR reaction was 25 μl, containing 12.5 μl Mix (Promega), 0.2 μl forward primer, 1 μl reverse primer, and 0.8 μl M-13 linked forward primer. The temperature conditions were set at 94°C for 10 min, and then 32 cycles at 94°C for 30 s, 56°C for 45 s, 72°C for 45 s, followed by eight cycles at 94°C for 30 s, 53°C for 45 s, 72°C for 45 s, and a final extension at 72°C for 10 min. PCR products were tested by gel electrophoresis, and reactions with no visible products were re-amplified with more PCR circles. All fluorescence-labelled PCR products were further scanned by ABI 3730 sequencer in Sangon Biotech (Shanghai). Sizes of the products were assigned by GeneMapper 4.1 (Applied Biosystems) and further checked manually.

Genetic variation and structure

We used GenAIEx6.5 (Peakall and Smouse, 2006) and fastat (Goudet, 2001) to calculate several metrics of population genetic variation, including number of alleles (*Na*), effective population size (*Ne*), Shannon index (*I*), observed heterozygosity (*Ho*), expected heterozygosity (*He*), and allelic richness (*Ar*). In addition, we treated populations of the same sampling time interval as a metapopulation and estimated genetic variation in each metapopulation by using number of alleles and number of private alleles based on the rarefaction method in ADZE (Szpiech et al., 2008). We also performed the pricipla coordinates analysis (PCoA) based on calculation of Nei's distance between populations in GenAIEx6.5 (Peakall and Smouse, 2006). The genetic structure of all sampled populations was further inferred using STRUCTURE (Pritchard et al., 2000) with 100,000 burn-in and 1,000,000 MCMC iterations. LOCPRIOR model was set with geographical information of populations (Hubisz et al., 2009). Each K from 1 to 5 was run with 10 replications, and the best K was determined by using STRUCTRE harvest (Earl and Vonholdt, 2012). The geographical distance of each sampling site pairs was measured by R package GEOSPHERE 1.5-10 (Fick and Hijmans, 2017) and used for regression analysis with Nei's distance in GraphPad Prism 8 (Swift, 1997).

Kinship assignment and contemporary gene flow analysis

Parentage inference based on individual genotypes of natural populations is important and has attracted research interests for decades (Jones et al., 2010). Genetic information of multiple offspring that share the same parent can provide important information for inferring parental genotype and is applied in joining sibship and parentage analysis based on likelihood (Wang and Santure, 2009) and Bayesian methods (Emery et al., 2001). Here, we used the Colony program (Jones and Wang, 2010) that could assign sibship and parentage jointly based on the full likelihood (FL) method. For populations in each sampling time interval, we conducted five long runs with settings including polygamy for both sexes, mating system with inbreeding, and high likelihood precision. To avoid bias due to small sampling size, populations with less than 10 individuals were excluded for this analysis. In addition, as the microsatellite loci can easily have genotyping errors, we further performed analysis based on different error rates (0.0001, 0.001, 0.01, 0.05) to check the marker informativeness (Wang, 2018). If these loci could provide essentially the same results under different error rates, then the markers were deemed informative and kinship assignment analysis accurate (Wang, 2018).

Based on the parentage relationships identified by Colony, we then investigated gene flow by calculating likelihood estimator of migration rates among the sampled populations in MigEst 1.0 (Wang, 2014). This software implements marker-based parentage assignments and jointly estimates contemporary migration rates (*m*) across the sampled populations from a metapopulation (Wang, 2014). Recent analysis indicates the coalescent-based method [e.g., Migrate-n (Beerli, 2006)] and disequilibrium estimates [e.g., BayesAss (Wilson and Rannala, 2002)] may lead to inaccurate estimates of gene flow under recent landscape change (Samarasin et al., 2017). Unlike BayesAss (Wilson and Rannala, 2002) that estimates recent gene flow in the last 5–15 generations and assumes low migration rates (Broquet et al., 2009; Martin et al., 2021), MigEst estimates gene flow based on the kinship (Wang, 2014) and therefore suitable for quantifying the contemporary and mass migration of insect species. A custom python script was used to count the number and percentage of parentage assignment in each population pair based on *.BestConfig_Ordered file generated by Colony and transform it into the input for MigEst. It is worth noting that the sex of the samples was not recorded before

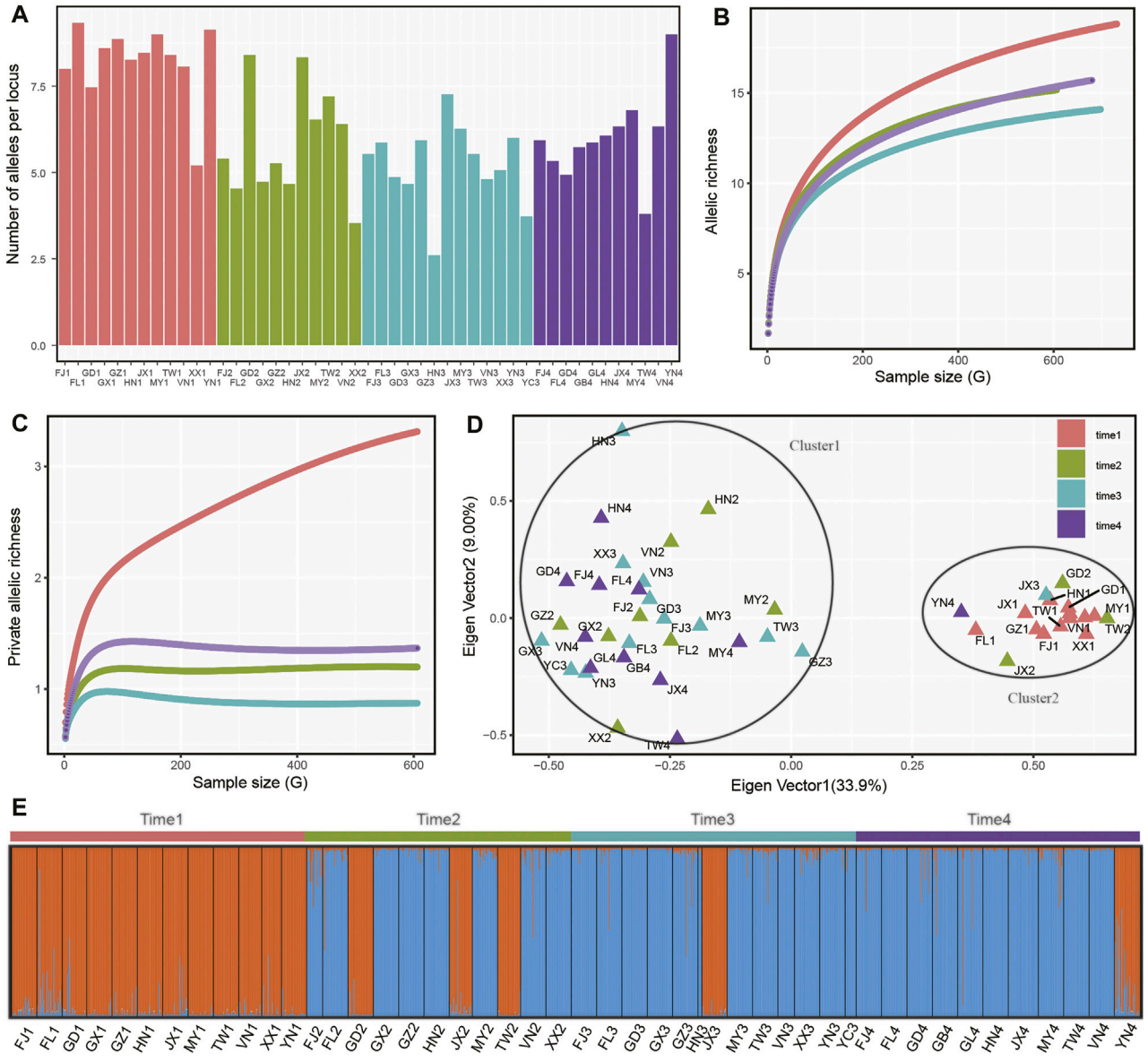


FIGURE 2
Genetic variation and differentiation of *P. xylostella* populations. **(A)** Average number of alleles per locus of each population. For other genetic parameters referred to Supplementary Table S2. **(B)** Accumulation of allelic richness in populations grouped by sampling time (i.e., a metapopulation). The sample size increased from two to the total number of alleles in each population. **(C)** Accumulation of private allelic richness in populations grouped by sampling time. **(D)** Principal coordinate analysis (PCoA) based on allele frequency of each sampling population. **(E)** Population genetic structure analysis based on STRUCTURE.

extracting the DNA. This might have limited effective quantification of migration among populations because the uncertainty of the paternity-offspring or maternity-offspring assignment (Jones and Wang, 2010). But as the parentage-offspring relationship was robust in Colony, the migration direction and linkage between populations remain effective and should provide useful information in investigating metapopulation dynamics (personal communication, Dr. Jinliang Wang).

Population network analysis and pivotal node identification

Using the contemporary gene flow matrix, we performed population network analysis based on igraph (Csardi and Nepusz, 2006). We constructed two population networks based on migration matrix. The first network employed gene flow between two populations as the non-weighted directed edge

10.3389/fgene.2022.986724

TABLE 2 Genetic diversity of *P. xylostella* populations.

Sampling time code	Pop	Na	*Ne*	*I*	*Ho*	*He*	*Ar*
time1	FJ1	8	3.558	1.437	0.546	0.654	8.43
	FL1	9.333	4.411	1.527	0.53	0.656	4.084
	GD1	7.467	3.754	1.443	0.518	0.665	4.503
	GX1	8.6	3.954	1.522	0.557	0.685	8.008
	GZ1	8.867	4.47	1.595	0.557	0.708	4.278
	HN1	8.267	4.127	1.512	0.565	0.686	4.83
	JX1	8.467	4.033	1.474	0.564	0.644	4.611
	MY1	9	4.367	1.563	0.596	0.683	8.158
	TW1	8.4	3.897	1.462	0.575	0.655	18.301
	VN1	8.067	4.129	1.501	0.579	0.681	13.174
	XX1	5.2	3.312	1.239	0.549	0.622	5.941
	YN1	9.133	4.214	1.575	0.572	0.693	9.174
	Mean	8.233	4.019	1.487	0.559	0.669	7.791
time2	FJ2	5.4	3.169	1.173	0.555	0.575	6.987
	FL2	4.533	2.913	1.084	0.459	0.567	2.747
	GD2	8.4	4.1	1.461	0.531	0.651	3.139
	GX2	4.733	2.932	1.042	0.354	0.535	5.616
	GZ2	5.267	2.62	1.038	0.362	0.508	3.478
	HN2	4.667	2.795	1.087	0.46	0.564	3.713
	JX2	8.333	3.466	1.437	0.468	0.637	4.016
	MY2	6.533	3.078	1.228	0.469	0.59	6.19
	TW2	7.2	3.797	1.426	0.515	0.655	9.368
	VN2	6.4	3.477	1.32	0.431	0.62	8.034
	XX2	3.533	1.685	0.634	0.289	0.342	4.66
	Mean	5.909	3.094	1.175	0.445	0.568	5.268
time3	FJ3	5.533	2.919	1.159	0.517	0.582	3.867
	FL3	5.867	2.855	1.183	0.471	0.583	2.192
	GD3	4.867	2.665	1.005	0.424	0.517	2.426
	GX3	4.667	2.33	0.906	0.361	0.455	2.745
	GZ3	5.933	3.243	1.28	0.471	0.625	2.358
	HN3	2.6	2.017	0.701	0.363	0.411	2.324
	JX3	7.267	3.734	1.434	0.538	0.668	2.808
	MY3	6.267	2.892	1.173	0.453	0.568	3.275
	TW3	5.533	2.852	1.181	0.488	0.6	4.291
	VN3	4.8	2.769	1.093	0.484	0.559	3.771
	XX3	5.067	2.875	1.124	0.466	0.567	2.772
	YN3	6	2.879	1.12	0.557	0.532	3.927
	YC3	3.733	2.131	0.88	0.51	0.476	3.449
	Mean	5.241	2.782	1.095	0.469	0.549	3.093
time4	FJ4	5.933	3.366	1.263	0.453	0.605	4.116
	FL4	5.333	3.179	1.242	0.446	0.631	2.45
	GD4	4.933	2.676	1.081	0.374	0.55	2.301
	GB4	5.733	3.295	1.25	0.446	0.608	3.979
	GL4	5.867	2.949	1.134	0.447	0.548	2.812
	HN4	6.067	3.017	1.148	0.438	0.554	3.005
	JX4	6.333	3.245	1.279	0.461	0.602	3.231
	JX5	2.6	1.827	0.618	0.3	0.347	4.001
	MY4	6.8	3.262	1.221	0.504	0.578	5.84

(Continued on following page)

TABLE 2 (*Continued*) Genetic diversity of *P. xylostella* populations.

Sampling time code	Pop	Na	*Ne*	*I*	*Ho*	*He*	*Ar*
	TW4	3.8	2.149	0.892	0.351	0.49	5.201
	VN4	6.333	3.391	1.279	0.575	0.604	3.005
	YN4	9	3.738	1.478	0.525	0.65	5.134
	Mean	5.728	3.008	1.157	0.443	0.564	3.756

Na, *Ne*, *I*, *Ho*,*He*, and *Ar* represent number of alleles, effective population size, Shannon index, observed heterozygosity, expected heterozygosity, and allelic richness, respectively.

and the degree centrality (Everett and Borgatti, 2005) of each site as the size of the node. We constructed the second population network by using betweenness centrality, a measure of centrality in a graph based on shortest paths (Everett and Borgatti, 2005), of each edge, and further performed the modularity clustering (Newman, 2006) based on fast greedy algorithm implemented in igraph (Csardi and Nepusz, 2006). When identifying pivotal nodes, we treated nodes that were important in population intermixing (e.g., sink populations that received immigration from different populations and mixed) or being the source of many other populations as key nodes. These key nodes generally had higher node degrees and/or being shared by two or more clusters in modularity clustering analysis.

Results

Genetic variation and structure

In four sampling time intervals, a total of 1,425 individuals from 48 *P. xylostella* populations were collected in South China and Southeast Asia (Table 1), with each time interval consisting of 12, 11, 13, 12 populations and 373, 334, 360, 358 individuals, respectively. Based on several metrics of genetic variation calculated in *P. xylostella* populations (Figure 2; Table 2), we found populations collected in time 1 (April–June 2016) generally had higher genetic variation than the populations from the other time intervals collected at the same or nearby site. There were some exceptions that had higher genetic variation compared with other populations collected at same sampling time interval: FJ2, TW2, and JX2 collected at time 2 (October 2016–January 2017), JX3 at time 3 (April–June 2017), and YN4 at time 4 (November 2017–January 2018) (Figure 2A; Table 2). By defining DBM populations collected from each time interval as a group (metapopulation), we further calculated mean number of distinct alleles and private alleles per locus in each group from a sampling size (i.e., G) from 2 to the maximum size of the group using iteration method. We found that time 1 had both higher mean number of distinct alleles and private alleles per locus compared with other groups (i.e., time 2 to time 4, Figures 2B,C). The saturation curve of private alleles in time 1 was not plateaued, but a clear platform was found for all other three time intervals (Figure 2C).

Further genetic structure analysis found that the populations with higher genetic variation clustered together (cluster1), while the other populations formed a second cluster (Figure 2B). We double checked the genotyping data and found that this pattern in our study was not due to genotyping error of SSR makers but due to coexistence of two DBM swarms. Populations from cluster1 were mainly collected from time 1, while some populations in this cluster also being found in isolated regions (i.e., Guangdong, Jiangxi, Taiwan and Yunan in China) collected in the subsequent sampling times (Figures 2D,E). Similarly, the genetic structure analysis with the best K value of 2 showed that cluster1 was the dominant insect swarm in time 1, while only three populations from cluster1 were found in time 2 (GD2, JX2, TW2). In time 3 (JX3) and time 4 (YN4), only one population from cluster1 was identified. Within each sampling time, we further performed correlation analysis based on Nei's genetic distance and geographical distance of each population pair and found no significant correlation in the four datasets (time 1: r = 0.063, p = 0.808; time 2 : r = −0.133, p = 0.502; time 3 : r = −0.026, p = 0.889; time 4 : r = −0.037, p = 0.853).

Source-sink populations and migration distance

The parentage analysis based on Colony identified the best configuration of offspring and parental individuals in each sampling time interval and was used to infer contemporary gene flow between source-sink populations. We found similar migration events among populations identified by markers using different error rates. This supports the accuracy of kinship assignment analysis (Wang, 2018). We thus used the estimation with an error rate of 0.0001 for further analysis.

The population network was constructed using the presence of gene flow between two populations as edge, with the arrow representing the direction and the size of node indicated the degree centrality (i.e., number of the connected nodes). In time 1, FJ1, FL1, GD1, GX1, GZ1, HN1, and JX1 were with high degree centrality and receiving immigrants from other populations (Figure 3A), while FJ2, FL2, GD2, GX2, and GZ2 in time

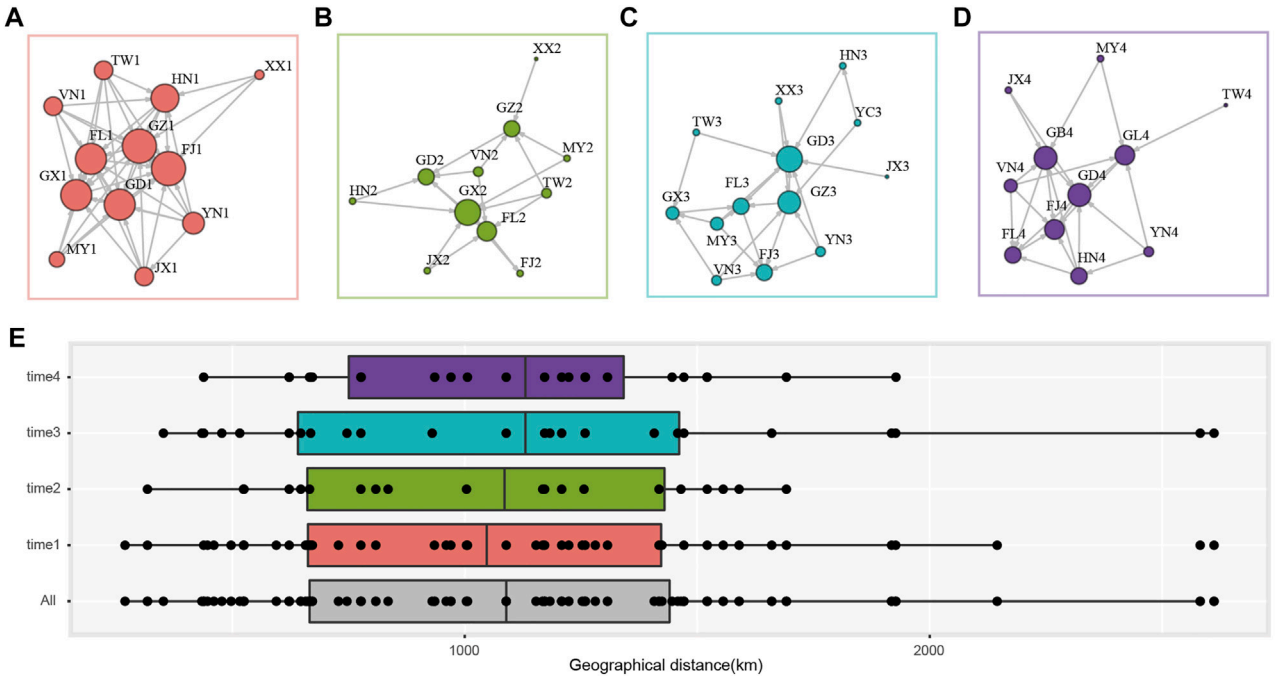


FIGURE 3
Gene flow and migration distance of sampling *P. xylostella* populations. Migration network of each sampling time. **(A)** time1; **(B)** time2; **(C)** time3; **(D)** time4. The presence of gene flow between populations was indicated with directed and non-weighted edge, and the degree centrality was represented by the size of each node. **(E)** The geographical distance (km) of population pairs with gene flow. time 1 to time 4 indicate population pair in each time interval, and "All" denotes all the population pairs. Boxes show the first and third quartile range (IQR) in each panel.

2 were sink populations and GX2 has the most connected populations (Figure 3B). In time 3 and time 4, FJ3, FL3, GD3, GX3, GZ3, HN3, and FJ4, FL4, GD4, GB4, GL4, HN4 received immigrants (Figures 3C,D). We found a pattern where DBM populations from Fujian (FJ), the Philippines (FL), Guangdong (GD, GB), Guangxi (GX), Guizhou (GZ) always received contemporary immigration and were the sink populations irrespective of the sampling time and wind direction. While regions such as Yunnan (YN) and Taiwan (TW), Vietnam (VN), and Myanmar (MY) located at peripheral regions and encircled the sink populations, they were always the source populations in four sampling times (Figure 3). Several populations [e.g., Hainan (HN)] were the source and sink that varied across sampling times.

We further summarized the geographical distance of each population pair with gene flow. A total of 115 population pairs with contemporary gene flow in the overall 249 pairs were identified (Figure 3). The data showed that DBM could migrate more than 2,500 km although there were only few successful immigration events at this distance (Figure 3E). Most of the migration events occurred at around 1,000 km, with a 95% confidence interval of 1,003.45 km–1,196.77 km. We found similar migration distance in each time interval, but only in time 1 and time 3 (during the spring of each year), we found migration events over 2,500 km (Figure 3E).

Migration network clustering and key nodes for temporal metapopulation dynamics

We further used Newman's modularity clustering in igraph to construct the migration network. To distinguish from the former network, we used betweenness as undirected and weighted edges. Overall, we found a highly differentiated pattern of clustering across sampling time intervals, and clustering analysis found samples from spring (time 1 and time 3) have lower clustering number compared with that in winter (time 2 and time 4) (Figure 4). At time 1, all the populations formed one cluster (Figure 4A). At time 2, four clusters were identified, with Myanmar (MY2) and Hunan (XX2) each forming independent clusters. The third cluster consisted of Guizhou (GZ2) and Vietnam (VN2) while the rest populations formed the fourth cluster (Figure 4B). At time 3, five clusters without overlapping were identified. At time 4, we further found six clusters with Jiangxi (JX4), Guizhou (GL4), Yunnan (YN4), Taiwan (TW4) and Myanmar (MY4) forming individual independent clusters.

Combining network clustering (Figure 4) and centrality (Figure 3), we were able to identify some key nodes that could be important in metapopulation dynamics. Basically, we

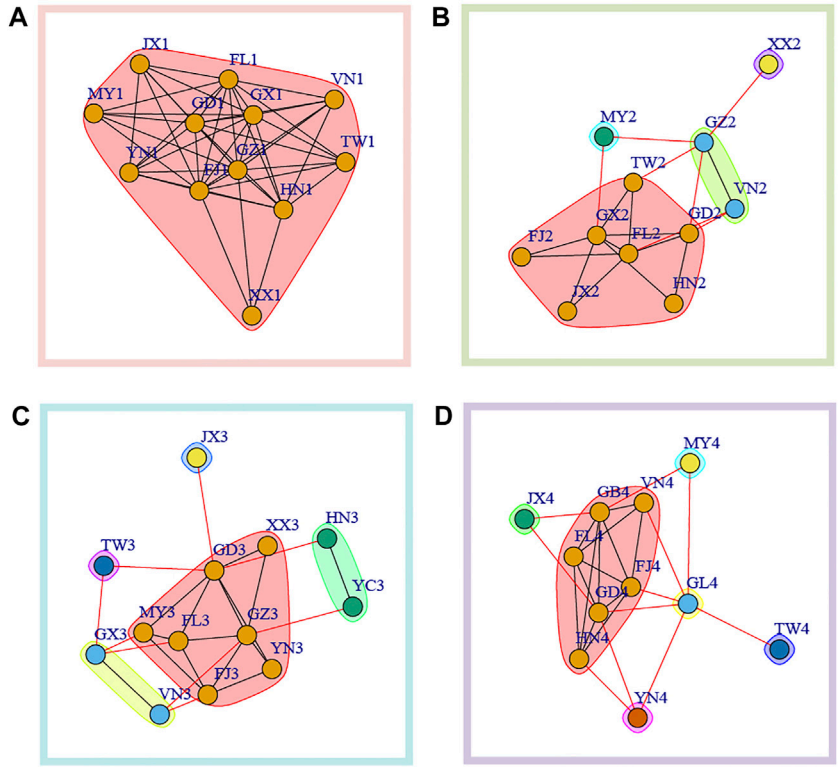


FIGURE 4
Modularity clustering analysis based on betweenness between populations in each sampling time interval. **(A)** time1; **(B)** time2; **(C)** time3; **(D)** time4.

considered the populations with higher centrality were key nodes serving as the sink for genetic intermixing (mixing of standing variation and/or *de novo* mutations) or as source population that could mediated gene flow into the other populations. In time 1, highest degree centrality was found in Fujian (FJ1) and Guizhou (GZ1). Across four clusters of time 2, Guangxi (GX2) was the population with the highest node degree (Figure 4B). Of five clusters at time3, we found that Guangdong (GD3) had the highest degree centrality (Figure 4C). In time4, Guangdong (GD4) and Guizhou (GB4) within the same cluster in the modularity clustering analysis had highest degree centrality (Figure 4D).

Discussion

Windborne migration mediates rapid population turnover in overwintering regions

For many insect pests, resources and habitats are ephemeral across time and space (Osborne and Woiwod, 2002) due to fast turnover of modern agriculture. Windborne migration of agricultural pests is important in escaping unfavorable habitats and sustaining succession of insect populations. In the sampling region, we found two *P. xylostella* swarms across all sampling populations, with one replaced by the other over time. This result suggested that windborne migration of this pest has facilitated a rapid population turnover across the metapopulation of *P. xylostella* in South China and Southeast Asia. Similar pattern of replacement has also been identified in other studies of insect pests. For example, Dinsdale et al. (2011) report rapid population turnover of *Bemisia tabaci* over time when investigating the impact of agricultural landscapes on its population genetic pattern. Working on temporal variation in genetic composition of *Helicoverpa zea* (Boddie), Perera et al. (2020) find that the pest populations tend to differentiate over time, but suggest that this can be a transient or temporary phenomenon. We thus proposed that the rapid population turnover of DBM should be taken into consideration for effective pest management such as temporal insecticide resistance monitoring under modern agricultural regime, as it might be for other species.

Contemporary population network clustering and key nodes for source-sink dynamics

Based on modularity clustering analysis, we found a differentiated clustering pattern across different sampling times, which supported a dynamic metapopulation of *P. xylostella* in overwintering regions. This could be explained by low level of autonomous control over migration direction (Gao et al., 2020) and the variation of regional wind directions and strength (Coulson et al., 2002). We found most of the migration distances between populations were about 1,000 km, which were similar to previous analysis based on genome-wide SNPs (Chen et al., 2021). The relatively "short" migration distance over large geographical scale of *P. xylostella* metapopulation in China not only support step-stone migration (Chen et al., 2021) that forms migration circuits of this pest, but could also contribute to temporal population clustering, likely regulated by seasonal winds. For example, we find a differentiated pattern of population clustering between seasons in two consecutive years (i.e., time 1 vs. time 2, and time 3 vs. time 4). To be specific, less clusters in spring (time 1 and time 3) compared with that in winter of the same year (time 2 and time 4) were observed. This pattern is likely due to more active windborne migration in spring that contributed to the linkage across regional populations (Fu et al., 2014). More active migration in the spring was also evidenced by migration events over 2,500 km were only identified in time 1 and time 3, while not in winter populations (time 2 and time 3). A combination of strain-specific characteristics with regional wind strength and directions (Coulson et al., 2002) as well as particular "weather window" thus could explain the temporal variation of population network and dynamics of this pest. Further research (e.g., energy reserve of *P. xylostella* individuals) should be conducted for a comprehensive understanding of migration at certain geographical distances and its contribution to regional metapopulation dynamics.

In addition, we found that populations from Southwest of China (e.g., Yunnan) as well as Indo-China Peninsula (ICP, e.g., Myanmar and Vietnam) were always the source for populations in our four samplings. This pattern has also been identified in a recent analysis based on genome-wide SNPs (Chen et al., 2021), where populations of Yunnan are the source in two consecutive years. While the populations from ICP were not sampled in Chen et al. (2021)'s study, previous analysis of many insect pests have found this region to be the source of insect immigration in China based on trajectory modeling [e.g., *Nilaparvata lugens* (Stål) (Hu et al., 2014); *Sogatella furcifera* (Sun et al., 2017); *Spodoptera frugiperda* (Li et al., 2020)] and genetic analysis [e.g., *Nilaparvata lugens* (Stål) (Hu et al., 2019)]. The year-round suitable temperature in Yunnan and Indo-China Peninsula (e.g., northern and central Vietnam, Laos, and northeastern Thailand) may be the reason for maintaining a large number of populations ready for migration. We thus suggest that these regions should not only be important for effective control of rice pests (Hu et al., 2014; Sun et al., 2017), but also for the management of other important vegetable pests, such as *P. xylostella*. We additionally found several regions such as Fujian and Guangdong were always being the receivers in these regions, a pattern that was consistent over 2 years. These regions may be in the migration corridor for *P. xylostella*'s migration, similar to Beihuang Island in Bohai gulf (Fu et al., 2014). Nevertheless, this pattern should be verified by more dense sampling of populations from both geographical and temporal scales as it was contradicted our the general understanding of monsoon circulation in East Asia (Hu et al., 2019).

Incorporation of contemporary gene flow and population network in pest management

Current insect ecology has witnessed a boom in population genetics/genomics studies of agricultural pests [www.mdpi.com/journal/insects/special_issues/population_genetics, e.g., Sethuraman et al. (2020)]. Research has advanced our understanding of insect genetic variation and differentiation under globalization. In insect ecology and pest management, choosing the proper scale and right parameters is important for understanding population dynamics and connectivity (Osborne and Woiwod, 2002). Genetic parameters, such as N_m, that represent historical gene flow are unsuitable for partially migrating populations of agricultural insects (Menz et al., 2019). Pedigree-based approaches that identify instant migration events should be a better choice (Martin et al., 2021).

Investigation of contemporary gene flow is important especially for insecticide resistant pest monitoring. For example, Wang et al. (2022) report higher frequency of insecticide resistant alleles in *P. xylostella* populations crossing the Bohai Strait during the spring compared with those in winter. This temporal differentiation of insecticide resistance strains is likely due to variation in insecticide pressure between Southern and Northern China (Wang et al., 2022). Further studies incorporating both neutral markers and selective loci should benefit disentangling factors (e.g. migration or adaptation) contributing to the formation of insecticide resistance. Nevertheless, populations that were consistent source populations, which were identified from population network, should be monitored for their insecticide resistance throughout the year to help define plans for proactive pest management. For sink populations to where insecticide resistant insects migrate, application of the spatially common chemicals should be delayed or avoided to reduce the potential of insecticide-resistance development. In addition, crop rotation such as cultivating non-brassicaceous

vegetables (Li et al., 2016; Ke et al., 2019) could be used to reduce source-sink connection among regions. Contemporary dynamics and population network analysis could be used to identify source-sink connections and regional metapopulation dynamics and thus should benefit effective management of highly dispersive insect pests.

Conclusion

In this research, we investigated the temporal variation in genetic makeup of *P. xylostella* across its overwintering region in Southern China and Southeast Asia. Our results depicted a rapid population turnover and dynamic metapopulation of this highly dispersive pest and highlighted the importance of temporal sampling and population network analysis in understanding contemporary genetic pattern and regional source-sink dynamics of agricultural insect pest. These results are expected to shed light on monitoring of the real-time dynamics of insecticide resistance and identify regional key populations towards more effective and sustainable management in the long term. More broadly, methods of this study could also be applied to other highly dispersive insect pests under modern agricultural regime.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary Material.

Author contributions

SY and MY designed this study. FK and JL performed the experiments and data analysis. FK, LV, SY, and MY prepared the manuscript. All authors have read and approved the manuscript.

Funding

This work was financially supported by the Natural Science Foundation of Fujian Province (2022J06013), the National Natural Science Foundation of China (No. 31972271), the Fujian Agriculture and Forestry University Science and Technology Innovation Fund Project (Nos. CXZX2019001G and CXZX2017206), and the Outstanding Young Scientific Research Talents Program of Fujian Agriculture and Forestry University (No. xjq201905).

Acknowledgments

We thanked Dr. Jinliang Wang from Institute of Zoology, Zoological Society of London, for his help and discussion during the data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.986724/full#supplementary-material

References

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22, 341–345. doi:10.1093/bioinformatics/bti803

Broquet, T., Yearsley, J., Hirzel, A. H., Goudet, J., and Perrin, N. (2009). Inferring recent migration rates from individual genotypes. *Mol. Ecol.* 18, 1048–1060. doi:10.1111/j.1365-294X.2008.04058.x

Cao, L. J., Gao, Y. F., Gong, Y. J., Chen, J. C., Chen, M., Hoffmann, A., et al. (2019). Population analysis reveals genetic structure of an invasive agricultural thrips pest related to invasion of greenhouses and suitable climatic space. *Evol. Appl.* 12, 1868–1880. doi:10.1111/eva.12847

Chapman, J. W., Reynolds, D. R., Smith, A. D., Riley, J. R., Pedgley, D. E., and Woiwod, I. P. (2002). High-altitude migration of the diamondback moth Plutella xylostella to the UK: A study using radar, aerial netting, and ground trapping. *Ecol. Entomol.* 27, 641–650. doi:10.1046/j.1365-2311.2002.00472.x

Chen, M. Z., Cao, L. J., Li, B. Y., Chen, J. C., Gong, Y. J., Yang, Q., et al. (2021a). Migration trajectories of the diamondback moth Plutella xylostella in China inferred from population genomic variation. *Pest Manag. Sci.* 77, 1683–1693. doi:10.1002/ps.6188

Chen, Y., Liu, Z., Régnière, J., Vasseur, L., Lin, J., Huang, S., et al. (2021b). Large-scale genome-wide study reveals climate adaptive variability in a cosmopolitan pest. *Nat. Commun.* 12, 7206–7211. doi:10.1038/s41467-021-27510-2

Chown, S., and Crafford, J. (1987). Plutella-xylostella l (lepidoptera, plutellidae) on marion island. *J. Entomological Soc. South. Afr.* 50, 259–260.

Coulson, S., Hodkinson, I., Webb, N., Mikkola, K., Harrison, J., and Pedgley, D. (2002). Aerial colonization of high arctic islands by invertebrates: The diamondback moth Plutella xylostella (Lepidoptera: Yponomeutidae) as a potential indicator species. *Divers. Distributions* 8, 327–334. doi:10.1046/j.1472-4642.2002.00157.x

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex Syst.* 1695, 1–9.

Dinsdale, A., Schellhorn, N., De Barro, P., Buckley, Y., and Riginos, C. (2012). Rapid genetic turnover in populations of the insect pest Bemisia tabaci Middle East: Asia Minor 1 in an agricultural landscape. *Bull. Entomol. Res.* 102, 539–549. doi:10.1017/S0007485312000077

Dosdall, L., Mason, P., Olfert, O., Kaminski, L., and Keddie, B. (2001). "The origins of infestations of diamondback moth, Plutella xylostella (L.), in canola in western Canada," in *The management of diamondback moth and other crucifer pests* (Proceedings of the Fourth International Workshop), 26–29.

Dyer, R. J. (2015). Population graphs and landscape genetics. *Annu. Rev. Ecol. Evol. Syst.* 46, 327–342. doi:10.1146/annurev-ecolsys-112414-054150

Earl, D. A., and Vonholdt, B. M. (2012). Structure harvester: A website and program for visualizing STRUCTURE output and implementing the evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7

Elameen, A., Klütsch, C. F., Fløystad, I., Knudsen, G. K., Tasin, M., Hagen, S. B., et al. (2020). Large-scale genetic admixture suggests high dispersal in an insect pest, the apple fruit moth. *PloS one* 15, e0236509. doi:10.1371/journal.pone.0236509

Emery, A., Wilson, I., Craig, S., Boyle, P. R., and Noble, L. R. (2001). Assignment of paternity groups without access to parental genotypes: Multiple mating and developmental plasticity in squid. *Mol. Ecol.* 10, 1265–1278. doi:10.1046/j.1365-294x.2001.01258.x

Esselink, G., Den Belder, E., Elderson, J., and Smulders, M. (2006). Isolation and characterization of trinucleotide repeat microsatellite markers for Plutella xylostella L. *Mol. Ecol. Notes* 6, 1246–1248. doi:10.1111/j.1471-8286.2006.01504.x

Everett, M. G., and Borgatti, S. P. (2005). Extending centrality. *Models methods Soc. Netw. analysis* 35, 57–76. doi:10.1017/cbo9780511811395.004

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi:10.1002/joc.5086

Fu, X., Xing, Z., Liu, Z., Ali, A., and Wu, K. (2014). Migration of diamondback moth, Plutella xylostella, across the Bohai Sea in northern China. *Crop Prot.* 64, 143–149. doi:10.1016/j.cropro.2014.06.021

Furlong, M. J., Wright, D. J., and Dosdall, L. M. (2013). Diamondback moth ecology and management: Problems, progress, and prospects. *Annu. Rev. Entomol.* 58, 517–541. doi:10.1146/annurev-ento-120811-153605

Gao, B., Hedlund, J., Reynolds, D. R., Zhai, B., Hu, G., and Chapman, J. W. (2020). The 'migratory connectivity'concept, and its applicability to insect migrants. *Mov. Ecol.* 8, 48–13. doi:10.1186/s40462-020-00235-5

Goudet, J. (2001). FSTAT, a program to estimate and test gene diversities and fixation indices version 2.9. 3. AvaliableAt: http://www2.unil.ch/popgen/softwares/fstat.html.

Guo, J., Fu, X., Zhao, S., Shen, X., Wyckhuys, K. A., and Wu, K. (2020). Long-term shifts in abundance of (migratory) crop-feeding and beneficial insect species in northeastern Asia. *J. Pest Sci. (2004).* 93, 583–594. doi:10.1007/s10340-019-01191-9

Hu, G., Lu, F., Zhai, B.-P., Lu, M.-H., Liu, W.-C., Zhu, F., et al. (2014). Outbreaks of the brown planthopper Nilaparvata lugens (stål) in the yangtze river delta: Immigration or local reproduction? *PLoS One* 9, e88973. doi:10.1371/journal.pone.0088973

Hu, G., Lu, M.-H., Reynolds, D. R., Wang, H.-K., Chen, X., Liu, W.-C., et al. (2019a). Long-term seasonal forecasting of a major migrant insect pest: The brown planthopper in the lower yangtze river valley. *J. Pest Sci.* 92, 417–428. doi:10.1007/s10340-018-1022-9

Hu, Q.-L., Zhuo, J.-C., Ye, Y.-X., Li, D.-T., Lou, Y.-H., Zhang, X.-Y., et al. (2019b). Whole genome sequencing of 358 brown planthoppers uncovers the landscape of their migration and dispersal worldwide. bioRxiv. 798876.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi:10.1111/j.1755-0998.2009.02591.x

Jiang, X.-F., Cao, W.-J., Zhang, L., and Luo, L.-Z. (2010). Beet webworm (Lepidoptera: Pyralidae) migration in China: Evidence from genetic markers. *Environ. Entomol.* 39, 232–242. doi:10.1603/EN08315

Jones, A. G., Small, C. M., Paczolt, K. A., and Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Mol. Ecol. Resour.* 10, 6–30. doi:10.1111/j.1755-0998.2009.02778.x

Jones, O. R., and Wang, J. (2010). Colony: A program for parentage and sibship inference from multilocus genotype data. *Mol. Ecol. Resour.* 10, 551–555. doi:10.1111/j.1755-0998.2009.02787.x

Ke, F. S., You, S. J., Huang, S. M., Liu, T. S., Xie, D. D., and You, M. S. (2019). Spatiotemporal dynamics of genetic variation in populations of the diamondback moth, Plutella xylostella (Lepidoptera:Plutellidae), in China. *Acta Entomol. Sin.* 62, 624–633.

Ke, F., You, S., He, W., Liu, T., Vasseur, L., Douglas, C. J., et al. (2015). Genetic differentiation of the regional Plutella xylostella populations across the Taiwan Strait based on identification of microsatellite markers. *Ecol. Evol.* 5, 5880–5891. doi:10.1002/ece3.1850

Kirk, H., Dorn, S., and Mazzi, D. (2013). Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evol. Appl.* 6, 842–856. doi:10.1111/eva.12071

Kohno, K., Soemori, H., and Takahashi, K. (2004). Seasonal occurrence of Plutella xylostella (Lepidoptera: Yponomeutidae) on Ishigaki-jima Island, with special reference to their sudden occurrence associated with a typhoon. *Appl. Entomol. Zool.* 39, 119–125. doi:10.1303/aez.2004.119

Li, X.-J., Wu, M.-F., Ma, J., Gao, B.-Y., Wu, Q.-L., Chen, A.-D., et al. (2020). Prediction of migratory routes of the invasive fall armyworm in eastern China using a trajectory analytical approach. *Pest Manag. Sci.* 76, 454–463. doi:10.1002/ps.5530

Li, Z., Feng, X., Liu, S.-S., You, M., and Furlong, M. J. (2016a). Biology, ecology, and management of the diamondback moth in China. *Annu. Rev. Entomol.* 61, 277–296. doi:10.1146/annurev-ento-010715-023622

Li, Z. Y., Chen, H. Y., Bao, H. L., Zhen-Di, H. U., Yin, F., Lin, Q. S., et al. (2016b). Progress in research on managing regional pesticide resistance in the diamondback moth in China. *Chin. J. Appl. Entomology* 53, 247–255.

Ma, C. S., Peng, Y., Zhao, F., Chang, X., Xing, K., Zhu, L., et al. (2021). Climate warming promotes pesticide resistance through expanding overwintering range of a global pest. *Nat. Commun.* 12, 5351–5410. doi:10.1038/s41467-021-25505-7

Ma, C., Yang, P., Fau - Jiang, F., Jiang F Fau - Chapuis, M.-P., Chapuis Mp Fau - Shali, Y., Y Fau - Sword, ShaliG. A., et al. (2012). Mitochondrial genomes reveal the global phylogeography and dispersal routes of the migratory locust. *Mol. Ecol.* 21, 4344–4358. doi:10.1111/j.1365-294X.2012.05684.x

Martin, S., Lipps, G., and Gibbs, H. (2021). Pedigree-based assessment of recent population connectivity in a threatened rattlesnake. *Mol. Ecol. Resour.* 21, 1820–1832. doi:10.1111/1755-0998.13383

Menz, M. H. M., Reynolds, D. R., Gao, B., Hu, G., Chapman, J. W., and Wotton, K. R. (2019). Mechanisms and consequences of partial migration in insects. *Front. Ecol. Evol.* 7, 403. doi:10.3389/fevo.2019.00403

Miyata, A. (1983). *Handbook of the moth ecology-moth as an indicator of the environment*. NagasakiDiv: Showado Printing Publ.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8577–8582. doi:10.1073/pnas.0601602103

Osborne, J. L, L. H. D., and Woiwod, I. P. (2002). "Monitoring insect dispersal: Methods and approaches," in *Dispersal ecology: The 42nd symposium of the British ecological society* (University of Reading).

Peakall, R., and Smouse, P. E. (2006). Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi:10.1111/j.1471-8286.2005.01155.x

Perera, O. P., Fescemyer, H. W., Fleischer, S. J., and Abel, C. A. (2020). Temporal variation in genetic composition of migratory Helicoverpa zea in peripheral populations. *Insects* 11, 463. doi:10.3390/insects11080463

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi:10.1093/genetics/155.2.945

Samarasin, P., Shuter, B. J., Wright, S. I., and Rodd, F. H. (2017). The problem of estimating recent genetic connectivity in a changing world. *Conserv. Biol.* 31, 126–135. doi:10.1111/cobi.12765

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* 18, 233–234. doi:10.1038/72708

Sethuraman, A., Janzen, F. J., Weisrock, D. W., and Obrycki, J. J. (2020). Insights from population genomics to enhance and sustain biological control of insect pests. *Insects* 11, 462. doi:10.3390/insects11080462

Sun, S.-S., Bao, Y.-X., Wu, Y., Lu, M.-H., and Tuan, H.-A. (2017). Analysis of the huge immigration of Sogatella furcifera (Hemiptera: Delphacidae) to southern China in the spring of 2012. *Environ. Entomol.* 47, 8–18. doi:10.1093/ee/nvx181

Swift, M. L. (1997). GraphPad prism, data analysis, and scientific graphing. *J. Chem. Inf. Comput. Sci.* 37, 411–412. doi:10.1021/ci960402j

Szpiech, Z. A., Jakobsson, M., and Rosenberg, N. A. (2008). Adze: A rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24, 2498–2504. doi:10.1093/bioinformatics/btn478

Taylor, K. L., Hamby, K. A., Deyonke, A. M., Gould, F., and Fritz, M. L. (2021). Genome evolution in an agricultural pest following adoption of transgenic crops. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2020853118. doi:10.1073/pnas.2020853118

Tiilikkala, K. V. A., Koistinen, J., and Salonoja, M. (1996). *Remote sensing in AgricultureRemote sensing in agriculture*. Journal of Agricultural Research Centre of Finland, 78–81.

Wang, J. (2018). Estimating genotyping errors from genotype and reconstructed pedigree data. *Methods Ecol. Evol.* 9, 109–120. doi:10.1111/2041-210x.12859

Wang, J. (2014). Estimation of migration rates from marker-based parentage analysis. *Mol. Ecol.* 23, 3191–3213. doi:10.1111/mec.12806

Wang, J., and Santure, A. W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* 181, 1579–1594. doi:10.1534/genetics.108.100214

Wang, M., Zhu, B., Zhang, L., Xiao, Y., Liang, P., and Wu, K. (2022). Influence of seasonal migration on evolution of insecticide resistance in Plutella xylostella. *Insect Sci.* 29, 496–504. doi:10.1111/1744-7917.12987

Wei, S.-J., Shi, B.-C., Gong, Y.-J., Jin, G.-H., Chen, X.-X., and Meng, X.-F. (2013). Genetic structure and demographic history reveal migration of the diamondback moth *Plutella xylostella* (Lepidoptera: Plutellidae) from the southern to northern regions of China. *PLOS ONE* 8, e59654. doi:10.1371/journal.pone.0059654

Whitlock, M. C., and Mccauley, D. E. (1999). Indirect measures of gene flow and migration: F_{ST}≠1/(4Nm+1). *Heredity* 82, 117–125. doi:10.1038/sj.hdy.6884960

Wilson, G., and Rannala, B. (2002). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163, 1177–1191. doi:10.1093/genetics/163.3.1177

Yang, X.-M., Sun, J.-T., Xue, X.-F., Li, J.-B., and Hong, X.-Y. (2012). Invasion genetics of the western flower thrips in China: Evidence for genetic bottleneck, hybridization and bridgehead effect. *PLOS ONE* 7, e34567. doi:10.1371/journal.pone.0034567

You, M., Ke, F., You, S., Wu, Z., Liu, Q., He, W., et al. (2020). Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nat. Commun.* 11, 2321. doi:10.1038/s41467-020-16178-9

You, M. S. W., H. (2007). *The research of diamondback moth*. Beijing: China Agriculture Press.

Zhou, X.-Y., Wu, Q.-L., Jia, H.-R., and Wu, K.-M. (2021). Searchlight trapping reveals seasonal cross-ocean migration of fall armyworm over the South China Sea. *J. Integr. Agric.* 20, 673–684. doi:10.1016/s2095-3119(20)63588-2

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 21 September 2022
DOI 10.3389/fgene.2022.1013192

OPEN ACCESS

EDITED BY
Shengqian Xia,
The University of Chicago, United States

REVIEWED BY
Xi Xia,
Shenzhen Hospital, Peking University, China
Yue Zhao,
Peking University Third Hospital, China

*CORRESPONDENCE
Yaoyao Zhang,
yaoyaozhang@scu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 06 August 2022
ACCEPTED 22 August 2022
PUBLISHED 21 September 2022

CITATION
Guo W, Dai M, Zhong Z, Zhu S, Gong G, Chen M, Guo J and Zhang Y (2022), The association between vitamin D and uterine fibroids: A mendelian randomization study.
Front. Genet. 13:1013192.
doi: 10.3389/fgene.2022.1013192

The association between vitamin D and uterine fibroids: A mendelian randomization study

Weijie Guo[1,2†], Mengyuan Dai[1,2†], Zhuoling Zhong[1,2], San Zhu[2], Guidong Gong[1], Mei Chen[1], Junling Guo[1,3] and Yaoyao Zhang[1,2*]

[1]Department of Obstetrics and Gynecology of West China Second University Hospital, BMI Center for Biomass Materials and Nanointerfaces, College of Biomass Science and Engineering, Sichuan University, Chengdu, Sichuan, China, [2]Key Laboratory of Birth Defects and Related of Women and Children of Ministry of Education, West China Second University Hospital, Sichuan University, Chengdu, China, [3]State Key Laboratory of Polymer Materials Engineering, Sichuan University, Chengdu, Sichuan, China

Uterine fibroids (UFs), the most common benign gynecological tumor, can bring severe negative impacts on a woman's life quality. Vitamin D, is thought to play an important role in regulating cell proliferation and differentiation. In recent years, several studies suggested that higher level of vitamin D has a negative effect on the occurrence of UFs, but the results of studies on the relationship between them are conflicting and further evidence needs to be studied. Here in, we used a two-sample Mendelian Randomization (2SMR) study to explore the causal relationship between genetically predicted vitamin D levels and the risk of UFs. The exposure data comes from a genome-wide association study (GWAS) summary dataset consisting of 441,291 individuals, which includes datasets from United Kingdom Biobank, FinnGen Biobank and the corresponding consortia. Single-nucleotide polymorphisms (SNPs) associated with vitamin D at a significant level of $p < 5 \times 10^{-8}$ and low linkage disequilibrium (LD) level ($r^2 < 0.01$) were selected. The outcome data comes from a GWAS dataset of IEU analysis of United Kingdom Biobank phenotypes consisting of 7,122 UFs cases and 455,811 controls. Our inverse-variance weight (IVW) analysis results support the causal association of genetically predicted vitamin D with the risk of UFs (OR = 0.995,95% CI = 0.990-0.999, $p = 0.024$). In addition, heterogeneity and pleiotropy were not observed in statistical models. In summary, our results indicate that elevated serum vitamin D levels are in strong relationship with reduction of the risk of UFs, which indicates that the clinical treatment of UFs may have a new and excellent option.

KEYWORDS

vitamin D, uterine fibroids, SNP, mendelian randomization, GWAS

Introduction

Uterine fibroids (UFs), also known as uterine leiomyomas, are benign tumors that negatively affect the function of the uterus in women of childbearing age. The most common symptom of UFs is heavy menstrual bleeding and the resulting anemia and pain (Kempson and Hendrickson, 2000). Other symptoms include pelvic pressure and pain, urinary incontinence and retention, and bowel dysfunction also place a significant burden on patients. UFs may also cause several reproductive problems, such as impaired fertility, pregnancy complications, miscarriage and adverse pregnancy outcomes (Wallach et al., 1981; Zimmermann et al., 2012; Khan et al., 2014; Vercellini and Frattaruolo, 2017). A study pointed out that UFs account for 1/3 to 1/2 of the reasons for hysterectomy and are the most common reason of hysterectomy in the United States (Stewart et al., 2016). Risk factors for UFs have received increasing attention in order to better prevent the occurrence of UFs. Many studies indicated that vitamin D deficiency increases the risk of UFs (Baird et al., 2013; Sabry et al., 2013; Ciebiera et al., 2016). There is a clinical study indicating that vitamin D supplementation has a significant therapeutic effect in patients with small UFs (Ciavattini et al., 2016). However, another clinical trial suggested that vitamin D levels have no significant effect on the occurrence of UFs (Arjeh et al., 2020). Overall, the relationship between vitamin D levels and the risk of UFs is still ambiguous, and further studies are needed to be carried out.

Vitamin D, a fat-soluble vitamin, is a general term for a group of structurally similar sterol derivatives. The target organs of vitamin D are widely distributed, and different type of vitamin D binds to their specific receptors and then play different roles (Lips, 2006). Vitamin D3 and vitamin D2 are the most important members of vitamin D (Ciebiera et al., 2018), and vitamin D3 is the main form of vitamin D in the human body. The level of vitamin D3 in serum can represent the total content of vitamin D in the body and the strength of the effect of vitamin D on the human body (Lips, 2006). Therefore, in the work, we focus on the level of vitamin D3 in serum. vitamin D3 has an anti-proliferative effect and can accelerate the release of tumor necrosis factors from macrophages, which has a broad killing effect on tumor cells (Van Den Bemd et al., 2000; Lips, 2006). Several recent studies pointed out that vitamin D deficiency is an important risk factor for UFs (Baird et al., 2013; Sabry et al., 2013; Ciebiera et al., 2016), and animal experiments have also shown that high doses of vitamin D can decrease the size of UFs (Al-Hendy and Badr, 2014; Ali et al., 2020), but the relationship between vitamin D and the pathogenesis of UFs or a positive and significant treatment effect remains unclear.

Mendelian randomization (MR) studies are conducted based on Mendel's laws of inheritance and the use of instrumental variables (IVs). Mendelian laws of inheritance state that genes are randomly assigned and freely selected in the process of inheritance, instrumental variables are related to the risk factors we are interested in but not related to other confounding factors, and its effect on the outcome can only be determined by the exposure factor (Burgess and Thomson, 2015). Therefore, in the two-sample MR study, we use genetic variables to analyze the causal relationship between exposure factors and outcomes, typically single-nucleotide polymorphisms (SNPs). Due to the genes assigned during pregnancy, the direction of the causal relationship can also be determined. The association between serum vitamin D3 levels and the occurrence of UFs has not previously been studied using MR. In this study, we focused on exploring the causal relationship between serum vitamin D3 levels and the occurrence of UFs.

Materials and methods

Study design

In a Mendelian randomization study, to obtain reliable results, genetic variables as instrumental variables must satisfy three assumptions (Figure 1): 1) Genetic variation is associated with exposure factors; 2) Genetic variation is not associated with confounders; 3) Genetic variation only influences the outcome by exposure factors. The dashed line in the figure indicates that the pathway is not allowed, and the solid line indicates the ideal pathway. The second and third assumptions are collectively referred to as independence from pleiotropy. Pleiotropy refers to genetic variation that affects outcomes through pathways independent of risk factors. Investigators need to use sensitivity analysis to make judgments (Emdin et al., 2017).

Genetic association dataset for Vitamin D

SNPs associated with vitamin D were selected from a genome-wide association study (GWAS) summary dataset, including 441,291 samples from European populations (Hemani et al., 2018), and the data were derived from research data including the United Kingdom biobank, FinnGen Biobank and other consortia, which were manually collected and organized into a summary dataset for use in MR studies. Independent SNPs that were related to vitamin D at the genome-wide significance level ($p < 5 \times 10^{-8}$) and low linkage disequilibrium (LD) level ($r^2 < 0.01$) were selected. In order to avoid the possible bias caused by weak instrumental variables, we use the F-statistic to judge the strength of the instrumental variable (Burgess et al., 2011). According to experience, the F-statistic should be at least 10 (Staiger and Stock, 1997).

Genetic instrumental vasriables for uterine fibroids

In this MR study, the occurrence of UFs was our outcome. Data for the outcomes were derived from the

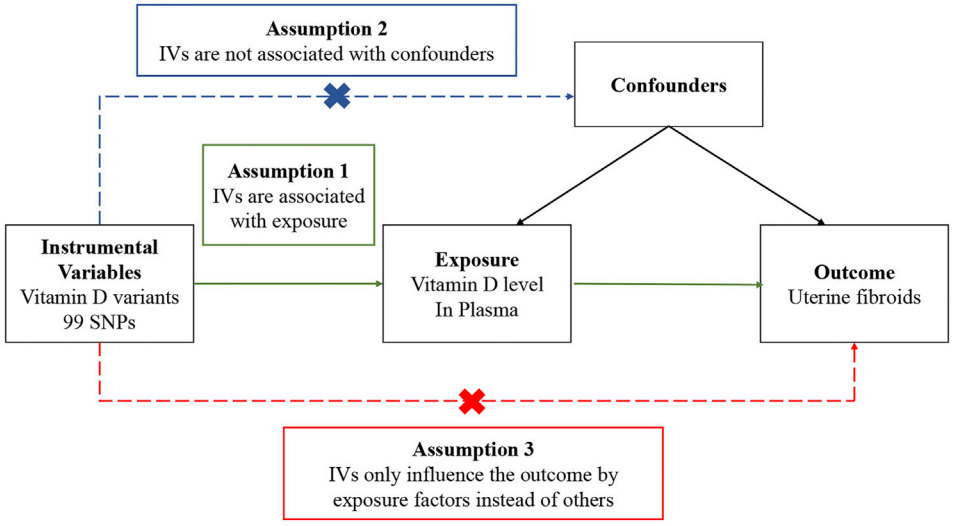


FIGURE 1
Overview of the design and three key assumptions of the Mendelian randomization study. IVs, instrument variables; SNPs, single-nucleotide polymorphisms.

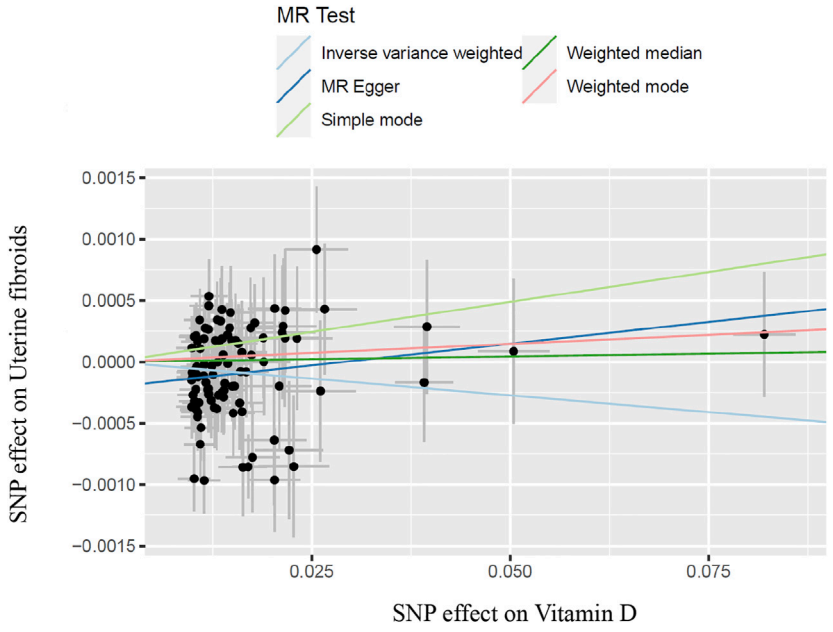


FIGURE 2
Scatter plot of SNPs with vitamin D and uterine fibroids and results of different test models; SNP, single-nucleotide polymorphism.

dataset provided to GWAS by the United Kingdom Biobank (Mitchell et al., 2019), containing 462,933 samples from European populations, of which 7122 were reported as UFs without other cancers. In this study, we extracted the effect estimates and standard errors for each of the 99 vitamin D–related SNPs from the GWAS summary statistics of UFs.

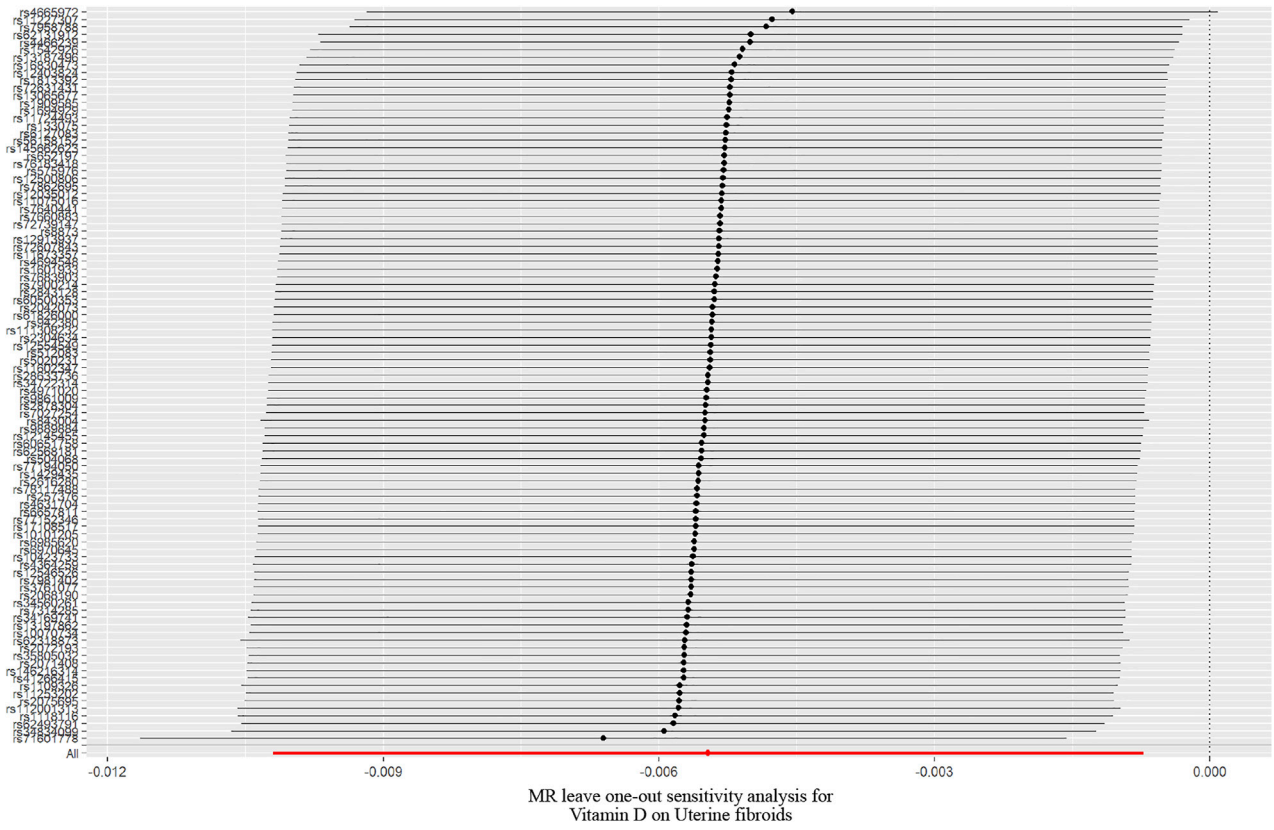


FIGURE 3
A leave one-out pleiotropy test was performed for each SNP to determine the robustness of the MR test; SNP, single-nucleotide polymorphism.

Statistical analysis

MR analysis of the association between vitamin D and the occurrence of UFs was performed using 99 SNPs associated with 25-hydroxyvitamin D levels as IVs. We primarily performed MR analyses using Inverse-variance Weights (IVW) with random effects to estimate odds ratios (OR) and 95% confidence intervals (CI) for the occurrence of UFs (Burgess and Bowden, 2015).

We then performed a sensitivity analysis to examine heterogeneity and pleiotropy among IVs. MR-Egger regression, weighted median, simple model, and weighted model methods were used to determine whether IVs affected UFs through their effect on vitamin D alone. The slope coefficients of the MR-Egger regressions provided estimates of causal effects, which were used to test for pleiotropic bias (Bowden et al., 2015). Simple medians provide consistent estimates of causal effects if at least 50% of the IVs are valid, but weighted medians provide consistent estimates if at least 50% of the weights come from valid IVs (Bowden et al., 2016). And the weighted mode requires that the largest subset of instruments identifying the same causal effect estimates is contributed by valid IVs so that the result is consistent (Hartwig et al., 2017). We applied MR- Pleiotropy Residual Sum and Outlier (MR-PRESSO) analysis method to analyze the pleiotropy of IVs and correct the possible outliers. In addition, we use Q test on the IVW and MR-Egger to estimate the heterogeneity of the IVs. We also used a leave-one-out sensitivity test to test whether the MR outcome was sensitive to its related IV. MR and sensitivity analyses were performed in R (version 4.2.0) using the Two-Sample MR package (version 0.5.6) and the MRPRESSO package (version 1.0).

Results

SNPs used as instrumental variables

Independent SNPs that were related to 25-hydroxyvitamin D serum levels at the genome-wide significance level ($p < 5 \times 10^{-8}$) and low linkage disequilibrium (LD) level ($r^2 < 0.01$) were selected from the GWAS dataset. Then SNPs with F > 10 were screened from these SNPs and the genes they are in were queried in Pubmed (https://www.ncbi.nlm.nih.gov/), and the genes of SNPs that did not belong to a specific gene were

TABLE 1 Vitamin D SNPs used to construct the instrument variable.

Chr	Position	SNP	Effect Allele	Other Allele	EAF	Beta	SE	Gene	*p* Value	F Statistics
1	41750648	rs12035012	A	C	0.222	0.014	0.002	HIVEP3	1.00E-200	882.53
1	62835936	rs12145455	C	T	0.095	0.019	0.003	ATG4C	1.00E-200	296.861
1	24609753	rs12403824	G	C	0.308	0.011	0.002	NULL	1.00E-200	1023.579
1	2315680	rs2843128	G	A	0.516	0.010	0.002	MORN1	1.00E-200	1319.309
1	151502427	rs34834099	T	C	0.066	0.026	0.004	CGN	1.00E-200	204.06
1	34684617	rs41266415	T	A	0.215	0.013	0.002	C1orf94	1.00E-200	790.3233
1	230293530	rs4631704	T	C	0.608	0.011	0.002	GALNT2	1.00E-200	1295.836
1	150522242	rs4971020	C	T	0.647	0.011	0.002	ADAMTSL4	1.00E-200	1162.396
1	46027355	rs512083	C	T	0.461	0.010	0.002	MAST2	1.00E-200	1309.729
1	62898984	rs60500353	T	C	0.156	0.015	0.003	LOC105378768	1.00E-200	541.4045
1	179423953	rs61826000	C	G	0.338	0.011	0.002	AXDND1	1.00E-200	1136.723
1	109807283	rs6657811	T	A	0.131	0.015	0.003	CELSR2	1.00E-200	425.2116
1	151037707	rs77152346	C	T	0.179	0.013	0.003	BNIPL	1.00E-200	583.828
2	21271707	rs34722314	A	T	0.136	0.017	0.003	NULL	1.00E-200	480.9317
2	27598097	rs4665972	C	T	0.606	0.017	0.002	ZNF512	1.00E-200	2042.473
3	52321788	rs13065677	T	C	0.051	0.023	0.004	DNAH1	1.00E-200	112.1106
3	153561145	rs1542926	C	T	0.117	0.016	0.003	NULL	1.00E-200	359.3695
3	173504091	rs16830473	C	T	0.091	0.018	0.003	NLGN1	1.00E-200	251.7777
3	85002871	rs1694929	T	C	0.430	0.010	0.002	CADM2	1.00E-200	1256.449
3	124687460	rs1909585	T	C	0.346	0.011	0.002	KALRN	1.00E-200	1109.83
3	47352998	rs76183418	C	T	0.179	0.013	0.003	KLHL18	1.00E-200	571.6435
3	125118082	rs7640441	A	C	0.241	0.013	0.002	SLC12A8	1.00E-200	922.0544
3	141654685	rs9861009	C	T	0.722	0.012	0.002	NULL	1.00E-200	1047.055
4	72501807	rs112001313	T	C	0.060	0.040	0.004	ADAMTS3	1.00E-200	260.385
4	71694520	rs11724493	C	T	0.056	0.022	0.004	GRSF1	1.00E-200	128.7095
4	69475763	rs12500806	T	C	0.613	0.012	0.002	NULL	1.00E-200	1064.943
4	100510550	rs145662623	A	G	0.061	0.020	0.004	EMCN	1.00E-200	144.4778
4	72179821	rs146216314	A	G	0.098	0.020	0.003	SLC4A4	1.00E-200	316.7743

(Continued on following page)

TABLE 1 (*Continued*) Vitamin D SNPs used to construct the instrument variable.

Chr	Position	SNP	Effect Allele	Other Allele	EAF	Beta	SE	Gene	*p* Value	F Statistics
4	70054650	rs28633736	C	T	0.149	0.016	0.003	HTN1	1.00E-200	531.4226
4	74458987	rs34169741	T	C	0.383	0.013	0.002	RASSF6	1.00E-200	1609.023
4	15892159	rs4364259	A	G	0.205	0.016	0.002	NULL	1.00E-200	879.1162
4	73681946	rs4694548	G	A	0.275	0.015	0.002	NULL	1.00E-200	1261.403
4	71489270	rs5020231	C	T	0.779	0.013	0.002	SLC4A4	1.00E-200	777.8031
4	72745430	rs62318873	T	C	0.052	0.050	0.005	NULL	1.00E-200	240.7028
4	72585683	rs71601778	C	T	0.074	0.082	0.004	NULL	1.00E-200	747.2279
4	73266887	rs72607843	T	C	0.329	0.012	0.002	ADAMTS3	1.00E-200	1214.048
4	87982876	rs7660883	G	C	0.378	0.012	0.002	SPP1	1.00E-200	1355.003
4	72396727	rs7683903	A	G	0.154	0.014	0.003	ADAMTS3	1.00E-200	506.9613
4	72599352	rs843004	G	A	0.076	0.039	0.004	NULL	1.00E-200	404.4502
5	87940026	rs10070734	C	T	0.704	0.012	0.002	LINC00461	1.00E-200	1097.487
5	118613707	rs13187496	G	T	0.345	0.011	0.002	LOC102467225	1.00E-200	1188.506
5	148007013	rs2068190	A	G	0.437	0.010	0.002	HTR4	1.00E-200	1313.18
6	80014585	rs13197862	A	G	0.130	0.015	0.003	TTK	1.00E-200	397.857
6	22755139	rs4466239	G	A	0.624	0.011	0.002	NULL	1.00E-200	1261.93
6	121854778	rs942380	G	A	0.595	0.011	0.002	NULL	1.00E-200	1407.937
7	106799997	rs257376	A	G	0.542	0.010	0.002	PRKAR2B	1.00E-200	1298.439
7	100798274	rs6970645	G	C	0.752	0.011	0.002	AP1S1	1.00E-200	835.7355
8	106470630	rs10101205	C	T	0.840	0.014	0.003	OXR1	1.00E-200	515.8738
8	143587121	rs12546526	C	T	0.859	0.015	0.003	EEF1D	1.00E-200	437.9549
8	30835535	rs2042073	G	A	0.604	0.010	0.002	TEX15	1.00E-200	1220.945
8	9168897	rs62493791	G	T	0.242	0.012	0.002	NULL	1.00E-200	841.1148
8	59370159	rs6985620	C	T	0.665	0.011	0.002	NULL	1.00E-200	1119.15
9	112239077	rs12554549	T	C	0.065	0.021	0.004	PTBP3	1.00E-200	159.146
9	107645674	rs62568181	C	T	0.104	0.016	0.003	ABCA1	1.00E-200	295.9726
9	125605840	rs7027254	C	T	0.147	0.014	0.003	MAPKAP1	1.00E-200	470.4589

(Continued on following page)

TABLE 1 (*Continued*) Vitamin D SNPs used to construct the instrument variable.

Chr	Position	SNP	Effect Allele	Other Allele	EAF	Beta	SE	Gene	*p* Value	F Statistics
9	35766116	rs7862695	T	C	0.411	0.010	0.002	NULL	1.00E-200	1245.768
10	5530385	rs11253202	C	T	0.211	0.012	0.002	NULL	1.00E-200	708.6447
10	118394551	rs2286779	C	G	0.519	0.010	0.002	PNLIPRP2	1.00E-200	1316.9
10	81965655	rs7900214	A	G	0.279	0.012	0.002	NRG3	1.00E-200	989.5996
11	15182597	rs1109326	T	G	0.542	0.014	0.002	INSC	1.00E-200	1859.66
11	13799056	rs111308232	A	G	0.054	0.026	0.004	NULL	1.00E-200	135.6992
11	70988410	rs1118116	G	A	0.205	0.018	0.002	SHANK2	1.00E-200	1018.199
11	65581135	rs11227307	A	G	0.647	0.012	0.002	EHBP1L1	1.00E-200	1280.435
11	2176852	rs11602347	G	C	0.407	0.010	0.002	INS-IGF2	1.00E-200	1227.656
11	117008946	rs504068	C	T	0.820	0.014	0.003	SIK3	1.00E-200	605.7437
11	75437630	rs575976	G	A	0.184	0.013	0.003	GDPD5	1.00E-200	612.6741
11	76469378	rs60651758	A	G	0.101	0.017	0.003	C11orf30	1.00E-200	305.5616
11	71849741	rs652197	T	C	0.864	0.016	0.003	FOLR3	1.00E-200	454.6313
12	38692203	rs1813392	C	T	0.537	0.010	0.002	CPNE8	1.00E-200	1368.937
12	111522026	rs7314285	G	T	0.068	0.022	0.004	ATXN2	1.00E-200	181.7633
12	96089917	rs78117488	T	C	0.062	0.021	0.004	NTN4	1.00E-200	145.9469
12	33692179	rs7958788	C	T	0.377	0.010	0.002	NULL	1.00E-200	1187.48
12	57979949	rs8873	A	G	0.223	0.012	0.002	KIF5A	1.00E-200	756.2545
13	60676803	rs7981402	A	G	0.340	0.010	0.002	LINC00378	1.00E-200	1112.565
14	39356756	rs17108517	A	G	0.056	0.021	0.004	LINC00639	1.00E-200	127.1741
14	103987078	rs2071408	A	G	0.366	0.012	0.002	TDRD9	1.00E-200	1306.471
15	77316131	rs12913937	A	G	0.354	0.010	0.002	PEAK1	1.00E-200	1137.209
15	58671559	rs1601933	T	C	0.468	0.014	0.002	ADAM10	1.00E-200	1793.52
15	90734426	rs34560261	T	C	0.170	0.014	0.003	BLM	1.00E-200	542.6187
15	58571401	rs72739147	T	A	0.130	0.016	0.003	NULL	1.00E-200	428.8216
16	11901557	rs11075016	G	A	0.285	0.012	0.002	GSPT1	1.00E-200	1053.374
16	72698702	rs1429435	G	C	0.942	0.022	0.004	LINC01572	1.00E-200	127.5589
16	4500544	rs2304634	T	C	0.686	0.011	0.002	HMOX2	1.00E-200	1032.636

(Continued on following page)

TABLE 1 (*Continued*) Vitamin D SNPs used to construct the instrument variable.

Chr	Position	SNP	Effect Allele	Other Allele	EAF	Beta	SE	Gene	*p* Value	F Statistics
16	30878366	rs2878304	T	C	0.726	0.011	0.002	BCL7C	1.00E-200	955.8311
16	84734147	rs56158152	T	G	0.355	0.011	0.002	USP10	1.00E-200	1112.916
16	89882826	rs72631431	T	C	0.313	0.011	0.002	TCF25	1.00E-200	1021.177
16	70687185	rs77194050	G	A	0.052	0.023	0.004	IL34	1.00E-200	115.9125
17	66394054	rs9889884	C	T	0.756	0.013	0.002	PRKCA	1.00E-200	928.8681
19	11185919	rs10423733	C	T	0.181	0.015	0.003	KANK2	1.00E-200	684.9655
19	54658102	rs11606	G	C	0.426	0.011	0.002	CNOT3	1.00E-200	1333.397
19	53700807	rs11673357	C	T	0.761	0.014	0.003	NULL	1.00E-200	797.2597
19	51518297	rs2075695	G	A	0.560	0.011	0.002	KLK10	1.00E-200	1401.33
19	58348570	rs35805032	T	C	0.155	0.014	0.003	A1BG	1.00E-200	501.5067
19	19325963	rs3761077	T	G	0.111	0.017	0.003	MAU2	1.00E-200	357.4002
19	48323130	rs62131912	T	G	0.101	0.020	0.003	NULL	1.00E-200	351.4042
20	52728499	rs2616280	A	G	0.076	0.019	0.004	NULL	1.00E-200	188.9118
20	52687181	rs6127083	C	G	0.152	0.015	0.003	BCAS1	1.00E-200	531.8351
22	41081164	rs133075	T	G	0.553	0.010	0.002	NULL	1.00E-200	1306.933
22	31533796	rs2072193	C	G	0.061	0.027	0.004	SFI1|	1.00E-200	184.4545

Chr, chromosome; SNP, single-nucleotide polymorphism; EAF, effect allele frequency; SE, standard error.

defined as NULL. The remaining 99 SNPs were included to establish the genetic IVs for vitamin D (Table 1).

Mendelian randomization test results and data visualization

In the work, the IVW method was used to test for causal effects firstly. We found that a one-SD increase in vitamin D levels was associated with a decreased risk of UFs [odds ratio (OR): 0.995, 95% CI: 0.990-0.999, $p = 0.024$]. The result reveals the causal relationship between vitamin D levels and the risk of UFs in the European population. Then we adopted four different models to test and verify the causal relationship between serum vitamin D3 levels and UFs. All of the MR-Egger regression, weighted median, simple model, and weighted model results were opposite to IVW analysis (Table 2). Nevertheless, according to the judgment method of MR test results (Burgess and Thomson, 2015), our results are still able to draw the same conclusion. In addition, there has no bias value between our IVs in the scatter plot of correlation analysis (Figure 2) and the results of the leave-one-out sensitivity test (Figure 3), it illustrates that

TABLE 2 Associations between genetically predicted vitamin D and risk of uterine fibroids.

Methods	OR (95% CI)	*p* Value
Inverse-variance Weighted	0.995 (0.999–0.990)	0.024
MR-Egger	1.007 (0.996–1.019)	0.216
Simple mode	1.010 (0.994–1.026)	0.225
Weighted median	1.001 (0.993–1.008)	0.814
Weighted mode	1.003 (0.993–1.013)	0.541

OR, odds ratio; CI, confidence interval.

TABLE 3 Heterogeneity testing of instrumental variables for vitamin D.

Method	Q	df	*p* Value
Inverse-variance weights	112.125	95	0.111
MR-egger	119.101	96	0.055

OR, odds ratio; CI, confidence interval.

TABLE 4 Pleiotropy testing of instrumental variables for vitamin D.

Method	Intercept	SE	*p* Value
MR-egger	2.05e-4	8.45e-5	0.017

SE, standard error.

the causal relationship between vitamin D levels and the risk of UFs is highly reliable.

Heterogeneity and pleiotropy test results

To remove the possible bias of instrumental variables, heterogeneity test and horizontal pleiotropy test was conducted in the MR study. In sensitivity analysis, there have no heterogeneity was detected in the IVW method or the MR-egger method between IVs ($p > 0.05$) (Table 3). This means that our results are not confounded by other factors between populations grouped by IVs. But the intercept obtained by the MR-egger method was too far from 0, suggesting that there may be horizontal pleiotropy between the IVs ($p < 0.05$) (Table 4). Therefore, we performed multiple operations with MR-PRESSO and found no offset in IVs and no pleiotropy ($p > 0.05$). Furthermore, there have no outliers and horizontal pleiotropy were found after 2000 simulations using MR-PRESSO (Table 5). This indicates that the IVs used in this work impacted the risk of UFs only by affecting serum vitamin D3 levels.

Discussion

The Mendelian randomization study performed an analysis of the causal relationship between vitamin D and UFs based on a summary dataset from the United Kingdom Biobank including 462,933 individuals using multiple SNPs as instrumental variables. Our results revealed a causal relationship between serum vitamin D3 levels and the occurrence of UFs that the reduction of vitamin D levels will increase the risk of UFs.

Our findings are consistent with numerous previous studies. Early animal experiments found that vitamin D supplementation can significantly reduce the volume of UFs (Halder et al., 2010). A subsequent observational experiment showed that differences in serum vitamin D levels were significantly associated with the risk of developing UFs (Sabry et al., 2013), patients with lower vitamin D levels having a higher risk of developing UFs. These studies provide a potentially excellent therapeutic approach for the clinical treatment of UFs, which inspired researchers to further explore the phenomenon.

In the early 2000s, several studies revealed the mechanism of UFs—excessive secretion of compounds from extracellular matrix (ECM) such as collagen and fibers can cause UFs—and recent studies have also confirmed this (Sozen I and Arici, 2002; Rafique et al., 2017). The symptoms of ECM accumulation during the occurrence of UFs is similar to inflammation, which manifested as massive exudation of intracellular material and accumulation of ECM. Some researchers have proposed the possible involvement of inflammation in the development of UFs, and these processes are closely related to the function of vitamin D *in vivo* (Protic et al., 2016).

Vitamin D is a natural active substance, and its receptors are widely distributed *in vivo* and play different functions. Anti-inflammatory and anti-tumor effects are representative functions of Vitamin D (Van Den Bemd et al., 2000; Lips, 2006). Vitamin D generally exerts its biological function by regulating the level of growth factors through various signaling pathways. For example, Vitamin D is involved in the regulation of Wnt/β-catenin and TGF-β pathways (Ciebiera et al., 2017), which play important roles in the anti-inflammation and regulation of cell proliferation (Protic et al., 2016). Overexpression of TGF-β can lead to excessive secretion of ECM by stimulating the synthesis of collagen, proteoglycans, and other ECM compounds, which further induces the occurrence of UFs (Leppert et al., 2004; Ciebiera et al., 2017). Other studies also pointed out that increased vitamin D levels can suppress the cell proliferation and slow down the development of UFs by inhibiting Wnt/β-catenin and TGF-β pathways in the process of culturing UFs *in vitro* (Al-Hendy et al., 2016).

There are a series of factors participate in the regulation of cell proliferation and apoptosis, such as proliferating cell nuclear antigen (PCNA), cyclin-dependent kinase 1 (CDK1), M-phase promoting factor and catechol-O-methyltransferase (COMT), etc., Overexpression of these factors can promote the development of UFs by stimulating cell proliferation, which showing that vitamin D compounds can significantly inhibit the activation of enzymes that regulate factor expression and down-regulate the expression of them (Sharan et al., 2011). Furthermore, a study found that vitamin D can inhibit the expression of estrogen and progesterone receptors, and then suppress estrogen and progesterone perform endocrine functions (Al-Hendy et al., 2015). All of these evidences suggest that vitamin D play an important role in the development of UFs.

The researchers further conducted randomized clinical trials (RCT) to evaluate the effect of vitamin D supplementation on UFs. A clinical study, conducted on patients with small UFs and published in 2016, had reported a positive effect of vitamin D on

TABLE 5 Pleiotropy testing of instrumental variables for vitamin D using MR-PRESSO.

	Exposure	MR Analysis	Casual Estimate	Sd	T-Stat	*p* Value
Main MR	VD	Raw	−5.27e−3	2.39e−3	−2.20	0.03
	VD	Outlier-corrected	NA	NA	NA	NA
RSSobs of Global Test in MR-PRESSO results:	123.669					
p value of Global Test in MR-PRESSO results:	0.063					

Sd, standard deviation; T-stat, T-statistics; VD, Vitamin D.

fibroid volume reduction (Ciavattini et al., 2016). It indicated that vitamin D supplementation can significantly reduce the volume of UFs, but this is completely opposite to the results of another RCT (Arjeh et al., 2020). Interference of confounding factors and heterogeneity are also unavoidable in clinical trials. Thus, it is necessary to perform a MR study for further research. High-quality MR studies use Mendel's law of random assignment and use SNPs as instrumental variables to minimize the influence of confounding factors and ensure that there is no heterogeneity in the study subjects, which would make the findings more convincing. Our MR study showed that low levels of vitamin D is associated with the increasing risk of UFs, which is consistent with the results of the former RCT (Ciavattini et al., 2016). We speculate that this is because the exposure simulated by the MR study is lifetime exposure, and the outcome is caused by chronically low serum vitamin D3 levels. On the other hand, vitamin D treatment within the RCT period is hardly produce enough therapeutic effects on normal-sized UFs compared to the life cycle, but it has better therapeutic effect on smaller-sized UFs.

The mechanism of action of vitamin D in the body is complex (Lips, 2006). This study found that some genetic variants can affect the risk of UFs through vitamin D3 serum levels. Genetic factors can affect vitamin D through multiple pathways (Hoffman et al., 2004), some non-vitamin D-related genes and their signaling pathways have been shown to play a role in promoting the development of uterine fibroids (Leppert et al., 2006), and this study can only explain some of the effects of vitamin D-related genetic variants on UFs. Therefore, vitamin D-related variants can only explain part of the risk of uterine fibroids, and other signaling pathways need to be analyzed to better understand other risk factors for uterine fibroids.

In the present work, we provide a scientific basis for further research on whether insufficient vitamin D is a causative risk factor for UFs, which may have important public health implications. Vitamin D is a natural component with high safety, relatively small side effects, high economic feasibility, and great research value. Further investigation of randomized clinical trials is needed to be constructed to actively explore the potential role of vitamin D or combination with other drugs on the treatment of UFs. This may help scientists develop a new generation of UFs treatment option (Al-Hendy and Badr, 2014).

One of the major strengths of study is the use of the MR study design, which can reduce the interference of confounders and determine the direction of causality. In addition, the study had large sample size, allowing us to examine more reliable causal association. Furthermore, we evaluate the consistency of the association through different methods to support the robustness of our results. However, several limitations in the study are also worth considering. First, our analysis is based on GWAS data from European populations, and the genetic variation among different races did not satisfy Mendel's law of inheritance, so the obtained results may be difficult to extrapolate to the whole population. Second, our study did not investigate the therapeutic effect of vitamin D on UFs, although it established a causal relationship between vitamin D levels and UFs. Thirdly, The study examined serum vitamin D3 levels only through genetic pathways, and these genetic variants only play a role in specific contexts, given the complex biological role of the vitamin (Lips, 2006).

Conclusion

In the MR study, we found that a one-SD decrease in serum vitamin D levels was associated with higher risk of UFs, consistent with previous studies describing a critical biological role for vitamin D in the development of UFs. Our study also implies the importance of adequate daily intake of vitamin D, which has positive effects on the prevention of UFs. However, due to the limited availability of evidence from clinical studies, further clinical studies are needed to explore the utility of vitamin D for the treatment of UFs.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: JG and YZ; data curation: WG and ZZ; MR analysis: WG; funding acquisition: JG and YZ; software and

visualization: MD and MC; writing—original draft: WG; writing—review and editing: YZ and GG. WG and SZ have verified the underlying data. All the authors approved the final version of the manuscript.

Funding

National Natural Science Foundation of China (YZ, Grant No. 82001496), project of Chengdu Science and Technology Bureau, (YZ, Grant No. 2021-YF05-02110-SN), China Postdoctoral Science Foundation (YZ, Grant Nos 2020M680149, 2020T130087ZX). The Fundamental Research Funds for the Central Universities (SCU2020D4132).

Acknowledgments

Data in the European population on uterine fibroids and serum 25-hydroxyvitamin D levels are available through the United Kingdom Biobank and data analysis is available through the GWAS database. The authors thank these researchers for their selfless sharing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Hendy, A., Diamond, M. P., Boyer, T. G., and Halder, S. K. (2016). Vitamin D3 inhibits wnt/β-catenin and mTOR signaling pathways in human uterine fibroid cells. *J. Clin. Endocrinol. Metab.* 101 (4), 1542–1551. doi:10.1210/jc.2015-3555

Al-Hendy, A., Diamond, M. P., El-Sohemy, A., and Halder, S. K. (2015). 1, 25-dihydroxyvitamin D3 regulates expression of sex steroid receptors in human uterine fibroid cells. *J. Clin. Endocrinol. Metab.* 100 (4), E572–E582. doi:10.1210/jc.2014-4011

Al-Hendy, A. M. B., and Badr, M. (2014). Can vitamin D reduce the risk of uterine fibroids. *Women's Health* 10 (4), 353–358. doi:10.2217/whe.14.24

Ali, M., Prince, L., and Al-Hendy, A. (2020). Vitamin D and uterine fibroids: Preclinical evidence is in; time for an overdue clinical study. *Fertil. Steril.* 113 (1), 89–90. doi:10.1016/j.fertnstert.2019.10.015

Arjeh, S., Darsareh, F., Asl, Z. A., and Azizi Kutenaei, M. (2020). Effect of oral consumption of vitamin D on uterine fibroids: A randomized clinical trial. *Complement. Ther. Clin. Pract.* 39, 101159. doi:10.1016/j.ctcp.2020.101159

Baird, D. D., Hill, M. C., Schectman, J. M., and Hollis, B. W. (2013). Vitamin d and the risk of uterine fibroids. *Epidemiology* 24 (3), 447–453. doi:10.1097/EDE.0b013e31828acca0

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Burgess, S., and Bowden, J. (2015). *Integrating summarized data from multiple genetic variants in Mendelian randomization bias and coverage properties of inverse-variance weighted methods*. arXiv preprint arXiv:. Available from: http://151204486 (Accessed Nov 27, 2015).

Burgess, S., Thompson, S. G., and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.* 40 (3), 755–764. doi:10.1093/ije/dyr036

Burgess, S., and Thomson, S. (2015). *Mendelian randomization: Methods for using genetic variants in causal estimation*. Florida, United States: CRC Press.

Ciavattini, A., Delli Carpini, G., Serri, M., Vignini, A., Sabbatinelli, J., Tozzi, A., et al. (2016). Hypovitaminosis D and "small burden" uterine fibroids: Opportunity for a vitamin D supplementation. *Med. Baltim.* 95 (52), e5698. doi:10.1097/MD.0000000000005698

Ciebiera, M., Wlodarczyk, M., Ciebiera, M., Zareba, K., Lukaszuk, K., and Jakiel, G. (2018). Vitamin D and uterine fibroids-review of the literature and novel concepts. *Int. J. Mol. Sci.* 19 (7), E2051. doi:10.3390/ijms19072051

Ciebiera, M., Wlodarczyk, M., Slabuszewska-Jozwiak, A., Nowicka, G., and Jakiel, G. (2016). Influence of vitamin D and transforming growth factor β3 serum concentrations, obesity, and family history on the risk for uterine fibroids. *Fertil. Steril.* 106 (7), 1787–1792. doi:10.1016/j.fertnstert.2016.09.007

Ciebiera, M., Wlodarczyk, M., Wrzosek, M., Meczekalski, B., Nowicka, G., Lukaszuk, K., et al. (2017). Role of transforming growth factor beta in uterine fibroid biology. *Int. J. Mol. Sci.* 18 (11), E2435. doi:10.3390/ijms18112435

Emdin, C. A., Khera, A. V. S. K., and Kathiresan, S. (2017). Mendelian randomization. *JAMA J. Am. Med. Assoc.* 318 (19), 1925–1926. doi:10.1001/jama.2017.17219

Halder, S. K., Sharan, C., and Al-Hendy, A. (2010). Vitamin D treatment induces dramatic shrinkage of uterine leiomyomas growth in the Eker rat model. *Fertil. Steril.* 94 (4), S75–S76. doi:10.1016/j.fertnstert.2010.07.293

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46 (6), 1985–1998. doi:10.1093/ije/dyx102

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7, e34408. doi:10.7554/eLife.34408

Hoffman, P. J., Milliken, D. B., Gregg, L. C., Davis, R. R., and Gregg, J. P. (2004). Molecular characterization of uterine fibroids and its implication for underlying mechanisms of pathogenesis. *Fertil. Steril.* 82 (3), 639–649. doi:10.1016/j.fertnstert.2004.01.047

Kempson, R. L, R. H. M., and Hendrickson, M. R. (2000). Smooth muscle, endometrial stromal, and mixed Müllerian tumors of the uterus. *Mod. Pathol.* 13 (3), 328–342. doi:10.1038/modpathol.3880055

Khan, A. T., Shehmar, M., and Gupta, J. K. (2014). Uterine fibroids: Current perspectives. *Int. J. Womens Health* 6, 95–114. doi:10.2147/IJWH.S51083

Leppert, P. C., Baginski, T., Prupas, C., Catherino, W. H., Pletcher, S., and Segars, J. H. (2004). Comparative ultrastructure of collagen fibrils in uterine leiomyomas and normal myometrium. *Fertil. Steril.* 82, 1182–1187. doi:10.1016/j.fertnstert.2004.04.030

Leppert, P. C., Catherino, W. H., and Segars, J. H. (2006). A new hypothesis about the origin of uterine fibroids based on gene expression profiling with microarrays. *Am. J. Obstet. Gynecol.* 195 (2), 415–420. doi:10.1016/j.ajog.2005.12.059

Lips, P. (2006). Vitamin D physiology. *Prog. Biophys. Mol. Biol.* 92 (1), 4–8. doi:10.1016/j.pbiomolbio.2006.02.016

Mitchell, R. E., Elsworth, B. L., Mitchell, R., Raistrick, C. A., Paternoster, L., Hemani, G., et al. (2019). *MRC IEU UK Biobank GWAS pipeline version 2*. Bristol, UK: University of Bristol.

Protic, O., Toti, P., Islam, M. S., Occhini, R., Giannubilo, S. R., Catherino, W. H., et al. (2016). Possible involvement of inflammatory/reparative processes in the development of uterine fibroids. *Cell Tissue Res.* 364 (2), 415–427. doi:10.1007/s00441-015-2324-3

Rafique, S., Segars, J. H., and Leppert, P. C. (2017). Mechanical signaling and extracellular matrix in uterine fibroids. *Semin. Reprod. Med.* 35 (6), 487–493. doi:10.1055/s-0037-1607268

Staiger, D., and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586. doi:10.2307/2171753

Sabry, M., Halder, S. K., Allah, A. S., Roshdy, E., Rajaratnam, V., and Al-Hendy, A. (2013). Serum vitamin D3 level inversely correlates with uterine fibroid volume in different ethnic groups: A cross-sectional observational study. *Int. J. Womens Health* 5, 93–100. doi:10.2147/IJWH.S38800

Sharan, C., Halder, S. K., Thota, C., Jaleel, T., Nair, S., and Al-Hendy, A. (2011). Vitamin D inhibits proliferation of human uterine leiomyoma cells via catechol-O-methyltransferase. *Fertil. Steril.* 95 (1), 247–253. doi:10.1016/j.fertnstert.2010.07.1041

Sozen I, A. A., and Arici, A. (2002). Interactions of cytokines, growth factors, and the extracellular matrix in the cellular biology of uterine leiomyomata. *Fertil. Steril.* 78 (1), 1–12. doi:10.1016/s0015-0282(02)03154-0

Stewart, E. A., Laughlin-Tommaso, S. K., Catherino, W. H., Lalitkumar, S., Gupta, D., and Vollenhoven, B. (2016). Uterine fibroids. *Nat. Rev. Dis. Prim.* 2, 16043. doi:10.1038/nrdp.2016.43

Van Den Bemd, G. J., Pols, H. A., and Leeuwen Van, J. P. (2000). Anti-tumor effects of 1, 25-dihydroxyvitamin D3 and vitamin D analogs. *Curr. Pharm. Des.* 6 (7), 717–732. doi:10.2174/1381612003400498

Vercellini, P., and Frattaruolo, M. P. (2017). Uterine fibroids: From observational epidemiology to clinical management. *BJOG* 124 (10), 1513. doi:10.1111/1471-0528.14730

Wallach, E. E., Buttram, V. C., and Reiter, R. C. (1981). Uterine leiomyomata: Etiology, symptomatology, and management. *Fertil. Steril.* 36 (4), 433–445. doi:10.1016/s0015-0282(16)45789-4

Zimmermann, A., Bernuit, D., Gerlinger, C., Schaefers, M., and Geppert, K. (2012). Prevalence, symptoms and management of uterine fibroids an international internet-based survey of 21, 746 women. *BMC women's health* 12 (1), 6–11. doi:10.1186/1472-6874-12-6

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fevo.2022.943299

OPEN ACCESS

EDITED BY
Nikica Šprem,
University of Zagreb, Croatia

REVIEWED BY
Wenwu Zhou,
Zhejiang University, China
Ankita Gupta,
Indian Council of Agricultural Research (ICAR), India

*CORRESPONDENCE
Shijun You
sjyou@fafu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 13 May 2022
ACCEPTED 29 August 2022
PUBLISHED 23 September 2022

CITATION
Sun L, Li J, Chen J, Chen W, Yue Z, Shi J, Huang H, You M and You S (2022) An ensemble learning approach to map the genetic connectivity of the parasitoid *Stethynium empoasca* (Hymenoptera: Mymaridae) and identify the key influencing environmental and landscape factors. *Front. Ecol. Evol.* 10:943299. doi: 10.3389/fevo.2022.943299

An ensemble learning approach to map the genetic connectivity of the parasitoid *Stethynium empoasca* (Hymenoptera: Mymaridae) and identify the key influencing environmental and landscape factors

Linyang Sun[1,2,3†], Jinyu Li[1,4†], Jie Chen[1,2,3,5], Wei Chen[1,2,3,5], Zhen Yue[6], Jingya Shi[6], Huoshui Huang[7], Minsheng You[1,2,3,5] and Shijun You[1,2,3,5,6*]

[1]State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Institute of Applied Ecology, Fujian Agriculture and Forestry University, Fuzhou, China, [2]International Joint Research Laboratory of Ecological Pest Control, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, China, [3]Ministerial and Provincial Joint Innovation Centre for Safety Production of Cross-Strait Crops, Fujian Agriculture and Forestry University, Fuzhou, China, [4]Tea Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, China, [5]Key Laboratory of Integrated Pest Management for Fujian-Taiwan Crops, Ministry of Agriculture and Rural Affairs, Fuzhou, China, [6]BGI-Sanya, BGI Genomics Shenzhen Technology Co., Ltd., Sanya, China, [7]Comprehensive Technology Service Center of Quanzhou Customs, Quanzhou, China

The effect of landscape patterns and environmental factors on the population structure and genetic diversity of organisms is well-documented. However, this effect is still unclear in the case of Mymaridae parasitoids. Despite recent advances in machine learning methods for landscape genetics, ensemble learning still needs further investigation. Here, we evaluated the performance of different boosting algorithms and analyzed the effects of landscape and environmental factors on the genetic variations in the tea green leafhopper parasitoid *Stethynium empoasca* (Hymenoptera: Mymaridae). The *S. empoasca* populations showed a distinct pattern of isolation by distance. The minimum temperature of the coldest month, annual precipitation, the coverage of evergreen/deciduous needleleaf trees per 1 km^2, and the minimum precipitation of the warmest quarter were identified as the dominant factors affecting the genetic divergence of *S. empoasca* populations. Notably, compared to previous machine learning studies, our model showed an unprecedented accuracy (r = 0.87) for the prediction of genetic differentiation. These findings not only demonstrated how the landscape shaped *S. empoasca* genetics but also provided an essential basis

for developing conservation strategies for this biocontrol agent. In a broader sense, this study demonstrated the importance and efficiency of ensemble learning in landscape genetics.

KEYWORDS

landscape genetics, machine learning, parasitoid, climate change, biology conservation

Introduction

The field of landscape genetics quantifies how heterogeneous landscape features and environmental factors shape genetic variations in living organisms. It has been applied in many research areas, such as conservation biology, alien species invasion, and pest management (Bowman et al., 2016; Jonsson et al., 2017). Traditional landscape genetics studies are often restricted by the subjectivity of producing resistance surfaces and the difficulty of addressing inter-variable interactions (Pless et al., 2021). Compared to these traditional methods, machine learning algorithms can develop strong non-linear regression models (Elavarasan et al., 2018). Further, with increasing access to remote sensing and climate data, machine learning methods can be used to explore the effects of multiple environmental factors on genetic variations at any sampling scale. Some machine learning approaches include deep learning (Kittlein et al., 2022) and random forest (Murphy et al., 2010; Sylvester et al., 2018; Shanley et al., 2021) been recently developed for their application in landscape genetics.

Despite increasing applications, current machine learning methods still show some limitations for landscape genetics. For example, a convolutional neural network (CNN), a deep learning method first introduced in landscape genetics by Kittlein et al. (2022), usually performs poorly on small datasets (Elavarasan et al., 2018). This is a major limitation as the number of sample sites in most population-based landscape genetic studies is often <50. Additionally, CNN approaches are limited in their ability to identify features at different sampling scales, such as sampling over oceans. They show substantial distance disparities, resulting in high variance across samples. Consequently, extracting useful features from remote sensing images using CNN becomes challenging. Comparatively, ensemble learning methods are gaining attention in landscape genetics. Ensemble learning aims to improve predictive performance by aggregating predictions from many weak models (Opitz and Maclin, 1999; Polikar, 2006) and has some algorithms, such as bagging, boosting, and stacking. Ensemble methods have currently shown excellent performance in various fields, including production forecasting and gas emission forecasting (Bossavy et al., 2013; Liu et al., 2020; Chen K. et al., 2021). Random forest is the most frequently used ensemble learning algorithm in landscape ecology and genetics. It is compatible with multiple variables and can effectively extract the relative importance of features. However, the recently developed algorithm, iterative random forest, have low prediction accuracy (Pless et al., 2021); therefore, evaluating the performance of different ensemble algorithms and identifying the appropriate methods that can be used in fine-scale landscape genetics studies is necessary.

Biological control is a sustainable pest management strategy to reduce the application and adverse effects of chemical pesticides (Cranham, 1966; Nakai, 2009; Zhuang et al., 2009; Yue et al., 2010; Carvalho, 2017) and genetically modified crops (Rodriguez-Saona, 2018). To develop effective biological control strategies, investigating the effects of landscape features and environmental factors on the genetic variations in wild natural enemies is necessary using landscape genetics methods. *Stethynium empoasca* Subba Rao (fairy wasp; Hymenoptera: Mymaridae) is an egg parasitic natural enemy (Huber, 1986; Mills, 1994) of *Empoasca onukii* Matsuda, the most destructive insect pest of tea plantations in East Asia. Some members of the Mymaridae family, such as *Gonatocerus ashmeadi* Girault in Tahiti (Grandgirard et al., 2007) and *Paranagrus optabilis* Perkins in Hawaii (Funasaki et al., 1988), have been used in biological control since a long time. However, only a few studies have analyzed their genetics (de León and Jones, 2005; De Leon et al., 2009; Nadel et al., 2012; Li et al., 2021). *S. empoasca*, having a high rate of parasitism (up to 30%; Li et al., 2021) in the field, is the most promising candidate for conservative biological control of *E. onukii*. Examining its population genetic variation and its relationship with the landscape features and environmental factors will provide a better understanding of its survival requirements and the influence of environmental factors; moreover, this may further assist in developing conservative strategies for better biological control of *E. onukii*.

In this study, we determined which ensemble model performed best on the collected data and then identified the environmental and landscape factors that could affect the genetic differentiation and diversity of *S. empoasca*. Our findings could provide practical suggestions for conserving *S. empoasca* parasitoids that serve as biocontrol agents. To the best of our knowledge, our research is the first to demonstrate a practical

empirical method for exploring the landscape genetics of the Mymaridae family.

Materials and methods

Sample collection and microsatellite genotyping

The study was conducted in Fujian Province, China. Twenty tea plantations with different ambient landscape patterns and latitudes were selected for the study. To minimize the influence of recent tea seedling transportation and differences in pesticide use, only conventional tea plantations planted many years ago were selected for this study. In total, 506 *S. empoasca* individuals were collected from 20 sample sites (17 sites in Wuyishan city and 1 site each in Anxi, Fuzhou, and Fuding cities) in 2019 (**Table 1**). After sample collection, all individuals were confirmed by morphological identification according to previous studies (Triapitsyn et al., 2019).

The habitus image of *S. empoasca* is shown in **Supplementary Figures 1**, **2**. Ten microsatellite loci developed by Li et al. (2021) were tested and selected to genotype *S. empoasca*. To improve the amplification efficiency of these loci and to reduce their cost, a primer tail C was added to the 5′ end of the candidate forward primer, and a fluorescent marker was added to identify the genotypes of the various loci (Blacket et al., 2012). A polymerase chain reaction (PCR) reaction was conducted in a 10 μL. After amplifying the microsatellite marker listed in **Supplementary Table 1** and the PCR procedures shown in **Supplementary Table 2**, the PCR products were analyzed using an ABI 3730 xl DNA Analyzer (Thermo Fisher Scientific, Waltham, MA, USA) and a GeneScan^{TM} 500 LIZ® Size Standard (Thermo Fisher Scientific). The microsatellite loci were manually determined using GeneMapper v. 3.2 (Lemonick., 2000) and checked for stuttering and large allele dropout by MICROCHECKER v. 2.2.3 (Van Oosterhout et al., 2004). Finally, microsatellite genotype data were obtained from 506 individuals and used in the subsequent landscape genetics analysis.

Climate and landscape data

Two datasets were used to evaluate the effect of the environment and landscape on genetic differentiation in *S. empoasca*. We selected 19 bioclimatic variables with 1 km^2 resolution in Woldclim (Fick and Hijmans, 2017) and 12 landscape variables with the same resolution in EarthEnv (Tuanmu and Jetz, 2014). In the EarthEnv datasets, some variables, such as open water, snow/ice, barren, deciduous broadleaf trees, and regularly flooded vegetation, which rarely exist in our study region, were finally not included in the model (**Supplementary Table 3**). Each pixel on the map in this EarthEnv dataset represents the percentage of one land-cover class in a 1 km^2 area. The ensemble learning method is based on decision trees, which have been proven highly efficient in dealing with redundant variables. Therefore, we did not remove multicollinearity variables. All datasets were cropped according to the extent of our study region. The straight-line (STR) method was applied to construct the resistance surface to calculate the resistance distance among the selected sample sites. All resistance distances in this study were calculated using the mean value of each pixel on the path between pairwise sampling sites, aiming to avoid some distance-based bias that potentially resulted from sampling site selection/distribution.

The land use raster map for each of the three regions was downloaded from the 2018 National Standard Land Use Type Classification on the Geospatial Data Cloud platform.[1] The downloaded raster maps were classified into four land cover types: forest, tea plantation, crop, and non-vegetation area (e.g., water body, built-up, and empty area). Further, 1,000 and 2,000-m radius buffers were drawn for each site using the "rgeos" package (Bivand et al., 2017) in R. To measure the fragmentation levels in each study region, four class-level and two landscape-level indexes were computed at the allocated buffers. At the class level, the number of patches (NP), edge density (ED), patch density (PD), and patch cohesion index (COHESION) were used to describe fragmentation and connectivity. Shannon's diversity index (SHDI) and Shannon's evenness index (SIEI) were used to illustrate the landscape-level fragmentation. The R package "landscape metrics" computed these indexes (Hesselbarth et al., 2019).

Data analysis

Population genetic differentiation and genetic structure analysis

Seven parameters, including allele number, allele proportion, allele richness (A_R), expected heterozygosity (H_e), observed heterozygosity (H_o), inbreeding coefficient within the population (F_{is}), and Hardy–Weinberg equilibrium (HWE), were selected to illustrate *S. empoasca* genetic diversity within a population. All variables were calculated using the R package "diveRsity" (Keenan et al., 2013). Population pairwise genetic differentiation (F_{ST}) was calculated between a population pair using the R package "adegenet" (Jombart, 2008).

Discriminant analysis of principal components (DAPC) was then performed using the R package "adegenet" to deduce the spatial pattern of population structure. DAPC is a low computational-cost method that performs a k-mean algorithm after the transformation of principal component analysis (PCA)

1 http://www.gscloud.cn/

TABLE 1 Genetic diversity and geographic information of 20 *S empoasca* populations based on 10 microsatellite loci.

Pop ID	Longitude	Latitude	Sample size	Allele number	Alleles proportion	A_R	H_o	H_e	$P_{(HWE)}$	F_{is}
XC	117.915436	27.638767	17	29	45.53	2.5	0.44	0.43	0.566	−0.0237
QLC	117.955289	27.609211	23	30	48.17	2.61	0.41	0.45	0.940	0.0886
FPC	118.027957	27.508488	30	36	55.42	2.84	0.46	0.45	0.904	−0.0138
XD	118.040168	27.755573	27	33	51.53	2.69	0.42	0.45	0.311	0.0627
HXZ	117.990902	27.767785	27	35	51.75	2.58	0.4	0.42	0.157	0.0611
TXC	117.986111	27.684444	22	30	47.61	2.46	0.4	0.41	0.951	0.0281
TM	117.93197	27.70965	25	31	48.81	2.56	0.43	0.44	0.26	0.0362
BYQ	117.935532	27.688545	26	31	49.5	2.55	0.45	0.45	0.086	0.0028
CD	117.857567	27.719317	32	30	48.17	2.45	0.45	0.42	0.847	−0.0794
JLS	117.941715	27.715562	31	31	50.08	2.51	0.42	0.42	0.26	0.004
DHP	117.961667	27.671231	28	32	51.67	2.68	0.49	0.46	0.937	−0.0644
WYX	117.99246	27.719104	18	33	52.3	2.7	0.43	0.46	0.578	0.0748
YZC	117.94686	27.802791	34	36	56.08	2.75	0.46	0.45	0.996	−0.0269
WX	118.005577	27.747276	30	33	53.33	2.65	0.43	0.46	0.717	0.0651
PKK	117.742787	27.684942	13	29	46.42	2.61	0.46	0.47	0.903	0.0102
SPC	117.985014	27.634001	24	37	57.58	2.84	0.48	0.5	0.003	0.0593
HXC	117.822625	27.673072	43	33	52.64	2.68	0.47	0.46	0.992	−0.0239
FD	120.388308	27.145455	20	39	62.86	2.9	0.51	0.5	0.038	−0.0109
FZ	119.22779	26.088089	28	37	56.42	3.02	0.5	0.51	0.418	0.0216
AX	117.873972	25.002417	8	26	42.17	2.45	0.48	0.47	0.851	−0.01

Pop ID, population abbreviations; allele number, the mean value of alleles observed across loci per population sample; alleles proportion, the mean percentage of total alleles observed across loci per population sample; A_R, allele richness; H_o and H_e, observed and expected heterozygosity, respectively; $P_{(HWE)}$, *p*-value from Fisher's exact test in Hardy–Weinberg equilibrium ($p < 0.05$); F_{is}, inbreeding coefficient.

(Jombart et al., 2010). We used the "find cluster" function with 10^7 iterations to determine the best genetic cluster. A linear regression model and Pearson's correlation analysis were performed to detect the pattern of isolation by distance (IBD).

Model comparison and construction

All models were run in Python 3.8. In the preliminary analysis, we evaluated the performance of eight commonly used ensemble learning algorithms: the Adaboost algorithm with decision tree and random forest classifier, the eXtreme Gradient Boosting (XGBoost) algorithm C decision tree and random forest classifier, GradientBoosting algorithm with decision tree classifier, the light gradient boosting (lightGBM) algorithm with decision tree classifier, the goss algorithm, and the cat boosting algorithm. In all eight models, environmental resistance distance was used as an explanatory variable, and the fixation index (F_{ST}) was used as a response variable. A Scikit-learn test train split function was used with a 0.3 test set split, and MinMaxScaler was used for data normalization. Four metrics, namely, Pearson's correlation coefficient (r), R-squared (R^2) value, root mean square error (RMSE), and mean absolute percentage error (MAPE), were used to evaluate and compare the performance of these models. Subsequently, a model with the best performance was selected for further analysis.

The best model was tuned by GridSearchCV in Scikit-learn. Specifically, all 28 environmental variables were used to predict the STR-based resistance surface by the fitted model. Resistance distance was then calculated using the STR-based resistance surface by the least cost path (LCP) method. A new model was tuned again using the new LCP datasets. We also use the permutation importance to visualize our machine learning model. Compared to other feature importance ranking methods, permutation importance reconstructs the relationship between the target and the feature through multiple permutation calculations to explore the model's dependence on the feature. Subsequently, the predicted resistance surface was transformed into a connectivity surface by taking the inverse of each pixel value.

Effect of landscape pattern on genetic diversity

Considering the small number of datasets of 20 sample sites, we performed Pearson's correlation analysis in the R package "Hmisc" (Harrell and Dupont, 2006) to evaluate the relationship between the *S. empoasca* population's genetic diversity and landscape features around sampled tea plantations. Three parameters, allele richness (A_R), expected heterozygosity (H_e), and observed heterozygosity (H_o), were selected to test the relative relationship with landscape metrics by calculating the relative coefficient and *p*-value in R.

Results

Population's genetic diversity and differentiation

The estimates of genetic diversity determined by analyzing the ten microsatellites in 506 individuals are shown in **Table 1**. All genetic diversity indices showed extremely strong and narrow ranges of changes (A_R: 2.45–3.02; H_o: 0.4–0.51; and H_e: 0.41–0.51). Except for Xingcun (XC), Fengpo (FPC), Chengdun (CD), Dahongpao (DHP), Yangzhuang (YZC), Hongxing (HXC), Fuding (FD), and Anxi (AX) populations, the inbreeding coefficients (F_{is}) were more than zero for all populations. The *p*-value of Fisher's exact test showed deviation from Hardy–Weinberg equilibrium (HWE) in Shangpu (SPC) and FD populations. Higher genetic differentiation ($F_{ST} > 0.07$) was found in Pikengkou (PKK), SPC, Fuzhou (FZ), and AX populations (**Figure 1**) than in other pair populations ($F_{ST} < 0.04$).

The DAPC cluster results showed that the best cluster number of all 506 individuals was three (**Figures 2A,B**), with no clear boundary identified between clusters. When the cluster number was five and seven, all clusters are closed to each other (**Figures 2C,D**). However, when the cluster number was greater than seven, one cluster was distinctly differentiated (**Figures 2E,F**). Both the linear regression model ($y = -0.0826 + 0.0133x$; $R^2 = 0.36$; $p < 0.001$; **Figure 3**) and Pearson's correlation analysis ($r = 0.60$; $p < 2.2e^{-16}$) showed a significantly positive relationship between log-transformed geographic distance and F_{ST}.

Effect of environmental factors and landscape features on genetic variation

The model comparison results showed that the eXtreme gradient boosting algorithm with a random forest regression model (XGBoost-RFR) performed better than the other seven boosting algorithms (**Figure 4**), with the highest R^2 and *r*-value and the lowest RMSE value. The MAPE value of the XGBoost-RFR model is not the lowest. However, this model still has an overall advantage over others. The Adaboost algorithm with random forest regression model (Adaboost-RFR) showed similar evaluation metrics, including r, R^2, and MSE, but

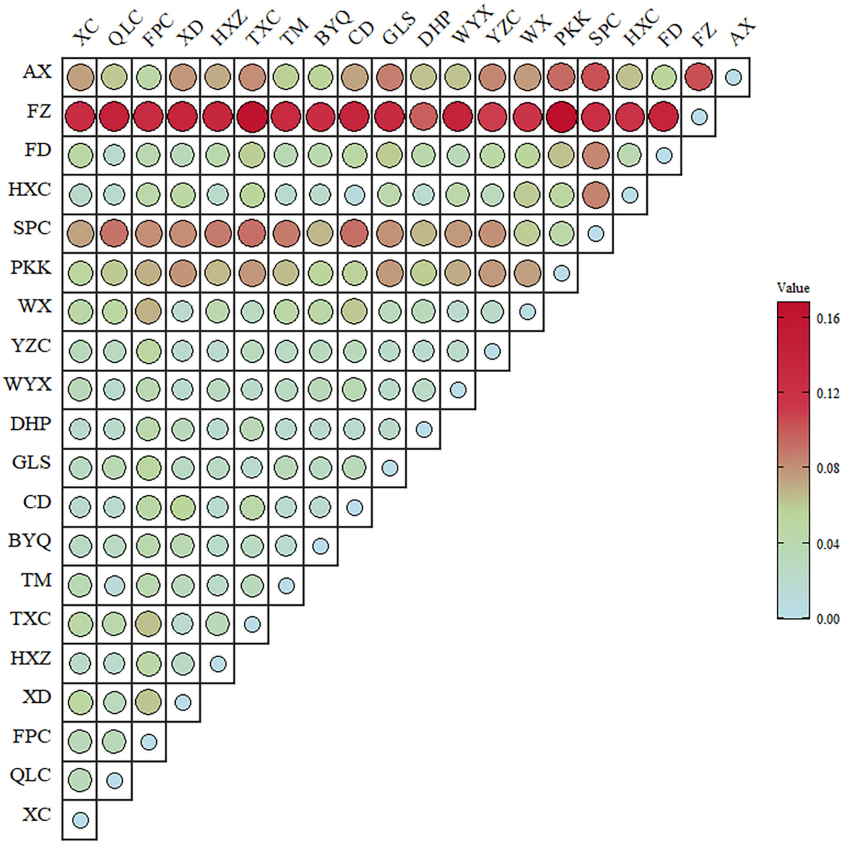


FIGURE 1
Pairwise F_{ST} values among the twenty *Stethynium empoasca* populations.

showed a higher MAPE value compared with the XGBoost-RFR model (Figure 4). For both STR and LCP models, the XGBoost-RFR model performed slightly worse on the test set than on the train set (Table 2). The permutation importance results showed that in the STR-based model, the top four important factors were annual precipitation (bio_12), temperature seasonality (bio_4), precipitation of the driest quarter (bio_17), and precipitation of the driest month (bio_14), while other factors, such as cultivated and managed vegetation (class_7), evergreen/deciduous needleleaf trees (class_1), and min temperature of the coldest month (bio_6), explained only a small fraction of the prediction of genetic differentiation (Figure 5). In the LCP-based model, the top four important factors were bio_6, bio_12, class_1, and bio_18. Moreover, in the test set, there is a strong correlation between the predicted values produced by the XGBoost-RFR model with LCP distance and the true values, but poor predictive power was found for low values in the F_{ST} dataset (Figure 6). The predicted genetic connectivity (Figure 7) and the map of the top four important environmental factors (Figure 8) showed that less precipitation and higher minimum temperature could block the genetic connectivity of *S. empoasca*.

Most landscape metrics showed no significant relationship with the three population's genetic diversity metrics (i.e.,

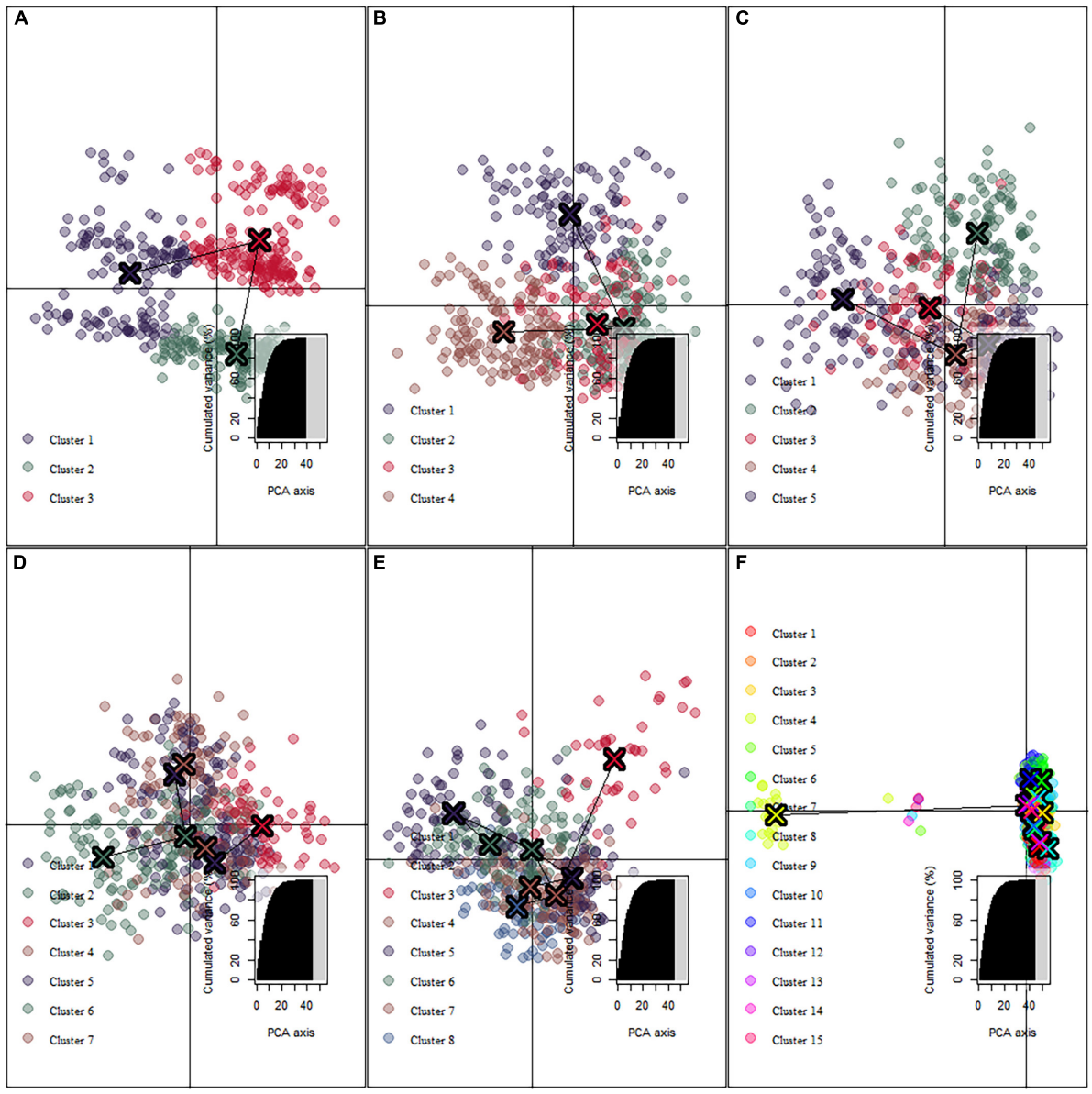


FIGURE 2
Estimated population genetic cluster of *S. empoasca* populations. **(A–F)** DAPC cluster results at K = 3, 4, 5, 7, 8, and 15 separately.

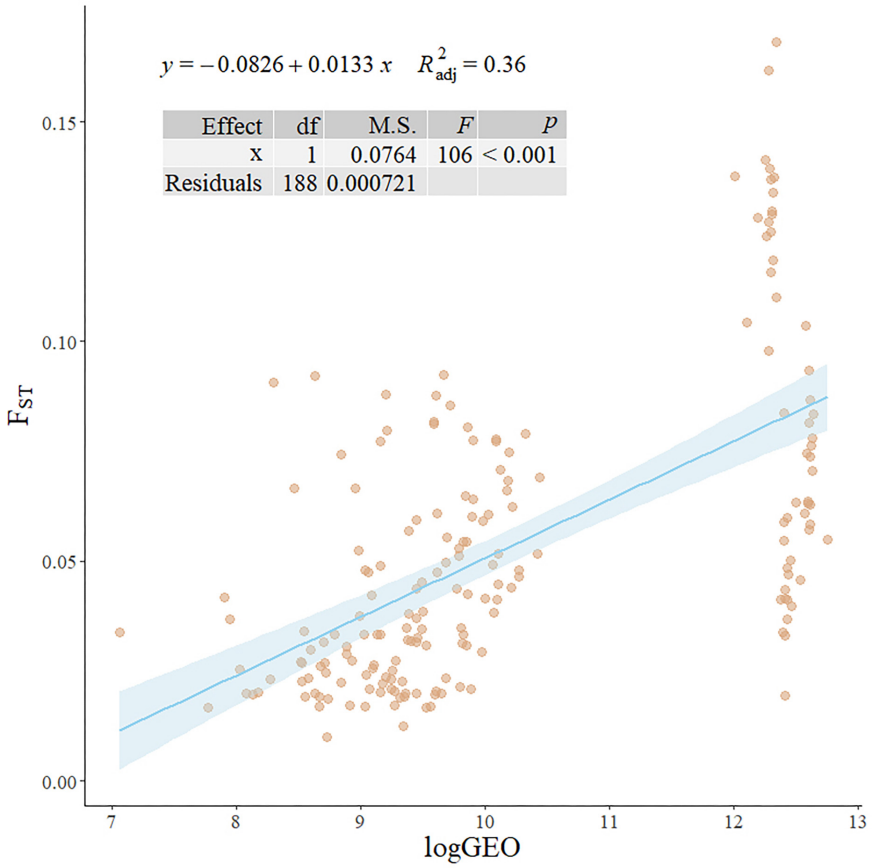


FIGURE 3
The linear regression model between log-transformed geographic distance and F_{ST}.

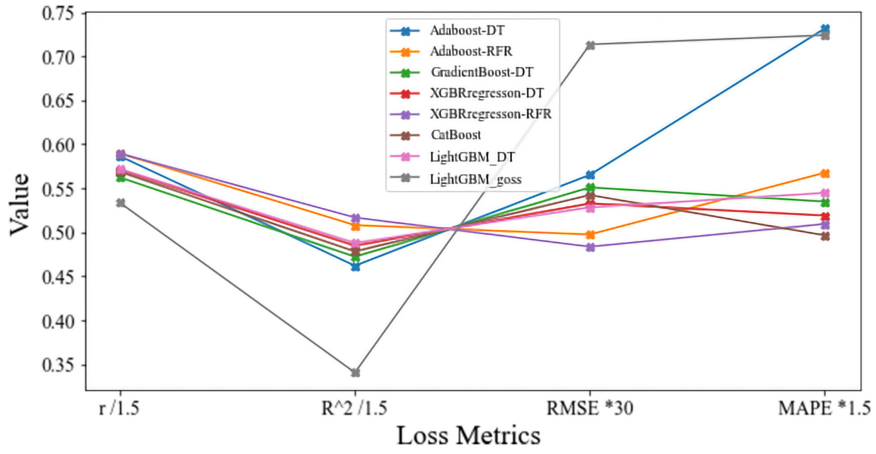


FIGURE 4
Model comparison of eight boosting algorithms with four loss metrics. "Adaboost-DT" and "Adaboost-RFR" are Adaboost algorithms with the decision tree and the random forest classifier. "XGBoost-DT" and "XGBoost-RFR" are XGBoost algorithms with the decision tree and the random forest classifier. "LightGBM-DT" and "LightGBM-goss" are light gradient boosting algorithms with the decision tree classifier and goss algorithm. "GBM" is the gradient boost algorithm with the decision tree.

TABLE 2 XGBoost-RFR model performance for train, test, and full sets using straight-line and least cost path methods.

	Str-train	Str-test	Str-full	LCP-train	LCP-test	LCP-full
Mean squared error (MSE)	0.00019	0.00031	0.00022	0.00012	0.00032	0.00018
Coefficient of determination (R2)	0.83063	0.73601	0.80143	0.88606	0.72528	0.83619
Mean absolute percentage error (MAPE)	0.29984	0.39686	0.32895	0.23631	0.42828	0.29390

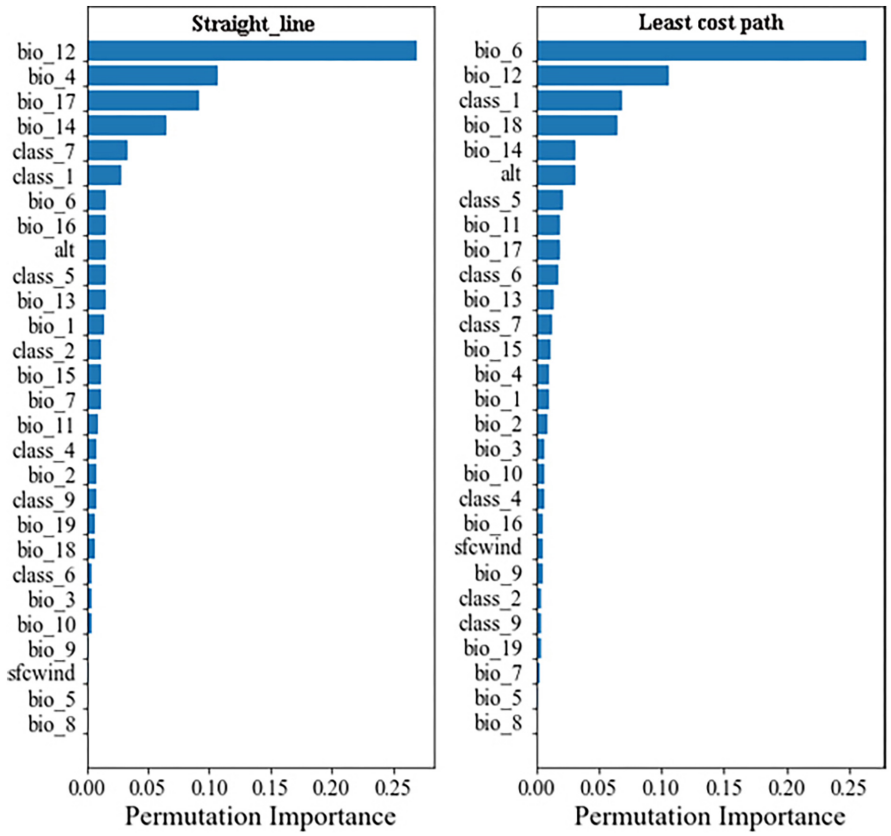


FIGURE 5
Permutation feature the importance of the straight-line-based and LCP-based XGBoost-RFR model. The *X*-axis represents the permutation importance score for each environmental factor.

A_R, H_e, and H_o). At the 1,000 and 2,000-m radius buffers (**Supplementary Table 4**), H_e and H_o showed a significantly negative relationship with the cohesion index of the cropland cover [1,000 m: r (H_e) = −0.57, r (H_o) = −0.56; 2,000 m: r (H_e) = −0.47, r (H_o) = −0.51]. At the 2,000-m radius buffer, A_R was significantly positively correlated to the cohesion index of non-vegetation land cover [r (A_R) = 0.53]. In contrast, Ho was significantly negatively correlated to the cohesion index of non-vegetation land cover [r (H_o) = −0.53].

Discussion

Mapping genetic connectivity and exploring the relationship between environmental variables and genetic differentiation in a species are critical preliminary aspects of landscape genetics (Manel et al., 2003; Manel and Holderegger, 2013). In this study, based on the population genetic differentiation analyses using DAPC and F_{ST} estimation, we proposed an ensemble learning method that uses XGBoost-RFR to map landscape connectivity and identify the most critical landscape variables associated with the population genetic variations in *S. empoasca*, an essential parasitic natural enemy in tea plantations.

Regarding population genetic differentiation, DAPC showed an unclear genetic cluster of *S. empoasca* populations, while F_{ST} estimation proved significant differences between the PKK, SPC, FZ, and AX populations. Previous studies on other parasitoids conducted at large scales (Mitrović et al., 2013; Tait et al., 2017) or both large and small scales (Zepeda-Paulo et al., 2016; Garba et al., 2019) showed a distinct population genetic

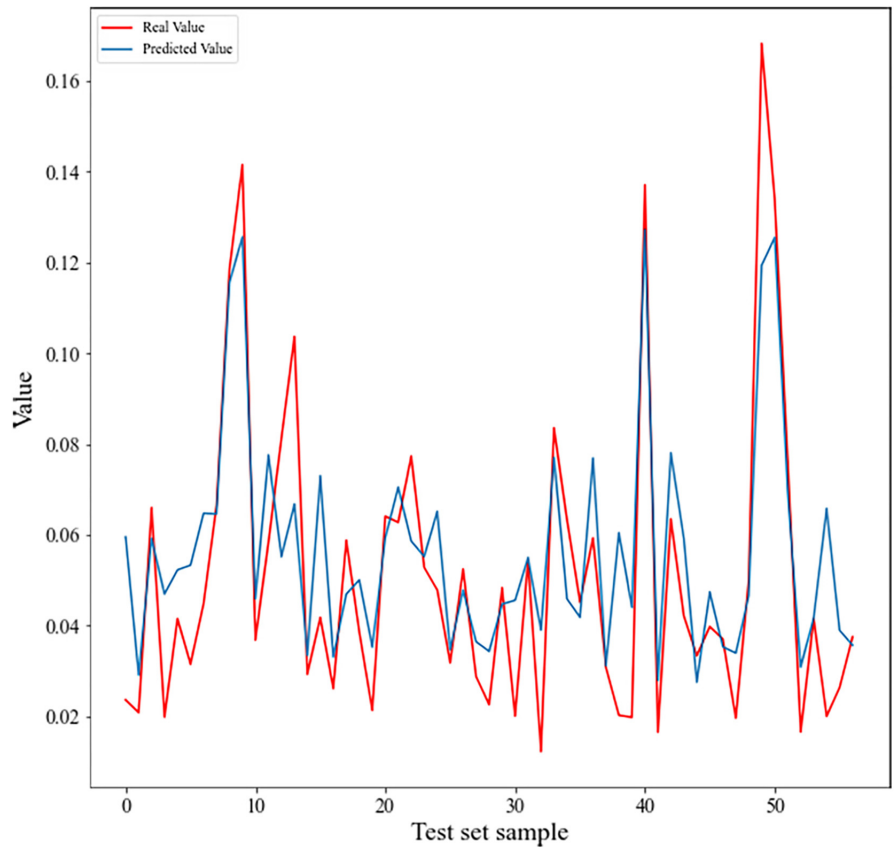


FIGURE 6
Line chart showing F_{ST} vs. predicted genetic differentiation value for the F_{ST} test set. Pearson's correlation coefficient is 0.87 ($p < 2.2e^{-16}$).

structure considering the large scale and an unclear population genetic structure considering the small scale. Based on the substantial differences between *S. empoasca* and other taxa in terms of body size and dispersal capacity, we interpreted that the genetic structure of parasitoids could be remarkably influenced by dispersal capacity (Kankare et al., 2005). Our results demonstrated that IBD strongly affected the genetic differentiation of *S. empoasca* populations, with population genetic distance increasing linearly with the log of geographic distance, as previously detected in most arthropod species (Silva-Brandão et al., 2015; Wright et al., 2015).

The results of primary model selection revealed that booting strategy evaluation results were similar for all models, except for the light gradient boost algorithm and the goss algorithm. The XGBoost-RFR model showed the best metrics. Therefore, XGBoost-RFR model algorithms can be a useful tool in landscape genetics, especially at small sampling scales. Our model exhibited a high correlation ($r = 0.87$) between the final predicted value and actual genetic differentiation data, which contrasts compared to the results of previous studies (Murphy et al., 2010; Hether and Hoffman, 2012; Sylvester et al., 2018; Pless et al., 2021; Shanley et al., 2021) with comparatively fewer computing resources and workload. Although the accuracy of the prediction depends on many aspects, such as data size, data quality (Farooqi et al., 2018), feature number, model selection, and parameter tuning (Deiss et al., 2020), the boosting algorithm has been proven to be more efficient than the bagging algorithm (Kotsiantis and Kanellopoulos, 2012). Conversely, the error metrics of the XGBoost-RFR model indicated that no overfitting was detected on the test set. Therefore, we believe that our model is a good representation of genetic status.

We use the inverse of each pixel value of each predicted map to represent genetic connectivity in our study region, which means an area with high resistance capacity will exhibit a low genetic connectivity value. All regions with distinctive features (high or low genetic connectivity value regions) in the connectivity map have a similar pattern to the original map (**Figures 7**, **8**) but are not the same. In other words, this connectivity map can be seen as a comprehensive result of all input factors. Individuals from high genetic connectivity regions (light blue and light orange regions on the map) may encounter more resistance when they move to the dark color region.

When we first used STR methods to build the resistance surface, each pixel on this surface contained information about

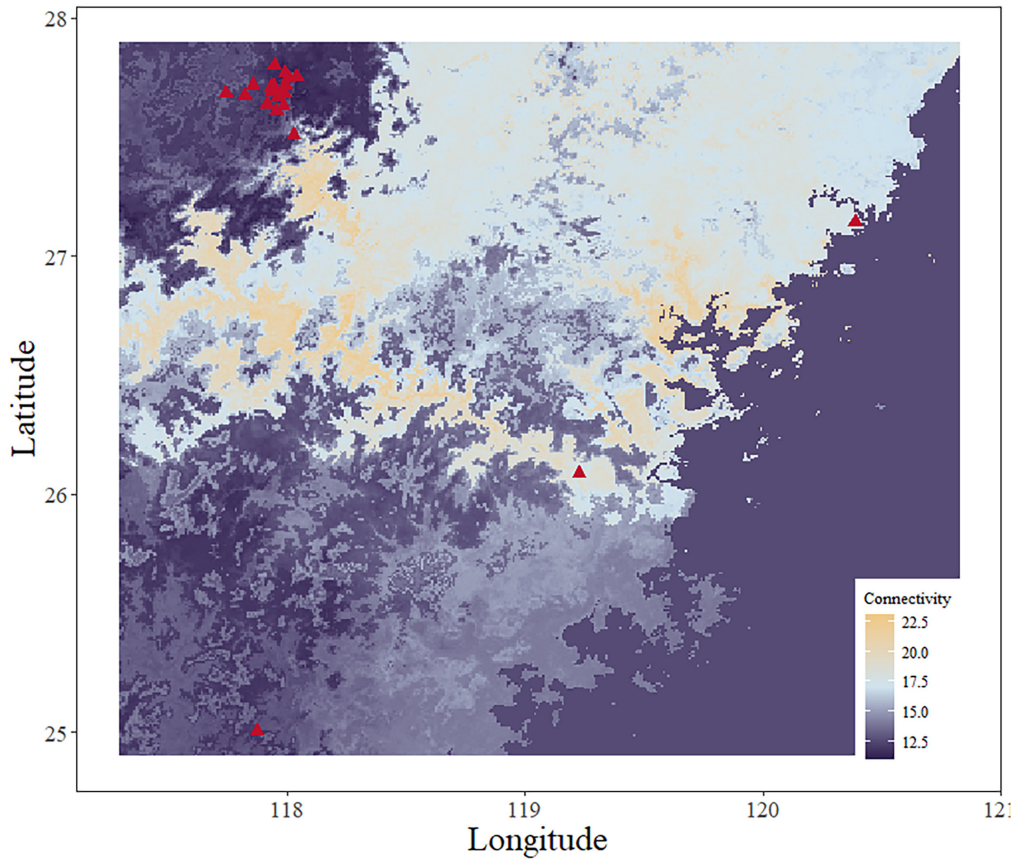


FIGURE 7
Genetic connectivity map using the F_{ST} full set. The red triangle shows the collection sites for *S. empoasca* (genetic data).

the environment and genetic data. When we constructed the LCP using the STR-based resistance surface, each pixel in the map represented a comprehensive result of 19 bioclimatic factors and was used for subsequent analysis. Annual precipitation is considered the critical factor, given that it has the highest importance score in both two models. The effect of precipitation on genetic differentiation has been frequently detected in plant and virus species (Avolio et al., 2013; Palinski et al., 2021) but seldom in arthropods (Du et al., 2009; Wellenreuther et al., 2011; French et al., 2022). This could be attributed to some reasons. For example, the precipitation variance across different seasons may be inconsistent with the life cycle of *S. empoasca*.

On the other hand, precipitation in the warmest quarter (bio_18) contains similar information to that of minimum precipitation in the driest quarter and month (bio_17 and bio_14) but deeper. As we know, higher temperatures always accompany lower precipitation amounts. Moreover, temperature seasonality was also implied in bio_18 and bio_6 in the LCP model; this could indicate that temperature seasonality across a year affects genetic differentiation, but lower temperature contributes more. Some other factors, such as wind speed (sfcwind), mean temperature of the wettest quarter (bio_8), the max temperature of the warmest month (bio_5), and others show less contribution to genetic differentiation, therefore can be seen as irrelevant factors. Regarding minimum temperature, many studies have shown that temperature, especially cold weather (Chen Y. et al., 2021), has a significant effect on the genetic differentiation and evolution process of living species (Lamb, 1992; Sinclair et al., 2003; Soderberg, 2021), and our results further confirmed this relationship. In contrast to the connectivity map and the original map, in the LCP model, areas with higher coverage of evergreen/deciduous needleleaf trees always occur with a lower genetic connectivity value. Previous studies on other species have shown a positive relationship between elevation and population genetic differentiation (Bowman et al., 2018; Mushegian et al., 2021). Our results showed that elevation has a marginal effect on the *S. empoasca* population's genetic differentiation, although it is not a decisive factor.

Furthermore, genetic diversity indexes and most landscape metrics were not significantly associated—only a few landscape metrics, for example, the cohesion of cropland, significantly negatively affected the *S. empoasca*

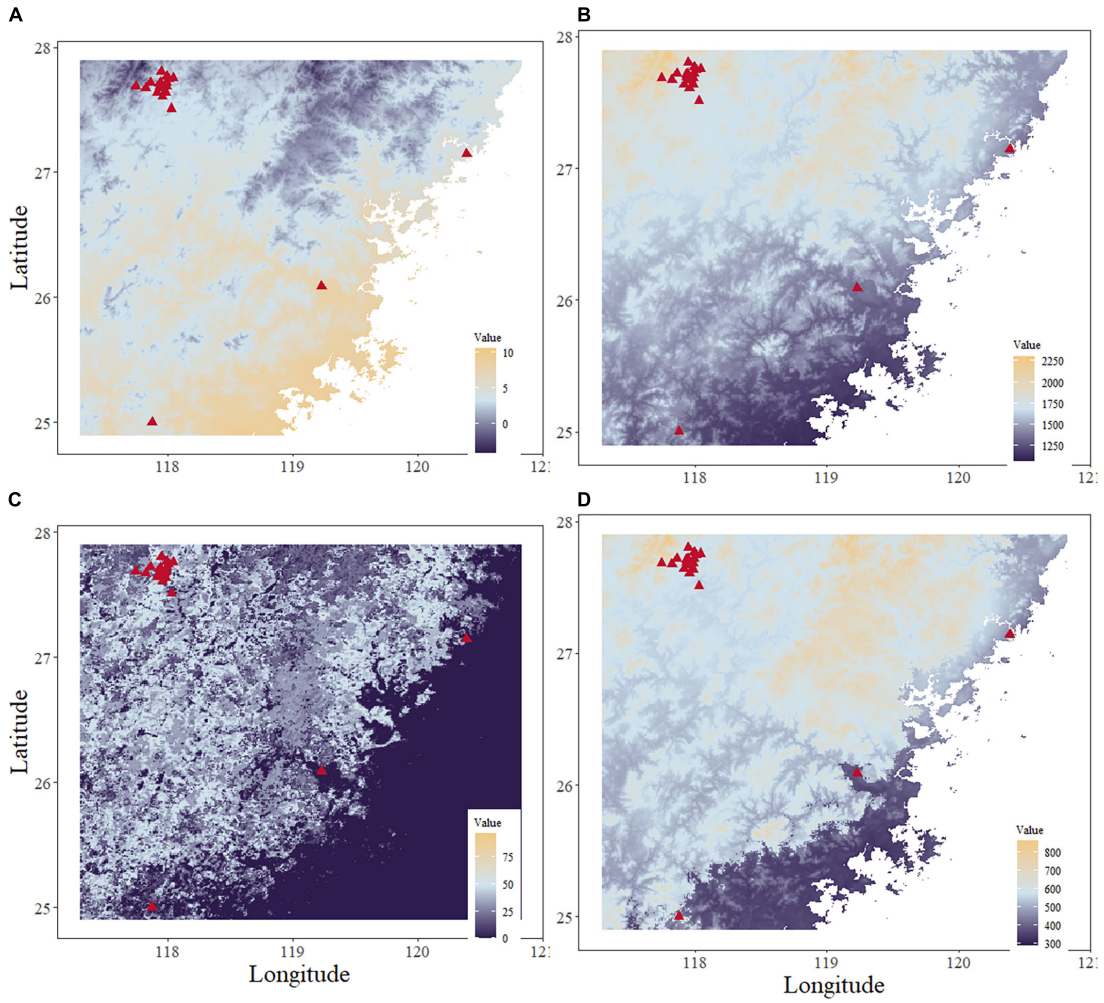


FIGURE 8
Maps of the top four environmental variables. **(A)** Minimum temperature of the coldest month. **(B)** Annual precipitation. **(C)** Evergreen/deciduous needleleaf trees. **(D)** Precipitation in the warmest quarter.

population's genetic diversity. This could be attributed to anthropogenic factors, such as pesticide utilization and farming practices, which may decrease the genetic diversity in *S. empoasca* (Dong et al., 2018; Mushegian et al., 2021).

Conclusion and implications for conservation

Our study indicated that annual precipitation, minimum precipitation in the warmest quarter, and minimum temperature in the coldest quarter are key climate factors in shaping the genetic differentiation of *S. empoasca*; moreover, evergreen/deciduous needleleaf tree land cover is the only key landscape factor that was related to the genetic differentiation of *S. empoasca*. The genetic connectivity map showed that *S. empoasca* populations in our sampling regions are genetically isolated. Therefore, the increasing occurrence of extreme weather events is unfavorable for the growth and development of *S. empoasca* populations, particularly those with a slight pattern of IBD. Our analyses also demonstrated a significant pattern of isolation by geographical distance in *S. empoasca* and a significantly negative effect of cropland on its population's genetic diversity. These findings indicate that reductions in anthropogenic activities may be one of the strategies to ensure better conservation strategies for *S. empoasca* populations. Further, to better promote the natural control of *S. empoasca* on the tea green leafhopper, a relatively stable environment should be considered when managing tea plantations, with lower temperature variation and appropriate precipitation. Besides, our study demonstrated that the XGBoost algorithm could be helpful in mapping genetic connectivity and identifying key environmental factors at fine spatial scales

for living species. From a broader perspective, we believe that the proposed method can be applied to other species at any scale.

To the best of our knowledge, this study is the first to practically explore the landscape genetics of a member of the Mymaridae family. We believe that the findings of this work may facilitate the development of more efficacious strategies for employing these natural enemies in biological control. Future studies could focus on expanding the study scale of landscape genetics.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

JL, HH, and WC: material collection and preparation. JC, LS, ZY, and JS: experiments and data analysis. LS: writing—original draft preparation. JL, SY, and MY: writing—review and editing. MY and SY: supervision. All authors have agreed to be accountable for all aspects of the work, read, and agreed to the published version of the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (Grant Number: 2019YFD1002100), Agricultural "Five New" Program of the Development and Reform Commission of Fujian, China [Minfa Reform Agriculture, Grant Number: (2017) 410], the Natural Science Foundation of Fujian Province, China (Grant Number: 2022J05080), and the Technology Research and Development Program of Quanzhou, China (Grant Number: 2020N008s).

Conflict of interest

ZY, JS, and SY were employed by BGI Genomics Shenzhen Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.943299/full#supplementary-material

References

Avolio, M. L., Beaulieu, J. M., and Smith, M. D. (2013). T diversity of a dominant C4 grass is altered with increased precipitation variability. *Oecologia* 171, 571–581. doi: 10.1007/s00442-012-2427-4

Bivand, R., Rundel, C., Pebesma, E., Stuetz, R., Hufthammer, K. O., and Bivand, M. R. (2017). *Package 'rgeos'. The Comprehensive R Archive Network (CRAN).*

Blacket, M., Robin, C., Good, R., Lee, S., and Miller, A. (2012). Universal primers for fluorescent labelling of PCR fragments—an efficient and cost-effective approach to genotyping by fluorescence. *Mole. Ecol. Resour.* 12, 456–463. doi: 10.1111/j.1755-0998.2011.03104.x

Bossavy, A., Girard, R., and Kariniotakis, G. (2013). Forecasting ramps of wind power production with numerical weather prediction ensembles. *Wind Energ.* 16, 51–63. doi: 10.1002/we.526

Bowman, J., Greenhorn, J. E., Marrotte, R. R., Mckay, M. M., Morris, K. Y., Prentice, M. B., et al. (2016). On applications of landscape genetics. *Conserv. Genet.* 17, 753–760. doi: 10.1007/s10592-016-0834-5

Bowman, L. L., Kondrateva, E. S., Timofeyev, M. A., and Yampolsky, L. Y. (2018). Temperature gradient affects differentiation of gene expression and SNP allele frequencies in the dominant Lake Baikal zooplankton species. *Mole. Ecol.* 27, 2544–2559. doi: 10.1111/mec.14704

Carvalho, F. P. (2017). Pesticides, environment, and food safety. *Food Energy Secur.* 6, 48–60. doi: 10.1002/fes3.108

Chen, K., Peng, Y., Lu, S., Lin, B., and Li, X. (2021). Bagging based ensemble learning approaches for modeling the emission of PCDD/Fs from municipal solid waste incinerators. *Chemosphere* 274:129802. doi: 10.1016/j.chemosphere.2021.129802

Chen, Y., Liu, Z., Regniere, J., Vasseur, L., Lin, J., Huang, S., et al. (2021). Large-scale genome-wide study reveals climate adaptive variability in a cosmopolitan pest. *Nat. Commun.* 12:7206. doi: 10.1038/s41467-021-27510-2

Cranham, J. (1966). Tea pests and their control. *Annu. Rev. Entomol.* 11, 491–514. doi: 10.1146/annurev.en.11.010166.002423

de León, J., and Jones, W. (2005). Genetic differentiation among geographic populations of *Gonatocerus ashmeadi* (Hymenoptera: Mymaridae), the predominant egg parasitoid of Homalodisca coagulata (Homoptera: Cicadellidae). *Insect. Sci.* 5:9. doi: 10.1673/031.005.0201

De Leon, J., Triapitsyn, S., Matteucig, G., and Viggiani, G. (2009). Molecular and morphometric analyses of Anagrus erythroneurae S. Trjapitzin and Chiappinni and A. ustulatus Haliday (Hymenoptera: Mymaridae). *Boll. Entomol. Agrar.* 62, 75–88.

Deiss, L., Margenot, A. J., Culman, S. W., and Demyan, M. S. (2020). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma* 365:114227. doi: 10.1016/j.geoderma.2020.114227

Dong, Z., Li, Y., and Zhang, Z. (2018). Genetic diversity of melon aphids Aphis gossypii associated with landscape features. *Ecol. Evolut.* 8, 6308–6316. doi: 10.1002/ece3.4181

Du, J., Gao, B.-J., Zhou, G.-N., and Miao, A.-M. (2009). Genetic diversity and differentiation of fall webworm (Hyphantria cunea Drury) populations. *Forest. Stud. China* 11, 158–163. doi: 10.1007/s11632-009-0034-1

Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., Srinivasan, K. J. C., and Agriculture, E. I. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Comput. Electr. Agricult.* 155, 257–282. doi: 10.1016/j.compag.2018.10.024

Farooqi, M. M., Khattak, H. A., and Imran, M. (2018). "Data quality techniques in the internet of things: Random forest regression," in *2018 14th International Conference on Emerging Technologies (ICET)*, (Netherland: IEEE), 1–4. doi: 10.1109/ICET.2018.8603594

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

French, C. M., Bertola, L. D., Carnaval, A. C., Economo, E. P., Kass, J. M., Lohman, D. J., et al. (2022). Global determinants of the distribution of insect genetic diversity. *bioRxiv* [Preprint]. doi: 10.1101/2022.02.09.479762

Funasaki, G. Y., Lai, P.-Y., Nakahara, L. M., Beardsley, J. W., and Ota, A. K. (1988). "A review of biological control introductions in Hawaii: 1890 to 1985," in *Proceedings, Hawaiian Entomological Society*, (Netherland: IEEE).

Garba, M., Loiseau, A., Tatard, C., Benoit, L., and Gauthier, N. J. B. O. E. R. (2019). Patterns and drivers of genetic diversity and structure in the biological control parasitoid Habrobracon hebetor in Niger. 109, 794–811. doi: 10.1017/S0007485319000142

Grandgirard, J., Hoddle, M. S., Petit, J. N., Roderick, G. K., and Davies, N. (2007). Engineering an invasion: classical biological control of the glassy-winged sharpshooter, Homalodisca vitripennis, by the egg parasitoid *Gonatocerus ashmeadi* in Tahiti and Moorea, French Polynesia. *Biol. Invas.* 10, 135–148. doi: 10.1007/s10530-007-9116-y

Harrell, F. E. Jr., and Dupont, M. C. (2006). *The Hmisc Package. R package version 3*. 3.

Hesselbarth, M. H., Sciaini, M., With, K. A., Wiegand, K., and Nowosad, J. (2019). landscapemetrics: An open-source R tool to calculate landscape metrics. *Ecography* 42, 1648–1657. doi: 10.1111/ecog.04617

Hether, T., and Hoffman, E. (2012). Machine learning identifies specific habitats associated with genetic connectivity in Hyla squirella. *J. Evolut. Biol.* 25, 1039–1052. doi: 10.1111/j.1420-9101.2012.02497.x

Huber, J. T. (1986). Systematics, biology, and hosts of the Mymaridae and Mymarommatidae (Insecta: Hymenoptera): 1758–1984. *Entomography* 4:185.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Jonsson, M., Kaartinen, R., and Straub, C. S. (2017). Relationships between natural enemy diversity and biological control. *Curr. Opin. Insect. Sci.* 20, 1–6. doi: 10.1016/j.cois.2017.01.001

Kankare, M., Van Nouhuys, S., Gaggiotti, O., and Hanski, I. (2005). Metapopulation genetic structure of two coexisting parasitoids of the Glanville fritillary butterfly. *Oecologia* 143, 77–84. doi: 10.1007/s00442-004-1782-1

Keenan, K., Mcginnity, P., Cross, T. F., Crozier, W. W., and Prodöhl, P. A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol. Evol.* 4, 782–788. doi: 10.1111/2041-210X.12067

Kittlein, M. J., Mora, M. S., Mapelli, F. J., Austrich, A., Gaggiotti, O. E., and Evolution. (2022). Deep learning and satellite imagery predict genetic diversity and differentiation. *Methods Ecol. Evolut.* 13, 711–721. doi: 10.1111/2041-210X.13775

Kotsiantis, S., and Kanellopoulos, D. (2012). Combining bagging, boosting and random subspace ensembles for regression problems. *Int. J. Innov. Comput. Inform. Control* 8, 3953–3961.

Lamb, R. (1992). Developmental rate of Acyrthosiphon pisum (Homoptera: Aphididae) at low temperatures: implications for estimating rate parameters for insects. *Environ. Entomol.* 21, 10–19. doi: 10.1093/ee/21.1.10

Li, J., Shi, L., Chen, J., You, M., and You, S. (2021). Development and characterization of novel microsatellite markers for a dominant parasitoid *Stethynium empoasca* (Hymenoptera: Mymaridae) in tea plantations using high-throughput sequencing. *Appl. Entomol. Zool.* 56, 41–50. doi: 10.1007/s13355-020-00704-8

Liu, W., Liu, W. D., and Gu, J. (2020). Forecasting oil production using ensemble empirical model decomposition based Long Short-Term Memory neural network. *J. Petrol. Sci. Engin.* 189:107013. doi: 10.1016/j.petrol.2020.107013

Manel, S., and Holderegger, R. (2013). Ten years of landscape genetics. *Trends Ecol. Evolut.* 28, 614–621. doi: 10.1016/j.tree.2013.05.012

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evolut.* 18, 189–197. doi: 10.1016/S0169-5347(03)00008-9

Mills, N. J. (1994). Parasitoid guilds: defining the structure of the parasitoid communities of endopterygote insect hosts. *Environ. Entomol.* 23, 1066–1083. doi: 10.1093/ee/23.5.1066

Mitrović, M., Petrović, A., Kavallieratos, N. G., Starý, P., Petrović-Obradović, O., Tomanović, Ž, et al. (2013). Geographic structure with no evidence for host-associated lineages in European populations of Lysiphlebus testaceipes, an introduced biological control agent. *Biol. Control* 66, 150–158. doi: 10.1016/j.biocontrol.2013.05.007

Murphy, M. A., Evans, J. S., and Storfer, A. J. E. (2010). Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology* 91, 252–261. doi: 10.1890/08-0879.1

Mushegian, A. A., Neupane, N., Batz, Z., Mogi, M., Tuno, N., Toma, T., et al. (2021). Ecological mechanism of climate-mediated selection in a rapidly evolving invasive species. *Ecol. Lett.* 24, 698–707. doi: 10.1111/ele.13686

Nadel, R. L., Wingfield, M. J., Scholes, M. C., Lawson, S. A., Noack, A., Neser, S., et al. (2012). Mitochondrial DNA diversity of Cleruchoides noackae (Hymenoptera: Mymaridae): a potential biological control agent for Thaumastocoris peregrinus (Hemiptera: Thaumastocoridae). *Biol. Control* 57, 397–404. doi: 10.1007/s10526-011-9409-z

Nakai, M. (2009). Biological control of tortricidae in tea fields in Japan using insect viruses and parasitoids. *Virol. Sin.* 24, 323–332. doi: 10.1007/s12250-009-3057-9

Opitz, D., and Maclin, R. (1999). Popular ensemble methods: An empirical study. *J. Artif. Intellig. Res.* 11, 169–198. doi: 10.1613/jair.614

Palinski, R., Pauszek, S. J., Humphreys, J. M., Peters, D. P., Mcvey, D. S., Pelzel-Mccluskey, A. M., et al. (2021). Evolution and expansion dynamics of a vector-borne virus: 2004–2006 vesicular stomatitis outbreak in the western USA. *Ecosphere* 12:e03793. doi: 10.1002/ecs2.3793

Pless, E., Saarman, N. P., Powell, J. R., Caccone, A., and Amatulli, G. (2021). A machine-learning approach to map landscape connectivity in Aedes aegypti with genetic and environmental data. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2003201118. doi: 10.1073/pnas.2003201118

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Magaz.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Rodriguez-Saona, C. (2018). Biological Control: Ecology and Applications. *Am. Entomol.* 64, E2–E2. doi: 10.1093/ae/tmy017

Shanley, C. S., Eacker, D. R., Reynolds, C. P., Bennetsen, B. M., Gilbert, S. L., and Management. (2021). Using LiDAR and Random Forest to improve deer habitat models in a managed forest landscape. *Forest Ecol. Manag.* 499:119580. doi: 10.1016/j.foreco.2021.119580

Silva-Brandão, K. L., Santos, T. V., Cônsoli, F. L., and Omoto, C. (2015). Genetic diversity and structure of Brazilian populations of Diatraea saccharalis (Lepidoptera: Crambidae): Implications for pest management. *J. Econ. Entomol.* 108, 307–316. doi: 10.1093/jee/tou040

Sinclair, B. J., Vernon, P., Klok, C. J., and Chown, S. L. (2003). Insects at low temperatures: an ecological perspective. *Trends Ecol. Evolut.* 18, 257–262. doi: 10.1016/S0169-5347(03)00014-4

Soderberg, D. N. (2021). *Susceptibility of High-Elevation Forests to Mountain Pine Beetle (Dendroctonus ponderosae Hopkins) Under Climate Change*. United States: Utah State University.

Sylvester, E. V., Bentzen, P., Bradbury, I. R., Clément, M., Pearce, J., Horne, J., et al. (2018). Applications of random forest feature selection for fine-scale genetic population assignment. *Evol. Appl.* 11, 153–165. doi: 10.1111/eva.12524

Tait, G., Vezzulli, S., Sassù, F., Antonini, G., Biondi, A., Baser, N., et al. (2017). Genetic variability in Italian populations of Drosophila suzukii. *BMC Genet.* 18:87. doi: 10.1186/s12863-017-0558-7

Lemonick. (2000). *Gene Mapper. The bad boy of science has jump-started a biological revolution*. New York: Lemonick, 17.

Triapitsyn, S. V., Adachi-Hagimori, T., Rugman-Jones, P. F., Barry, A., Abe, A., Matsuo, K., et al. (2019). Egg parasitoids of the tea green leafhopper Empoascaonukii (Hemiptera, Cicadellidae) in Japan, with a description of a new species of Anagrus (Hymenoptera, Mymaridae). *ZooKeys* 836, 93–112. doi: 10.3897/zookeys.836.32634

Tuanmu, M. N., and Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Glob. Ecol. Biogeogr.* 23, 1031–1045. doi: 10.1111/geb.12182

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P., and Shipley, P. (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mole. Ecol. Notes* 4, 535–538.

Wellenreuther, M., Sanchez-Guillen, R. A., Cordero-Rivera, A., Svensson, E. I., and Hansson, B. (2011). Environmental and climatic determinants of molecular diversity and genetic population structure in a coenagrionid damselfly. *PLoS One* 6:e20440. doi: 10.1371/journal.pone.0020440

Wright, D., Bishop, J. M., Matthee, C. A., Von Der, and Heyden, S. (2015). Genetic isolation by distance reveals restricted dispersal across a range of life histories: implications for biodiversity conservation planning across highly variable marine environments. *Div. Distrib.* 21, 698–710. doi: 10.1111/ddi.12302

Yue, N., Kuang, H., Sun, L., Wu, L., and Xu, C. (2010). An empirical analysis of the impact of EU's new food safety standards on China's tea export. *Int. J. Food Sci. Technol.* 45, 745–750. doi: 10.1111/j.1365-2621.2010.02189.x

Zepeda-Paulo, F., Dion, E., Lavandero, B., Maheo, F., Outreman, Y., Simon, J.-C., et al. (2016). Signatures of genetic bottleneck and differentiation after the introduction of an exotic parasitoid for classical biological control. *Biol. Invas.* 18, 565–581. doi: 10.1007/s10530-015-1029-6

Zhuang, J., Fu, J., Su, Q., Li, J., and Zhan, Z. (2009). The regional diversity of resistance of tea green leafhopper, Empoasca vitis (GÖthe), to insecticides in Fujian Province. *J. Tea Sci.* 29, 154–158.

frontiers | Frontiers in Genetics

TYPE Opinion
PUBLISHED 09 November 2022
DOI 10.3389/fgene.2022.1047100

OPEN ACCESS

EDITED BY
Yongjie Wu,
Sichuan University, China

REVIEWED BY
Jialiang Li,
Sichuan University, China
Dezhi Zhang,
Institute of Zoology (CAS), China

*CORRESPONDENCE
Yong-Peng Ma,
mayongpeng@mail.kib.ac.cn

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 17 September 2022
ACCEPTED 26 October 2022
PUBLISHED 09 November 2022

CITATION
Tian X-L and Ma Y-P (2022), Horticultural applications of natural hybrids as an accelerating way for breeding woody ornamental plants. *Front. Genet.* 13:1047100. doi: 10.3389/fgene.2022.1047100

Horticultural applications of natural hybrids as an accelerating way for breeding woody ornamental plants

Xiao-Ling Tian[1] and Yong-Peng Ma[2]*

[1]Guiyang Institute of Humanities and Technology, Guiyang, China, [2]Yunnan Key Laboratory for Integrative Conservation of Plant Species With Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

KEYWORDS

hybrid swarm, natural hybrids, woody ornamental plants, rhododendron (Ericaceae), SNP data

China has abundant woody ornamental plant resources comprising a large diversity of families, genera, and species in the wild. There is also a great demand for woody ornamental plants on the Chinese market. It has been estimated that China has the largest production area worldwide (740,954 ha) for the production of woody ornamentals, with a value of 8051 million EUR (AIPH, 2015). Despite Chinese traditional woody cultivars (e.g., roses, *Osmanthus*, and *Camellias*), imported woody ornamental plants have become dominant on the Chinese horticulture market. It should be noted that, in this study, we exclude any woody ornamental plants found on the market without accurate ancestral and/or breeding information, such as many azaleas on the Chinese market (Chang et al., 2020). In contrast to the large diversity of wild resources, horticultural applications of woody ornamental plants remain insufficient, due to long vegetative cycles, making breeding challenging. Moreover, the unavailability of whole-genome information and the difficulties in establishing genetic transformation systems for most woody ornamental plants mean that potential applications of genome editing in these plants have also been limited at present (Van Laere et al., 2018).

Most existing frameworks for breeding of woody ornamental plants have focused on creation of new morphological variations, like classical/traditional cross breeding and artificial polyploidization (Van Huylenbroeck and Van Laere, 2008). Currently, there are few options to accelerate the breeding process, particularly for woody plants with long vegetative cycles. We therefore raise natural hybrids as valuable candidates for solving the problem and facilitating selection of new cultivars.

It is well recognized that natural hybridization occurs frequently in >25% of known plant species (Mallet, 2005) and plays important roles in several aspects of evolution, including the origins of new ecotypes or species, the origin and transfer of genetic adaptations, and the reinforcement or breakdown of reproductive barriers (Rieseberg and Carney, 1998). Due to the characteristics of different genetic combinations from parental species in natural hybrids, various phenotypes for horticultural usage can be created, including different leaf types, flower shapes, colors, and scents. However, this great potential of natural hybrids to act as a source for the breeding of new cultivars has been often ignored, probably owing to the big gap in focus between plant taxonomists and horticulturalists worldwide.

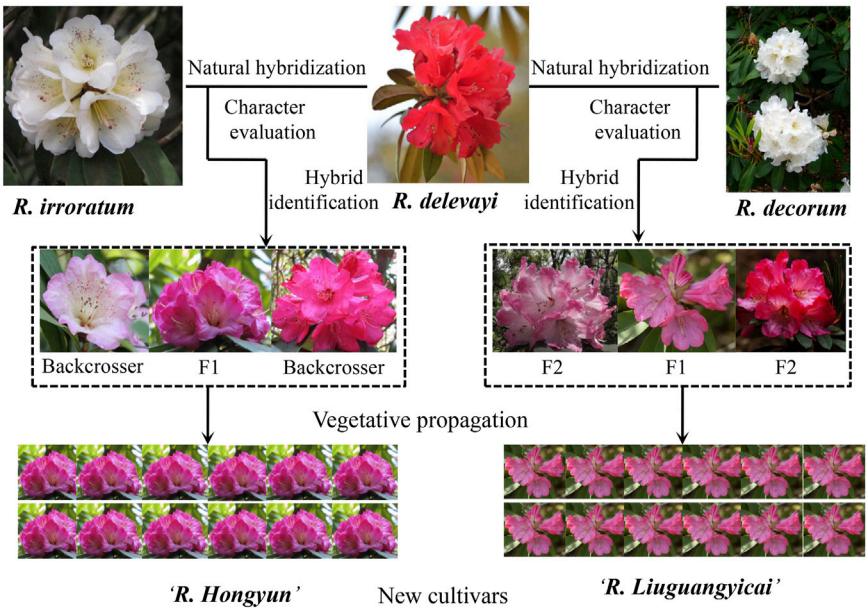


FIGURE 1
Framework with a step-by-step approach for the rapid production of new cultivars from natural hybrids, using case studies from alpine rhododendrons.

Whether these natural hybrids are a stable, long-term phenomenon or a relatively recent occurrence, certain hybrid genotypes will have been selected out because of genetic incompatibilities and other reasons (e.g., the BDM incompatibility model, Coyne and Orr, 2004). Thus, those natural hybrids with poor adaptability are effectively removed by nature and the remaining hybrids that we find growing in the field may be adapted to the local environment. With respect to this aspect, using natural hybrids for the breeding of new cultivars is not only time saving but may also result in fitter and more locally adapted cultivars than those created by traditional open or controlled cross-pollinations, which can often produce unfit hybrids due to hybrid sterility and/or inviability.

Here, we develop a framework with a step-by-step approach for the rapid production of new cultivars from natural hybrids (Figure 1). We use case studies from alpine rhododendrons, which have a much higher horticultural value than those of azaleas (e.g., larger leaves, improved floral shapes, scents, and various colors), but which take a longer time than azaleas to flower (some azaleas can flower from seedlings within 2 years). 1) Confirmation of natural hybrids by integrating morphological and genetic evidence. For early generation hybrids, intermediate morphological characteristics are often displayed (Ma et al., 2010a,b; Ma et al., 2014), and we have demonstrated that most natural hybrids between *Rhododendron delavayi* and *R. irroratum* in the Baili Scenic Reserve (in the western Guizhou province) seem to have been detectable with adequate morphological data alone (detailed statistical analysis of morphological data can be found in Marczewski et al., 2016). Further confirmation of hybrids can be carried out using genetic data (e.g., ISSR, AFLP, SSR, and SNPs). For example, principal component analysis (PCA) implemented in GenAlEx 6.0 (Peakall and Smouse, 2006) and structural analysis of genetic data can reveal genetic relationships between hybrids and parents (Hubisz et al., 2009), and the genetic compositions of hybrids from parental species, whereas NewHybrids (Anderson and Thompson, 2002), together with data simulations, can clarify six genotypes (two parents, F_1, F_2, and two backcrosses to F_1) to which natural hybrids may belong (Ma et al., 2014, 2019). If sufficient molecular markers (e.g., SNP data) can be obtained, up to 45 natural hybrid genotypes can be assessed (Milne and Abbott, 2008). 2) Characteristic evaluation and introduction of natural hybrids into botanical gardens or nurseries. Through this step, natural hybrids with good ornamental value and adaptations can be targeted. However, the horticultural success of these natural hybrids also depends on their survival and easy vegetative propagation in alternate habitats. Some work with regards to the introduction, and domestication of these natural hybrids into a horticultural environment needs to be performed. 3) Vegetative propagation to obtain uniformity and stability in the horticultural characteristics of natural hybrids. In general, stability in horticultural characteristics can be achieved through rooting and grafting; although if these methods are unsuccessful, tissue culture can be employed to fix the ornamental characteristics. 4) New cultivar application

following guidelines of the DUS (distinctness, uniformity, and stability) test and subsequent commercial demonstrations.

Although in the genomic era, integration of the genomic technologies (e.g. GWAS associated trait selection) for ornamental woods with availability of whole-genome information will enhance selection efficiency, we are confident that our framework can provide a reference and be beneficial to colleagues, particularly those working toward breeding new cultivars of most woody ornamental plants without whole-genome information and establishment of genetic transformation systems. Several new cultivars of Chinese indigenous plants generated according to our proposed framework will be emerging in the near future, and we hope that some of them will be commercially competitive against the imported cultivars on the Chinese flower market even before the successful application of genome editing in woody ornamental plants in China.

Author contributions

X-LT and Y-PM conceived and wrote the manuscript.

Funding

This opinion was supported by the National Natural Science Foundation of China (No. 31901237) and the Key Basic Research Program of Yunnan Province, China (No.202101BC070003).

Acknowledgments

The authors would like to thank Dr. Marczewski Jane for her help with English modification of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer DZ declared a shared affiliation with the author YM to the handling editor at the time of review.

Publisher's note

References

Aiph (2015). *International statistics flowers and plants 2015*. UK: Reading.

Anderson, E. C., and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160, 1217–1229. doi:10.1093/genetics/160.3.1217

Chang, Y., Tian, X., Zhang, C., and Ma, Y. (2020). Problems and thoughts on the classification of Rhododendron cultivars in China. *Word For. Res.* 33, 60–65.

Coyne, J. A., and Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi:10.1111/j.1755-0998.2009.02591.x

Ma, Y., Marczewski, T., Xue, D., Wu, Z., Liao, R., Sun, W., et al. (2019). Conservation implications of asymmetric introgression and reproductive barriers in a rare primrose species. *BMC Plant Biol.* 19, 286. doi:10.1186/s12870-019-1881-0

Ma, Y., Milne, R. I., Zhang, C., and Yang, J. B. (2010a). Unusual patterns of hybridization involving a narrow endemic *Rhododendron* species in Yunnan (Ericaceae), China. *Am. J. Bot.* 97, 1749–1757. doi:10.3732/ajb.1000018

Ma, Y., Zhang, C., Zhang, J., and Yang, J. (2010b). Natural hybridization between *Rhododendron delavayi* and *R. cyanocarpum* (Ericaceae), from morphological, molecular and reproductive evidence. *J. Integr. Plant Biol.* 52, 844–851. doi:10.1111/j.1744-7909.2010.00970.x

Ma, Y., Xie, W., Tian, X., Sun, W., Wu, Z., and Milne, R. I. (2014). Unidirectional hybridization and reproductive barriers between two heterostylous primrose species in NW Yunnan, China. *Ann. Bot.* 113, 753–761.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi:10.1016/j.tree.2005.02.010

Marczewski, T., Ma, Y., Zhang, X., Sun, W., and Marczewski, A. J. (2016). Why is population information crucial for taxonomy? A case study employing a hybrid swarm and related described varieties. *AOB Plants* 8. doi:10.1093/aobpla/plw070

Milne, R. I., and Abbott, R. J. (2008). Reproductive isolation among two interfertile *Rhododendron* species: Low frequency of post-F1 hybrid genotypes in alpine hybrid zones. *Mol. Ecol.* 17, 1108–1121. doi:10.1111/j.1365-294X.2007.03643.x

Peakall, R., and Smouse, P. E. (2006). Genalex 6: Genetic analysis in excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi:10.1111/j.1471-8286.2005.01155.x

Rieseberg, L. H., and Carney, S. E. (1998). Plant hybridization. *New Phytol.* 140, 599–624. doi:10.1046/j.1469-8137.1998.00315.x

Van Huylenbroeck, J., and Van Laere, K. (2008). Breeding strategies for woody ornamentals. *Int. Symposium Woody Ornamentals Temp. Zone* 885, 391–401.

Van Laere, K., Hokanson, S. C., Contreras, R., and Van Huylenbroeck, J. (2018). "Woody ornamentals of the temperate zone,". Editor J. Van Huylenbroeck, Ornamental crops, handbook of plant breeding, 11. doi:10.1007/978-3-319-90698-0_29

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fgene.2022.1020789

OPEN ACCESS

EDITED BY
Yongjie Wu,
Sichuan University, China

REVIEWED BY
Ming Li,
University of Konstanz, Germany
Lei Chen,
Sichuan University, China

*CORRESPONDENCE
Wenjuan Shan,
swj@xju.edu.cn
Deyan Ge,
gedy@ioz.ac.cn

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 16 August 2022
ACCEPTED 10 November 2022
PUBLISHED 23 November 2022

CITATION
Mu D, Wu X, Feijó A, Wu W, Wen Z, Cheng J, Xia L, Yang Q, Shan W and Ge D (2022), Transcriptome analysis of pika heart tissue reveals mechanisms underlying the adaptation of a keystone species on the roof of the world. *Front. Genet.* 13:1020789. doi: 10.3389/fgene.2022.1020789

Transcriptome analysis of pika heart tissue reveals mechanisms underlying the adaptation of a keystone species on the roof of the world

Danping Mu[1,2], Xinlai Wu[2,3], Anderson Feijó[2], Wei Wu[4], Zhixin Wen[2], Jilong Cheng[2], Lin Xia[2], Qisen Yang[2], Wenjuan Shan[1]* and Deyan Ge[2]*

[1]Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Science and Technology, Xinjiang University, Urumqi, China, [2]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, [3]Key Laboratory of Zoological Systematics and Application, School of Life Science, Institute of Life Science and Green Development, Hebei University, Baoding, Hebei, China, [4]CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, Sichuan, China

High-altitude environments impose intense stresses on living organisms and drive striking phenotypic and genetic adaptations, such as hypoxia resistance, cold tolerance, and increases in metabolic capacity and body mass. As one of the most successful and dominant mammals on the Qinghai-Tibetan Plateau (QHTP), the plateau pika (*Ochotona curzoniae*) has adapted to the extreme environments of the highest altitudes of this region and exhibits tolerance to cold and hypoxia, in contrast to closely related species that inhabit the peripheral alpine bush or forests. To explore the potential genetic mechanisms underlying the adaptation of *O. curzoniae* to a high-altitude environment, we sequenced the heart tissue transcriptomes of adult plateau pikas (comparing specimens from sites at two different altitudes) and Gansu pikas (*O. cansus*). Differential expression analysis and weighted gene co-expression network analysis (WGCNA) were used to identify differentially expressed genes (DEGs) and their primary functions. Key genes and pathways related to high-altitude adaptation were identified. In addition to the biological processes of signal transduction, energy metabolism and material transport, the identified plateau pika genes were mainly enriched in biological pathways such as the negative regulation of smooth muscle cell proliferation, the apoptosis signalling pathway, the cellular response to DNA damage stimulus, and ossification involved in bone maturation and heart development. Our results showed that the plateau pika has adapted to the extreme environments of the QHTP *via* protection against cardiomyopathy, tissue structure alterations and improvements in the blood circulation system and energy metabolism. These adaptations shed light on how pikas thrive on the roof of the world.

KEYWORDS

high-altitude adaptation, transcriptome analysis, heart tissues, pika, gene expressions, cold tolerance, Qinghai-Tibetan Plateau

Introduction

The Qinghai-Tibetan Plateau (QHTP) in China has an average altitude of more than 4,000 m, the average annual temperature is less than 10°C, and the oxygen concentration is only approximately 50% of the average index on Earth, making it one of the harshest places for animals to live. As a result, species inhabiting the QHTP have evolved distinctive morphological, behavioural, and physiological mechanisms to cope with severe selection pressure, including a lack of oxygen, low temperatures, and intense ultraviolet radiation (Qu et al., 2013b; Zhu et al., 2017; Feijo et al., 2020). The QHTP is therefore an area of great interest to many scholars, and these species have become invaluable models for the comparative analysis of local adaptations. Studies have shown that phenotypic and physiological adaptations for living in extreme environments are recorded at the level of gene transcription (Lan et al., 2018). In particular, studies of high-altitude adaptation have revealed numerous rapidly evolving genes that have undergone natural selection or show different expression patterns in high-altitude species compared to low-altitude inhabitants (Marfell et al., 2013). However, how high-altitude environments shape gene expression patterns remains largely unknown (Tang et al., 2017).

To ensure adequate oxygen to meet the basic needs of life, animals have evolved an elaborate physiological system consisting of respiratory organs (lungs), transport vehicles (erythrocytes), transport channels (blood vessels), and a circulatory power system (heart) (Azad et al., 2017). The role of the heart is to promote blood flow, provide sufficient oxygen and various nutrients to organs and tissues, and remove the final products of metabolism (such as carbon dioxide, inorganic salts, urea, and uric acid), allowing cells to maintain normal metabolism and function (Ream et al., 2008; Azad et al., 2017). Increases in heart rate, blood pressure, and other reactions to altitude can result in changes in cardiac structure and function (Ai et al., 2014). Studies have concluded that an increased heart rate and elevated blood pressure can occur in high-altitude environments, and severe altitude hypoxia can cause myocardial interstitial oedema, degeneration, necrosis, and scar formation (Sahota and Panwar, 2013). The subsequent compensatory increase in cardiac output results in a greater ventricular load, ventricular hypertrophy, a slow heart rate, conduction block, and other arrhythmias (Dor et al., 2001; Zhang et al., 2017).

As adaptations to the harsh high-altitude environment, the ability to promote the expression of genes related to adaptation to hypoxic conditions, protect cardiac muscle cells and tissue structure, and increase blood circulation has been observed in the indigenous people of the QHTP (Zhang et al., 2017). Moreover, Guan et al. (2017) found that yak (*Bos grunniens*) has larger lung and heart tissues than cattle (*Bos taurus*), and the differential miRNAs between these two tissues show synergistic effects playing a significant role in high-altitude adaptation. Zhang et al. (2017) found that the *HIF-1* signalling pathway, cardiovascular development, and *VEGF* signalling pathway may play critical roles in hypoxia adaptation. Overall, these studies reveal genes and pathways related to high-altitude adaptation and demonstrate that cardiopulmonary organs show adaptive transcriptional changes in high-altitude environments. Nevertheless, the heart gene expression patterns of animals at high altitudes are not completely clear, and additional systematic and in-depth studies of the adaptive strategies and regulation mechanisms of animal cardiopulmonary organs in this environment are needed.

Pika (*Ochotona* spp.) is a small nonhibernating herbivorous mammal distributed on the QHTP, in its vicinity and in mountains of Central Asia, Northeast Asia, and North America (Gureev, 1964; Su and Liu, 2000; Melo Ferreira et al., 2015; Koju et al., 2017; Solari and Hadly, 2018). They are highly adapted to cold and alpine environments (Wang et al., 2020b). In addition, the plateau pika (*O. curzoniae*) is one of the most dominant keystone mammal species on the QHTP (Wilson and Smith, 2015; Jukes, 2018; Smith et al., 2019). Studies have demonstrated that plateau pikas have evolved several adaptations to low-temperature and low-oxygen conditions (Montuire, 2001; Liu et al., 2012; Wang et al., 2020b). Qi et al. (2008) showed that plateau pikas can improve their adaptation to anoxic environments by increasing the density of heart mitochondria and the microvasculature and myoglobin contents. Zhu et al. (2021) measured the resting metabolic rate and transcriptional expression levels of adipose tissue, liver and skeletal muscle in pikas along different elevation gradients and proved that plateau pikas could adapt to the extreme environment of the QHTP through increases in the expression of thermogenesis genes and energy metabolism. However, current research on the high-altitude adaptation of pikas is mainly focused on nonshivering thermogenesis (Wang et al., 1999; Wang et al., 2006), the regulation of skeletal muscle lactate dehydrogenase-c, liver hypoxia and pulmonary circulation (Liu et al., 2012; Melo Ferreira et al., 2015; Pichon et al., 2015; Zhu et al., 2021). The role and adaptive mechanisms of cardiac tissue in extreme environments are less well studied, and most of the relevant previous studies have been focused on single species of pikas, without intraspecific and interspecific comparisons.

The life history of a species is the product of natural selection and can reflect the continuous adaptation of a species to its habitat. It is important to further explore the adaptive evolution of species by analysing the differences in closely related species in

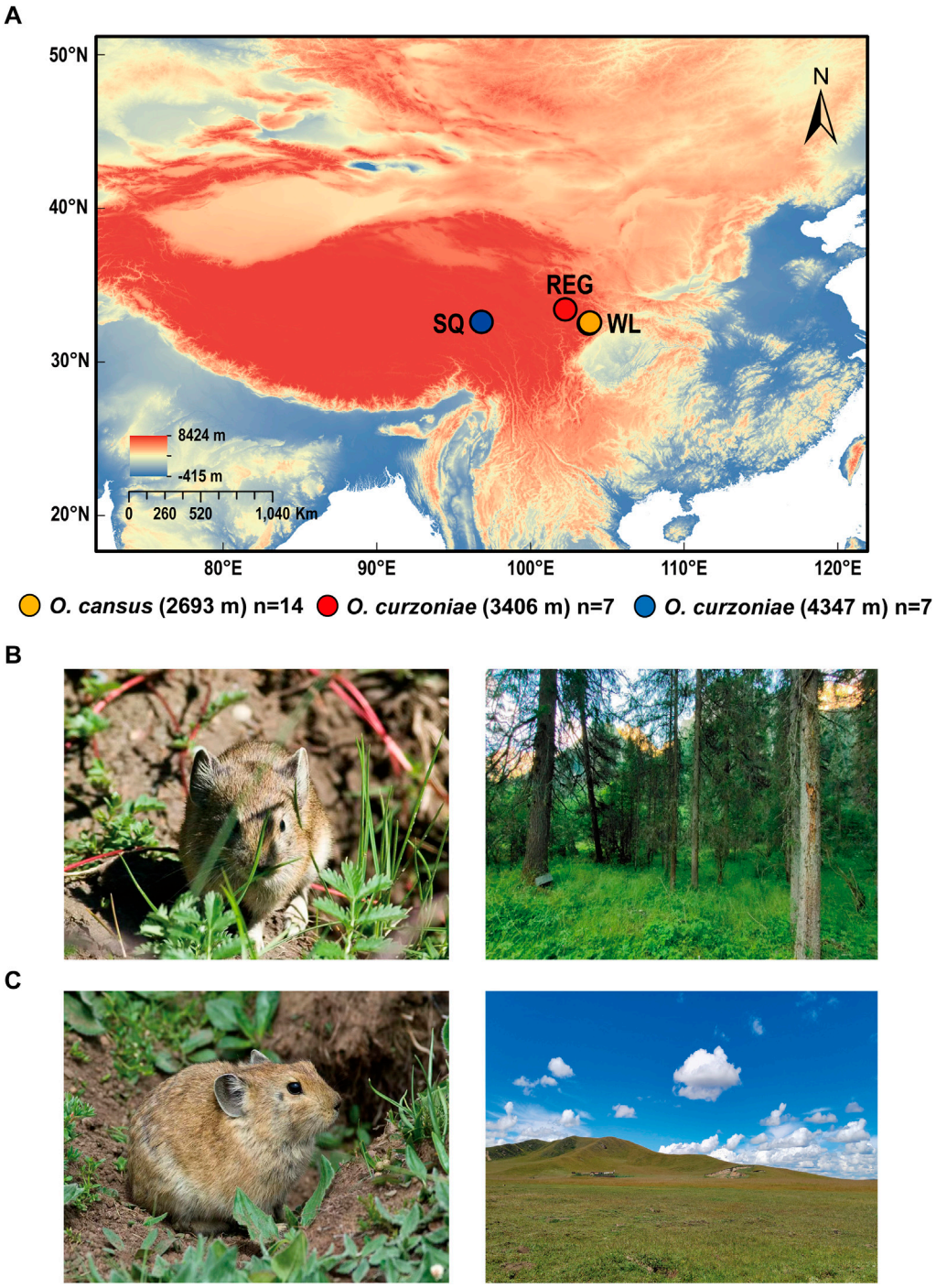


FIGURE 1
Diagram of pika sample collection sites. **(A)** Collection of pika samples at WL, REG and SQ. **(B)** *O. cansus* (From Andrey Lissovsky) and collection site in the Wanglang National Nature Reserve (From Danping Mu). **(C)** *O. curzoniae* (From Andrey Lissovsky) and collection site in the Zoige Wetland National Nature Reserve (From Danping Mu).

distinct habitats (Qu et al., 2013a). Previous studies have shown that pikas originated on the QHTP and gradually expanded to Eurasia and North America (Wang et al., 2020b). Among extant species, the Gansu pika (*O. cansus*) mainly inhabits the alpine bushes and forest margins of the mountainous coniferous and broad-leaved mixed forest belt from altitudes of 2,200–4,000 m.

This species excavates burrows near roots, ridges in grass fields and piles of rocks (Su et al., 2005). However, the plateau pika mainly lives in alpine meadows and open grasslands from altitudes of 3,100–5,100 m, and its distribution on the QHTP is the most extensive among pika species, obviously exceeding that of its congeneric species (Smith and Foggin, 1999). Therefore, in this study, cardiac tissues of Gansu pika and plateau pika, two closely related species, were selected for a comparative transcriptome analysis.

Here, we explore the potential genetic mechanisms underlying adaptation to high-altitude environments in pikas by analysing cardiac transcriptomic differences. Heart tissues of plateau pika and Gansu pika were collected for transcriptome sequencing, and the differentially expressed genes (DEGs) and their main functions were identified through differential expression analysis and weighted gene co-expression network analysis (WGCNA). The results showed that there were considerable differences in the expression patterns between plateau pikas and Gansu pikas. In addition, we identified slight differences between plateau pikas from different altitudes. The genes identified *via* differential expression analysis and WGCNA were mainly enriched in signal transduction, energy metabolism, material transport, negative regulation of smooth muscle cell proliferation and heart development.

Materials and methods

Sequencing material collection

Fourteen *O. cansus* adults were collected in Wanglang National Nature Reserve (WL, N = 32.9050, E = 104.0540, 2,693 m), seven *O. curzoniae* adults were collected in Zoige Wetland National Nature Reserve (REG, N = 33.6499, E = 102.8204, 3,406 m), and seven *O. curzoniae* were collected in Shiqu (SQ, N = 32.9950, E = 98.4420, 4,347 m) (Figure 1; Supplementary Table S1). We collected pika heart tissue, cut it into small pieces and preserved the obtained tissue in RNA preservation solution for 4 h at room temperature. The tissue was then frozen in liquid nitrogen for 10 min and stored at −80°C for subsequent RNA extraction. Voucher specimens were preserved in the National Zoological Museum, Institute of Zoology, Chinese Academy of Sciences.

RNA extraction and transcriptome sequencing

Transcriptome sequencing was conducted on 28 heart tissue samples from two pika species (Supplementary Table S1). Total RNA was extracted from heart tissues using the RNeasy Fibrous Tissue Mini Kit (Tiangen Biotech, Beijing, China) according to the manufacturer's instructions (Hao et al., 2019). We used 2 μg of RNA from each sample as the input material for sequencing. Before library preparation, we checked RNA integrity, purity, and concentrations. RNA purity was checked with a Nanodrop 1000 UV instrument (Thermo Fisher Technologies, United States). RNA degradation and contamination were monitored by agarose gel electrophoresis. A Qubit™ 4.0 Fluorometer (Invitrogen ABI, Palo Alto, CA, United States) was used to measure the concentration of RNA, and an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States) and associated RNA 6000 Nano kits were used to assess and quantify the quality of RNA. Subsequently, the mRNA was isolated with Oligo (dT) magnetic beads, and the fragments were used to synthesize cDNA. After purification and terminal repair, an A (adenine) nucleotide was added. The mRNA fragments were linked with an adapter and amplified by PCR. Sequencing was conducted on the Illumina NovaSeq HiSeq platform to generate 150 bp paired-end reads. Sequencing was conducted by Berry Genomics (Beijing).

Transcriptome quality controls and quantitative Salmon analysis

The quality control and preprocessing of the data were carried out by using FastQC V0.11 (Andrews, 2010). Fastp V0.23.1 (Chen et al., 2018) was used to filter adaptors and poor-quality reads. Then, Salmon V1.5.2 (Patro et al., 2017) was used to quickly quantify transcript expression, and the genome assembly of *O. curzoniae* (RefSeq GCF_017591425.1, NIBS_Ocur_1.0 from NCBI) was used as a reference. The expression of transcripts was estimated from dual-end sequencing data, and the "quant.sf" quantification file was the output file. The transcripts per million (TPM) method was chosen to represent the expression level of each individual gene.

Differential expression analysis of RNA-Seq datas

Salmon was utilized to calculate gene expression levels in 28 individuals, and the Salmon results were then normalized and log transformed. We obtained 32,693 genes to construct the dds matrix with the DESeq2 V1.34.0 R package (Love et al., 2014) to analyse differential expression in the heart tissues. In particular, we used the DESeq2 package to conduct differential expression analysis based on the negative binominal distribution. Prior to the differential expression analysis, a PCA was performed on the standardized expression results of 14 samples (five females and nine females) of Gansu pikas to examine the influence of sexual dimorphism.

The *p*-value generated by Benjamin and Hochberg's calculation method (false discovery rate, FDR) (Storey, 2002;

Storey et al., 2004; Bin and Steve, 2005), was used as the parameter for screening DEGs (Benjamini and Hochberg, 1995; Maurano et al., 2012). DEGs were screened according to an FDR < 0.01 and a |fold change (FC)| > 2. DEGs were visualized in a volcano plot by using the R package "ggplot2 (V3.36)" (Yang et al., 2021) and in a heatmap plot by using the R package pheatmap V1.0.12 (Kucukural et al., 2019). PCA was performed after the dds matrix was normalized in the above steps. The first two PCA axes were used to visualize the differentiation among individuals (Bao et al., 2021). The selected DEGs were used for subsequent functional enrichment analysis.

Weighted gene coexpression network analysis

We calculated gene expression levels from the raw counts of each sample, and TPM values calculated by Salmon were used as a measure of transcript abundance to construct a gene expression matrix for 28 samples. The transformation of the gene expression matrix was normalized using the variance-stabilizing transformation (VST) included in DESeq2 (Bin and Steve, 2005). Herein, the top 20% of upregulated genes were selected for WGCNA (Zhu et al., 2019; Bai et al., 2020; Liu et al., 2020; Gong et al., 2022; Wang et al., 2022). A weighted gene coexpression network was constructed using the WGCNA V.1.66 R package (Langfelder and Horvath, 2008) to ensure that the maximum number of genes related to altitude were obtained. A network based on the approximate scale-free topology was constructed by selecting the most suitable soft threshold, which resulted in a scale-free R^2 fit. We then performed average linkage hierarchical clustering with TOM-based dissimilarity to construct a dendrogram, setting 0.3 at the height cut-off and 25 as the minimum module size. Modules with a higher correlation were merged (r < 0.25). Each module was assigned different colours for visualization. We calculated the first principal components as a measure of module expression. Genes with similar expression patterns were grouped into a coexpression module with a specific molecular mechanism. Then, the expression levels were accompanied by the trait data, and association analysis between module genes and traits was conducted. Spearman's correlation was used to analyse the correlation between characteristic module genes and altitude, and the module with the highest correlation was selected as the module related to altitude adaptation (Langfelder and Horvath, 2008). The hub genes in the elevation-related module were extracted based on the criteria of a kME value greater than 0.8 and a significance of genes and traits greater than 0.2 (Wu et al., 2016). Scatter plots were generated to illustrate the maximum module (MM) and gene significance (GS) genetically related to altitude (Pei et al., 2017) based on the cut-off criteria of |MM| > 0.8 and |GS| > 0.2, which were extracted from the centers in Gene Modules of Interest (Tang et al., 2018). Hub genes are those that show high connectivity in the network and play crucial roles in biological processes and influence the regulation of other genes in related pathways (Bao et al., 2021).

Gene functional enrichment and pathway analysis

To assess the gene functions and metabolic pathways of the screened hub genes, we used ShinyGO V0.75 (Ge et al., 2020) for the functional enrichment analysis of the selected genes obtained from the above analyses. ShinyGO is a Shiny application developed based on several R/Bioconductor packages and a large annotation and pathway database compiled from many sources, with graphical visualization of enrichment, pathways, gene characteristics and protein interactions. The American pika was chosen as a background species with a *p*-value cut-off (FDR) < 0.05 as the major criterion for selecting the significantly enriched biological pathways, and the $p < 0.05$ value corrected using the Benjamini & Hochberg algorithm was set as the threshold for identifying biological processes and pathways. In addition, the intersection dataset of DEGs and hub genes related to altitude was retrieved, and the genes screened *via* the two methods were analysed to determine the mechanism whereby pika heart tissue has adapted to the plateau environment.

Results

Results of transcriptome sequencing and Salmon analysis

We obtained a total of 106 Gb of clean read files for the 28 heart samples. The average number of clean reads was 23,415,709.86, and the average amount of data was approximately 7.0 GB. The average GC content of the 28 samples was 50.7%, indicating that the transcriptome sequencing data were of high quality and could be further analysed (Supplementary Figure S1; Supplementary Table S2). The quantitative results were filtered to remove genes with read counts of less than half the total number of samples, and the expression results of 32,693 standardized genes were obtained.

Gene expression and cluster analysis

PCA on samples of O. cansus from Wanglang revealed insignificant differences between female and male (Supplementary Figure S2). Then, we combined both gender in the following analyses.

In the analysis of transcriptome expression results of different pika hearts, the results of PCA showed that WL samples were separated from REG and SQ samples and that

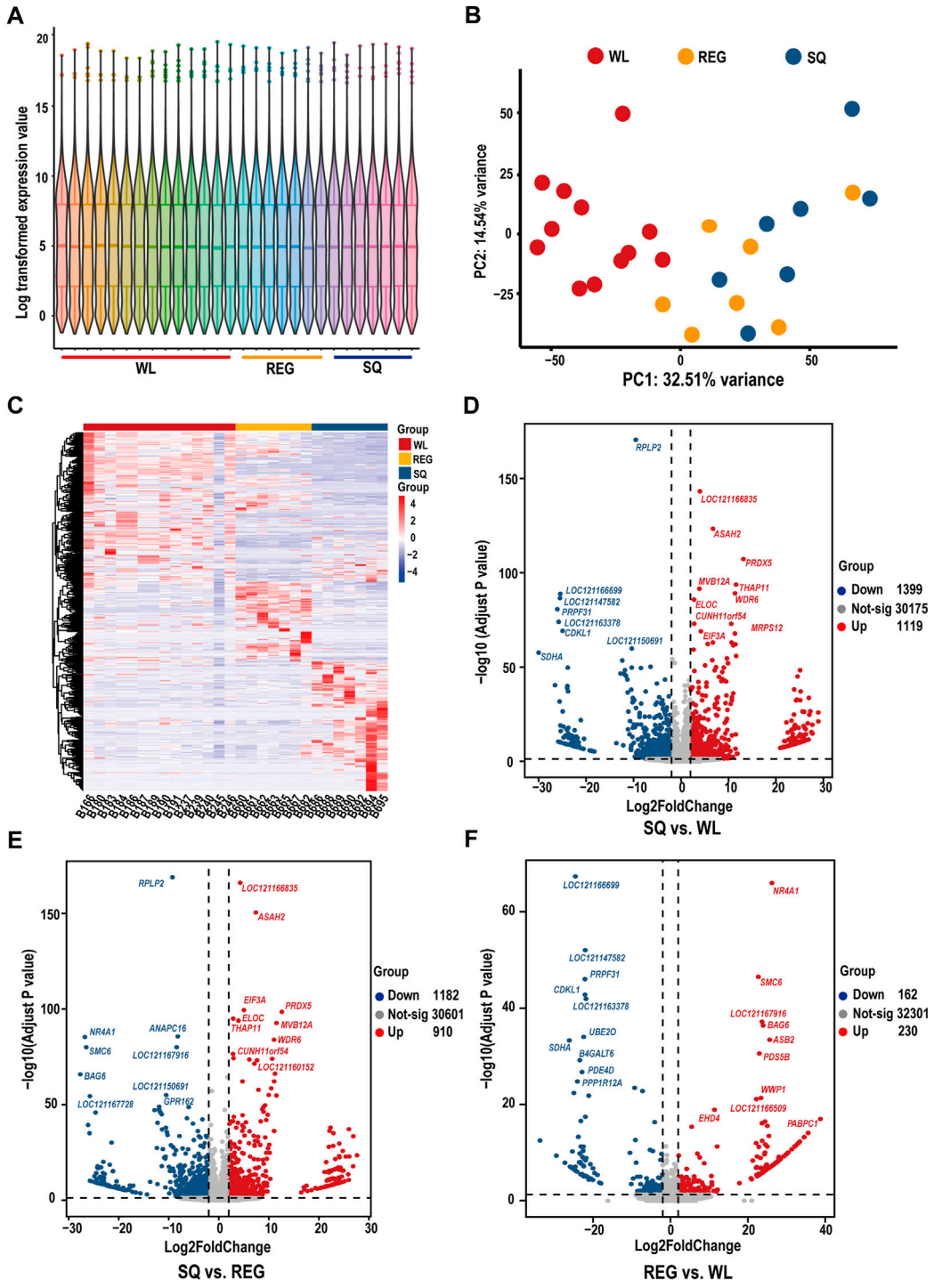


FIGURE 2
Expression levels, principal component analysis, differential gene expression heatmap and pair-to-pair-differential gene analysis based on the heart transcriptome of pikas. **(A)** Violin diagram of gene expression levels in the 28 samples. **(B)** PCA of the expression data of heart tissue samples from three populations of pikas. WL indicates Gansu pikas collected at 2,693 m, REG refers to plateau pikas collected at 3,406 m, and SQ refers to plateau pikas collected at 4,347 m. **(C)** Heatmap of the top 1% of genes according to their expression levels in 28 samples. **(D)** Volcano plot of differentially expressed genes in SQ vs. WL (note the ten most expressed genes). **(E)** Volcano plot of differentially expressed genes in SQ vs. REG (note the ten most expressed genes). **(F)** Volcano plot of differentially expressed genes in REG vs. WL (note the ten most expressed genes).

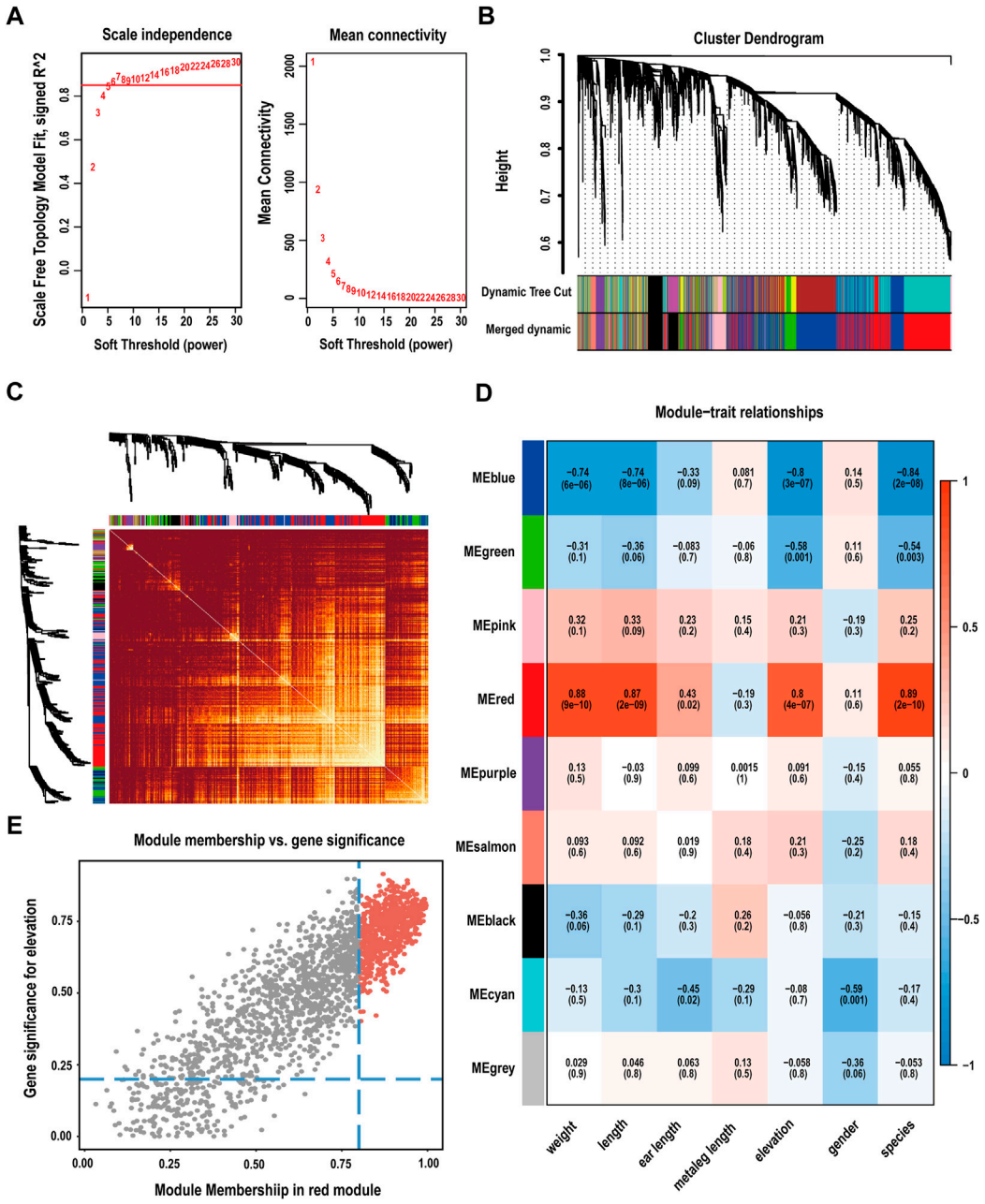


FIGURE 3
Construction of a weighted gene coexpression network. **(A)** Analysis of soft-thresholding power based on scale independence (left) and means connectivity (right). **(B)** Cluster dendrogram of genes. Modules considered to have high similarity were mixed. **(C)** Network heatmap plots of all genes in the WGCNA. **(D)** Heatmap of the correlations between modules and phenotypes. Altitude-related genes were mainly enriched in the red module. **(E)** Scatter plot of correlations between modules and phenotypes, screening of genes as hub genes according to |MM| > 0.8 and |GS| > 0.2 (red area).

there may be expression pattern differences between Gansu pika and plateau pika heart tissues (Figure 2B). According to a $|\log2FoldChange| \geq 1$ and padj < 0.05, 2518 DEGs were identified between plateau pikas from SQ and WL, and 1,119 genes were found to be significantly upregulated and 1,399 genes to be significantly downregulated (Figure 2D). Similarly, a total of 2093 DEGs were identified in the heart tissues of plateau pikas from REG vs. WL, among which 910 genes were significantly upregulated, and 1,182 genes were significantly downregulated (Figure 2F). On the other hand, only

392 DEGs were detected between the two populations of plateau pikas from SQ and REG, among which 230 genes were significantly upregulated, and 162 genes were significantly downregulated (Figure 2E). The results of volcano map analysis were consistent with the results of PCA, which showed significant differences in the expression patterns of cardiac tissues of different pikas. In other words, the expression patterns of plateau pikas and Gansu pikas were quite different. In addition, we found subtle differences among plateau pikas at different altitudes.

Weighted gene co-expression network

In the rapid analysis of expression, 32,693 genes in 28 samples were quantitatively analysed. For a more precise analysis, we selected the top 20% of genes and eliminated outlier data to perform WGCNA. Module expression was summarized using the first principal component of gene expression for each module and regressed against the altitude of the sample. A soft threshold ($\beta = 6$, scale-free $R^2 = 0.85$) was used to guarantee a scale-free network (Figure 3A; Supplementary Figure S3B), which identified thirteen modules and met the conditions required for standard scale-free network construction. By obtaining the feature vector of each module and merging similar modules, nine gene coexpression modules were obtained (Figures 3B,C). The number of genes per module ranged from a minimum of 10 (grey module) to a maximum of 2,457 (blue module) (Supplementary Tables S7, S8). Genes that could not be split into any modules were placed in the grey module and identified as non-coexpressed genes.

Significantly correlated modules and enrichment of hub genes

To identify the modules that were most relevant to phenotype, we performed correlation analysis between modules and traits. The results showed that the red module (R = 0.8, P = 4E-0.7) presented the highest correlation with altitude and contained 2015 genes (Figure 3D; Supplementary Table S9). Hub genes are those with high connectivity in the network; these genes play crucial roles in biological processes and affect the regulation of other genes in related pathways. Using a kME value greater than 0.8 and gene and trait similarity values greater than 0.2 as screening criteria, 706 hub genes associated with high-altitude adaptability were screened from the red module (Figure 3E).

Enrichment analysis of differentially expressed genes

Through the differential expression analysis of WL, REG and SQ samples, we found significant differences between the plateau pika and Gansu pika, while only small differences were observed between plateau pikas from SQ and REG. GO enrichment analysis showed that compared with Gansu pikas from low altitudes, the upregulated genes identified in plateau pikas from SQ were mainly enriched in the monocarboxylic acid metabolic process, protein catabolic process, organonitrogen compound catabolic process, nitrogen compound metabolic process, positive regulation of biological process, response to chemical, positive regulation of cellular process, cellular component organization or biogenesis, endomembrane system and cellular component organization terms (Figure 4A; Supplementary Table S3). The downregulated genes identified in plateau pikas from SQ were mainly enriched in the regulation of signal transduction, cell communication, positive regulation of biological process, positive regulation of cellular process and intracellular membrane-bounded organelle terms (Supplementary Figure S4A; Supplementary Table S7).

When we compared differential gene expression between plateau pikas from REG and Gansu pikas, we found that the upregulated genes identified in plateau pikas were mainly enriched in the negative regulation of smooth muscle cell proliferation, positive regulation of cell projection organization, carboxylic acid metabolic process, oxoacid metabolic process, organic acid metabolic process, neurogenesis, nervous system development, regulation of cellular component organization, cell differentiation, system development, organic substance transport, and cellular developmental process categories (Figure 4B; Supplementary Table S4). The downregulated genes were mainly enriched in the regulation of cell communication, establishment of localization, regulation of signalling, negative regulation of biological process and negative regulation of cellular process categories (Supplementary Figure S4B; Supplementary Table S8).

There were 559 upregulated genes identified in SQ vs. WL and REG vs. WL (Figure 4E), and these upregulated genes were mainly enriched in the positive regulation of neuron projection development, nervous system development, cellular response to DNA damage stimulus, protein metabolic process, organonitrogen compound metabolic process and nitrogen compound metabolic process categories (Supplementary Figure S4F; Supplementary Table S6). There were 569 identical downregulated genes between plateau pikas and Gansu pikas at different altitudes (Supplementary Figure S4E). The enriched biological functions included phosphatidylinositol-mediated signalling, inositol lipid-mediated signalling, establishment of localization and cellular magnesium ion homeostasis (Supplementary Figure S4D; Supplementary Table S10).

Pairwise analysis of plateau pikas from different altitudes revealed that the upregulated genes of SQ samples were mainly enriched in the cellular nitrogen compound catabolic process, apoptosis signalling pathway, cellular response to DNA damage stimulus, DNA repair, and response to stress terms (Figure 4C;

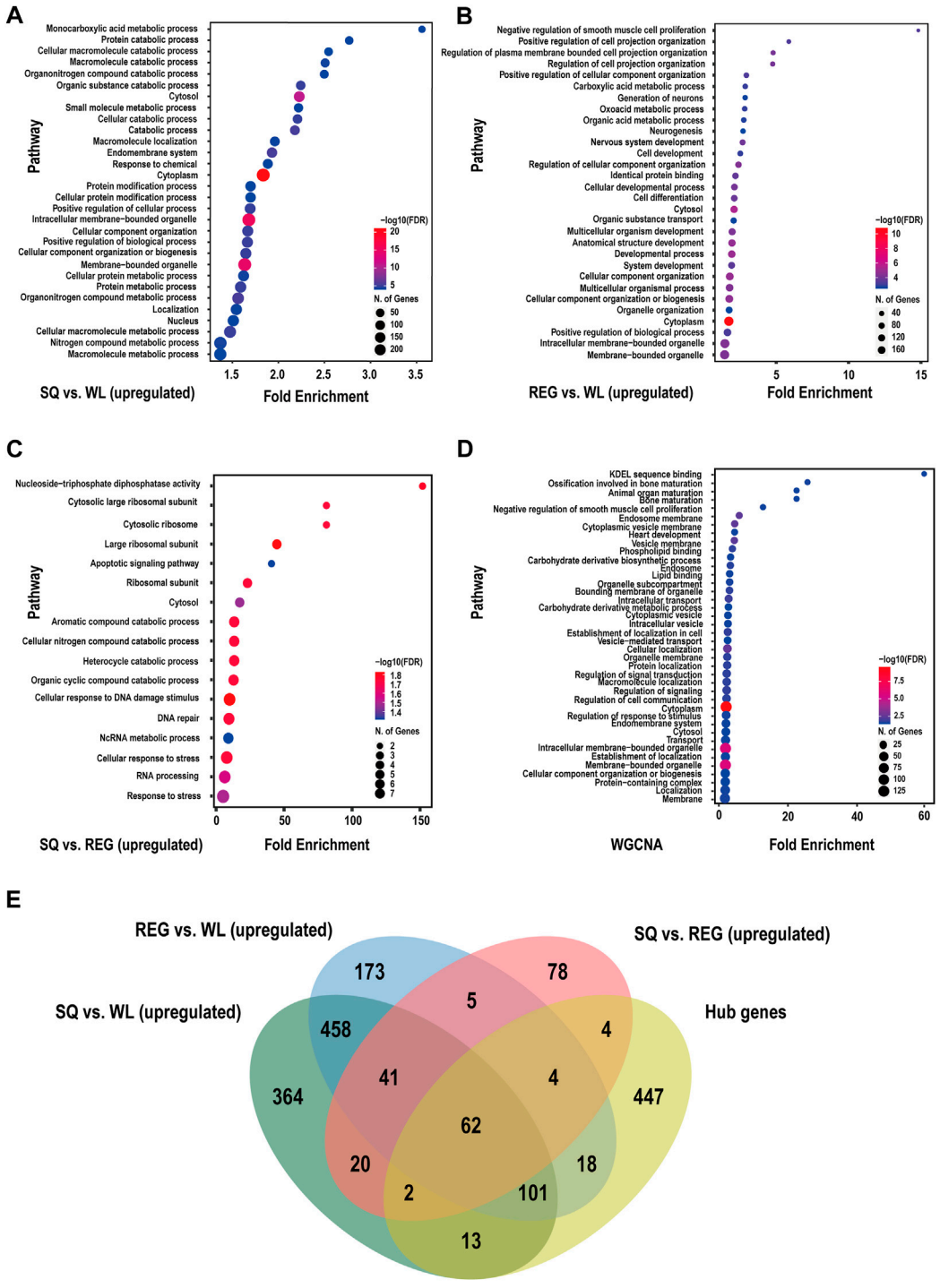


FIGURE 4
Functional enrichment analysis of DEGs (upregulated) and hub genes in plateau pika. **(A)** Functional analysis results of genes upregulated in SQ vs. WL. **(B)** Functional analysis results of genes upregulated in REG vs. WL. **(C)** Functional analysis results of genes upregulated in SQ vs. REG. **(D)** Functional analysis results of altitude-related hub genes screened by WGCNA. **(E)** Venn diagram presenting the results of interspecific and intraspecific variance analysis and WGCNA.

Supplementary Table S5). The downregulated genes identified in SQ samples were mainly enriched in the regulation of cell cycle, mitotic cell cycle, trans-Golgi network transport vesicle and microtubule cytoskeleton organization terms (Supplementary Figure S4D; Supplementary Table S9).

Enrichment analysis of hub genes and related genes

To further investigate the biological functions of hub genes associated with altitude, we performed a GO term analysis (Figure 4B). Among the identified GO terms, the hub genes were mainly enriched in the intracellular membrane-bounded organelle, carbohydrate derivative biosynthetic process, phospholipid binding, lipid binding, cellular localization, regulation of signal transduction, ossification involved in bone maturation, animal organ maturation, regulation of signalling and heart development categories (Supplementary Table S6).

Discussion

In this study, considering that gender differences will affect the results, we used the results of heart tissue expression of Gansu pika (including five females and nine males) to analyse the differences between different genders, and the results showed that there was no significant difference. Other transcription studies of pika heart and lung tissue have also not considered the effect of gender on the results (Zhang et al., 2022; Zhu et al., 2022). Therefore, instead of sex typing of plateau pikas and Gansu pikas, we directly analysed the transcriptome of heart tissues of 28 samples in order to explore the potential genetic mechanism of plateau pikas adapting to high altitude environment. We obtained normalized gene expression results for 32,693 genes in pika heart tissue. Then, we performed DEG and WGCNA to identify DEGs and their main functions. Our results showed that the plateau pika has adapted to the extreme environments of the QHTP *via* the protection of cardiomyocytes, tissue structure alterations and improvements in the blood circulation system and energy metabolism.

Interspecies difference between *O. curzoniae* and *O. cansus*

Transcriptome analysis of the heart tissue of plateau pikas showed that the changes in several biological metabolic processes were significantly different from those in Gansu pikas. For example, the comparison of SQ and WL samples showed that upregulated genes were enriched in the monocarboxylic acid metabolism process, protein catabolism process, organic nitrogen compound catabolism process, and nitrogen compound metabolism process categories. The comparison of samples from REG vs. WL showed that upregulated genes were enriched in the carboxylic acid metabolism process, oxidative acid metabolism process, and organic acid metabolism process categories, indicating that plateau pikas have higher requirements for the Krebs cycle and energy metabolism than Gansu pikas. At high altitude, hypoxia causes a decrease in adenosine triphosphate (ATP), the formation of lactic acid, and alterations in many other metabolites (Cao et al., 2017). Chronic exposure to hypoxia induces adaptive cardiopulmonary changes to guarantee adequate oxygen intake and efficient delivery to tissues under conditions of limited oxygen utilization. Since oxygen is required in tissues to support oxidative metabolism, the process of adapting to hypoxia is likely to play a role the adjustment of relevant metabolic pathways and corresponding metabolites (Serkova et al., 2008). Studies on high-altitude adaptability have shown that the metabolic rates and body temperatures (Tb) of some species can be significantly reduced in response to cold or low-oxygen environments (Ostadal and Kolar, 2007). Hibernating animals show special adaptations in this regard, involving the reduction of their basal metabolic rate and energy consumption (Geiser, 2004; Long et al., 2005; Larson et al., 2014), and protection against ischaemic injury after cardiac arrest (Dave et al., 2006). Although the pika is a small mammal that does not hibernate, its central nervous system can perceive the downregulation of its metabolic rate and body temperature under hypoxia, and the pika can cope with challenges such as a high resting metabolic rate and low oxygen consumption rate through increases in the expression of thermogenic genes and energy metabolism (Wang et al., 2008). Therefore, to adapt to the extreme environment, plateau pikas require an efficient energy metabolism system to maintain the basic needs of life and compensatory mechanisms, such as the rational utilization of metabolic substrates, improvement of respiratory chain efficiency, and closer coupling of ATP supply and demand pathways. These findings are similar to those of studies of Tibetan Plateau aborigines showing that a major trend of plateau adaptation is an increased rate of evolution and positive selection for genes involved in energy metabolism (Simonson et al., 2010; Qiu et al., 2012; Qu et al., 2013b; Ai et al., 2014).

In addition to the differences in metabolic regulation identified between plateau pikas and Gansu pikas, neurogenesis, nervous system development, and system regulation were found to be enriched biological processes. Similarly, studies of the particular tolerance of Tibetans and Sherpas to hypoxia have found that Sherpas show downregulation of sympathetic nervous system (SNS) adrenergic receptors over time in response to hypoxic stimulation. Tibetans, on the other hand, show marked vagal innervation, which persists even after migration to lower elevations (Gilbert-Kawai et al., 2014; Gnecchi-Ruscone et al., 2018). Pichon et al. (2013) studied the cardiac adaptability of

plateau pikas and found that the cardiac function of plateau pikas is affected by parasympathetic and sympathetic nerve regulation. Hypoxia can cause enhanced sympathetic nerve excitation, accelerate the heart rate, strengthen cardiac contraction, and improve cardiac output and blood oxygen transport capacity. These results are consistent with the results of the present study indicating that the cardiovascular adaptability of the plateau pika can be improved to cope with high-altitude environments *via* cardiac nerve regulation.

Intraspecific differences in plateau pika

Pairwise analysis of plateau pikas from different altitudes showed that the DEGs of SQ samples were mainly enriched in the cellular response to DNA damage stimulus (*PAXX*, *MPG*, *ATM*, *NUDT1*, *ALKBH7* and *ZFYVE26*) and DNA repair (*PAXX*, *ATM*, *DCTPP1*, *MPG* and *ZFYVE26*) (Supplementary Table S5). Mammals living on the Tibetan Plateau face not only the harsh threat of hypoxia and cold but also apoptosis, DNA damage, inflammation and cancer caused by intense solar radiation and prolonged hypoxia (Scheinfeldt and Tishkoff, 2010; Bartels et al., 2013; Stoecklein et al., 2015). Dead or damaged cells release endogenous signals that activate inflammation to influence immune responses (Rock and Kono, 2008). Thereafter, the signals released by the damaged cells of eukaryotes are received, the expression of immune pathway components is activated, various cellular structural mechanisms are enhanced, and the bodily damage caused by hypoxia and ultraviolet radiation is resisted (Svobodová et al., 2012; Sklar et al., 2013).

In addition, the DEGs enriched in the above GO pathways (*NUDT1*, *PAXX*, *ATM* and *DCTPP1*) were highly expressed in SQ. PAXX nonhomologous end joining factor (*PAXX*) has been shown to be a key helper gene in nonhomologous end joining (NHEJ), the most prominent DNA double-strand break (DSB) repair pathway in mammalian cells, defining the molecular function of *PAXX* in KU accumulation at DNA ends (Liu et al., 2017). Ataxia-telangiectasia mutated (*ATM*) plays a key role in regulating the cellular response to ionizing radiation, and studies suggest that *ATM* may interact with, or be recruited to DNA-damaged chromatin (Andegeko et al., 2001; Bakkenist and Kastan., 2003). Nudix hydrolase 1 (*NUDT1*) and PAH-PASMCs hijack persistent oxidative stress and prevent the incorporation of oxidized nucleotides into DNA, thereby allowing cells to escape apoptosis, proliferate and to some extent participate in vascular remodelling in pulmonary hypertension (Mur et al., 2018; Vitry et al., 2021). A protein with NTP-PPase activity (*DCTPP1*) and the biological pathway in which it is involved (nucleoside triphosphate diphosphatase activity) were identified among the DEGs between samples from REG and SQ. This pathway can hydrolyse abnormal nucleotides in cells without affecting newly synthesized DNA or RNA, reduce the mutation rate, guarantee the stability of the genome, and play a "gatekeeping" role (Requena et al., 2014). Taken together, these results reflect the process of adaptation to hypoxia and UV radiation in plateau pikas at high altitudes. If the limits of respiratory motility, capillary uptake, and species transport have been reached, then maintaining homeostasis, improving metabolic efficiency, and immune regulation to reduce disease risk may be key to long-term adaptation to high altitudes. However, these aspects still need further study to elucidate and better understand the underlying mechanisms.

Adaptive evolution of the heart at high altitude

Through the analysis of the cardiac transcriptome data of Gansu pikas and plateau pikas from SQ and REG, 2518, 2093 and 392 DEGs were obtained, and genes and pathways related to plateau adaptation were identified. These results proved that the heart tissue of plateau pikas shows gene expression changes induced by a high-altitude environment, similar to the results of other studies of plateau species adaptation. It is likely that these genes and tissues have been subject to natural selection and thus tend to confer beneficial adaptations to ongoing environmental pressure (Gilbert-Kawai et al., 2014).

In addition, we used WGCNA to identify modules with similar expression patterns, analyse the associations between modules and sample phenotypes (Pei et al., 2017), and comprehensively consider the related genes and pathways involved in high-altitude adaptation in heart tissue. Through the WGCNA method, we found one specific module related to high-altitude adaptation (Figure 3D, the red module). A total of 706 hub genes were screened from the red module (Figure 3E). GO enrichment analysis revealed some functions that were in accord with the results of DESeq2 differential expression analysis, such as the negative regulation of smooth muscle cell proliferation, regulation of signal transduction, intracellular transport, carbohydrate derivative biosynthetic process, and phospholipid binding (Figure 4D). These findings highlight that plateau pikas can adapt to extreme environments by upgrading material transport and metabolic pathways. Second, some of the identified genes (*DCHS1*, *PLXNB1* and *PHOSPHO1*) are involved in organ development, ossification and bone maturation as well as heart development (involving genes with *CASP7*, *NDRG4*, *IFT20*, *DCHS1*, *NDST1*, *RB1*, *SNX17*, *CDKN1B* and *MYL3*) and other biological functions (Supplementary Table S6). These are genes related to the maturation of heart organs and the formation of fibrous skeletal structures. For example, *DCHS1* plays an important role in mitral valve formation, and the mutation of this gene causes changes in the mitral valve similar to changes observed in human diseases (Durst et al., 2015). Similarly, to adapt to high-altitude environments, plateau yaks and Tibetan pigs have evolved larger heart tissues to provide

greater blood oxygen delivery (Wang et al., 2020a; Ge et al., 2021; Tian et al., 2021; Wang et al., 2022). Pichon et al. (2013) identified left ventricular hypertrophy in pikas and showed that the heart exhibits increased angiogenesis mediated by the *VEGF* pathway. These results are similar to our findings and show that the heart of pikas has adaptively evolved under long-term exposure to high altitudes by strengthening the myocardial connective tissue band to provide structure and support to the heart, ensuring that the ventricle can drive blood flow and intermittently deliver blood to various parts of the body under high-altitude pressure (Saremi et al., 2017).

In this study, we revealed the genes and pathways related to cardiac adaptability through transcriptome analysis and assessed the basic characteristics of the pika cardiomyocyte population. However, it is still unknown how cardiac development and the regulation of myocardial components are controlled in pikas. Future studies combining single-cell transcriptomic and metabolic group analyses can shed further light on the evolutionary mechanism underlying the high-altitude adaptation of pika heart tissue.

Conclusion

To explore the potential genetic mechanism underlying the high-altitude adaptation of pikas, we performed a comparative transcriptomic analysis of heart tissues of plateau pikas and Gansu pikas using the DESeq2 and WGCNA methods. We identified key genes and pathways related to high-altitude adaptation. In particular, these key genes are involved in cardiac organ maturation and the formation of fibrous skeletal structures (*DCHS1*, *PLXNB1*, *PHOSPHO1*) and in the response to hypoxia and ultraviolet radiation (*ATM*, *PAXX*, *MPG*, *NUDT1*, *ZFYVE26*). Moreover, *PDGF* family genes are involved in the regulation of the platelet-derived growth factor receptor signalling pathway (*PDGFRA* and *PBGFRB*). These results indicate that plateau pika cardiac structure and gene expression are regulated by natural selection in long-term high-altitude environments. Plateau pikas have adapted to extreme environments on the QHTP by enhancing the functional strength of the heart (myocardial connective tissue band, blood vessel wall, and blood supply capacity), maintaining the stability of the internal environment, and reducing bodily damage *via* immune regulation. Through interspecies difference analysis, we found that metabolic efficiency improvement and immune regulation to reduce disease risk may be critical to long-term adaptation to high-altitude environments. Our findings demonstrate the adaptability of the heart of plateau pikas to the extreme environments of the QHTP and provide new clues for further understanding the molecular mechanisms and characteristics of pika adaptation to high altitudes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://db.cngb.org/search/project/CNP0003389/.

Ethics statement

The animal study was reviewed and approved by Institute of Zoology, Chinese Academy pf Sciences.

Author contributions

DG and DM designed this research. DG, AF, DM, and XW led the research team to collect samples. DM, DG, WS, ZW, JC, LX, and QY conducted data analysis and wrote the manuscript, and WW assisted with the WGCNA. All the authors were instrumental in producing the first draft.

Funding

This work was sponsored by the Second Tibetan Plateau Scientific Expedition and Research Program (No. 2019QZKK0402/2019QZKK0501) and the National Nature Science Fund of China (31872958, 32170426 to DG).

Acknowledgments

We appreciate Xinyuan Cui, Siyuan Xu, Jinglan Sun, and Yuyao Qin for assisting in field collection. We thank all the members of the research team for discussing this paper. We also appreciate Andrey Lissovsy for providing pictures of the plateau pika and Gansu pika.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their

affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1020789/full#supplementary-material

References

Ai, H., Yang, B., Li, J., Xie, X., Chen, H., and Ren, J. (2014). Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics* 15. doi:10.1186/1471-2164-15-834

Andegeko, Y., Moyal, L., Mittelman, L., Tsarfaty, I., Shiloh, Y., and Rotman, G. (2001). Nuclear retention of ATM at sites of DNA double strand breaks. *J. Biol. Chem.* 276 (41), 38224–38230. doi:10.1074/jbc.M102986200

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Azad, P., Stobdan, T., Zhou, D., Hartley, I., Akbari, A., Bafna, V., et al. (2017). High-altitude adaptation in humans: From genomics to integrative physiology. *J. Mol. Med.* 95 (12), 1269–1282. doi:10.1007/s00109-017-1584-7

Bai, K., He, S., Shu, L., Wang, W., Lin, S., Zhang, Q., et al. (2020). Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by WGCNA analysis of transcriptome stemness index. *Cancer Med.* 9 (12), 4290–4298. doi:10.1002/cam4.3047

Bakkenist, C. J., and Kastan, M. B. (2003). DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* 421 (6922), 499–506. doi:10.1038/nature01368

Bao, Q., Zhang, X., Bao, P., Liang, C., Guo, X., Chu, M., et al. (2021). Using weighted gene co-expression network analysis (WGCNA) to identify the hub genes related to hypoxic adaptation in yak (*Bos grunniens*). *Genes Genomics* 43 (10), 1231–1246. doi:10.1007/s13258-021-01137-5

Bartels, K., Grenz, A., and Eltzschig, H. K. (2013). Hypoxia and inflammation are two sides of the same coin. *Proc. Natl. Acad. Sci. U. S. A.* 110 (46), 18351–18352. doi:10.1073/pnas.1318345110

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bin, Z., and Steve, H. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4 (1), Article17. doi:10.2202/1544-6115.1128

Cao, X., Bai, Z., Ma, L., Ma, S., and Ge, R. (2017). Metabolic alterations of qinghai–tibet plateau pikas in adaptation to high altitude. *High. Alt. Med. Biol.* 18 (3), 219–225. doi:10.1089/ham.2016.0147

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–i890. doi:10.1093/bioinformatics/bty560

Dave, K. R., Prado, R., Raval, A. P., Drew, K. L., and Perez-Pinzon, M. A. (2006). The arctic ground squirrel brain is resistant to injury from cardiac arrest during euthermia. *Stroke* 37 (5), 1261–1265. doi:10.1161/01.Str.0000217409.60731.38

Djan, M., Stefanovic, M., Velickovic, N., Lavadinovic, V., Alves, P. C., and Suchentrunk, F. (2017). Brown hares (Lepus europaeus pallas, 1778) from the balkans: A refined phylogeographic model. *Hystrix-Italian J. Mammal.* 28 (2), 186–193. doi:10.4404/hystrix-28.2-12202

Dor, Y., Camenisch, T. D., Itin, A., Fishman, G. I., McDonald, J. A., Carmeliet, P., et al. (2001). A novel role for VEGF in endocardial cushion formation and its potential contribution to congenital heart defects. *Development* 128 (9), 1531–1538. doi:10.1242/dev.128.9.1531

Durst, R., Sauls, K., Peal4, D. S., deVlaming, A., Toomer, K., Leyne, M., et al. (2015). Mutations in *DCHS1* cause mitral valve prolapse. *Nature* 525 (7567), 109–113. doi:10.1038/nature14670

Feijo, A., Ge, D., Wen, Z., Xia, L., and Yang, Q. (2020). Divergent adaptations in resource-use traits explain how pikas thrive on the roof of the world. *Funct. Ecol.* 34 (9), 1826–1838. doi:10.1111/1365-2435.13609

Ge, Q., Guo, Y., Zheng, W., Zhao, S., Cai, Y., and Qi, X. (2021). Molecular mechanisms detected in yak lung tissue via transcriptome-wide analysis provide insights into adaptation to high altitudes. *Sci. Rep.* 11 (1), 7786. doi:10.1038/s41598-021-87420-7

Ge, S. X., Jung, D., and Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36 (8), 2628–2629. doi:10.1093/bioinformatics/btz931

Geiser, F. (2004). Metabolic rate and body temperature reduction during hibernation and daily torpor. *Annu. Rev. Physiol.* 66 (1), 239–274. doi:10.1146/annurev.physiol.66.032102.115105

Gilbert-Kawai, E. T., Milledge, J. S., Grocott, M. P. W., and Martin, D. S. (2014). King of the mountains: Tibetan and sherpa physiological adaptations for life at high altitude. *Physiol. (Bethesda)* 29 (6), 388–402. doi:10.1152/physiol.00018.2014

Gnecchi-Ruscone, G. A., Abondio, P., Fanti, S. D., Sarno, S., Sherpa, M. G., Sherpa, P. T., et al. (2018). Evidence of polygenic adaptation to high altitude from Tibetan and Sherpa genomes. *Genome Biol. Evol.* 10 (11), 2919–2930. doi:10.1093/gbe/evy233

Gong, G., Fan, Y., Yan, X., Li, W., Yan, X., Liu, H., et al. (2022). Identification of genes related to hair follicle cycle development in inner Mongolia cashmere goat by WGCNA. *Front. Vet. Sci.* 9, 894380. doi:10.3389/fvets.2022.894380

Guan, J., Long, K., Ma, J., Zhang, J., He, D., Jin, L., et al. (2017). Comparative analysis of the microRNA transcriptome between yak and cattle provides insight into high-altitude adaptation. *PeerJ* 5, e3959. doi:10.7717/peerj.3959

Gureev, A. A. (1964). The phylogeny of the hares, rabbits and pikas (Lagomorpha Mammalia), in the light of new data paleontology and comparative morphology. *Doklady Proc. Acad. Sci. USSR Biol. Sci. Sect.* 155, 319–321.

Hao, Y., Xiong, Y., Cheng, Y., Song, G., Jia, C., Qu, Y., et al. (2019). Comparative transcriptomics of 3 high-altitude passerine birds and their low-altitude relatives. *Proc. Natl. Acad. Sci. U. S. A.* 116 (24), 11851–11856. doi:10.1073/pnas.1819657116

Jukes, E. (2018). Lagomorphs: Pikas, rabbits and hares of the world. *Ref. Rev.* 32 (6), 25–27. doi:10.1108/rr-05-2018-0082

Koju, N. P., He, K., Chalise, M. K., Ray, C., Chen, Z., Zhang, B., et al. (2017). Multilocus approaches reveal underestimated species diversity and inter-specific gene flow in pikas (*Ochotona*) from southwestern China. *Mol. Phylogenet. Evol.* 107, 239–245. doi:10.1016/j.ympev.2016.11.005

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019). DEBrowser: Interactive differential expression analysis and visualization tool for count data. *BMC Genomics* 20 (1), 6. doi:10.1186/s12864-018-5362-x

Lan, D., Xiong, X., Ji, W., Li, J., Mipam, T. D., Ai, Y., et al. (2018). Transcriptome profile and unique genetic evolution of positively selected genes in yak lungs. *Genetica* 146 (2), 151–160. doi:10.1007/s10709-017-0005-8

Langfelder, P., and Horvath, S. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Larson, J., Drew, K. L., Folkow, L. P., Milton, S. L., and Park, T. J. (2014). No oxygen? No problem! Intrinsic brain tolerance to hypoxia in vertebrates. *J. Exp. Biol.* 217 (7), 1024–1039. doi:10.1242/jeb.085381

Liu, J., Wu, Z., Sun, R., Nie, S., Meng, H., Zhong, Y., et al. (2020). Using mRNAsi to identify prognostic-related genes in endometrial carcinoma based on WGCNA. *Life Sci.* 258, 118231. doi:10.1016/j.lfs.2020.118231

Liu, M., Qu, J., Wang, Z., Wang, Y., Zhang, Y., and Zhang, Z. (2012). Behavioral mechanisms of male sterilization on plateau pika in the Qinghai-Tibet plateau. *Behav. Process.* 89 (3), 278–285. doi:10.1016/j.beproc.2011.12.009

Liu, X., Shao, Z., Jiang, W., Lee, B. J., and Zha, S. (2017). PAXX promotes KU accumulation at DNA breaks and is essential for end-joining in XLF-deficient mice. *Nat. Commun.* 8, 13816. doi:10.1038/ncomms13816

Long, M. Y., Xiongwei, Z., Rivera, P. M., Øivind, T., Barnes, B. M., LaManna, J. C., et al. (2005). Absence of cellular stress in brain after hypoxia induced by arousal from hibernation in Arctic ground squirrels. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 289 (5), R1297–R1306. doi:10.1152/ajpregu.00260.2005

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Marfell, B. J., O'Brien, R., and Griffin, J. F. T. (2013). Global gene expression profiling of monocyte-derived macrophages from red deer (*Cervus elaphus*) genotypically resistant or susceptible to *Mycobacterium avium* subspecies paratuberculosis infection. *Dev. Comp. Immunol.* 40 (2), 210–217. doi:10.1016/j.dci.2013.02.004

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337 (6099), 1190–1195. doi:10.1126/science.1222794

Melo Ferreira, J., de Matos, A. L., Areal, H., Lissovsky, A. A., Carneiro, M., and Esteves, P. J. (2015). The phylogeny of pikas (Ochotona) inferred from a multilocus coalescent approach. *Mol. Phylogenet. Evol.* 84, 240–244. doi:10.1016/j.ympev.2015.01.004

Montuire, S. (2001). "Lagomorpha rabbits, hares and pikas," in *Encyclopedia of life sciences*. New York, NY: John Wiley & Sons, Ltd.

Mur, P., Jemth, A. S., Bevc, L., Amaral, N., Navarro, M., Valdes-Mas, R., et al. (2018). Germline variation in the oxidative DNA repair genes NUDT1 and OGG1 is not associated with hereditary colorectal cancer or polyposis. *Hum. Mutat.* 39 (9), 1214–1225. doi:10.1002/humu.23564

Ostadal, B., and Kolar, F. (2007). Cardiac adaptation to chronic high-altitude hypoxia: Beneficial and adverse effects. *Respir. Physiol. Neurobiol.* 158 (2-3), 224–236. doi:10.1016/j.resp.2007.03.005

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14 (4), 417–419. doi:10.1038/nmeth.4197

Pei, G., Chen, L., and Zhang, W. (2017). WGCNA application to proteomic and metabolomic data analysis. *Methods Enzymol.* 585, 135–158. doi:10.1016/bs.mie.2016.09.016

Pichon, A. e., Voituron, N., Bai, Z., Jeton, F., Tana, W., Marchant, D., et al. (2015). Comparative ventilatory strategies of acclimated rats and burrowing plateau pika (*Ochotona curzoniae*) in response to hypoxic-hypercapnia. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* 187, 103–110. doi:10.1016/j.cbpa.2015.05.004

Pichon, A., Zhenzhong, B., Marchant, D., Jin, G., Voituron, N., Haixia, Y., et al. (2013). Cardiac adaptation to high altitude in the plateau pika (*Ochotona curzoniae*). *Physiol. Rep.* 1 (2), e00032. doi:10.1002/phy2.32

Qi, X., Wang, X., Zhu, S., Rao, X., Wei, L., and Wei, D. (2008). Hypoxic adaptation of the hearts of plateau zokor (Myospalax baileyi) and plateau pika (Ochotona curzoniae). *Sheng Li Xue Bao* 60 (3), 348–354.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., et al. (2012). The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44 (8), 946–949. doi:10.1038/ng.2343

Qu, J., Li, W., Yang, M., Ji, W., and Zhang, Y. (2013a). Life history of the plateau pika (*Ochotona curzoniae*) in alpine meadows of the Tibetan Plateau. *Mamm. Biol.* 78 (1), 68–72. doi:10.1016/j.mambio.2012.09.005

Qu, Y., Zhao, H., Han, N., Zhou, G., Song, G., Gao, B., et al. (2013b). Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nat. Commun.* 4, 2071. doi:10.1038/ncomms3071

Ream, M., Ray, A. M., Chandra, R., and Chikaraishi, D. M. (2008). Early fetal hypoxia leads to growth restriction and myocardial thinning. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 295 (2), R583–R595. doi:10.1152/ajpregu.00771.2007

Requena, C. E., Pérez-Moreno, G., Ruiz-Pérez, L. M., Vidal, A. E., and González-Pacanowska, D. (2014). The NTP pyrophosphatase DCTPP1 contributes to the homoeostasis and cleansing of the dNTP pool in human cells. *Biochem. J.* 459 (1), 171–180. doi:10.1042/BJ20130894

Rock, K. L., and Kono, H. (2008). The inflammatory response to cell death. *Annu. Rev. Pathol.* 3 (1), 99–126. doi:10.1146/annurev.pathmechdis.3.121806.151456

Sahota, I. S., and Panwar, N. S. (2013). Prevalence of chronic mountain sickness in high altitude districts of Himachal Pradesh. *Indian J. Occup. Environ. Med.* 17 (3), 94–100. doi:10.4103/0019-5278.130839

Saremi, F., Sánchez-Quintana, D., Mori, S., Muresian, H., Spicer, D. E., Hassani, C., et al. (2017). Fibrous skeleton of the heart: Anatomic overview and evaluation of pathologic conditions with CT and MR imaging. *Radiographics* 37 (5), 1330–1351. doi:10.1148/rg.2017170004

Scheinfeldt, L. B., and Tishkoff, S. A. (2010). Living the high life: High-altitude adaptation. *Genome Biol.* 11 (9), 133–3. doi:10.1186/gb-2010-11-9-133

Serkova, N. J., Reisdorph, N. A., and Tissot van Patot, M. C. (2008). Metabolic markers of hypoxia: Systems biology application in biomedicine. *Toxicol. Mech. Methods* 18 (1), 81–95. doi:10.1080/15376510701795769

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., et al. (2010). Genetic evidence for high-altitude adaptation in tibet. *Science* 329 (5987), 72–75. doi:10.1126/science.1189406

Sklar, L. R., Almutawa, F., Lim, H. W., and Hamzavi, I. (2013). Effects of ultraviolet radiation, visible light, and infrared radiation on erythema and pigmentation: A review. *Photochem. Photobiol. Sci.* 12 (1), 54–64. doi:10.1039/c2pp25152c

Smith, A. T., BadingqiuyingWilson, M. C., and Hogan, B. W. (2019). Functional-trait ecology of the plateau pika *Ochotona curzoniae* in the Qinghai-Tibetan Plateau ecosystem. *Integr. Zool.* 14 (1), 87–103. doi:10.1111/1749-4877.12300

Smith, A. T., and Foggin, J. M. (1999). The plateau pika (*Ochotona curzoniae*) is a keystone species for biodiversity on the Tibetan plateau. *Anim. Conserv.* 2 (4), 235–240. doi:10.1111/j.1469-1795.1999.tb00069.x

Solari, K. A., and Hadly, E. A. (2018). Evolution for extreme living: Variation in mitochondrial cytochrome c oxidase genes correlated with elevation in pikas (genus ochotona). *Integr. Zool.* 13 (5), 517–535. doi:10.1111/1749-4877.12332

Stoecklein, V. M., Osuka, A., Ishikawa, S., Lederer, M. R., Wanke-Jellinek, L., and Lederer, J. A. (2015). Radiation exposure induces inflammasome pathway activation in immune cells. *J. Immunol.* 194 (3), 1178–1189. doi:10.4049/jimmunol.1303051

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (3), 479–498. doi:10.1111/1467-9868.00346

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. B* 66 (1), 187–205. doi:10.1111/j.1467-9868.2004.00439.x

Su, J., Lian, X., Zhang, T., Cui, Q., and Liu, J. (2005). Hay-pile caches as winter food by Gansu Pikas and its biological signif icance. *Acta Theriol. Sin.* 24 (1), 23.

Su, J., and Liu, J. (2000). Overwinter of small herbivorous mammals inhabiting Alpine area. *Acta Theriol. Sin.* 20 (3), 186–192.

Svobodová, A. R., Galandáková, A., Šianská, J., Doležal, D., Lichnovská, R., Ulrichová, J., et al. (2012). DNA damage after acute exposure of mice skin to physiological doses of UVB and UVA light. *Arch. Dermatol. Res.* 304 (5), 407–412. doi:10.1007/s00403-012-1212-x

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol.* 8, 374. doi:10.3389/fonc.2018.00374

Tang, Q., Gu, Y., Zhou, X., Jin, L., Guan, J., Liu, R., et al. (2017). Comparative transcriptomics of 5 high-altitude vertebrates and their low-altitude relatives. *Gigascience* 6 (12), 1–9. doi:10.1093/gigascience/gix105

Tian, X., Ma, J., Wu, Y., Zhang, P., Li, Q., Zhang, H., et al. (2021). Functional analysis of the brain natriuretic peptide gene for high-altitude adaptation in Tibetan pigs. *Gene* 768, 145305. doi:10.1016/j.gene.2020.145305

Vitry, G., Paulin, R., Grobs, Y., Lampron, M.-C., Shimauchi, T., Lemay, S.-E., et al. (2021). Oxidized DNA precursors cleanup by NUDT1 contributes to vascular remodeling in pulmonary arterial hypertension. *Am. J. Respir. Crit. Care Med.* 203 (5), 614–627. doi:10.1164/rccm.202003-0627OC

Wang, C., Zhao, X., Liu, Z., Lippert, P. C., Graham, S. A., Coe, R. S., et al. (2008). Constraints on the early uplift history of the Tibetan Plateau. *Proc. Natl. Acad. Sci. U. S. A.* 105 (13), 4987–4992. doi:10.1073/pnas.0703595105

Wang, D., Sun, R., and Wang, Z. (1999). Effects of photoperiod and temperature on browh adipose tissue thermogenic properties in plateau pika. *Zoological Res.* 20 (5), 347–351.

Wang, H., Wang, X., Li, M., Wang, S., Chen, Q., and Lu, S. (2022). Identification of key sex-specific pathways and genes in the subcutaneous adipose tissue from pigs using WGCNA method. *BMC Genom. Data* 23 (1), 35. doi:10.1186/s12863-022-01054-w

Wang, J., Zhang, Y., and Wang, D. (2006). Seasonal thermogenesis and body mass regulation in plateau pikas (*Ochotona curzoniae*). *Oecologia* 149 (3), 373–382. doi:10.1007/s00442-006-0469-1

Wang, Q., Li, D., Guo, A., Li, M., Li, L., Zhou, J., et al. (2020a). Whole-genome resequencing of Dulong Chicken reveal signatures of selection. *Br. Poult. Sci.* 61 (6), 624–631. doi:10.1080/00071668.2020.1792832

Wang, X., Liang, D., Jin, W., Tang, M., Liu, S., and Zhang, P. (2020b). Out of tibet: Genomic perspectives on the evolutionary history of extant pikas. *Mol. Biol. Evol.* 37 (6), 1577–1592. doi:10.1093/molbev/msaa026

Wilson, M. C., and Smith, A. T. (2015). The pika and the watershed: The impact of small mammal poisoning on the ecohydrology of the Qinghai-Tibetan Plateau. *Ambio* 44 (1), 16–22. doi:10.1007/s13280-014-0568-x

Wu, Y., Cheng, T., Liu, C., Liu, D., Zhang, Q., Long, R., et al. (2016). Systematic identification and characterization of long non-coding RNAs in the

silkworm, *Bombyx mori*. *PloS One* 11 (1), e0147147. doi:10.1371/journal.pone.0147147

Yang, J., Chen, C., Jin, X., Liu, L., Lin, J., Kang, X., et al. (2021). Wfs1 and related molecules as key candidate genes in the Hippocampus of depression. *Front. Genet.* 11, 589370. doi:10.3389/fgene.2020.589370

Zhang, B., Chamba, Y., Shang, P., Wang, Z., Ma, J., Wang, L., et al. (2017). Comparative transcriptomic and proteomic analyses provide insights into the key genes involved in high-altitude adaptation in the Tibetan pig. *Sci. Rep.* 7 (1), 3654. doi:10.1038/s41598-017-03976-3

Zhang, X.-Z., Fu, L., Zou, X.-Y., Li, S., Ma, X.-D., Xie, L., et al. (2022). Lung transcriptome analysis for the identification of genes involved in the hypoxic adaptation of plateau pika (Ochotona curzoniae). *Comp. Biochem. Physiol. Part D. Genomics Proteomics* 41, 100943. doi:10.1016/j.cbd.2021.100943

Zhu, H., Zhong, L., Li, J., Wang, S., and Qu, J. (2021). Differential expression of metabolism-related genes in plateau pika (*Ochotona curzoniae*) at different altitudes of Qinghai–Tibet Plateau. *Front. Genet.* 12, 784811. doi:10.3389/fgene.2021.784811

Zhu, H., Zhong, L., Li, J., Wang, S., and Qu, J. (2022). Differential expression of metabolism-related genes in plateau pika (ochotona curzoniae) at different altitudes on the Qinghai-Tibet Plateau. *Front. Genet.* 12, 784811. doi:10.3389/fgene.2021.784811

Zhu, M., Xie, H., Wei, X., Dossa, K., Yu, Y., Hui, S., et al. (2019). WGCNA analysis of salt-responsive core transcriptome identifies novel hub genes in rice. *Genes* 10 (9), E719. doi:10.3390/genes10090719

Zhu, W., Hou, D., Sun, S., and Wang, Z. (2017). White adipose tissue undergoes 'browning'in tree shrews (*Tupaia belangeri*) during cold acclimation. *Mamm. Study* 42 (4), 1–8. doi:10.3106/041.042.0405

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 December 2022
DOI 10.3389/fgene.2022.1018632

OPEN ACCESS

EDITED BY
Yongjie Wu,
Sichuan University, China

REVIEWED BY
Yongshuang Xiao,
Institute of Oceanology (CAS), China
Hung-Du Lin,
National Tainan First Senior High School, Taiwan
Deyan Ge,
Institute of Zoology (CAS), China

*CORRESPONDENCE
Wenjuan Shan,
swj@xju.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 03 September 2022
ACCEPTED 14 November 2022
PUBLISHED 12 December 2022

CITATION
Mamat M, Shan W, Dong P, Zhou S, Liu P, Meng Y, Nie W, Teng P and Zhang Y (2022), Population genetics analysis of Tolai hares (*Lepus tolai*) in Xinjiang, China using genome-wide SNPs from SLAF-seq and mitochondrial markers. *Front. Genet.* 13:1018632. doi: 10.3389/fgene.2022.1018632

Population genetics analysis of Tolai hares (*Lepus tolai*) in Xinjiang, China using genome-wide SNPs from SLAF-seq and mitochondrial markers

Miregul Mamat†, Wenjuan Shan*†, Pengcheng Dong†, Shiyu Zhou, Peng Liu, Yang Meng, Wenyue Nie, Peichen Teng and Yucong Zhang

Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Science and Technology, Xinjiang University, Urumqi, China

The main topic of population genetics and evolutionary biology is the influence of the ecological environment, geographical isolation, and climatic factors on population structure and history. Here, we estimated the genetic diversity, genetic structure, and population history of two subspecies of Tolai hares (*Lepus tolai* Pallas, 1778), *L. t. lehmanni* inhabiting Northern and Northwest Xinjiang and *L. t. centrasiaticus* inhabiting Central and Eastern Xinjiang using SNP of specific-length amplified fragment sequencing (SLAF-seq) and four mitochondrial DNA (mtDNA). Our results showed a relatively high degree of genetic diversity for Tolai hares, and the diversity of *L. t. lehmanni* was slightly higher than that of *L. t. centrasiaticus*, likely due to the more favorable ecological environment, such as woodlands and plains. Phylogenetic analysis from SNP and mtDNA indicated a rough phylogeographical distribution pattern among Tolai hares. Strong differentiation was found between the two subspecies and the two geographical groups in *L. t. centrasiaticus*, possibly due to the geographical isolation of mountains, basins, and deserts. However, gene flow was also detected between the two subspecies, which might be attributed to the Tianshan Corridor and the strong migration ability of hares. Tolai hare population differentiation occurred at approximately 1.2377 MYA. Population history analysis based on SNP and mtDNA showed that the Tolai hare population has a complex history and *L. t. lehmanni* was less affected by the glacial event, possibly because its geographic location and terrain conditions weaken the drastic climate fluctuations. In conclusion, our results indicated that the joint effect of ecological environment, geographic events, and climatic factors might play important roles in the evolutionary process of *L. t. lehmanni* and *L. t. centrasiaticus*, thus resulting in differentiation, gene exchange, and different population history.

KEYWORDS

Lepus tolai, SLAF-seq, mtDNA, genetic diversity, genetic structure

Introduction

Climate factors and habitat environment will profoundly impact the evolution of biological populations, thus leaving historical traces in the genetic diversity, population structure, distribution pattern, and other aspects of today's populations. For example, climate changes profoundly affected phylogeographic structure and the evolutionary history of brown hares (*L. europaeus*) by isolating populations in the distinct refugia where they adapted and differentiated in allopatry, leading to genome incompatibilities (Djan et al., 2017; Minoudi et al., 2018; Giannoulis et al., 2019). The southwest of Tarim Basin in Xinjiang, China, as the origin of rivers in the basin, was a glacial refugia for Yarkand hares (*L. yarkandensis*) during the Quaternary climate oscillations, providing a suitable environment for maintaining the relatively high genetic diversity of this species (Shan et al., 2011). In addition, genome-wide SNPs have confirmed differentiation between the southwest and northern populations (Ababaikeri et al., 2021) and higher diversity in the former population.

Lepus species have a wide distribution and can survive in various complex terrestrial habitats (Ben Slimen et al., 2018). Hence, they are indispensable for local ecosystems as part of the food chain. The Tolai hare (*Lepus tolai* Pallas, 1778) has sandy yellow, brownish gray, or gray dorsal pelage with a dark ripple (Smith et al., 2018). It occupies in various habitats, including desert, semi-desert, mountain steppe, forest-steppe, rocky habitats, and grasslands, and ranges from low to high elevations. Their strong ability to survive in diverse habitats makes them a good model for studying animal adaptation to the environment. However, since the low differences in morphology, and hybridization among hares, their classification has been challenging (Liu et al., 2011). Tolai hares were once classified as Cape hares (*L. capensis*) or Brown hares, but studies based on molecular biology (Wu et al., 2005; Wang and Yang, 2012; Shan et al., 2020b) and skull measurements (Cheng et al., 2012) did not support the classification of "*L. capensis*" in China as *L. capensis*. Therefore, the original Xinjiang's "*L. capensis*" population was divided into *L. tibetanus* and *L. tolai*, and the population distributed in the north of the Tianshan Mountains was classified as *L. tolai* (Shan et al., 2020a).

Xinjiang, China, is in the hinterland of the Eurasian continent, with a dry climate and unique geological structure called "three mountains nip two basins". Tolai hare distributes in the vast area of northern, central, and northwest in Xinjiang (Smith and Xie, 2008; Smith et al., 2018) including Altai Mountains and Junggar Basin in the north, Turpan-Hami Basin in the east, and the Tianshan Mountains crossing the central-eastern part of Xinjiang. Recent studies have shown two subspecies of Tolai hare in Xinjiang, with only slight different appearances (Smith et al., 2018; Shan et al., 2020a), among which *L. t. lehmanni* mainly inhabits the northern and northwestern regions. Except for Junggar Basin, these areas have many rivers, abundant water resources, and relatively humid climates. *L. t. centrasiaticus* mainly distributes in the central and eastern Tianshan Mountains. These areas are relatively dry with little rainfall and are dominated by arid and desert habitats.

Early molecular biological studies of Tolai hares were based on the misclassification of *L. tolai* into *L. capensis*. For example, Wu et al. (Wu et al., 2005) studied the phylogenetic relationships, biogeographic distribution, and species origin patterns between Chinese hare groups, including "*L. capensis*" in central Xinjiang. Wang et al. (Wang and Yang, 2012) sequenced the entire mitochondrial genome of Cape hares. They reconstructed phylogenetic relationships in genus *Lepus* based on *CYTB*, including so-called *"L. capensis"* distributed in Xinjiang. Liu et al. (Liu et al., 2011) used four mitochondrial DNA (mtDNA) fragments and nuclear gene to demonstrate that frequent introgression occurred through historical and recent interspecific hybridization among six Chinese hare species, including those in northern Xinjiang. Wu et al. (Wu et al., 2011) found extensive bidirectional mitochondrial DNA and SRY gene introgression in hybrids of Yarkand hare and Xinjiang "*L. capensis*". Recently, some studies have mapped the full mitochondrial genome of the Xinjiang Tolai hare (Shan et al., 2020b). However, there has been no comprehensive evaluation of their genetic diversity and structure related to habitat and climate changes, and research on the biology of this species is scarce.

Specific-Locus Amplified Fragment sequencing (SLAF-seq) is a high throughput, high-resolution, and low-cost marker development technology that has emerged recently (Sun et al., 2013). This technique focuses on finding single nucleotide polymorphisms (SNPs), an abundant form of genetic variation, in an economical way (Zhang and Zhang, 2005; Wang et al., 2016; Ali et al., 2018; Qin et al., 2019). SNPs are found throughout the genome, and their distribution can reflect the population's genetic variation. SLAF-seq has been used to analyze several species' genetic diversity and phylogenetic structure (Li et al., 2017; Chang et al., 2019; Qin et al., 2019; Zhang J. et al., 2020a; Fang et al., 2020; Ababaikeri et al., 2021). In addition, mtDNA provides a different perspective of population genetic structure because it is maternally inherited and generally lacks intermolecular recombination (Allendorf, 2017).

This study used SLAF-seq to identify genome-wide SNP markers and combined them with four mtDNA markers

TABLE 1 Description of analyzed Tolai hare samples from Xinjiang.

Subspecies	Geographical grouping	Sampling site	Samples for SLAF-seq	Samples for mtDNA
L. t. lehmanni	Northern group	Altay (ALT)	19	24
		Burqin (BRJ)	—	3
		Fuhai (FH)	—	26
		Habahe (HBH)	—	4
		Qinghe (QH)	—	2
	Northwest group	Tarbagatay (TC)	2	3
		Jinghe (JH) and Wenquan (WQ)	—	13
		Ili (YL)	2	6
L. t. centrasiaticus	The central group	Dabancheng (DBC)	11	11
		Tuokexun (TKX)	2	2
	Eastern group	Kumul (HM)	—	12
Total	4	12	36	106

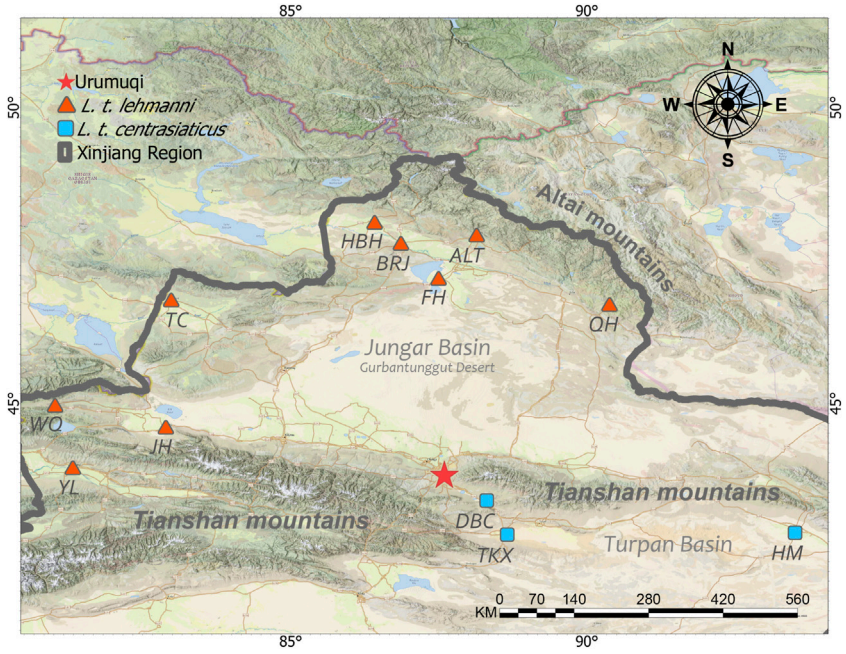


FIGURE 1
Approximate sampling sites of Tolai hare populations in Xinjiang, China.

(*COI*, *ND4*, *CYTB*, and D-LOOP) to evaluate the genetic diversity of two subspecies of Tolai hare in different habitats and reveal the effects of ecological environment, geographical isolation, and climate change on population structure characteristics and population demography history. This study will help to understand the evolutionary history of this species and provide essential data to maintain the biodiversity and stability of the Xinjiang ecosystem.

Materials and methods

Sampling and DNA extraction

Muscle or skin tissue samples were collected from 106 Tolai hares from 12 geographic populations in northern, northwestern, central, and eastern Xinjiang from 2008 to 2019 (Table 1 and Figure 1). The 81 *L. t. lehmanni* samples included 59 samples

from Altay Prefecture (24 from Altay, 26 from Fuhai, three from Burqin, four from Habahe, and two from Qinghe), three from Tarbagatay Prefecture, 13 from Bortala Mongol Autonomous Prefecture (including 12 from Jinghe and one from Wenquan), and six from Ili. Twenty-five *L. t. centrasiaticus* samples included 11 from Dabancheng, two from Tuokexun, and 12 from Kumul. The samples were divided into northern, northwest, central, and eastern groups, and the geographical details of the sampled populations are shown in Table 1. All Tolai hare samples were used for mitochondrial DNA analysis. For SLAF-seq analysis, samples that were stored for a long time, severely degraded, or whose DNA quality was too low to be sequenced were eliminated, and a total of 36 Tolai hare samples in two subspecies remained. These included samples from Altay, Tarbagatay Prefecture, Ili, Dabancheng, and Tuokexun (Table 1). Some samples in this study were confiscated from poachers and provided to us by local forestry bureaus, while others came from hares that died of natural causes. All experimental protocols involved in this study were approved by the Institutional Animal Care and Use Committee of the College of Life Science and Technology, Xinjiang University, Urumqi, China.

Muscle samples were preserved in sterile tubes with anhydrous alcohol at -80°C until total genomic DNA extraction using a DNA tissue extraction kit. Genomic DNA integrity was determined using 1.0% agarose gel electrophoresis.

Construction of SLAF-seq library and high-throughput sequencing

The domestic rabbit (*Oryctolagus cuniculus*) OryCun 2.0 genome (Lindblad-Toh et al., 2011) from the National Center for Biotechnology Information (NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/003/625/GCF_000003625.3_OryCun2.0/) served as the reference genome for genome electronic enzyme digestion prediction. The selection principle for the enzyme digestion scheme was as follows: the proportion of restriction fragments located in the repetitive sequence was as low as possible, the restriction fragments were distributed as evenly as possible across the genome, the length of restriction fragments was highly consistent with the experimental study system, and the number of restriction fragments obtained matched the expected number of tags. A RsaI-EcoRV-HF® restriction enzyme was used to digest the genomic DNA. From the constructed SLAF library (Sun et al., 2013; Zhang et al., 2013) DNA fragments 314–344 base pairs (bp) long were selected for sequencing on an Illumina HiSeq 2,500 system (Illumina, Inc., San Diego, CA, United States) at Beijing Biomarker Technologies Corporation (Beijing, China). The *Oryzasativa ssp. japonica* genome (http://rapdb.dna.affrc.go.jp/) was selected as the positive control for sequencing, and SOAP v2 software (Li R. et al., 2009b) was used to calculate the enzyme digestion efficiency, fragment insertion distribution, and alignment efficiency to evaluate the reliability of the enzyme digestion experiment and the accuracy of library construction.

Quality control and SNP calling

Reads were obtained for each sample, and after filtering the connectors of the sequencing reads, the sequencing and data volumes were evaluated. We calculated the ratio of high-quality reads based on two key quality indicators, Q30 (read quality score of 30, indicating a base sequencing error probability of 0.1%) and GC content. All sample reads were mapped to the OryCun 2.0 genome sequence using BWA v0.7.5a-r405 (Li and Durbin, 2010). Then, we mined the SLAF tags according to the restriction fragment size defined by the enzyme digestion scheme. SNP calling was performed using GATK v3.3.2 (McKenna et al., 2010) and SAM tools v0.1.18 (Li H. et al., 2009a), and the intersection of SNP markers called by the two packages was chosen to build the original SNP dataset. Plink v1.07 (Purcell et al., 2007) was used to filter high-quality SNPs according to the criteria of site information integrity (INT) ≥ 0.5 and minor allele frequency (MAF) ≥ 0.05. Finally, the selected high-quality SNPs were used for further analysis.

Mitochondrial DNA sequencing

The *COI*, *ND4*, *CYTB*, and D-LOOP fragments of mitochondrial DNA were amplified. *COI* PCR primers were 5′-AGGAACAGCCCTYAGTCT-3′ (Forward, F) and 5′-GGTGGGCTCAAACAATAA-3′ (Reverse, R) (Zhang Y. et al., 2020b). PCR primers for *CYTB* were 5′-GCAAAGAATCATTACTACGCAAA-3′ (F) and 5′-TTGCGACGATTACTAAGGCTA-3′ (R) (Zhang Y. et al., 2020b). PCR primers for *CYTB* were 5′-CGAACCCCACAAAACCAATTAC-3′ (F) and 5′-GGTGAGTTGATCTCCGTTTCTG-3′ (R), *CYTB* primers were designed by ourselves based on the published mitochondrial genome of Tolai hare. PCR primers for the D-LOOP were 5′-CAGAGATGGAGATYAACTC-3′ (F) and 5′-GCATGGGCTGATTAGTCAT-3′ (R) (Shan et al., 2011). The PCR amplification reaction comprised 13 μL Premix Taq (1.25U · 25 μL-1), 1 μL each forward and reverse primers (10 μmol·L-1), 1 μL DNA template, and 9 μL sterile deionized water. PCR amplification included denaturation at 95°C for 3–5 min, followed by 25–35 cycles of denaturation at 94°C for 30 s, annealing at 51°C for 30 s, and extension at 72°C for 1 min. A final extension was made at 72°C for 10 min. The PCR products were identified using 1.5% agarose gel electrophoresis, and PCR products with good amplification results were sequenced. The mtDNA gene sequences were aligned by MAFFT and combined using Phylosuite v1.2.2 (Zhang et al., 2018).

Data analysis

Summary statistics describing genetic diversity, including nucleotide diversity (π), observed heterozygosity (*Ho*), excepted heterozygosity (*He*), and the polymorphism information content (*PIC*) were calculated using Powermarker v3.25 (Liu and Muse, 2005) and Arlequin ver3.5 (Excoffier and Lischer, 2010).

Phylogenetic analysis based on SNPs was reconstructed using the Neighbor-Joining (NJ) method and Maximum-Likelihood (ML) method. The NJ phylogenetic tree was performed in Mega X (Kumar et al., 2018) with 1,000 bootstraps, and the ML phylogenetic tree was performed in IQ-TREE (Nguyen et al., 2015) with 3,000 bootstraps on the Cipres (https://www.phylo.org/index.php/). mtDNA-based phylogenetic analysis was conducted on the Cipres using the Bayesian Inference (BI) method by MrBayes 3.2.7a (Ronquist et al., 2012) with five million Bayesian Markov chain Monte Carlo (MCMC) generations and ML tree by RAxML-NG (Kozlov et al., 2019) with 1,000 bootstraps. The rabbit (*O. cuniculus*) genome was used as the outgroup (GenBank accession: GCF_000003625.3 for SNP and AJ001588.1 for mtDNA). ModelFinder was used to identify the best fit base-pair substitution model according to the Bayesian information criterion (BIC) using Phylosuite v1.2.2 (Zhang et al., 2018), and the general time-reversible (GTR) model was selected as the optimal model for phylogenetic analysis of the SNPs and mtDNA concatenated dataset. Principal component analysis (PCA) was conducted using Plink v1.07 (Purcell et al., 2007). The population structure among the Tolai hare populations was inferred using ADMIXTURE v1.3.0 (Alexander et al., 2009). Treemix v.1.13 (Pickrell and Pritchard, 2012) was used to infer multiple population splitting and mixing events using genome-wide allele frequency data. The median-joining network was conducted by Popart (Leigh and Bryant, 2015). We performed a hierarchical analysis of molecular variation (AMOVA) using Arlequin ver3.5 (Excoffier and Lischer, 2010).

Divergence times were estimated based on mtDNA using Beast v1.10.4 (Suchard et al., 2018). Four points were calibrated to build the time tree: the split of *Lepus* (approximately 11.57 million years ago, MYA), the divergence time of *L. americanus* (approximately 8.6 MYA), the divergence time of *L. europaeus* (approximately 1.84 MYA) (Ge et al., 2013), and Tolai hare fossil calibration (0.78 MYA) (Erbajeva and Alexeeva, 2000). Best substitution models according to BIC (HKY model for *COI*, *ND4*, and *CYTB* and GTR model for D-LOOP) were found using modelFinder in Phylosuite v1.2.2 (Zhang et al., 2018), and uncorrelated relaxed lognormal clock with a prior coalescent tree of constant size was used. The Markov chain Monte Carlo analysis was run thrice for 1×10^8 generations, sampling every 1,000 generations. Tracer v1.7.2 (Rambaut et al., 2018) was used to check the log files and ensure that the effective sample size (ESS) for all parameters exceeded 200. TreeAnnotator.v1.10.4 (Suchard et al., 2018) was used to summarize the tree data, and the first 10% of trees were discarded as a burn-in. The tree and divergence times were displayed and edited in Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). SMC++ v1.15.5 (Terhorst et al., 2017) was used to check the population history of the two subspecies based on SNPs. Tajima's D and Fu's *Fs* analyses were performed to test neutrality based on mtDNA in Arlequin ver3.5 (Excoffier and Lischer, 2010). The mismatch distributions of pairwise sequence differences for the Tolai hare populations were estimated using DnaSP6 (Rozas et al., 2017). To estimate changes in effective population size through evolutionary time, we explored population history by constructing Extended Bayesian Skyline Plots (EBSP) in Beast. v2.4.7 (Bouckaert et al., 2014) with mtDNA. The prior tree was set as the Coalescent Extended Bayesian Skyline and sampled every 1,000 steps for 1×10^8 steps. The ESS values for all the parameters were assessed using Tracer v1.7.2 (Rambaut et al., 2018).

Results

SLAF-seq and SNP discovery

The rabbit genome served as the reference genome for predicting electron enzymes and identified a restriction fragment length of 314–344 bp, ultimately defined as a SLAF tag. Sequencing results for the positive control indicated that the efficiency of paired-end comparison was 96.84%, and the enzyme digestion efficiency of the control was 94.74%, indicating that the process was normal and reliable.

High-throughput sequencing of the SLAF library yielded 226.85 Mb of high-quality clean data after strict filtration (Additional file 1: Supplementary Table S1). The average Q30 was 95.39%, indicating that the results of the tested sequences were reliable (Additional file 1: Supplementary Table S1). The average GC content was 41.80%, and the average mapping rate of our samples to the reference genome was 96.13% (Additional file 1: Supplementary Table S1).

A total of 2,205,716 SLAF tags were obtained, with an average sequencing depth of 16.93× (Additional file 1: Supplementary Table S1). A total of 2,005,461 SNPs were obtained from 36 samples, and SNP integrity ranged from 31.59% to 49.50%, averaging 42.72%. SNP heterozygosity ranged from 4.48% to 8.22%, with an average of 5.61% (Additional file 1: Supplementary Table S1). To reduce sequencing errors, eliminate baseline differentiation, and evaluate the accuracy, 473,241 consistent and high-confidence SNPs were selected for further analysis

TABLE 2 Summary statistics describing Xinjiang's Tolai hare genetic diversity based on SNPs.

Subspecies	Geographical grouping	Population (abbreviation)	Nucleotide diversity (π)	Expected heterozygosity (*He*)	Observed heterozygosity (*Ho*)	Polymorphism information content (*PIC*)
L. t. lehmanni	Northern	ALT	0.0590	0.3009	0.2410	0.2445
	Northwest	TC	0.0444	0.4203	0.5314	0.3302
		YL	0.0823	0.4220	0.4459	0.3311
		Mean	0.06335	0.42115	0.48865	0.33065
L. t. centrasiaticus	Central	DBC	0.0609	0.3166	0.2690	0.2564
		TKX	0.0497	0.4108	0.5240	0.3248
		Mean	0.0553	0.3637	0.3965	0.2906
Total	—	Mean	0.05926	0.37412	0.40226	0.2974

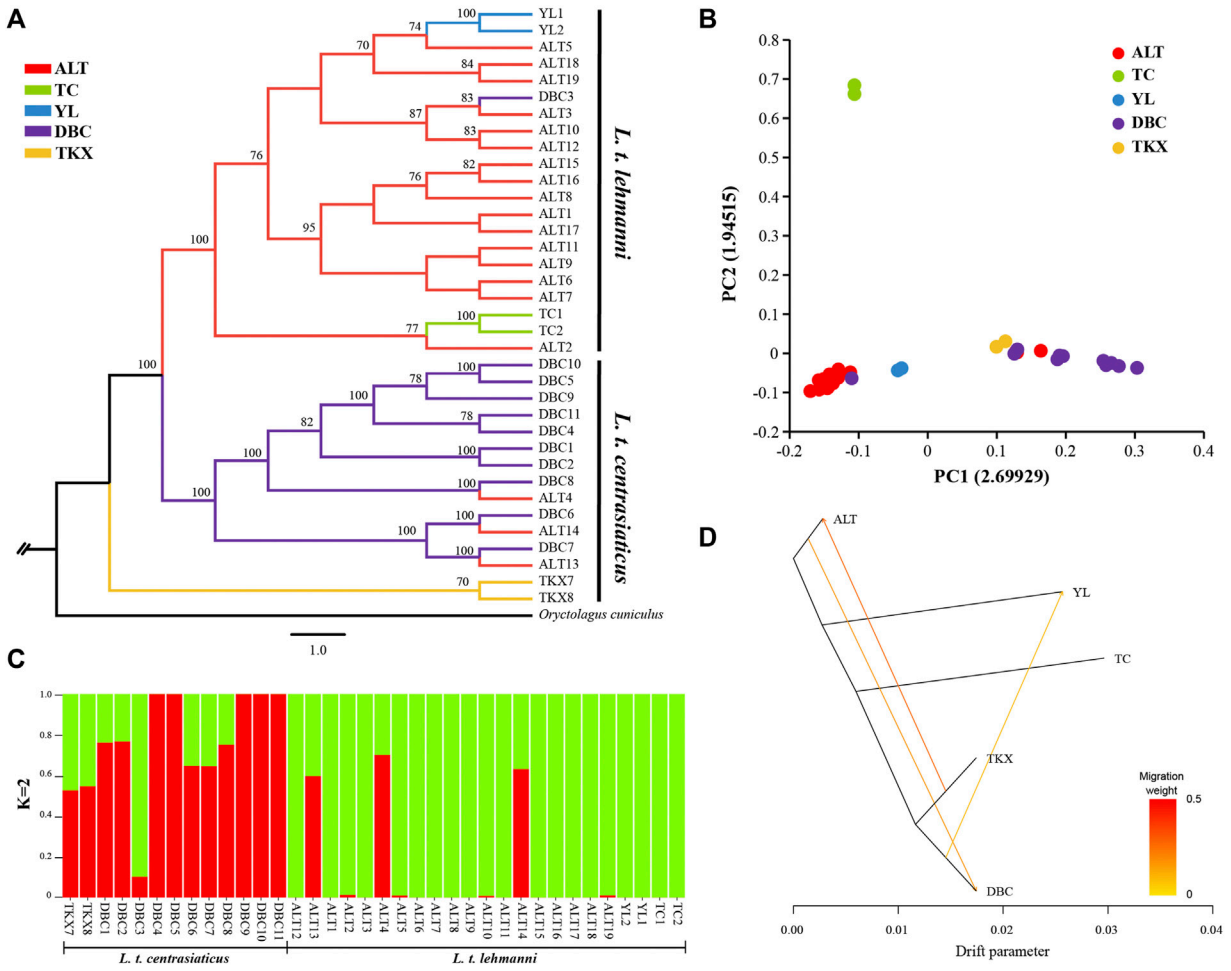


FIGURE 2
Phylogenetic analysis and population structure of Tolai hares in Xinjiang based on SNPs. **(A)**. Phylogenetic tree constructed using the ML method. **(B)**. Principal component analysis (PCA). **(C)**. Genetic Structure of Tolai hares. The number of ancestral populations (K=2) is indicated. **(D)**. Inferred Tolai hare phylogenetic tree with mixture events among populations. Arrows indicate migration events and are colored according to their migration weight.

TABLE 3 AMOVA of the Tolai hare groups in Xinjiang based on SNPs.

Groups	F_{ST}	Percentage of variation (%)	
		Among groups	Within group
[Northern] [Northwest] [Central]	0.0765**	7.65**	92.35
[Northern] [Northwest]	0.0657**	6.57**	93.43
[*L. t. lehmanni*] [*L. t. centrasiaticus*]	0.0657**	6.57**	93.43

*$p < 0.05$ **$p < 0.01$.

according to INT ≥0.5 and MAF ≥0.05. Accession number of SNP data for the Xinjiang Tolai hare provided in Additional File 2: Supplementary Table S2.

Genetic diversity and population structure analysis of SLAF-seq

Nucleotide diversity (π) ranged from 0.0444 (TC population) to 0.0823 (YL population) across the five geographic populations of Tolai hare in Xinjiang and averaged 0.05926 per population (Table 2). The average *He*, *Ho*, and *PIC* values of all populations were 0.37412, 0.40226, and 0.2974, respectively (Table 2). The northwest group had the highest genetic diversity (π = 0.06335, *He* = 0.42115, *Ho* = 0.48865, *PIC* = 0.33065). In general, the genetic diversity indices of *L. t. lehmanni* (π = 0.0619, *He* = 0.3811, *Ho* = 0.4061, *PIC* = 0.3019) were slightly higher than those of *L. t. centrasiaticus* (π = 0.0553, *He* = 0.3637, *Ho* = 0.3965, *PIC* = 0.2906) (Table 2).

We reconstructed phylogeny to explore genome-wide relationships among Tolai hare populations. The topology of ML (Figure 2A) and NJ trees (Additional file 3: Supplementary Figure S1) was consistent. The phylogenetic tree results showed that the Tolai hares analyzed in this study based on SNPs were divided into two main clusters with high confidence, including three clades. The first branch was located at the tree's root, containing two samples from the TKX population. The second branch consisted of samples from the DBC population and three individuals from the ALT population. These two clades were grouped into one cluster because it included most samples from the central group of *L. t. centrasiaticus*, except for three individuals from the ALT population of *L. t. lehmanni*. The third branch consisted mainly of samples from the northern and northwestern groups, including samples from the ALT, TC, and YL populations of *L. t. lehmanni*, and one sample from the DBC population. The clustering relationships among populations were also evident in the PCA, in which *L. t. centrasiaticus* individuals were clustered, and *L. t. lehmanni* individuals were clustered together except for individuals from the TC population (Figure 2B). The population relationships were largely consistent with the geographical distribution of the samples, consistent with the phylogenetic tree.

To evaluate population structure, the ADMIXTURE was performed (Figure 2C, Additional file 4: Supplementary Figure S2). Since the cross-validation errors at K = 2 is close to the lowest cross-validation errors at K = 1 (Additional file 5: Supplementary Figure S3), the assessment of the population structure indicated a clear subdivision of two main ancestral lineages when K = 2. Except for DBC3, most samples in *L. t. lehmanni* formed an ancestral cluster (shown in green in Figure). Most samples in *L. t. centrasiaticus*, and three ALT individuals formed another ancestral cluster (shown in red in the image). In addition, fifteen samples have mixed lineages.

Genetic differences among the geographical groups were examined using AMOVA (Table 3), and the results indicated moderate genetic differences among the geographical groups (7.65%, F_{ST} = 0.0765, $p < 0.01$). When individuals were pooled into the northern group and northwest group, only 6.57% ($p < 0.01$) of the variability was partitioned among groups, and differentiation among groups was 0.0657 ($p < 0.01$). When individuals were pooled into the *L. t. lehmanni* and *L. t. centrasiaticus*, only 6.57% ($p < 0.01$) of the variability was partitioned among populations, and differentiation among subspecies was 0.0657 ($p < 0.01$).

To further infer population gene flow, we performed a Treemix analysis based on SNPs (Figure 2D). Three migrations were detected. One event began from the TKX to the ALT population (migration weight 0.2775). The second began from the ALT population to the DBC population (migration weight 0.184,875). The third event is from the DBC to the YL population (migration weight 0.0999,906).

Genetic diversity and population structure of mtDNA

Sequencing of the four mitochondrial DNA segments (*COI*, *ND4*, *CYTB*, and D-LOOP) with an alignment length of 2,885 bp in 106 Tolai hare individuals resulted in 99 haplotypes. Nucleotide diversity ranged from 0.0295 (eastern group) to 0.0449 (northwest group) (Figure 3B). The π value of *L. t. lehmanni* (0.0410) was slightly higher than that of *L. t. centrasiaticus* (0.0355) (Figure 3B). The total π value of the Tolai hare in Xinjiang was 0.0442.

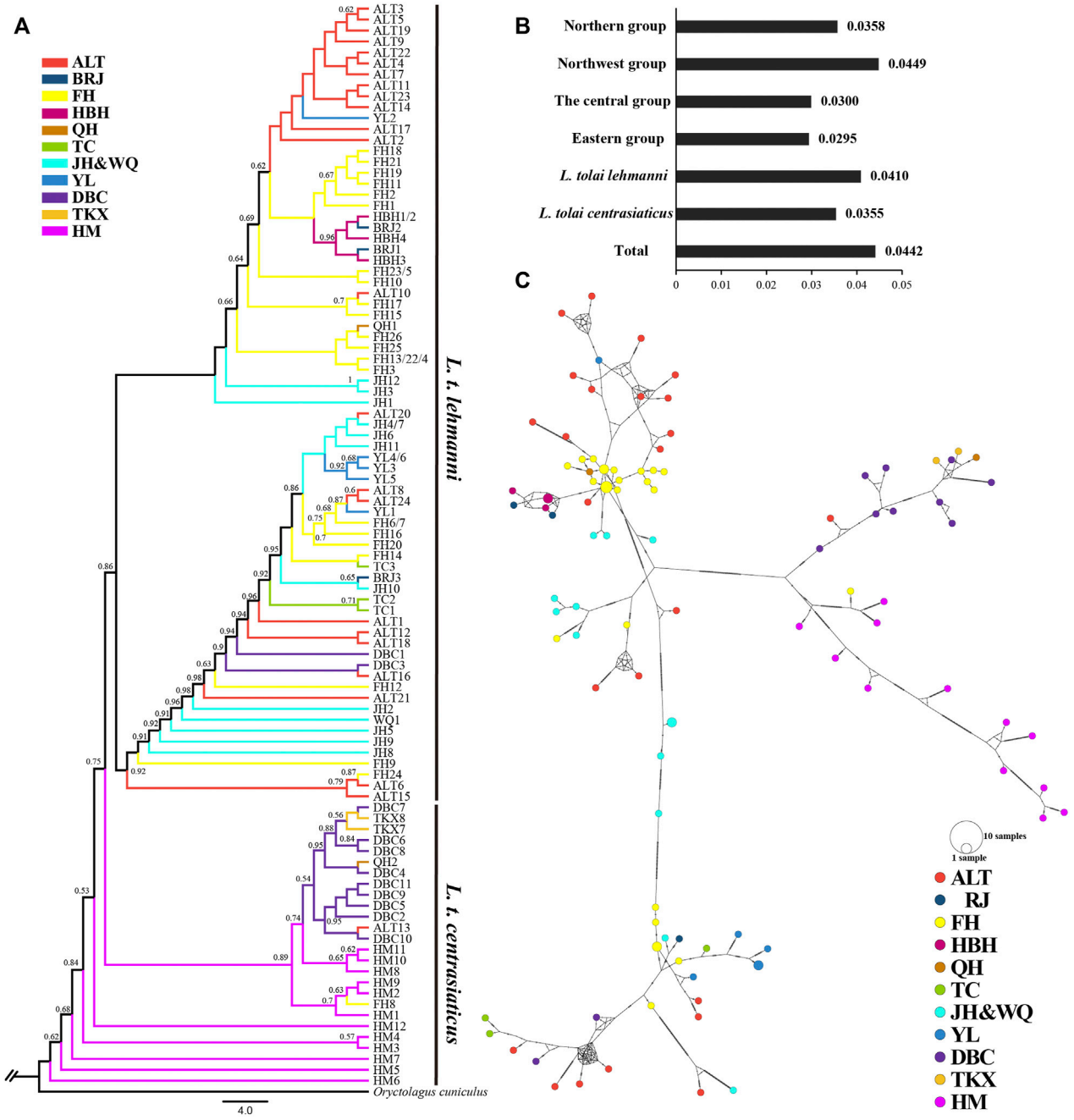


FIGURE 3
Genetic diversity and population structure of Tolai hares in Xinjiang based on mtDNA. **(A)**. Phylogenctic tree constructed using ML method. **(B)**. Nucleotide diversity (π). **(C)**. Median-joining network of Tolai hare haplotypes. The hatch marks on the line indictes the mutation numbers.

To explore the phylogenetic structure, ML (Figure 3A) and BI evolutionary trees (Additional file 6: Supplementary Figure S4) were constructed, and the topological structures of the two trees were consistent. The phylogenetic tree based on combined haplotype sequences of four mitochondrial genes showed that the Tolai hare samples analyzed in this study contained four branches and could be further divided into two main clusters (Figure 3A). The first cluster was located at the tree's root and predominantly comprised of *L. t. centrasiaticus* individuals from the central and eastern groups. Each individual from the QH, ALT, and FH populations were included. The second cluster was comprised of *L. t. lehmanni* individuals from the northern and

TABLE 4 AMOVA of Tolai hare groups in Xinjiang based on mtDNA.

Groups	F_{ST}	Percentage of variation (%)	
		Among groups	Within group
[Northern] [Northwest] [Central] [Eastern]	0.2600**	26.00**	74.00
[Northern] [Northwest]	0.1558**	15.58**	84.42
[Central] [Eastern]	0.2724**	27.24**	72.76
[*L. t. lehmanni*] [*L. t. centrasiaticus*]	0.2376**	23.76**	76.24

*$p < 0.05$; **$p < 0.01$.

northwest groups, and two individuals from the DBC population. In addition to the evolutionary tree, a median-joining network was constructed to further elucidate the phylogenetic relationships among haplotypes (Figure 3C). The haplotype grouping in the median-joining network was consistent with the clustering in the evolutionary tree. The network verified the existence of two distinct clades separated by several mutational steps.

Genetic differentiation among the groups was next examined using AMOVA. The results showed that when pooling individuals into northern, northwest, central, and eastern groups, the genetic variation among groups was 26.00% ($p < 0.01$), and differentiation was 0.2600 ($p < 0.01$) (Table 4). Genetic variation and differentiation among groups was lower when individuals were pooled into the northern and northwest group (15.58%, $F_{ST} = 0.1558$, $p < 0.01$) than when individuals were pooled into the central and eastern group (27.24%, $F_{ST} = 0.2724$, $p < 0.01$) (Table 4). In addition, when individuals pooled into *L. t. lehmanni* and *L. t. centrasiaticus*, genetic variation among populations was 23.76% ($p < 0.01$), and F_{ST} was 0.2376 ($p < 0.01$).

Divergence time estimation and population history

We used mtDNA data to construct a Tolai hare differentiation time-merging tree. The tree showed that Tolai hare population differentiation occurred at approximately 1.2377 MYA, and the eastern HM population showed the earliest differentiation. The divergence between *L. t. lehmanni* and *L. t. centrasiaticus* occurred at approximately 1.1898 MYA (Figure 4A).

To estimate the historical population dynamics of Tolai hare in Xinjiang, we performed SMC++ analysis to track changes in effective population sizes (*Ne*) over time based on SNPs. The dynamics of historic *Ne* for *L. t. lehmanni* and *L. t. centrasiaticus* are shown in Figure 4B. The effective population size of *L. t. centrasiaticus* decreased in the last glacial maximum (LGM), while *L. t. lehmanni* population remained relatively stable. Both subspecies expanded during the interglacial period after the last glacial period (LGP), especially *L. t. lehmanni*. About 0.001–0.0015 MYA, however, populations of both subspecies declined. In addition, the Tajima's D, Fu's *Fs*, mismatch distribution, and EBSP based on mtDNA were used to analyze the population history of Tolai hares. The neutrality test results showed that Tajima's D and Fu's *Fs* values for *L. t. lehmanni* were 2.06761 ($p = 0.9800 > 0.05$) and -7.73931 ($p = 0.0480 < 0.05$), and -0.00167 ($p = 0.5720 > 0.05$) and -3.95338 ($p = 0.0310 < 0.05$) for *L. t. centrasiaticus*. The mismatch distribution analysis revealed that *L. t. lehmanni* and *L. t. centrasiaticus* showed multi-peak curves (Additional file 7: Supplementary Figure S5). The demographic scenario for Tolai hare populations determined through EBSP analysis suggested a more pronounced population expansion of *L. t. lehmanni* beginning from 0.02 MYA, and the *L. t. centrasiaticus* population showed a less distinct tendency to expand (Figure 4C).

Discussion

The relatively high genetic diversity of Tolai hare populations

The amount of genetic diversity reflects the evolutionary potential of a species, and populations with more genetic diversity are expected to adapt better to environmental changes such as climate change, habitat loss, over-harvesting, invasive species, and disease than populations with low genetic diversity (Kardos, 2021). In this study, we used a set of SLAF-seq genome-wide SNP and mtDNA markers to estimate the genetic diversity of Tolai hares in Xinjiang. At the genome-wide level, the genetic diversity of Tolai hares based on SNPs was higher (Table 2) ($\pi = 0.05926$, $He = 0.37412$, $Ho = 0.40226$, $PIC = 0.2974$) than that of other reported species, including rabbits ($PIC = 0.2–0.2281$, $He = 0.2511–0.2857$, $Ho = 0.3072–0.3418$, (Ren et al., 2019), and Yarkand hares ($\pi = 0.0655$, $He = 0.3130$, $Ho = 0.2852$, $PIC = 0.2543$ (Ababaikeri et al., 2021). The estimated nucleotide diversity of Tolai hares ($\pi = 0.0442$) based on mtDNA was also relatively high compared to other *Lepus* taxa, such as brown hares ($\pi_{D\text{-}LOOP} = 0.030$, (Minoudi et al., 2018), Yarkand hares ($\pi_{D\text{-}LOOP} = 0.033$, $\pi_{CYTB} = 0.008$, (Shan et al., 2011), and Italian hares (*L. corsicanus*, $\pi_{D\text{-}LOOP} = 0.018$, (Pierpaoli et al.,

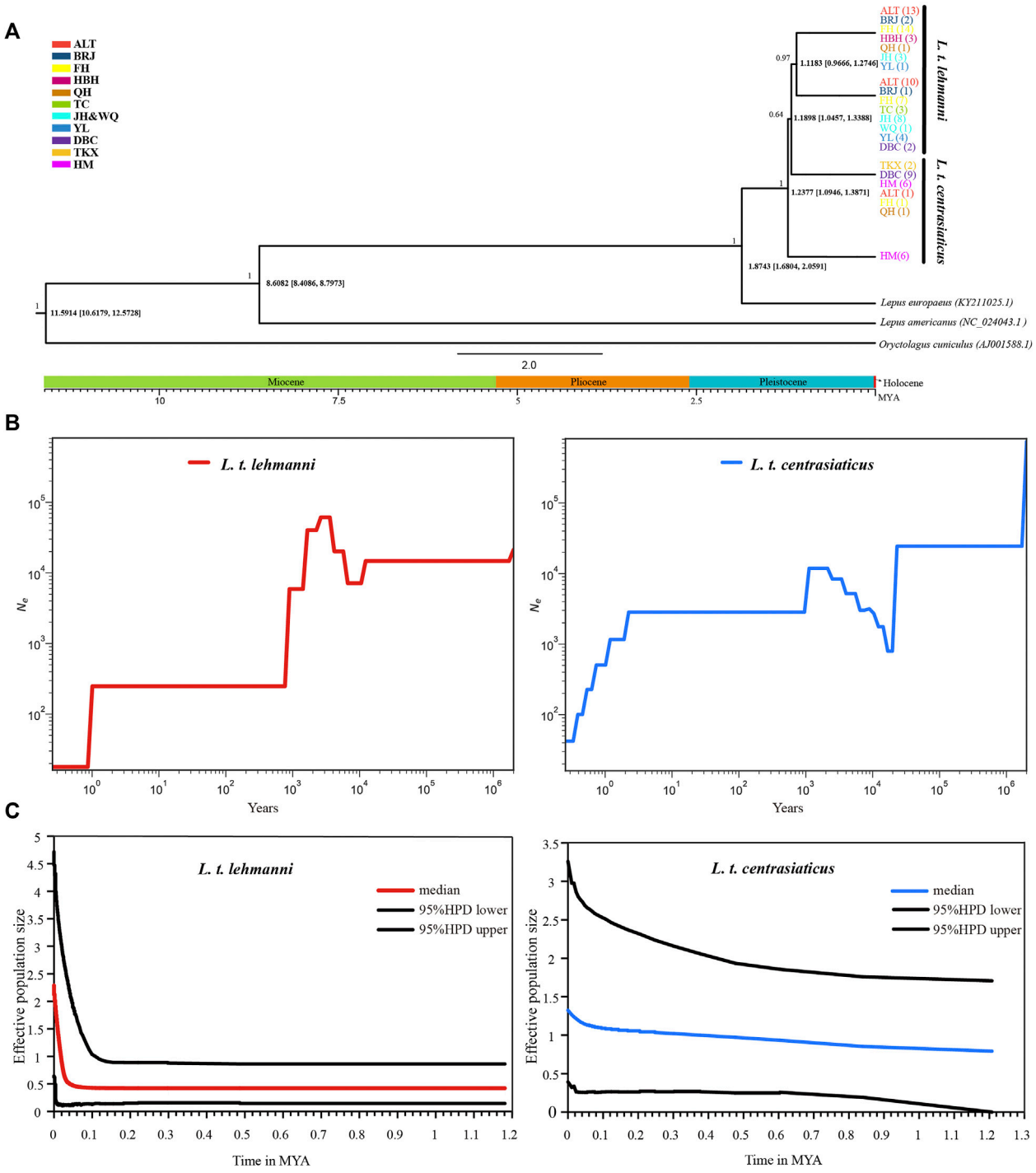


FIGURE 4
Divergence time and population history of Tolai hare populations. **(A)**. Time scale of Tolai hare populations divergence in Xinjiang based on mtDNA. Posterior probabilities are shown above nodes, divergence time with 95% highest posterior densities (HPD) are shown below nodes. **(B)**. SMC++ for *L. t. lehmanni* and *L. t. centrasiaticus*. **(C)**. Historical demographic trends represented by Extended Bayesian Skyline Plot (EBSP) for *L. t. lehmanni* and *L. t. centrasiaticus*.

2003). These higher genetic diversity values indicated that Tolai hare populations evolved under long-term favorable environmental conditions and remained relatively stable and genetically variable (Pironon et al., 2017). Larger effective population sizes may have also contributed to this (Kimura, 1983; Hague and Routman, 2016).

SNPs and mtDNA-based genetic diversity analysis of the subspecies showed that *L. t. lehmanni* polymorphism level was higher than that of *L. t. centrasiaticus*, likely due to a more favorable living environment and distribution (Table 2 and Figure 3B). In Xinjiang, *L. t. lehmanni* is mainly distributed in the northern and northwestern regions, which are more humid and include vast forests, grasslands, and croplands, providing more water and food for survival. While *L. t. centrasiaticus* is mainly distributed in the central and eastern Xinjiang, including the Turpan-Hami basin, which is primarily the Gobi desert; thus, survival there is likely more challenging.

The coexistence of genetic differentiation and gene flow in Tolai hare populations

Geological formation events can cause physical barriers to disperse species, such as mountains and rivers, leading to divergence and speciation (Chaves et al., 2011). Thus, geographical isolation is an important factor in genetic differentiation (Duncan et al., 2015; Gao et al., 2021). The effects of geographic isolation on genetic structure have been reported in several species, such as *Sarcophanops* (Campillo et al., 2020), *Orestias ascotanensis* (Cruz-Jofré et al., 2016), *Weigela coraeensis* (Yamada and Maki, 2012), *L. capensis,* and *L. europaeus* (Ben Slimen et al., 2008). In this study, phylogenetic trees based on SNP and mtDNA markers (Figures 2A, 3A) showed that Tolai hares were divided into two main clusters (*L. t. lehmanni* and *L. t. centrasiaticus*). Most samples from the same or adjacent regions clustered together, and their clustering relationships were consistent with their geographic distribution, indicating a rough phylogeographic distribution pattern. This was also supported by the network analysis (Figure 3C) and the PCA (Figure 2B). In addition, combined with the F_{ST} and significant *p*-values in AMOVA based on SNPs and mtDNA (Tables 3 and 4), differentiation occurred between subspecies *L. t. lehmanni* and *L. t. centrasiaticus*. These two subspecies distribution areas span across the Tianshan Mountains in Xinjiang. We speculate that the high mountains, the barren basins and deserts, and long geographical distances serve as geographical barriers that limit the dispersal and communication between two subspecies, leading to genetic divergence. In addition, the different populations may have evolved different ecotypes with different feeding habits or habitats, hindering gene exchange and leading to differentiation (Sarabia et al., 2021).

A strong differentiation was also found between the Central and Eastern groups in *L. t. centrasiaticus* according to the phylogenetic analysis and AMOVA results from mtDNA (Table 4). Although the Dabancheng (DBC) and Tuokexun (TKX) of the Central Group are adjacent to the Kumul (HM) of the Eastern Group, they are separated by the huge Turpan-Hami Basin and the Gobi Desert and are relatively distant geographically. Moreover, the food shortage caused by drought, little rain, and hot habitat further prevent the migration, spread, and exchange of hares, thus promoting differentiation.

In this study, the genetic structure analysis showed that geographical isolation caused by geographical barriers, such as distance, mountains, basins, and deserts, as well as the ecological environment of the habitat, affected the genetic structure of hare populations. This also occurs in *Yarkand hare* in Xinjiang. The population genetics of Yarkand hares based on SNPs (Ababaikeri et al., 2021) and mtDNA (Shan et al., 2011) showed a systematic geographical distribution pattern and genetic differentiation between the north and southwest populations, which may have been caused by geographical isolation and different living environment. It is concluded that geographical isolation and complex habitats can affect hare populations' genetic structure and differentiation.

On the other hand, parapatric and sympatric subspecies or populations are accessible to gene flow, thus affecting the genetic structure and evolutionary history. Despite clear population differentiation and significant variation among subspecies and geographical groups, our phylogenetic analyses, PCA plot, ADMIXTURE, and Treemix results revealed a certain degree of lineage admixture and gene flow between subspecies. This is likely due to the strong adaptability to environmental changes, large effective population sizes, and relatively strong migration capability, promoting gene exchange among populations (Ababaikeri et al., 2021). Moreover, since ancient times, the Tianshan Corridor in the Silk Road has connected the northern and southern Xinjiang (Chen, 2014). Although Tianshan Mountains lie across Xinjiang, the biological corridors may contribute to admixture. Gene flow also has an important impact on population differentiation. Generally speaking, differentiation occurs when there is limited gene flow between populations or strong natural selection is sufficient to overcome the dilution effect of gene flow between populations on population differentiation (Nosil, 2008; Li et al., 2014). Many studies reported gene flow coexisting in species or population differentiation (Nadachowska and Babik, 2009; Ababaikeri et al., 2021). In short, the geographical factors and the properties of hares might contribute to the coexistence of genetic differentiation and exchange in Tolai hares.

The complex population history of Tolai hare populations

Climate-driven environmental changes during the Pleistocene influenced the evolution of many terrestrial species (Meiri et al., 2020). However, the influence of climate fluctuation on the Tolai hares population history has not been studied. In this study, neutrality tests, nucleotide mismatch distribution, EBSP analysis based on mtDNA, and SMC++ analysis based on SNP were used to explore the population history of Tolai hares in Xinjiang. The integrated results indicated that the Tolai hare population has a complex history. SMC++ shows that *L. t. centrasiaticus* effective population size decreased during the LGM while *L. t. lehmanni* population remained stable. *L. t. lehmanni* was less affected by LGM might because it is distributed between the mountains, which cushion drastic climate fluctuations (Yuan et al., 2019). In the interglacial period after the LGP, the population of both subspecies had a small increase with the growth of *L. t. lehmanni* being more evident, consistent with the

EBSP analysis based on mtDNA. The population of the two subspecies decreased from about 0.001 to 0.0015 MYA. Previous studies have shown that early humans in Xinjiang expanded into the present region during this period (Tan et al., 2022), and that the early diet of humans in northwest China included wild animals (Ren et al., 2017). Meanwhile, previous researchers believed that *Lepus* could be influenced by human activities (Ge et al., 2013). Therefore, We hypothesize that human activity during this period led to population decline in both subspecies.

Unlike other species in Europe and eastern China, such as the Brown hare (Minoudi et al., 2018), Panda (*Ailuropoda melanoleuca*) (Zhao et al., 2013), and Lizards (*Shinisaurus crocodilurus*) (Xie et al., 2022), our comprehensive indexes of two markers showed that *L. t. lehmanni* population remained stable, even in the LGM (0.0265–0.019 MYA) (Clark et al., 2009; Liang et al., 2017). Other studies have shown that among the Tibetan Plateau species, population size did not decline in the LGM (Liang et al., 2017), implying that the geographical distribution of species and climate oscillation indeed play an important role in population history.

Tolai hare differentiated at 1.2377MYA, and the divergence between subspecies occurred between 1.1183 and 1.1898MYA, later than the formation time of the Tianshan Mountains and the Turpan-Hami Basin (Li et al., 2006). Therefore, mountains and basins act as geographical barriers for gene exchange between hare populations, resulting in apparent differentiation between the two subspecies and within *L. t. centrasiaticus*.

However, the limitations of the SLAF-seq and the mtDNA fragments used in population history analysis should also be considered. In the future, whole genome resequencing and complete mitogenome sequencing will be performed to obtain more reliable results on the Tolai hare population history.

Conclusion

This paper used SNP and mitochondrial markers to explore the effects of geological formation and environmental and climate factors on the Tolai hare. We found that Tolai hares have a high genetic diversity due to their strong adaption and migration ability. In addition, geological events such as the formation of mountains, basins, and other geographical factors led to differentiation and gene flow between subspecies *L. t. lehmanni* and *L. t. centrasiaticus*. Thus, it is due to the different geographical and geological conditions resulting in a relatively steady climate, making *L. t. lehmanni* stable and less affected by LGM compared to *L. t. centrasiaticus*.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA850843. The Accession number of mtDNA data for the Xinjiang Tolai hare in this study see the Additional file 2: Supplementary Table S2.

Ethics statement

All experimental protocols involved in this study were approved by the Institutional Animal Care and Use Committee of the College of Life Science and Technology, Xinjiang University, Urumqi, China.

Author contributions

WS, MM and PD conceived and designed the study. MM, SZ, and CZ sorted out the samples. MM, PD, SZ, PL, YM, WN, and PT participated in the experiment. WS, MM and PD performed data analysis, interpreted the results and wrote the manuscript. All authors reviewed and approved the manuscript before submission.

Funding

This work was funded by the National Natural Science Foundation of China (No. 31860599, 32260116).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1018632/full#supplementary-material

References

Ababaikeri, B., Zhang, Y., Dai, H., and Shan, W. (2021). Revealing the coexistence of differentiation and communication in an endemic hare, *Lepus yarkandensis* (Mammalia, Leporidae) using specific-length amplified fragment sequencing. *Front. Zool.* 18, 50. doi:10.1186/s12983-021-00432-x

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109

Ali, I., Teng, Z., Bai, Y., Yang, Q., Hao, Y., Hou, J., et al. (2018). A high density SLAF-SNP genetic map and QTL detection for fibre quality traits in Gossypium hirsutum. *BMC Genomics* 19, 879. doi:10.1186/s12864-018-5294-5

Allendorf, F. W. (2017). Genetics and the conservation of natural populations: Allozymes to genomes. *Mol. Ecol.* 26, 420–430. doi:10.1111/mec.13948

Ben Slimen, H., Suchentrunk, F., Stamatis, C., Mamuris, Z., Sert, H., Alves, P. C., et al. (2008). Population genetics of cape and Brown hares (*Lepus capensis* and *L. europaeus*): A test of petter's hypothesis of conspecificity. *Biochem. Syst. Ecol.* 36, 22–39. doi:10.1016/j.bse.2007.06.014

Ben Slimen, H., Awadi, A., Tolesa, Z. G., Knauer, F., Alves, P. C., Makni, M., et al. (2018). Positive selection on the mitochondrial ATP synthase 6 and the NADH dehydrogenase 2 genes across 22 hare species (genus *Lepus*). *J. Zool. Syst. Evol. Res.* 56, 428–443. doi:10.1111/jzs.12204

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., et al. (2014). Beast 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537. doi:10.1371/journal.pcbi.1003537

Campillo, L. C., Manthey, J. D., Thomson, R. C., Hosner, P. A., and Moyle, R. G. (2020). Genomic differentiation in an endemic Philippine genus (Aves: *Sarcophanops*) owing to geographical isolation on recently disassociated islands. *Biol. J. Linn. Soc. Lond.* 131, 814–821. doi:10.1093/biolinnean/blaa143

Chang, Y., He, P., Wang, H., Li, H., Wang, S., and Li, L. (2019). Application of high-throughput sequencing to evaluate the genetic diversity among wild apple species indigenous to shandong, China, and introduced cultivars. *Plant Mol. Biol. Rep.* 37, 63–73. doi:10.1007/s11105-019-01138-5

Chaves, J. A., Weir, J. T., and Smith, T. B. (2011). Diversification in Adelomyia hummingbirds follows Andean uplift. *Mol. Ecol.* 20, 4564–4576. doi:10.1111/j.1365-294X.2011.05304.x

Chen, T. (2014). "The Silk Road: The initial section and tianshan corridor network" highlights the study of universal values". *China Cult. Herit.* (3), 72–81. (in Chinese)

Cheng, C., Ge, D., Xia, L., Zhou, C., and Yang, Q. (2012). Morphometrics study on the so called 'cape hare' (lagomorpha: Leporidae: *Lepus*) in China. *Acta Theriol. Sin.* 32, 275–286. doi:10.16829/j.slxb.2012.04.001 (in Chinese)

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., et al. (2009). The last glacial maximum. *Science* 325, 710–714. doi:10.1126/science.1172873

Cruz-Jofré, F., Morales, P., Vila, I., Esquer-Garrigos, Y., Hugueny, B., Gaubert, P., et al. (2016). Geographical isolation and genetic differentiation: The case of *Orestias ascotanensis* (teleostei: Cyprinodontidae), an andean killifish inhabiting a highland salt pan. *Biol. J. Linn. Soc.* 117, 747–759. doi:10.1111/bij.12704

Djan, M., Stefanovi, M., Velikovi, N., Lavadinovi, V., Paulo, C. A., and Suchentrunk, F. (2017). Brown hares (*Lepus europaeus* Pallas, 1778) from the balkans: A refined phylogeographic model. *Hystrix-italian J. Mammal.* 28, 186–193doi:10.4404/hystrix–28.2-12202

Duncan, C. J., Worth, J. R. P., Jordan, G. J., Jones, R. C., and Vaillancourt, R. E. (2015). Genetic differentiation in spite of high gene flow in the dominant rainforest tree of southeastern Australia, *Nothofagus cunninghamii*. *Heredity* 116, 99–106. doi:10.1038/hdy.2015.77

Erbajeva, M. A., and Alexeeva, N. V. (2000). Pliocene and Pleistocene biostratigraphic succession of Transbaikalia with emphasis on small mammals. *Quat. Int.* 68, 67–75. doi:10.1016/s1040-6182(00)00033-1

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under linux and windows. *Mol. Ecol. Resour.* 10, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Fang, H., Liu, H., Ma, R., Liu, Y., Li, J., Yu, X., et al. (2020). Genome-wide assessment of population structure and genetic diversity of Chinese Lou onion using specific length amplified fragment (SLAF) sequencing. *PLoS One* 15, e0231753. doi:10.1371/journal.pone.0231753

Gao, Y., Wang, D.-j., Wang, K., Cong, P.-h., Li, L.-w., and Piao, J.-c. (2021). Analysis of genetic diversity and structure across a wide range of germplasm reveals genetic relationships among seventeen species of *Malus* Mill. native to China. *J. Integr. Agric.* 20, 3186–3198. doi:10.1016/s2095-3119(20)63421-9

Ge, D., Wen, Z., Xia, L., Zhang, Z., Erbajeva, M., Huang, C., et al. (2013). Evolutionary history of lagomorphs in response to global environmental change. *PLoS One* 8, e59668. doi:10.1371/journal.pone.0059668

Giannoulis, T., Plageras, D., Stamatis, C., Chatzivagia, E., Tsipourlianos, A., Birtsas, P., et al. (2019). Islands and hybrid zones: Combining the knowledge from "natural laboratories" to explain phylogeographic patterns of the European brBrownare. *BMC Evol. Biol.* 19, 17. doi:10.1186/s12862-019-1354-y

Hague, M. T., and Routman, E. J. (2016). Does population size affect genetic diversity? A test with sympatric lizard species. *Heredity* 116, 92–98. doi:10.1038/hdy.2015.76

Kardos, M. (2021). Conservation genetics. *Curr. Biol.* 31, R1185–R1190. doi:10.1016/j.cub.2021.08.047

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kozlov, A. M., Diego, D., Tomáš, F., Benoit, M., and Alexandros, S. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi:10.1093/bioinformatics/btz305

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Leigh, J. W., and Bryant, D. (2015). PopART: Full-Feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi:10.1111/2041-210x.12410

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698

Li, J., Wang, K., Li, Y., Sun, G., Chu, C., Li, L., et al. (2006). Geomorphological features, crustal composition and geological evolution of the Tianshan Mountains. *Geol. Bull. China* 25, 895–909. doi:10.3969/j.issn.1671-2552.2006.08.001 (in Chinese)

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi:10.1093/bioinformatics/btp336

Li, Z., Liu, Z., Wang, M., Qian, Z., Zhao, P., Zhu, Z., et al. (2014). A review on studies of speciation in the presence of gene flow: Evolution of reproductive isolation. *Biodivers. Sci.* 22, 88–96. doi:10.3724/sp.j.1003.2014.13143 (in Chinese)

Li, Z., Wei, S., Li, H., Wu, K., Cai, Z., Li, D., et al. (2017). Genome-wide genetic structure and differentially selected regions among Landrace, Erhualian, and Meishan pigs using specific-locus amplified fragment sequencing. *Sci. Rep.* 7, 10063. doi:10.1038/s41598-017-09969-6

Liang, Y., He, D., Jia, Y., Sun, H., and Chen, Y. (2017). Phylogeographic studies of schizothoracine fishes on the central Qinghai-Tibet Plateau reveal the highest known glacial microrefugia. *Sci. Rep.* 7, 10983. doi:10.1038/s41598-017-11198-w

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482. doi:10.1038/nature10530

Liu, J., Yu, L., Arnold, M. L., Wu, C. H., Wu, S. F., Lu, X., et al. (2011). Reticulate evolution: Frequent introgressive hybridization among Chinese hares (genus *lepus*) revealed by analyses of multiple mitochondrial and nuclear DNA loci. *BMC Evol. Biol.* 11, 1–14. doi:10.1186/1471-2148-11-223

Liu, K., and Muse, S. V. (2005). PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* 21, 2128–2129. doi:10.1093/bioinformatics/bti282

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Meiri, M., Lister, A., Kosintsev, P., Zazula, G., and Barnes, I. (2020). Population dynamics and range shifts of moose (*Alces alces*) during the Late Quaternary. *J. Biogeogr.* 47, 2223–2234. doi:10.1111/jbi.13935

Minoudi, S., Papapetridis, L., Karaiskou, N., Chatzinikos, E., Triantaphyllidis, C., Abatzopoulos, T. J., et al. (2018). Genetic analyses of Brown hare (*Lepus europaeus*) support limited migration and translocation of Greek populations. *PLoS One* 13, e0206327. doi:10.1371/journal.pone.0206327

Nadachowska, K., and Babik, W. (2009). Divergence in the face of gene flow: The case of two newts (amphibia: Salamandridae). *Mol. Biol. Evol.* 26, 829–841. doi:10.1093/molbev/msp004

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300

Nosil, P. (2008). Speciation with gene flow could be common. *Mol. Ecol.* 17, 2103–2106. doi:10.1111/j.1365-294X.2008.03715.x

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967. doi:10.1371/journal.pgen.1002967

Pierpaoli, M., Riga, F., Trocchi, V., and Randi, E. (2003). Hare populations in Europe: Intra and interspecific analysis of mtDNA variation. *C. R. Biol.* 326, 80–84. doi:10.1016/s1631-0691(03)00042-8

Pironon, S., Papuga, G., Villellas, J., Angert, A. L., Garcia, M. B., and Thompson, J. D. (2017). Geographic variation in genetic and demographic performance: New insights from an old biogeographical paradigm. *Biol. Rev. Camb. Philos. Soc.* 92, 1877–1909. doi:10.1111/brv.12313

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Qin, M., Li, C., Li, Z., Chen, W., and Zeng, Y. (2019). Genetic diversities and differentially selected regions between shandong indigenous pig breeds and western pig breeds. *Front. Genet.* 10, 1351–1361. doi:10.3389/fgene.2019.01351

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi:10.1093/sysbio/syy032

Ren, L., Li, X., Kang, L., Brunson, K., Liu, H., Dong, W., et al. (2017). Human paleodiet and animal utilization strategies during the Bronze Age in northwest Yunnan Province, southwest China. *PLoS One* 12, e0177867. doi:10.1371/journal.pone.0177867

Ren, A., Du, K., Jia, X., Yang, R., Wang, J., Chen, S. Y., et al. (2019). Genetic diversity and population structure of four Chinese rabbit breeds. *PLoS One* 14, e0222503. doi:10.1371/journal.pone.0222503

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi:10.1093/sysbio/sys029

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi:10.1093/molbev/msx248

Sarabia, C., vonHoldt, B., Larrasoana, J. C., Urios, V., and Leonard, J. A. (2021). Pleistocene climate fluctuations drove demographic history of African golden wolves (*Canis lupaster*). *Mol. Ecol.* 30, 6101–6120. doi:10.1111/mec.15784

Shan, W., Liu, J., Yu, L., Robert, W. M., Mahmut, H., and Zhang, Y. (2011). Genetic consequences of postglacial colonization by the endemic Yarkand hare (*Lepus yarkandensis*) of the arid Tarim Basin. *Chin. Sci. Bull.* 56, 1370–1382. doi:10.1007/s11434-011-4460-9

Shan, W., Dai, H., and Zhang, Y. (2020a). Classification and genetic diversity of three hare species in Xinjiang based on mitochondrial DNA. *Acta Veterinaria Zootechnica Sinica* 51, 80–89. doi:10.11843/j.issn.0366-6964.2020.10.008 (in Chinese)

Shan, W., Tursun, M., Zhou, S., Zhang, Y., and Dai, H. (2020b). The complete mitochondrial genome sequence of *Lepus tolai* in Xinjiang. *Mitochondrial DNA Part B* 5, 1336–1337. doi:10.1080/23802359.2020.1735267

Smith, A. T., Johnston, C. H., Alves, P. C., and Hacklander, K. (2018). *Lagomorphs: Pikas, rabbits, and hares of the world.* Baltimore: Johns Hopkins University Press.

Smith, A. T., and Xie, Y. (2008). *A guide to the mammals of China* (in Chinese).

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4, vey016. doi:10.1093/ve/vey016

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-Seq: An efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8, e58700. doi:10.1371/journal.pone.0058700

Tan, B., Wang, H., Wang, X., Yi, S., Zhou, J., Ma, C., et al. (2022). The study of early human settlement preference and settlement prediction in Xinjiang, China. *Sci. Rep.* 12, 5072. doi:10.1038/s41598-022-09033-y

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* 49, 303–309. doi:10.1038/ng.3748

Wang, J., and Yang, G. (2012). Complete mitogenome of cape hare *Lepus capensis* (Lagomorpha: Leporidae) and its phylogenetic considerations. *Acta Theriol. Sin.* 32, 1–11. doi:10.16829/j.slxb.2012.01.001

Wang, W., Zhang, T., Wang, J., Zhang, G., Wang, Y., Zhang, Y., et al. (2016). Genome-wide association study of 8 carcass traits in Jinghai Yellow chickens using specific-locus amplified fragment sequencing technology. *Poult. Sci.* 95, 500–506. doi:10.3382/ps/pev266

Wu, C., Wu, J., Bunch, T. D., Li, Q., Wang, Y., and Zhang, Y. P. (2005). Molecular phylogenetics and biogeography of *Lepus* in Eastern Asia based on mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 37, 45–61. doi:10.1016/j.ympev.2005.05.006

Wu, Y., Xia, L., Zhang, Q., Yang, Q., and Meng, X. (2011). Bidirectional introgressive hybridization between *Lepus capensis* and *Lepus yarkandensis*. *Mol. Phylogenet. Evol.* 59, 545–555. doi:10.1016/j.ympev.2011.03.027

Xie, H. X., Liang, X. X., Chen, Z. Q., Li, W. M., Mi, C. R., Li, M., et al. (2022). Ancient demographics determine the effectiveness of genetic purging in endangered Lizards. *Mol. Biol. Evol.* 39, msab359. doi:10.1093/molbev/msab359

Yamada, T., and Maki, M. (2012). Impact of geographical isolation on genetic differentiation in insular and mainland populations of *Weigela coraeensis* (Caprifoliaceae) on Honshu and the Izu Islands. *J. Biogeogr.* 39, 901–917. doi:10.1111/j.1365-2699.2011.02634.x

Yuan, J., Ye, Z., and Bu, W. (2019). Phylogeography of widespread species in Eurasia: Current progress and future prospects. *Sci. Sin. -Vitae.* 49, 1155–1164. doi:10.1360/ssv-2019-0163 (in Chinese)

Zhang, D., and Zhang, Z. (2005). Single nucleotide polymorphisms (SNPs) discovery and linkage disequilib-rium (LD) in forest trees. *For. Stud. China* 7, 1–14. doi:10.1007/s11632-005-0024-x

Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., et al. (2013). Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biol.* 13, 141–152. doi:10.1186/1471-2229-13-141

Zhang, D., Gao, F., Li, W., Jakovlić, I., Zou, H., Zhang, J., et al. (2018). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* 20, 348–355. doi:10.1111/1755-0998.13096

Zhang, J., Sun, B., Li, C., Chen, W., Jiang, L., Lv, S., et al. (2020a). Molecular diversity and genetic structure of wild rice accessions (*Oryza rufipogon* Griff.) in Guangdong Province, China, as revealed by SNP markers. *Genet. Resour. Crop Evol.* 68, 969–978. doi:10.1007/s10722-020-01038-8

Zhang, Y., Zeng, W., Xu, P., Alemujiang, G., and Shan, W. (2020b). The screening of DNA barcode for hares in Xinjiang. *Acta Veterinaria Zootechnica Sinica* 51, 270–278. doi:10.11843/j.issn.0366-6964.2020.02.008 (in Chinese)

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., et al. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* 45, 67–71. doi:10.1038/ng.2494

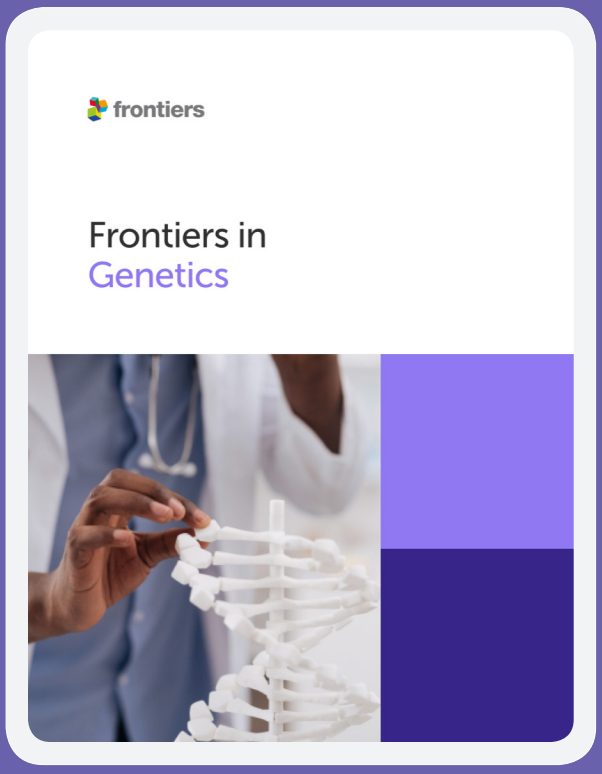
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact