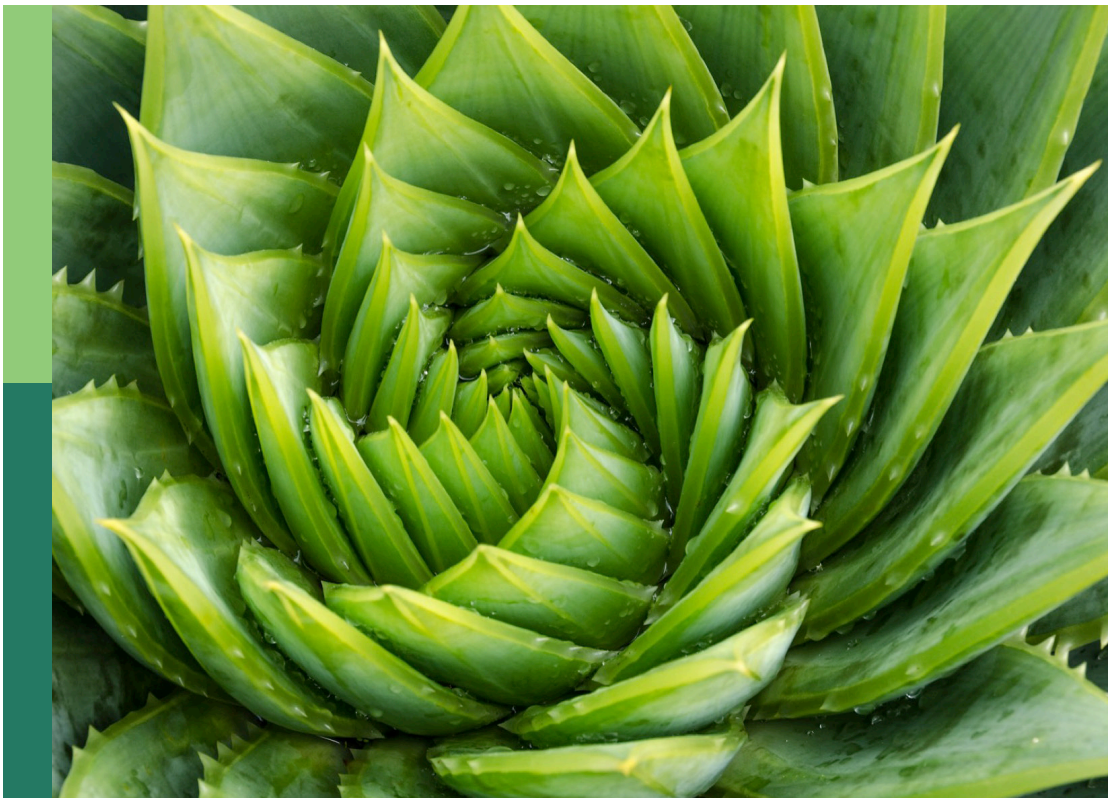
frontiers | Research Topics

Economic plant genome and database construction and research

Edited by
Gao Jihai, Mark Chapman, Shuangyang Wu,
Wei Xu and Zhichao Xu

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4679-6
DOI 10.3389/978-2-8325-4679-6

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

Economic plant genome and database construction and research

Topic editors

Gao Jihai — Chengdu University of Traditional Chinese Medicine, China
Mark Chapman — University of Southampton, United Kingdom
Shuangyang Wu — Gregor Mendel Institute of Molecular Plant Biology (GMI), Austria
Wei Xu — Kunming Institute of Botany, Chinese Academy of Sciences (CAS), China
Zhichao Xu — Northeast Forestry University, China

Citation

Jihai, G., Chapman, M., Wu, S., Xu, W., Xu, Z., eds. (2024). *Economic plant genome and database construction and research*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4679-6

Table of contents

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 30 January 2023
DOI 10.3389/fpls.2023.1131977

OPEN ACCESS

EDITED BY
Zhichao Xu,
Northeast Forestry University, China

REVIEWED BY
Xuan Ma,
Tianjin Normal University, China
Chao Yang,
South China Botanical Garden, Chinese Academy of Sciences (CAS), China
Minting Liang,
South China Botanical Garden, Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Yan Chen
yanchen@gzhu.edu.cn
Lei Zhu
lei_zhu@scu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 26 December 2022
ACCEPTED 11 January 2023
PUBLISHED 30 January 2023

CITATION
Wang C, Chen W, Aili M, Zhu L and Chen Y (2023) tRNA-derived small RNAs in plant response to biotic and abiotic stresses. *Front. Plant Sci.* 14:1131977. doi: 10.3389/fpls.2023.1131977

tRNA-derived small RNAs in plant response to biotic and abiotic stresses

Chaojun Wang[1†], Weiqiang Chen[2,3†], Maimaiti Aili[3], Lei Zhu[4*] and Yan Chen[5*]

[1]Institute of Education Science, Leshan Normal University, Leshan, China, [2]Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [3]Xinjiang Institute of Traditional Uyghur Medicine, Urumqi, China, [4]Institute of Thoracic Oncology and Department of Thoracic Surgery, West China Hospital, Sichuan University, Chengdu, China, [5]Guangzhou Key Laboratory of Crop Gene Editing, Innovative Center of Molecular Genetics and Evolution, School of Life Sciences, Guangzhou University, Guangzhou, China

tRNA-derived small RNAs (tsRNAs) represent a novel category of small non-coding RNAs and serve as a new regulator of gene expression at both transcriptional and post-transcriptional levels. Growing evidence indicates that tsRNAs can be induced by diverse stimuli and regulate stress-responsive target genes, allowing plants to adapt to unfavorable environments. Here, we discuss the latest developments about the biogenesis and classification of tsRNAs and highlight the expression regulation and potential function of tsRNAs in plant biotic and abiotic stress responses. Of note, we also collect useful bioinformatics tools and resources for tsRNAs study in plants. Finally, we propose current limitations and future directions for plant tsRNAs research. These recent discoveries have refined our understanding of whether and how tsRNAs enhance plant stress tolerance.

KEYWORDS

plant, tsRNAs, stress response, expression regulation, biological function

Introduction

Being sessile, plants are continuously exposed to a variety of biotic and abiotic stresses, e.g., salt, drought, cold or heat stress as well as fungal or virus infection, which are major constrains for the growth, productivity and quality of all kinds of agricultural and horticultural plants. To cope with these extreme situations and resist the resulting adverse effects, plants have evolved sophisticated response strategies based on multiple gene regulatory mechanisms, including transcriptional regulation by changing epigenetic modifications (Chang et al., 2020) and post-transcriptional regulation through miRNAs induced gene silencing (Jones-Rhoades et al., 2006). With the rapid development of the next-generation sequencing technologies and bioinformatics approaches, tRNA-derived small RNAs (tsRNAs), first considered as byproducts of tRNAs random degradation, have been characterized in all three kingdoms of life as a new class of regulatory small non-coding

RNAs involved in a wide range of biological processes, such as growth, development, diseases as well as stress responses (Maute et al., 2013; Goodarzi et al., 2015; Chen et al., 2016; Kim et al., 2017; Zhu et al., 2018a; Zhu et al., 2018b; Yamasaki et al., 2009; Zhu et al., 2019). tsRNAs, depending on the length, cleavage site and precursor type, can be divided into three major types: 5' or 3' tRHs (tRNA-derived halves) derived from the cleavage of mature tRNAs at the anticodon loop, 30-35 nt in length; 5', 3' or inter tRFs (tRNA-derived fragments) derived from the cleavage of mature tRNAs at the D and/or TψC loop, 10-30 nt in length; 3'U tRF derived from the cleavage of pre-tRNAs by RNase Z during processing (Zhu et al., 2018a; Zhu et al., 2018b; Lyons et al., 2018; Zhu et al., 2019; Ma et al., 2021a).

Growing evidence shows that the expression of several specific tsRNAs in plant are changed obviously under certain stress conditions like oxidative, drought or heat stress, as well as phosphate (Pi) starvation (Thompson et al., 2008; Hsieh et al., 2009; Wang et al., 2011; Loss-Morais et al., 2013; Hackenberg et al., 2013). Functional analyses have demonstrated that those stress-regulated tsRNAs play vital roles in plant response to both biotic and abiotic stresses, often by regulating the expression of stress-related genes (Asha and Soniya, 2016; Gu et al., 2022; Sun et al., 2022). Therefore, characterizing these stress-responsive tsRNAs and understanding tsRNA-guided stress regulatory networks could provide new ways to enhance stress tolerance in plants, which is of great value in sustainable agricultural and horticultural production. In this review, we comprehensively summarize the current progresses in the diversity, biogenesis and function of tsRNAs in plants, and highlight the expression regulation and potential function of plant tsRNAs in biotic and abiotic stress responses. In addition, we also collect the relevant information about useful bioinformatics tools and resources for tsRNAs study in plants. At present, research into tsRNAs still faces tough challenges as how to accurately and efficiently interfere or quantify their expression and thus interpret their exact functions and mechanisms, which require future efforts to develop new and efficient approaches.

Roles of ribonucleases in RNA metabolism

As ribonucleases (RNases) are responsible for tsRNAs production, we first give a brief introduction of the types and functions of RNases. Primary transcripts are synthesized by RNA polymerases, while subsequent RNA processing to generate shorter functional RNA species (mRNA, rRNA, tRNA or regulatory RNAs) or degradation to eliminate aberrant RNAs are mainly catalyzed by RNases (Deshpande and Shankar, 2002). RNases are present in almost all organisms including bacteria, virus, yeast, plants, and animals, and play vital roles in RNA metabolism (Irie, 1999). They come in two categories namely endoribonucleases and exoribonucleases on the basis of their mechanism of action (Condon, 2009; Matos et al., 2011). Endoribonucleases cut RNA molecules internally like a pair of scissors while exoribonucleases remove terminal nucleotides from either the 3' end or the 5' end of the RNA molecules as a "Pacman". RNases can act on single-stranded RNAs, double-stranded RNAs and DNA-RNA hybrids hydrolytically or phosphorolytically (Irie, 1999; Condon, 2009; Luhtala and Parker, 2010; Matos et al., 2011).

In plants, the majority of RNases cleave RNAs *via* the formation of 2',3'-cyclic phosphate (cP) intermediates, ultimately generating oligo- or mononucleotides with a 3'-phosphate group (Eun, 1996; Irie, 1999). The 2',3'-cyclizing RNases, also known as transferase-type RNases, include three groups corresponding to RNase T1, RNase A and RNase T2 families ((Irie, 1999; MacIntosh, 2011). These RNases are usually secreted or targeted to organelles associated with the secretory system such as the lysosome or vacuole (Deshpande and Shankar, 2002; MacIntosh, 2011). Thus, they are localized in a space normally without the presence of RNA substrates. Enzymes from RNase T1 family are guanylic acid specific alkaline RNases with optimal pH7-8 and distributed in certain species of fungi and bacteria. The vertebrate-specific RNase A family is weakly acidic (pH6.5-7) or alkaline (pH7-8), with pyrimidine base specificity. First purified from the fungal *Aspergillus orzae*, RNase T2 family proteins are acidic transferase-type endoribonucleases without base-specificity, present in almost all organisms and highly conserved in eukaryotes (Luhtala and Parker, 2010). Phylogenetic analyses have defined three subclasses of the RNase T2 family in plants (MacIntosh and Castandet, 2020). Class I enzymes are diversified, tissue-specific and often regulated by stresses. Class II proteins are highly conserved in plant genomes and carry out a housekeeping role in rRNA recycling, the ancestral function of eukaryotic RNase T2 enzymes. Class III, the first identified plant RNase T2 proteins, were initially cloned in *Nicotiana alata* as self-incompatibility genes (S genes) encoding style-specific glycoproteins, and subsequently shown to be ribonucleases (S-RNases). In *Arabidopsis*, five members of the RNase T2 family have been identified (RNS1-5), among which RNS1, RNS3, RNS4 and RNS5 are categorized as Class I and RNS2 belongs to Class II (MacIntosh et al., 2010). Clearly, understanding the types and mechanisms of RNases are essential for studying the biogenesis and function of different types of RNAs in plant.

Biogenesis and classification of tsRNAs in plants

In addition to their well-known function in protein synthesis, tRNAs can be cleaved at specific sites by different endoribonucleases to produce tsRNAs, varying in length, sequence and functions (Zhu et al., 2018a; Zhu et al., 2018b; Ma et al., 2021a). Broadly, tsRNAs can be classified as three main categories: 5' or 3' tRHs, 5', 3' or inter tRFs and 3'U tRF (Zhu et al., 2018a; Zhu et al., 2018b; Lyons et al., 2018; Zhu et al., 2019; Ma et al., 2021a) (Figure 1). Notably, most of the known tsRNAs are derived from mature tRNAs and no 3'U tRFs have been reported in plants so far. In mammals, 5' or 3' tRHs are 30-35 nt fragments generated as a result of the tRNA cleavage at the anticodon loop by Angiogenin belonging to the RNase A superfamily (Fu et al., 2009). In *Saccharomyces cerevisiae* and *tetrahymena thermophilus*, tRHs are cleaved by Rny1p and Rnt2, respectively, both of which are from RNase T2 family (Thompson and Parker, 2009; Andersen and Collins, 2012). Recent studies from Megel et al. show that RNase T2, but not Dicer-like proteins (DCLs), are key players for tRHs generation in *Arabidopsis* (Megel et al., 2019). For tRFs biogenesis,

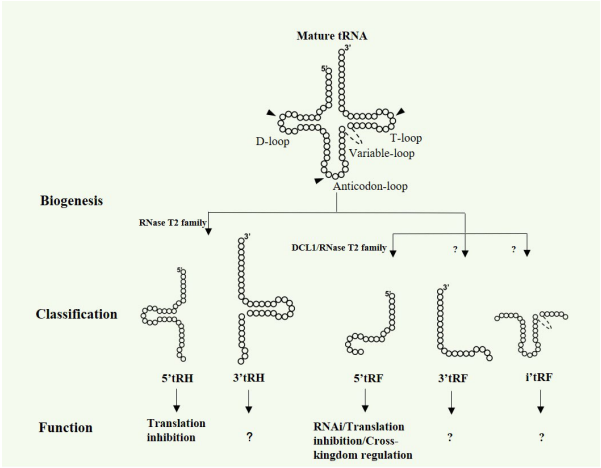


FIGURE 1
Biogenesis, classification and function of tsRNAs in plants. Arrow heads show the cleavage sites in mature tRNAs, generating 5 types of tsRNAs.

it is still somewhat controversial and requires further clarification. Early studies in human HeLa cells reveal that the abundance of a 20 nt tRF derived from $tRNA^{Gln}$ is markedly decreased when the Dicer expression is suppressed by siRNA, indicating the requirement of Dicer for tRFs biogenesis (Cole et al., 2009). However, subsequent small RNA sequencing data show that the mutation of DICER1 does not result in the decrease of tRFs expression in mouse, *Drosophila* and yeast (Kumar et al., 2014). In *Arabidopsis*, Dicer-like 1 (DCL1) is proposed to *be responsible for* the 19 nt tRFs generation in pollen grains (Martinez et al., 2017). Nevertheless, two independent studies indicate that DCLs are not essentially involved in tRFs biogenesis in *Arabidopsis* flower tissue and seedling (Alves et al., 2017; Megel et al., 2019). Megel and collaborators demonstrate that RNS1 and RNS3 are the main endoribonucleases to produce both tRFs and tRHs in siliques, whereas RNS2 is implicated in the tRFs biogenesis in leaf (Megel et al., 2019). Recently, RNS1 and RNS3 are reported to produce 5'tRFs and 5'tRHs from specific mature tRNAs, while the three prime ends of these tsRNAs are *2',3'-cP*, which further demonstrate the diversity and heterogeny of tsRNAs in plants. Based on the above work, tsRNAs production is quite different from miRNAs that are processed almost exclusively by DCLs. RNase T2 proteins, rather than DCLs, are the main players in plant tsRNAs biogenesis.

Molecular functions of tsRNAs

Mounting evidence in animals show that tsRNAs are abundantly expressed small non-coding RNAs that can regulate gene expression at multi-dimensional layers, such as transcription inhibition (Zhang et al., 2016), RNA degradation (Maute et al., 2013) and translation regulation (Yamasaki et al., 2009; Ivanov et al., 2011). Although functional studies on plant tsRNAs are relatively limited, a few results show that tsRNAs in plants share similar modes of action with that in animals, such as RNA silencing and translation inhibition (Martinez et al., 2017; Lalande et al., 2020) (Figure 1). The functional conservation of tsRNAs between plants and animals could be partially due to their similarity in biogenesis processes, inferring their key role in evolution. Interestingly, cross-kingdom regulation by tsRNAs has also been discovered in plant recently (Ren et al., 2019; Cao et al., 2022) (Figure 1).

RNA interfering

Increasing studies in animals demonstrate that several types of tsRNAs are miRNAs-like for their Dicer-dependent biogenesis and Argonaute (AGO)-associated functional mechanism (Maute et al., 2013; Megel et al., 2019). tsRNAs in animals can recognize RNA targets through sequence complementarity and induce RNA silencing (Maute et al., 2013). However, whether tsRNAs also act in an AGO-dependent manner and involve in the RNAi pathway are still obscure in plants. AGO-associated tRFs were first identified in *Arabiopdsis* by analyzing small RNAs co-immunoprecipitated with AGO proteins, indicating the possible contribution of plant tsRNAs in suppressing gene expression through RNAi pathway (Loss-Morais et al., 2013). Then, based on sequence complementarity between tsRNAs and mRNAs, four possible targets of the AGO-associated tRFs were predicted using a well-known plant small RNA target analysis server, namely psRNATarget (Dai and Zhao, 2011). Further degradome analyses, generally used to identify miRNAs cleavage sites and ta-siRNAs (trans-acting siRNAs) targets, were applied to confirm the possible cleavage of the four predicted tRFs targets, which can lower the false positive rate in the tsRNAs target prediction. Besides, AGO1-immunoprecipitated (AGO1-IP) small RNAs sequencing data from roots and flowers of *Arabidopsis* show that tsRNAs are derived from both nucleus- and plastid-encoded tRNAs. Further analyses of total RNAs and AGO1-IP small RNAs from *Arabidopsis* leaf treated with or without UV reveal that the amounts of 5' tRF from plastid-encoded tRNAs in total RNAs and/or AGO1-associated small RNAs are both decreased under UV treatment, while a 5' tRF from nucleus-encoded $tRNA^{GlyTCC}$ is significantly increased in total RNAs and strongly enriched in AGO1, suggesting their potential role in UV stress response (Cognat et al., 2017).

Bioinformatics analyses proved initially that tsRNAs may well associate with AGO system and their potential targets were predicted in several studies. However, standard confirmative experiments, such as Northern blot analyses for AGO-IP small RNAs and target RNA cleavage products, are still lacking. Plant tsRNAs are first indicated to be processed by DCL1 and mediate the target RNA degradation through AGO1 pathway by Martinez et al. (Martinez et al., 2017). A 19 nt pollen enriched 5' tRF^{AlaAGC} was shown to be decreased in *dcl1* mutant and enriched in AGO1, and the accumulation of this tRF in AGO1-IP small RNAs disappeared in *dcl1* mutant. These results were further confirmed by Northern blot and demonstrate the specific AGO1 loading of this DCL1 generated tRF. In addition, the target cleavage mediated by another 19 nt pollen enriched 5' tRF^{MetCAT} is dependent almost completely on DCL1 and partially on AGO1, which was verified through 5' RLM RACE (5' RNA Ligase-Mediated Rapid Amplification of cDNA Ends) that can capture the degraded products of target RNAs. Moreover, knockdown of 5' tRF^{MetCAT} with STTM (short tandem target mimic), a method initially designed for miRNAs silencing, can inhibit the target RNA cleavage (Yan et al., 2012). Thus, this study provides strong evidence that tRFs can regulate target gene expression in an

AGO-dependent manner. Most recently, another independent study in *Arabidopsis* raise again that the 19 nt 5' tRF^{AlaAGC} can suppress target gene expression through AGO1 pathway, which was validated by both Northern blot and 5' RACE (Gu et al., 2022). Using 5' RLM RACE, the cleavage of predicted tsRNAs targets was also testified in non-model plant organisms, including wheat and black pepper, whereas there is no evidence for the AGO association of these tsRNAs (Asha and Soniya, 2016; Sun et al., 2022).

Translation inhibition

It is well-known that amino acid charged tRNAs cooperate with rRNA and involve in protein synthesis. Under amino-acids starvation, uncharged tRNAs can suppress protein translation (Phizicky and Hopper, 2010). Intriguingly, increasing evidence in multiple organisms reveal that tsRNAs can repress or promote translation in an AGO-dependent or independent manner. (Ivanov et al., 2011; Kim et al., 2017; Shi et al., 2019). 5' tRHs, but not 3' tRHs, can suppress protein translation in human cells, and the terminal oligoguanine motif containing 4 Gs at the five prime end of 5' tRHs are required for displacing translation initiation factors engaged in both capped and uncapped mRNAs (Yamasaki et al., 2009; Ivanov et al., 2011). On the other hand, a 3' tRF derived from $tRNA^{LeuCAG}$ was proved to be able to unfold the duplexed RPS mRNAs at the targeting site, thus facilitating ribosome protein biogenesis and enhancing translation (Kim et al., 2017).

In plants, two studies suggest that plant tsRNAs can inhibit protein translation *in vitro*, while the exact mechanisms are still unclear (Zhang et al., 2009; Lalande et al., 2020). Fragments from non-coding RNAs, such as tRNAs, ribosomal RNAs, and spliceosomal RNAs, were found to be present in the phloem of pumpkin, and total RNAs extracted from phloem sap (PS) can suppress translation *in vitro* (Zhang et al., 2009). To prove the translation inhibition effect is caused by tsRNAs in PS, *in vitro* translation assay was performed using tRNA fragments produced from yeast tRNAs, as it is not feasible technically to isolate pure PS tRNA fragments in high amounts. Indeed, protein translation is inhibited by yeast tRNA fragments *in vitro*, while whether PS tRNA fragments are the principal agents of the translation inhibition remain non-conclusive and in controversy. The other study suggest that *Arabidopsis* tRNA fragments can repress translation in an unspecific manner. A series of oligo ribonucleotides mimicking natural tRFs were analyzed and only two, derived from the 5' ends of $tRNA^{AlaAGC}$ and $tRNA^{AsnGTT}$, can strongly attenuate translation. Unlike the mechanism in human, the G18 and G19 residues of *Arabidopsis* tRF^{Ala}, but not the 4 Gs present at the 5' ends, are essential for the translation inhibition. Furthermore, the 5' tRF^{AlaAGC} or 5' tRF^{AsnGTT} needs to associate with polyribosomes to induce translation inhibition, while sequence complementarity between tRFs and mRNAs is not required, suggesting that tRFs may act as general modulation factors of the translation process in plants (Lalande et al., 2020). Nevertheless, more efforts are needed to elucidate the precise mechanisms of tsRNAs in plant translation inhibition.

Cross-kingdom regulation

Exogenous plant miRNAs were first detected in the serum and plasma of human and animals by Zhang et al. (Zhang et al., 2012). Food-derived MIR168a, a miRNA highly expressed in rice, was further proved to be able to resist gastrointestinal tract and reach the serum and organs like liver, where it inhibits LDLRAP1 (low-density lipoprotein receptor adapter protein 1) expression and eventually suppresses the removal of LDL from the plasma. Since then, increasing studies revealed the cross-kingdom regulation by plant-derived miRNAs, while several negative evidences of miRNAs transference between kingdoms were also reported (Del Pozo-Acebo et al., 2021). Recently, SIDT1 (SID-1 transmembrane family member 1) expressed on gastric pit cells in the stomach was suggested to be required for the absorption of dietary miRNAs, which not only confirmed the phenomenon of cross-kingdom regulation, but also indicated the great potential of plant small RNAs for therapeutic purposes (Chen et al., 2021). Based on this, a 5' tRF derived from $tRNA^{HisGUG}$ of Chinese yew, namely tRF-T11, was found to display comparable anti-cancer effects with taxol on ovarian cancer cell A2780 and its xenograft animal model (Cao et al., 2022). It was further proved that tRF-T11 can interact with AGO2 to directly target oncogene TRPA1 and suppress its expression through the RNAi pathway in ovarian cancer cells. This study uncovers a novel role of plant-derived tRFs in regulating endogenous cancer-related genes, showing great promise for exploiting natural RNA drugs for therapeutics. There are no data, however, to indicate whether tRF-T11 from Chinese yew can transfer to another kingdom through diet, which may probably be the case given that plant tsRNAs have similar properties with plant miRNAs in some ways. Remarkably, the cross-kingdom communication of tsRNAs was observed between rhizobial and its host soybean. Rhizobial tRFs can transfer to soybean roots and hijack the host RNAi machinery to silence key host genes, thus enhancing nodulation in soybean (Ren et al., 2019).

Expression of tsRNAs under stresses

tsRNAs, initially reported as tRNA-derived stress-induced small RNAs in different organisms, can be up-regulated under a variety of stresses including oxidative stress (Thompson et al., 2008), heat (Wang et al., 2016) and drought (Hackenberg et al., 2015). Later studies demonstrated that the up-regulation of tsRNAs is not a general effect of all stresses, as only specific tsRNAs are induced under certain stress conditions. These observations further suggested that tsRNAs are not random degradation products, but potential regulators during stress responses. Accumulating data showed that the expression of plant tsRNAs can be regulated by Pi starvation (Hsieh et al., 2009; Hackenberg et al., 2013), heat stress (Wang et al., 2016), UV treatment (Cognat et al., 2017) and fungal infection (Zahra et al., 2021; Gu et al., 2022; Sun et al., 2022), indicating the possible function of these tsRNAs in plant stress responses (Table 1).

TABLE 1 List of tsRNAs studies in plant biotic and abiotic stresses.

	Plant Species	Stress	Stress responsive tsRNAs identified by small RNA-seq or Northern blot	Identification method	Confirmation method	Function	Target characterization method	Reference
Abiotic	*Arabidopsis*	Oxidative	5' $tRH^{HisGTG,GluCTC,ArgCCT,TrpCCA}$	Northern	\	\	\	Thompson et al., 2008
	Arabidopsis	Pi starvation	5' $tRF^{AspGTC,GlyTCC}$	Small RNA-seq	Northern	\	\	Hsieh et al., 2009
	Brassica rapa	Heat	5' $tRF^{Ala,Gly}$	Small RNA-seq	Northern	\	\	Wang et al., 2011
	Barley	Pi starvation	\	Small RNA-seq	\	\	\	Hackenberg et al., 2013
	Arabidopsis	Cold, Drought and Salt	5' $tRF^{AlaAGC,ArgCCT,ArgTCG,GlyTCC}$	Small RNA-seq	\	RNAi	Prediction	Loss-Morais et al., 2013
	Wheat	Heat	3' $tRF^{ThrTGT,TyrGTA}$, 5' tRF^{SerTGA}	Small RNA-seq	qPCR	\	\	Wang et al., 2016
	Barley	Drought	i' tRF^{ValAAC}	Small RNA-seq	Northern	\	\	Hackenberg et al., 2015
	Arabidopsis	Oxidative	5' tRF^{ArgTCG}, 3' tRF^{TyrGTA}	Small RNA-seq	qPCR	\	\	Alves et al., 2017
	Arabidopsis	UV	5' $tRF^{GlyGCC,GlyTCC,ProTGG,ValAAC}$	Small RNA-seq	\	\	\	Cognat et al., 2017
Biotic	Black Pepper	Phytophthora capsici	5′tRF^{AlaCGC}	Small RNA-seq	qPCR	RNAi	Prediction, 5' RLM RACE	Asha and Soniya, 2016
	Tomato	Tomato mosaic virus	\	Small RNA-seq	\	\	\	Zahra et al., 2021
	Arabidopsis	Botrytis cinerea	5'tRF^{AlaAGC}	Small RNA-seq	Northern	RNAi	Prediction, 5' RLM RACE, STTM	Gu et al., 2022
	Wheat	Fusarium head blight	\	Small RNA-seq	\	RNAi	Prediction, 5' RLM RACE	Sun et al., 2022

Under abiotic stresses

A number of stress-regulated tsRNAs have been identified in different plant species. Northern blot analyses indicated that 5' tRHs from $tRNA^{HisGTG}$, $tRNA^{ArgCCT}$ and $tRNA^{TrpGTA}$, but not $tRNA^{TyrGTA}$, were induced under oxidative stress in *Arabidopsis*. In addition, tsRNAs can also be up-regulated in yeast and human Hela cells under oxidative stress, implying that the up-regulation of tsRNAs might be a conserved response to oxidative stress (Thompson et al., 2008).

The introduction of the next-generation sequencing technology has enabled high-throughput detection and evaluation of tsRNAs expression in both model and non-model plants under different stress conditions (Zahra et al., 2021). In *Arabidopsis*, a novel peak at 19 nt was uncovered in root, but not in shoot, by deep sequencing of small RNAs responsive to Pi deficiency (Hsieh et al., 2009). Further analyses revealed that the majority of the 19 nt small RNAs are 5' tRFs originated from $tRNA^{AspGTC}$ and $tRNA^{GlyTCC}$. The percentage of these two types of 5' tRFs sharply increased under Pi deficiency, which were further verified by Northern blot. Subsequently, small RNAs were profiled in shoot of barley under Pi sufficiency and deficiency conditions. Six nuclear-derived and four chloroplast-derived tsRNAs were significantly up-regulated in Pi-deficient shoot, whereas four nuclear-derived and one chloroplast-derived tsRNAs were down-regulated under the same condition (Hackenberg et al., 2013). However, this study did not provide sequences of these Pi starvation responsive tsRNAs or confirm their expression levels through RT-qPCR or Northern blot, so no conclusion can be drawn about whether or not Pi deficiency responsive tsRNAs were conserved between barley and *Arabidopsis*. Next, drought responsive small RNAs were investigated in barley and results showed that tsRNAs had the tendency to be up-regulated under drought stress. Similarly, sequence information for these drought responsive tsRNAs was not available (Hackenberg et al., 2015).

Also, a series of tsRNAs responsive to heat stress were characterized through small RNA-seq in different plant species including *Arabidopsis*, *Brassica rapa* and wheat (Wang et al., 2011; Wang et al., 2016; Zahra et al., 2021). In *Arabidopsis*, three tsRNAs exhibited dysregulation after 0.5 hour of heat stress, while the number rose to 42 after 6 hours treatment, suggesting that heat stress induced tsRNAs generation is time-dependent (Zahra et al., 2021). In *Brassica rapa*, a variety of heat responsive chloroplast-derived tsRNAs were uncovered (Wang et al., 2011). Consistent with the deep sequencing result, Northern blot analyses indicated that a 29 nt 5' tRF^{Ala} was

declined under heat stress, while a 17 nt 5' tRF^{Ala} and a 23 nt 5' tRF^{Gly} are remarkably increased. In wheat seedlings, 292 tsRNAs were significantly increased and 41 are decreased under heat stress. Besides, most of these heat responsive tsRNAs were classified as 3' tRFs (67%), suggesting that the increased cleavage of tRNAs was preferentially induced at 3' ends under heat stress. Furthermore, the expression patterns of four tRFs derived from $tRNA^{ValCAC}$, $tRNA^{ThrTGT}$, $tRNA^{TyrGTA}$ and $tRNA^{SerTGA}$ were tested in wheat under heat stress by real-time RT-PCR. Results showed that stRNA0011d (3' tRF^{TyrGTA}) and stRNA0015 (5' tRF^{SerTGA}) were up-regulated by high temperature, which well coincides with the bioinformatics analyses (Wang et al., 2016).

The above studies unveiled that some tsRNAs only respond to specific stresses. Several studies, on the other hand, indicated that certain types of tsRNAs can be induced by different abiotic stresses. For example, in *Arabidopsis*, the 19 nt 5' tRF^{ArgCCT} can be up-regulated by both drought and oxidative stresses (Alves et al., 2017), and the salt-induced 5' tRF^{GlyGCC} also increased under UV treatment (Cognat et al., 2017). Moreover, in wheat seedlings, stRNA0011d (3' tRF^{TyrGUA}) was found to respond to heat, salt and drought stresses (Wang et al., 2016). Besides, the 19 nt 5' tRF^{ArgCCT} induced by drought in *Arabidopsis* displayed no change in rice under drought stress (Alves et al., 2017), suggesting that the stress response of the same tsRNAs may be varied in different plant species.

Under biotic stresses

The expression pattern of tsRNAs can also be altered under biotic stresses, for example, fungi or virus infection, indicating their potential role in biotic stress response (Asha and Soniya, 2016; Zahra et al., 2021; Gu et al., 2022). To elucidate the functional role of tsRNAs during *Peronosporales capsica* (*P. capsica*) infection, small RNAs in black pepper were systematically analyzed and a 23 nt 5' tRF^{AlaCGC} was found to be up-regulated in leaf and root of black pepper infected by *P. capsici* (Asha and Soniya, 2016). Fusarium head blight (FHB) that occurs in wheat is a devastating fungal disease caused by *Fusarium graminearum* (*F. graminearum*). Recently, small RNAs from the spikelets of an FHB-susceptible variety Chinese Spring (CS) and an FHB-resistant variety Sumai3 (SM) with *F. graminearum* infection and mock inoculation were analyzed, respectively. As the first report on tRFs response to FHB in wheat, different responsive patterns of tRFs to *F. graminearum* infection were observed between CS and SM. 1249 putative tRFs were identified, among which 15 tRFs were CS-specific and 12 were SM-specific. 39 tRFs were significantly increased in both wheat varieties after *F. graminearum* challenge and only nine tRFs were down-regulated. The expression patterns of tRF^{Glu}, tRF^{Lys} and tRF^{Thr}, three highly induced tRFs with significantly higher fold changes in CS than in SM, were further validated by stem-loop qRT-PCR. It is worth mentioning that RNase T2 family members were also induced by *F. graminearum* infection, to which the accumulation of tRFs were closely related (Sun et al., 2022). In *Arabidopsis*, 137 5' tsRNAs were down-regulated and 13 were up-regulated in *Botrytis cinerea* (*B. cinerea*) inoculated plants compared to mock inoculation, suggesting that *B. cinerea* infection led to the down-regulation of a significant proportion of 5' tsRNAs (Gu et al., 2022). In addition, 757 differentially expressed tsRNAs were characterized in tomato plant subjected to *Tomato Mosaic Virus* (TMV) infection, of which the majority were categorized as 15 nt tRFs (Zahra et al., 2021).

Potential roles of tsRNAs under stresses

The regulation of tsRNAs expression under various types of abiotic and biotic stresses have been well documented, while their functional roles during stress response are still poorly understood. Given that tsRNAs are similar with miRNAs regarding the length and AGO-association, several studies applied the mechanism and characteristics of target recognition for miRNAs to tsRNAs. Thus, a substantial portion of the current functional studies were based on one assumption that tsRNAs act like miRNAs. These studies can be classified into three groups: 1) Only predict tsRNAs target using miRNAs target prediction tools. 2) Further validate the cleavage site of target RNAs through 5' RLM RACE. 3) Test the association between tsRNAs-mediated RNAi and AGO system (Table 1).

Under abiotic stresses

Previous studies in mammals and yeast demonstrated that some tsRNAs induced by abiotic stress can suppress protein translation (Yamasaki et al., 2009), whereas it has not been systematically investigated and remains largely unknown in plants. In *Arabidopsis*, drought induced tRFs were substantially enriched in AGO and the targets of these tsRNAs were characterized using psRNATarget coupled with degradome analyses (Addo-Quaye et al., 2008; Dai and Zhao, 2011). Four putative targets for the drought responsive tRFs were identified, which involve in wounding response (AT3G61060.1), protein phosphorylation (AT3G05050.1), photomorphogenesis (AT2G24790.1) and unknown processes (AT3G57280.1), respectively (Loss-Morais et al., 2013). However, further experiments are needed to prove the authenticity of these tsRNAs targets.

Under biotic stresses

Research on tsRNAs under biotic stresses is relatively less than that under abiotic stresses, but the biological function of tsRNAs under biotic stresses is much better deciphered. In several studies, tsRNAs targets are predicted and further validated through 5' RLM RACE and/or AGO-IP assay. For example, to reveal the potential role of 5' tRF^{AlaCGC}, which is induced in black pepper during *Phytophthora capsici* infection, two mRNA homologs of NPR1, a key regulator of salicylic acid-dependent gene expression during systemic acquired resistance, were predicted as its putative targets. Moreover, the 5' tRF^{AlaCGC} mediated cleavage on the target mRNAs was validated by the modified 5' RLM RACE experiment (Asha and Soniya, 2016).

To reveal the role of tsRNAs induced by *F. graminearum* infection, targets of all identified tRFs were predicted in wheat. Gene ontology enrichment analyses showed that these targets play pivotal roles in stress response, energy metabolism and protein digestion. Furthermore, transcriptome analyses unveiled that the expression levels of the tRFs

targets are negatively associated with those of the corresponding tRFs. qRT-PCR was performed to validate the expression of the putative tRFs target genes and the results are highly consistent with the transcriptome data. What's more, the inhibitory effect of *F. graminearum* induced tRFs on their target genes was confirmed *in vivo* through 5' RLM RACE (Sun et al., 2022). The above analyses suggested that tRFs induced by *F. graminearum* infection might inhibit the expression of the disease resistance-related targets and consequently contribute host susceptibility to *F. graminearum*.

A recent study in *Arabidopsis* showed that the expression of *CYP71A13* (*At2g30770*), which is involved in camalexin biosynthesis and critical for plant defense against *Botrytis cinerea*, is negatively correlated with that of 5'-tsR-Ala (5' tRF^{AlaAGC}), the most abundant 5' tsRNAs identified by RtcB sRNA-seq (Gu et al., 2022). Furthermore, 5'-tsR-Ala was detected as the most abundant 5' tsRNAs in AGO1 immunoprecipitates (IPs). Northern blot analyses confirmed that 5'-tsR-Ala accumulation was significantly decreased in *ago1* mutants, wherein the expression of *CYP71A13* was increased. In addition, the 5'-tsR-Ala mediated cleavage of *CYP71A13* mRNA was proved by 5' RACE. These findings indicate that 5'-tsR-Ala may function as a miRNA and repress *CYP71A13* expression through associating with AGO1. What's more, the negative regulation by 5'-tsR-Ala of *CYP71A13* expression and anti-fungal defense was again borne out *in vivo* through knocking down 5'-tsR-Ala using the STTM method. Thus, this study unraveled the important role of a 5' tRF in regulating anti-fungal defense by modifying gene expression through direct target cleavage.

Bioinformatics tools and resources for tsRNAs study

tsRNAs identification pipelines

With the fast development and wide application of high-throughput sequencing technology, a considerable body of small RNA-seq datasets have emerged, covering different biological or pathological processes in various plant species. These publicly available data provide valuable resources for the characterization, expression analysis and functional exploration of tsRNAs. Accordingly, increasing pipelines for tsRNAs characterization are developed (Shi et al., 2018; Zahra et al., 2021; Donovan et al., 2021; Ma et al., 2021b; Rawal et al., 2022), which has greatly facilitated tsRNAs research. For example, SPORTS1.0 is a tool for annotating and profiling non-coding RNAs optimized for rRNA and tRNA derived small RNAs and available for a wide range of 68 species across bacteria, yeast, plant and animal kingdoms (Shi et al., 2018). Afterwards, an improved methodology for predicting miRNAs and tsRNAs in both model and non-model organisms were developed, which have expanded the tsRNAs study in more plants without genome reference (Rawal et al., 2022).

tsRNAs database

Several plant tsRNAs expression database have been developed, making tsRNAs expression analysis much easier for those researchers without bioinformatic background. tRex is the first on line resource dedicated to tsRNAs in *Arabidopsis thaliana* (http://combio.pl/trex). tRex collates the in-house-generated and publicly available small RNA-seq data from various tissues, ecotypes, genotypes and stress conditions, as well as provides web-based tools for tsRNAs identification, RNA structure analyses, modification predictions and target predictions (Thompson et al., 2018). Later, a plant tsRNAs database named PtRFdb was introduced based on the analyses of 1344 small RNA-seq datasets from 10 different plant species, and 5607 unique tRFs, represented by 487,765 entries, were identified (*http://www.nipgr.res.in/PtRFdb/*). Besides, the information of experimentally identified tsRNAs available in literatures from *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, *Piper nigum* and *Triticum aestivum* were collected, which can be downloaded as an excel sheet (Gupta et al., 2018). PtncRNAdb, another plant tsRNAs web resource, consists of 4,809,503 tsRNAs entries identified from ~2500 small RNA-seq libraries generated in six plants including *Arabidopsis thaliana*, *Cicer arietinum*, *Zea mays*, *Oryza sativa*, *Medicago truncatula* and *Solanum lycopersicum* (https://nipgr.ac.in/PtncRNAdb). The 'DE tncRNAs' is a feature module in PtncRNAdb for differential expression analysis of tsRNAs under various conditions. Apart from the basic information about tsRNAs, the modification, secondary structure, putative targets, interactive networks of target enrichment and related publications can also be obtained for further interpretation of their biological functions (Zahra et al., 2022). Recently, we developed a comprehensive tsRNAs database named tsRBase. tsRBase covers 20 species and 6 of them are plants, viz., *Arabidopsis thaliana*, *Glycine max*, *Oryza sativa*, *Physcomitrella patens*, *Vitis vinifera* and *Zea mays* (http://www.tsrbase.org). tsRBase not only provides differential expression analysis, but also incorporates experimentally validated targets of tsRNAs (Zuo et al., 2021).

Target prediction tools

Target identification is central for defining the biological function of tsRNAs, whereas it is unrealistic to characterize the targets for all tsRNAs experimentally and there have been few studies on the relationship between tsRNAs and mRNAs, especially in plants. Therefore, researchers have to predict the targets based on algorithms. miRNAs target prediction tools, such as psRNATarget and PsRobot, have been broadly applied to predict tsRNAs targets given that tsRNAs may also suppress gene expression through sequence complementarity (Addo-Quaye et al., 2008; Megel et al., 2019; Zahra et al., 2021). Several computation tools have also been developed specifically for predicting tsRNAs targets in mammals, including tRFTars, tRFTar and tRForest. tRFTars is the first database for tsRNAs target prediction (http://trftars.cmuzhenninglab.org:3838/tar/). First, features that influence tsRNAs targeting were screened. Then, tsRNA-mRNA pairs identified by crosslinking, ligation and sequencing of hybrids (CLASH) and covalent ligation of endogenous AGO-bound RNAs (CLEAR)-CLIP were used to select key features through a genetic algorithm (GA). Finally, support vector machine (SVM) was applied to construct tsRNAs prediction models with the selected key features (Shi et al., 2018). tRFTar is a resource for predicting tRF target gene interactions (TGIs) based on the fact that tsRNAs can be loaded onto AGO family proteins to perform

post-transcriptional regulations (http://www.rnanut.net/tRFTar/). 146 cross-linking immunoprecipitation and high-throughput sequencing (CLIP-seq) datasets were systematically reanalyzed and 920,690 TGIs between 12,102 tRFs and 5,688 target genes were identified. tRFTar enables various functions like custom searching, co-expressed TGI filtering, genome browser and TGI-based tRF functional enrichment analysis (Rawal et al., 2022). Recently, using cross-linking, Ligation, and Sequencing of Hybrids (CLASH) data as the training and testing dataset, a novel tsRNAs target prediction tool, tRForest, was developed based on the random forest machine learning algorithm (Parikh et al., 2022) (https://trforest.com). However, no specific target prediction tools are currently available for plant tsRNAs, so the development of tsRNAs target prediction tools is an urgent issue for plant tsRNAs study.

Conclusion and future perspective

With the help of improved high-throughput sequencing technologies, a large body of tsRNAs have been identified in various organisms and the numbers are still expanding (Zhu et al., 2018a; Zhu et al., 2018b; Zuo et al., 2021). tsRNAs are thought as a heterogeneous class of small RNAs because of their multitudinous sources and lengths (Zuo et al., 2021). The spatially and temporally regulated expression pattern of tsRNAs has been proposed to play important roles in plant development and stress response. However, direct and in-depth functional analyses of tsRNAs are still missing, especially in plants. Conventional methods for dissecting gene function relied much on genetic mutants. However, this approach is not feasible for the study of tsRNAs due to their small sizes, non-coding property, multiple members and overlapping sequences with tRNAs that is indispensable for protein translation and normal cellular processes. In fact, relevant technologies and achievements regarding tsRNAs study in plants still lag far behind those in animals. Antisense oligonucleotides (ASOs) are widely applied to specifically bind target tsRNAs in mammalian cells (Goodarzi et al., 2015; Kim et al., 2017), which can efficiently knock-down the abundance of corresponding tsRNAs and testify their involvement or functional role in certain physiological and pathological processes more straightforwardly, thus offering a promising alternative to therapies. In plants, there are piecemeal applications of STTM for tsRNAs block (Martinez et al., 2017; Gu et al., 2022). As is known, miRNAs are generally 21 nt in length, and the three-nucleotide bulge that prevent the cleavage of the target mimic (TM) stuck out between the 10th and 11th nucleotide of the targeted miRNAs (Yan et al., 2012). Therefore, it remains to be seen whether STTM is applicable or just as efficient for tsRNAs with other lengths. Besides, results in different organisms showed that specific tsRNAs are associated with protein translation machinery or AGO system and regulate gene expression post-transcriptionally (Ivanov et al., 2011; Maute et al., 2013). In mammals, methods to identify tsRNAs associated proteins have been applied, which allows a more comprehensive exploration of the mechanism and characteristics of tsRNAs (Keam et al., 2014; Goodarzi et al., 2015; Cho et al., 2019). Further efforts are needed to develop new methods for characterizing tsRNAs associated proteins in plants.

Another point worth noting is that traditional small RNAs cloning methods applied by most studies can only capture those with 5'-OH and 3'-Pi. Actually, a large proportion of tsRNAs generated by endoribonucleases are ended with 2',3'-cP, so they cannot be ligated to the adaptors directly (Shi et al., 2021; Gu et al., 2022). Other internal modifications embedded in tsRNAs, such as methylation, will suppress the reverse transcription and consequently impact the cloning efficiency. Recently, several studies have improved the small RNA cloning methods through removing the end and internal modifications present in small RNAs, which will greatly benefit the tsRNAs research in plants (Shi et al., 2021; Wang et al., 2021; Gu et al., 2022).

Author contributions

Conceptualization: YC and LZ. Data curation: LZ and YC. Original draft preparation: CW, WC and MA. Supervision: YC and LZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (2022YFC3501703) and National Natural Science Foundation of China (81902819 and 32201799).

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

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. *Curr. Biol.* 18 (10), 758–762. doi: 10.1016/j.cub.2008.04.042

Alves, C. S., Vicentini, R., Duarte, G. T., Pinoti, V. F., Vincentz, M., and Nogueira, F. T. (2017). Genome-wide identification and characterization of tRNA-derived RNA fragments in land plants. *Plant Mol. Biol.* 93 (1-2), 35–48. doi: 10.1007/s11103-016-0545-9

Andersen, K. L., and Collins, K. (2012). Several RNase T2 enzymes function in induced tRNA and rRNA turnover in the ciliate *Tetrahymena*. *Mol. Biol. Cell* 23 (1), 36–44. doi: 10.1091/mbc.E11-08-0689

Asha, S., and Soniya, E. V. (2016). Transfer RNA derived small RNAs targeting defense responsive genes are induced during *Phytophthora capsici* infection in black pepper (*Piper nigrum* l.). *Front. Plant Sci.* doi: 10.3389/fpls.2016.00767

Cao, K. Y., Yan, T. M., Zhang, J. Z., Chan, T. F., Li, J., Li, C., et al. (2022). A tRNA-derived fragment from Chinese yew suppresses ovarian cancer growth *via* targeting TRPA1. *Mol. Ther. Nucleic Acids* 27, 718–732. doi: 10.1016/j.omtn.2021.12.037

Chang, Y. N., Zhu, C., Jiang, J., Zhang, H., Zhu, J. K., and Duan, C. G. (2020). Epigenetic regulation in plant abiotic stress responses. *J. Integr. Plant Biol.* 62 (5), 563–580. doi: 10.1111/jipb.12901

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., et al. (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science.* 351 (6271), 397–400. doi: 10.1126/science.aad7977

Chen, Q., Zhang, F., Dong, L., Wu, H., Xu, J., Li, H., et al. (2021). SIDT1-dependent absorption in the stomach mediates host uptake of dietary and orally administered microRNAs. *Cell Res.* 31 (3), 247–258. doi: 10.1038/s41422-020-0389-3

Cho, H., Lee, W., Kim, G. W., Lee, S. H., Moon, J. S., Kim, M., et al. (2019). Regulation of La/SSB-dependent viral gene expression by pre-tRNA 3′ trailer-derived tRNA fragments. *Nucleic Acids Res.* 47 (18), 9888–9901. doi: 10.1093/nar/gkz732

Cognat, V., Morelle, G., Megel, C., Lalande, S., Molinier, J., Vincent, T., et al. (2017). The nuclear and organellar tRNA-derived RNA fragment population in *Arabidopsis thaliana* is highly dynamic. *Nucleic Acids Res.* 45, 3460–3472. doi: 10.1093/nar/gkw1122

Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., Brown, J. W., et al. (2009). Filtering of deep sequencing data reveals the existence of abundant dicer-dependent small RNAs derived from tRNAs. *RNA.* 15 (12), 2147–2160. doi: 10.1261/rna.1738409

Condon, C. (2009). "RNA Processing," in *Encyclopedia of microbiology (Third edition)*. Ed. M. Schaechter (Oxford: Academic Press), 395–408.

Dai, X., and Zhao, P. X. (2011). psRNATarget: A plant small RNA target analysis server. *Nucleic Acids Res.* 39, W155–W159. doi: 10.1093/nar/gkr319

Del Pozo-Acebo, L., Lopez de Las Hazas, M. C., Margolles, A., Davalos, A., and Garcia-Ruiz, A. (2021). Eating microRNAs: Pharmacological opportunities for cross-kingdom regulation and implications in host gene and gut microbiota modulation. *Br. J. Pharmacol.* 178 (11), 2218–2245. doi: 10.1111/bph.15421

Deshpande, R. A., and Shankar, V. (2002). Ribonucleases from T2 family. *Crit. Rev. Microbiol.* 28 (2), 79–122. doi: 10.1080/1040-840291046704

Donovan, P. D., McHale, N. M., Veno, M. T., and Prehn, J. H. M. (2021). A pipeline for the identification of tRNA and ncRNA fragments from small RNA sequencing data. *Bioinformatics* 2021:btab515. doi: 10.1093/bioinformatics/btab515

Eun, H.-M. (1996). "3-nucleases," in *Enzymology primer for recombinant DNA technology*. Ed. H.-M. Eun (San Diego: Academic Press), 145–232.

Fu, H., Feng, J., Liu, Q., Sun, F., Tie, Y., Zhu, J., et al. (2009). Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett.* 583 (2), 437–442. doi: 10.1016/j.febslet.2008.12.043

Goodarzi, H., Liu, X., Nguyen, H. C., Zhang, S., Fish, L., and Tavazoie, S. F. (2015). Endogenous tRNA-derived fragments suppress breast cancer progression *via* YBX1 displacement. *Cell.* 161 (4), 790–802. doi: 10.1016/j.cell.2015.02.053

Gu, H., Lian, B., Yuan, Y., Kong, C., Li, Y., Liu, C., et al. (2022). A 5' tRNA-ala-derived small RNA regulates anti-fungal defense in plants. *Sci. China Life Sci.* 65 (1), 1–15. doi: 10.1007/s11427-021-2017-1

Gupta, N., Singh, A., Zahra, S., and Kumar, S. (2018). PtRFdb: A database for plant transfer RNA-derived fragments. *Database (Oxford)* 2018:bay063. doi: 10.1093/database/bay063

Hackenberg, M., Gustafson, P., Langridge, P., and Shi, B. J. (2015). Differential expression of microRNAs and other small RNAs in barley between water and drought conditions. *Plant Biotechnol. J.* 13 (1), 2–13. doi: 10.1111/pbi.12220

Hackenberg, M., Huang, P. J., Huang, C. Y., Shi, B. J., Gustafson, P., and Langridge, P. (2013). A comprehensive expression profile of microRNAs and other classes of non-coding small RNAs in barley under phosphorous-deficient and -sufficient conditions. *DNA Res.* 20 (2), 109–125. doi: 10.1093/dnares/dss037

Hsieh, L. C., Lin, S. I., Shih, A. C., Chen, J. W., Lin, W. Y., Tseng, C. Y., et al. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in *Arabidopsis* by deep sequencing. *Plant Physiol.* 151 (4), 2120–2132. doi: 10.1104/pp.109.147280

Irie, M. (1999). Structure-function relationships of acid ribonucleases: Lysosomal, vacuolar, and periplasmic enzymes. *Pharmacol. Ther.* 81 (2), 77–89. doi: 10.1016/s0163-7258(98)00035-7

Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P., and Anderson, P. (2011). Angiogenin-induced tRNA fragments inhibit translation initiation. *Mol. Cell.* 243 (4), 613–623. doi: 10.1016/j.molcel.2011.06.022

Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53. doi: 10.1146/annurev.arplant.57.032905.105218

Keam, S. P., Young, P. E., McCorkindale, A. L., Dang, T. H., Clancy, J. L., Humphreys, D. T., et al. (2014). The human piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucleic Acids Res.* 42 (14), 8984–8995. doi: 10.1093/nar/gku620

Kim, H. K., Fuchs, G., Wang, S., Wei, W., Zhang, Y., Park, H., et al. (2017). A transfer-RNA-derived small RNA regulates ribosome biogenesis. *Nature.* 552 (7683), 57–62. doi: 10.1038/nature25005

Kumar, P., Anaya, J., Mudunuri, S. B., and Dutta, A. (2014). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol.* 12, 78. doi: 10.1186/s12915-014-0078-0

Lalande, S., Merret, R., Salinas-Giege, T., and Drouard, L. (2020). *Arabidopsis* tRNA-derived fragments as potential modulators of translation. *RNA Biol.* 17 (8), 1137–1148. doi: 10.1080/15476286.2020.1722514

Loss-Morais, G., Waterhouse, P. M., and Margis, R. (2013). Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biol. Direct. 12* 8, 6. doi: 10.1186/1745-6150-8-6

Luhtala, N., and Parker, R. (2010). T2 family ribonucleases: Ancient enzymes with diverse roles. *Trends Biochem. Sci.* 35 (5), 253–259. doi: 10.1016/j.tibs.2010.02.002

Lyons, S. M., Fay, M. M., and Ivanov, P. (2018). The role of RNA modifications in the regulation of tRNA cleavage. *FEBS Lett.* 592 (17), 2828–2844. doi: 10.1002/1873-3468.13205

MacIntosh, G. C. (2011). "RNase T2 family: Enzymatic properties, functional diversity, and evolution of ancient ribonucleases," in *Ribonucleases*. Ed. A. W. Nicholson (Berlin Heidelberg: Springer), 89–114.

MacIntosh, G. C., and Castandet, B. (2020). Organellar and secretory ribonucleases: Major players in plant RNA homeostasis. *Plant Physiol.* 183 (4), 1438–1452. doi: 10.1104/pp.20.00076

MacIntosh, G. C., Hillwig, M. S., Meyer, A., and Flagel, L. (2010). RNase T2 genes from rice and the evolution of secretory ribonucleases in plants. *Mol. Genet. Genomics* 283 (4), 381–396. doi: 10.1007/s00438-010-0524-9

Ma, X., Liu, C., and Cao, X. (2021a). Plant transfer RNA-derived fragments: Biogenesis and functions. *J. Integr. Plant Biol.* 63 (8), 1399–1409. doi: 10.1111/jipb.13143

Ma, X., Liu, C., Kong, X., Liu, J., Zhang, S., Liang, S., et al. (2021b). Extensive profiling of the expressions of tRNAs and tRNA-derived fragments (tRFs) reveals the complexities of tRNA and tRF populations in plants. *Sci. China Life Sci.* 64 (4), 495–511. doi: 10.1007/s11427-020-1891-8

Martinez, G., Choudury, S. G., and Slotkin, R. K. (2017). tRNA-derived small RNAs target transposable element transcripts. *Nucleic Acids Res.* 45 (9), 5142–5152. doi: 10.1093/nar/gkx103

Matos, R. G., Pobre, V., Reis, F. P., Andrade, J. M., and Arraiano, C. M. (2011). "Structure and degradation mechanisms of 3′ to 5′ exoribonucleases," in *Ribonucleases, Nucleic acids and molecular biology*. Ed. A. W. Nicholson (Berlin, Heidelberg: Springer). doi: 10.1007/978-3-642-21078-5_8

Maute, R. L., Schneider, C., Sumazin, P., Holmes, A., Califano, A., Basso, K., et al. (2013). tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in b cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* 110 (4), 1404–1409. doi: 10.1073/pnas.1206761110

Megel, C., Hummel, G., Lalande, S., Ubrig, E., Cognat, V., Morelle, G., et al. (2019). Plant RNases T2, but not dicer-like proteins, are major players of tRNA-derived fragments biogenesis. *Nucleic Acids Res.* 47 (2), 941–952. doi: 10.1093/nar/gky1156

Parikh, R., Wilson, B., Marrah, L., Su, Z., Saha, S., Kumar, P., et al. (2022). tRForest: a novel random forest-based algorithm for tRNA-derived fragment target prediction. *NAR Genom Bioinform.* 4 (2), lqac037. doi: 10.1093/nargab/lqac037

Phizicky, E. M., and Hopper, A. K. (2010). tRNA biology charges to the front. *Genes Dev.* 24 (17), 1832–1860. doi: 10.1101/gad.1956510

Rawal, H. C., Ali, S., and Mondal, T. K. (2022). miRPreM and tiRPreM: Improved methodologies for the prediction of miRNAs and tRNA-induced small non-coding RNAs for model and non-model organisms. *Brief Bioinform.* 23 (1):bbab448. doi: 10.1093/bib/bbab448

Ren, B., Wang, X., Duan, J., and Ma, J. (2019). Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. *Science* 365 (6456), 919–922. doi: 10.1126/science.aav8907

Shi, J., Ko, E. A., Sanders, K. M., Chen, Q., and Zhou, T. (2018). SPORTS1.0: A tool for annotating and profiling non-coding RNAs optimized for rRNA- and tRNA-derived small RNAs. *Genomics Proteomics Bioinf.* 16 (2), 144–151. doi: 10.1016/j.gpb.2018.04.004

Shi, J., Zhang, Y., Tan, D., Zhang, X., Yan, M., Zhang, Y., et al. (2021). PANDORA-seq expands the repertoire of regulatory small RNAs by overcoming RNA modifications. *Nat. Cell Biol.* 23 (4), 424–436. doi: 10.1038/s41556-021-00652-7

Shi, J., Zhang, Y., Zhou, T., and Chen, Q. (2019). tsRNAs: The Swiss army knife for translational regulation. *Trends Biochem. Sci.* 44 (3), 185–189. doi: 10.1016/j.tibs.2018.09.007

Sun, Z., Hu, Y., Zhou, Y., Jiang, N., Hu, S., Li, L., et al. (2022). tRNA-derived fragments from wheat are potentially involved in susceptibility to fusarium head blight. *BMC Plant Biol.* 22 (1), 3. doi: 10.1186/s12870-021-03393-9

Thompson, D. M., Lu, C., Green, P. J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA.* 14 (10), 2095–2103. doi: 10.1261/rna.1232808

Thompson, D. M., and Parker, R. (2009). The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in *Saccharomyces cerevisiae*. *J. Cell Biol.* 185 (1), 43–50. doi: 10.1083/jcb.200811119

Thompson, A., Zielezinski, A., Plewka, P., Szymanski, M., Nuc, P., Szweykowska-Kulinska, Z., et al. (2018). tRex: a web portal for exploration of tRNA-derived fragments in *Arabidopsis thaliana*. *Plant Cell Physiol.* 59 (1):e1. doi: 10.1093/pcp/pcx173

Wang, H., Huang, R., Li, L., Zhu, J., Li, Z., Peng, C., et al. (2021). CPA-Seq reveals small ncRNAs with methylated nucleosides and diverse termini. *Cell Discov.* 7 (1), 25. doi: 10.1038/s41421-021-00265-2

Wang, Y., Li, H., Sun, Q., and Yao, Y. (2016). Characterization of small RNAs derived from tRNAs, rRNAs and snoRNAs and their response to heat stress in wheat seedlings. *PloS One* 11 (3):e0150933. doi: 10.1371/journal.pone.0150933

Wang, L., Yu, X., Wang, H., Lu, Y. Z., de Ruiter, M., Prins, M., et al. (2011). A novel class of heat-responsive small RNAs derived from the chloroplast genome of Chinese cabbage (*Brassica rapa*). *BMC Genomics* 12, 289. doi: 10.1186/1471-2164-12-289

Yamasaki, S., Ivanov, P., Hu, G. F., and Anderson, P. (2009). Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J. Cell Biol.* 185 (1), 35–42. doi: 10.1083/jcb.200811106

Yan, J., Gu, Y., Jia, X., Kang, W., Pan, S., Tang, X., et al. (2012). Effective small RNA destruction by the expression of a short tandem target mimic in arabidopsis. *Plant Cell.* 24 (2), 415–427. doi: 10.1105/tpc.111.094144

Zahra, S., Bhardwaj, R., Sharma, S., Singh, A., and Kumar, S. (2022). PtncRNAdb: plant transfer RNA-derived non-coding RNAs (tncRNAs) database. *3 Biotech.* 12 (5), 105. doi: 10.1007/s13205-022-03174-7

Zahra, S., Singh, A., Poddar, N., and Kumar, S. (2021). Transfer RNA-derived non-coding RNAs (tncRNAs): Hidden regulation of plants' transcriptional regulatory circuits. *Comput. Struct. Biotechnol. J.* 19, 5278–5291. doi: 10.1016/j.csbj.2021.09.021

Zhang, X., He, X., Liu, C., Liu, J., Hu, Q., Pan, T., et al. (2016). IL-4 inhibits the biogenesis of an epigenetically suppressive PIWI-interacting RNA to upregulate CD1a molecules on monocytes/dendritic cells. *J. Immunol.* 196 (4), 1591–1603. doi: 10.4049/jimmunol.1500805

Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., et al. (2012). Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res.* 22 (1), 107–126. doi: 10.1038/cr.2011.158

Zhang, S., Sun, L., and Kragler, F. (2009). The phloem-delivered RNA pool contains small noncoding RNAs and interferes with translation. *Plant Physiol.* 150 (1), 378–387. doi: 10.1104/pp.108.134767

Zhu, L., Li, J., Gong, Y., Wu, Q., Tan, S., Sun, D., et al. (2019). Exosomal tRNA-derived small RNA as a promising biomarker for cancer diagnosis. *Mol. Cancer.* 18 (1), 74. doi: 10.1186/s12943-019-1000-8

Zhu, L., Liu, X., Pu, W., and Peng, Y. (2018a). tRNA-derived small non-coding RNAs in human disease. *Cancer Lett.* 419, 1–7. doi: 10.1016/j.canlet.2018.01.015

Zhu, L., Ow, D. W., and Dong, Z. (2018b). Transfer RNA-derived small RNAs in plants. *Sci. China Life Sci.* 61 (2), 155–161. doi: 10.1007/s11427-017-9167-5

Zuo, Y., Zhu, L., Guo, Z., Liu, W., Zhang, J., Zeng, Z., et al. (2021). tsRBase: a comprehensive database for expression and function of tsRNAs in multiple species. *Nucleic Acids Res.* 49 (D1), D1038–D1D45. doi: 10.1093/nar/gkaa888

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 February 2023
DOI 10.3389/fpls.2023.1134352

OPEN ACCESS

EDITED BY
Gao Jihai,
Chengdu University of Traditional Chinese Medicine, China

REVIEWED BY
Qi Tang,
Hunan Agricultural University, China
Chang-Jiang-Sheng Lai,
China Academy of Chinese Medical Sciences, China
Shuncang Zhang,
Yangzhou University, China

*CORRESPONDENCE
Rufeng Wang
wrffrw0801@shutcm.edu.cn
Xiaolin Yang
xiaolinysn@126.com

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 30 December 2022
ACCEPTED 31 January 2023
PUBLISHED 10 February 2023

CITATION
Pan J, Huang C, Yao W, Niu T, Yang X and Wang R (2023) Full-length transcriptome, proteomics and metabolite analysis reveal candidate genes involved triterpenoid saponin biosynthesis in *Dipsacus asperoides*. *Front. Plant Sci.* 14:1134352. doi: 10.3389/fpls.2023.1134352

Full-length transcriptome, proteomics and metabolite analysis reveal candidate genes involved triterpenoid saponin biosynthesis in *Dipsacus asperoides*

Jie Pan[1,2], Chaokang Huang[1,2], Weilin Yao[1,2], Tengfei Niu[1,2], Xiaolin Yang[1,2,3]* and Rufeng Wang[1,2,3,4]*

[1]Institute of Chinese Materia Medica, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]The SATCM Key Laboratory for New Resources and Quality Evaluation of Chinese Medicines, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Shanghai R&D Center for Standardization of Chinese Medicines, Shanghai, China, [4]The MOE Key Laboratory for Standardization of Chinese Medicines, Shanghai University of Traditional Chinese Medicine, Shanghai, China

Dipsacus asperoides is a traditional medicinal herb widely used in inflammation and fracture in Asia. Triterpenoid saponins from *D. asperoides* are the main composition with pharmacological activity. However, the biosynthesis pathway of triterpenoid saponins has not been completely resolved in *D. asperoides*. Here, the types and contents of triterpenoid saponins were discovered with different distributions in five tissues (root, leaf, flower, stem, and fibrous root tissue) from *D. asperoides* by UPLC-Q-TOF-MS analysis. The discrepancy between five tissues in *D. asperoides* at the transcriptional level was studied by combining single-molecule real-time sequencing and next- generation sequencing. Meanwhile, key genes involved in the biosynthesis of saponin were further verified by proteomics. In MEP and MVA pathways, 48 differentially expressed genes were identified through co-expression analysis of transcriptome and saponin contents, including two isopentenyl pyrophosphate isomerase and two 2,3-oxidosqualene β-amyrin cyclase, etc. In the analysis of WGCNA, 6 cytochrome P450s and 24 UDP-glycosyltransferases related to the biosynthesis of triterpenoid saponins were discovered with high transcriptome expression. This study will provide profound insights to demonstrate essential genes in the biosynthesis pathway of saponins in *D. asperoides* and support for the biosynthetic of natural active ingredients in the future.

KEYWORDS

***Dipsacus asperoides*, saponin distribution, biosynthesis, transcriptome, proteomics**

1 Introduction

Dipsacus asperoides belonging to the Dipsacaceae family is a kind of widely applied traditional Chinese medicinal crops (Wan et al., 2021). The dried root of *D. asperoides* known as "Xu Duan" is frequently prescribed for the treatments of fracture and impotence due to its beneficial health properties. Over the last decade, the wild resource of *D. asperoides* was over-

exploited and the demand for this medicinal plant has been progressively increasing (Wang et al., 2016). Therefore, the researches of botany, cultivation, molecular biology, and metabolic engineering in *D. asperoides* are indispensable for the effective production of bioactive secondary metabolites in natural medicinal plants or crops, which predominantly count on the elucidation of biosynthesis pathway in these secondary metabolites. Up to now, large amounts of research has been conducted and evaluated on chemical compositions (Yu et al., 2019) and pharmacological (Yu et al., 2012) activities of *D. asperoides*. Modern pharmacological research has verified that the saponin extract of *D. asperoides* had numerous significant biological activities, such as anti-inflammatory (Li et al., 2013; Lu et al., 2020), anti-oxidatant (Tran et al., 2008), Alzheimer's disease inhibitory (Ji et al., 2012; Yu et al., 2012; Wang et al., 2018), antifungal (Choi et al., 2017), anti-apoptotic (Lu et al., 2020), and anti-cancer (Jeong et al., 2008), etc. The studies of chemical analysis and isolation on *D. asperoides* showed that its chemical compositions mainly consisted of triterpenoid saponins (Jung et al., 1993), iridoid glycosides (Sun et al., 2015) and alkaloids (Li et al., 2013), etc. Triterpenoid saponins including asperosaponin VI, hederagenin and alpha-Hederin are the principal bioactive components of *D. asperoides* (Liu et al., 2011; Wang et al., 2020). Previous research showed that the content of asperosaponin VI was dissimilar in different tissues of *D. asperoides*, as well as in various habitats (Jin et al., 2020). Nevertheless, the content distributions of saponins in different tissues of *D. asperoides* have not been investigated.

Since triterpenoid saponins are the principal active components in *D. asperoides*, it is vital for revealing candidate genes involved in the biosynthetic pathways of triterpenoid saponins. Saponins are originally derived from isopentenyl diphosphate (IPP) in the cytosol mevalonic acid (MVA) pathway and plastid methylerythritol phosphate (MEP) pathway (Thimmappa et al., 2014). Two molecules of IPP and one molecule of dimethylallyl diphosphate (DMAPP) are catalyzed to form farnesyl pyrophosphate (FPP) through geranyl pyrophosphate synthase (GPS) and farnesyl pyrophosphate synthase (FPS) (Vranova et al., 2013). Then 2,3-oxidosqualene is derived from two molecules of FPP *via* squalene synthase (SS) and squalene epoxidase (SE), whereafter diverse oxidosqualene cyclase (OSC) enzymes catalyze 2,3-oxidosqualene to a series of triterpene backbones, such as β-amyrin, dammarane and phytosterol (Cheng et al., 2020). β-Amyrin and other products are further oxidated and hydroxylated by cytochromep450 (CYPs) monooxygenases and glycosylated *via* UDP- glycosyltransferases (UGTs) at the C-3 or C-28 positions to generate various triterpenoid saponins (Seki et al., 2015). Recently, researches have been certified the pivotal function of different enzymes in the synthesis of the triterpene skeleton (Wang Y. et al., 2022; Wang Z. L. et al., 2022). However, the genes related to the modification of saponins in *D. asperoides* remain to be comprehensively illuminated.

Currently, metabolomics and transcriptomics have been extensively performed to clarify the correlation of components and key genes involving saponin biosynthesis. Saponins as paramount pharmacological chemicals have various distribution patterns in medicinal plants of different tissues (Jia et al., 2013). In this study, ultra-performance liquid chromatography-quadrupole time-of-fight mass spectrometry (UPLC-Q-TOF-MS) was applied to explore the contents of triterpenoid saponins and distribution patterns of saponin in five different tissues from *D. asperoides*, including roots, leaves, flowers, stem, and fibrous roots. Meanwhile, single-molecule real-time (SMRT) sequencing and next-generation sequencing (NGS) techniques were jointly used to obtain an outright transcriptome dataset of *D. asperoides*. By analyzing the relationship of different triterpenoid saponins and sequencing data in five tissues, some tissue-specific patterns of specific genes and saponins were discovered in *D. asperoides*. Then the weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) was further applied to identify critical hub genes attached to the biosynthesis of triterpenoid saponins. Moreover, proteomics technology was used to study the discrepancies in protein levels of three *D. asperoides* tissues comprising roots, leaves, and flowers. Finally, the candidate genes involved triterpenoid saponin biosynthesis in *D. asperoides* were revealed by multiple omics strategy. This study will provide profound insights to get essential genes in saponin biosynthesis pathway and lay a foundation for biosynthetic natural ingredients in *D. asperoides*.

2 Materials and methods

2.1 Plant materials

The fresh samples of *D. asperoides* were collected from Baoshan, Yunnan, China (25°06′43″N, 99°09′42″E). The fresh specimens were carefully cleaned and immediately separated into five tissues (root, leaf, flower, stem, and fibrous root) to store for the following experiments.

2.2 Chemical compositional analysis

2.2.1 Sample preparation

Each tissue sample was dried in an oven at 50°C, and 500 mg powder of each sample was added to 25 mL of 70% methanol. Ultrasonication was conducted for 1 h at room temperature (100 W, 40 kHz). Then, all prepared samples were centrifuged at 14,000 rpm for 30 min, and corresponding supernatants were used for analysis by UPLC-Q-TOF-MS.

2.2.2 Standard preparation

Standards (loganin, sweroside, loganic acid, hederagenin, alpha-Hederin, dipsacoside B, asperosaponin VI and hederacoside C) were purchased from Chengdu MUST Biotechnology Co (Chengdu, China). The purity of each standard substance was above 98%. Pre-weighed standards were dissolved in methanol at the final concentration of 0.1 mg/mL, and all standard solutions were stored at 4°C.

2.2.3 UPLC-Q-TOF-MS analysis

The contents of chemicals were determined as described in the literature (Tao et al., 2019) with minor modifications. The analytical facility contained an UPLC system (Shimadzu, Japan) and a Q-TOF 5600^{+} mass spectrometer provided with Turbo V sources (AB sciex, USA). The chromatographic conditions were set as below: Waters ACQUITY UPLC HSS T3 (2.1 mm × 100 mm, 1.8 μm); sample injection volume, 5 μL; temperature of column oven, 35°C; flowrate, 0.4 mL/min; mobile phases, water with 0.1% formic acid (solvent A) and

acetonitrile (solvent B). A gradient programmer was employed as follows: 5% B (0 - 2 min), 5-30% B (2.0 - 8.0 min), 30-45% B (8.0 - 9.0 min), 45-60% B (9.0 - 10.0 min), 60-80% B (10.0 - 16.0 min), 80-95% B (16.0 - 21.0 min), 95-100% B (21.0 - 22.0 min). The operating parameters for Q-TOF-MS were set as below: full-scan data acquisition was performed from m/z 100 to 1,500 in the negative mode; ion spray voltage, - 4.5 kV; collision energy, - 35 eV.

2.3 Transcriptomic analysis

2.3.1 RNA preparation, illumina library preparation and sequencing

The tissues (root, leaf, flower, stem, and fibrous root) of *D. asperoides* were used for illumina library preparation and sequencing. In brief, the total RNA was extracted from each tissue using TRIzol® Reagent (Magen). Each total RNA sample was then used for NGS analysis, while equivalent amounts of RNA from roots, leaves, flowers, stems, and fibrous roots were mixed for SMRT analysis.

The first-strand cDNAs were synthesized with random hexamer primers and Reverse Transcriptase (RNase H) using mRNA fragments as templates, followed by second-strand cDNA synthesis using DNA polymerase I, RNAseH, buffer, and dNTPs. Adaptor-ligated cDNA was used for PCR amplification. PCR products were purified (AMPure XP system) and library quality was assessed on an Agilent Bioanalyzer 4150 system. Finally, sequencing was performed with an Illumina Novaseq 6000/MGISEQ-T7 instrument. Raw data obtained from the transcriptome sequencing by removing the adapter sequence and filtering out low-quality reads to gain high-quality clean reads was used for subsequent analysis. Clean data were used to do *de novo* assembly with Trinity. The assembled transcriptome sequences were compared with five databases (NR, SwissProt, Pfam, GO and KEGG databases) to obtain the annotation information in each database.

2.3.2 SMRT library construction, sequencing, and data analysis

The RNA extracted from five tissue types was mixed into one specimen to establish SMRT library. Full-length cDNA was produced using a SMARTer PCR cDNA Synthesis Kit (Clontech), and isoform sequencing (Iso-Seq) libraries were constructed using a SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). Sequencing was performed on a PacBio Sequel II instrument with a Sequel™ Sequencing Kit 2.0 (Pacific Biosciences). Functional annotations were conducted using BLAST (version 2.2.26) against different protein and nucleotide databases including the NR database, Swissprot database, Gene Ontology (GO) database, eggNOG (Evolutionary Genealogy of Genes: Non-super-vised Orthologous Groups) database, and KEGG (Kyoto Encyclopedia of Genes and Genomics) database. Principal component analysis (PCA) is an important analytical method that analyzes the multiple sets of data and interprets it with fewer principal components, while visualizing differences and interpreting most characteristics of the original data (Wang et al., 2012). Heatmap was plotted by an online platform for data analysis and visualization (https://www.bioinformatics.com.cn).

2.4 Label-free proteomic analysis

2.4.1 Protein extraction and LC-MS/MS analysis

SDT (4% SDS, 100 mM Tris-HCl, 1 mM DTT, pH 7.6) buffer was used for sample analysis and protein extraction. The amount of protein was quantified with the BCA Protein Assay Kit (Bio-Rad, USA). Protein digestion was performed according to filter-aided sample preparation (FASP) procedure described by Matthias Mann. The digest peptides of each sample were desalted on C18 Cartridges (Empore™ SPE Cartridges C18 (standard density), bed I.D. 7 mm, volume 3 mL, Sigma), concentrated by vacuum centrifugation and reconstituted in 40 µl of 0.1% (v/v) formic acid. The proteins were separated on 12.5% SDS-PAGE gel (constant current 14 mA, 90 min). LC-MS/MS analysis was performed on a Q Exactive mass spectrometer (Thermo Scientific) that was coupled to Easy nLC (Thermo Fisher Scientific) for 120 min. The peptides of each sample were re-separated using a reverse phase trap column (Thermo Scientific Acclaim PepMap 100, 100 µm × 2 cm, nanoViper C18), with the C18-reversed phase analytical column in buffer A (0.1% Formic acid) and separated with a linear gradient of buffer B (84% acetonitrile and 0.1% formic acid) at a flow rate of 300 µL/min controlled by IntelliFlow technology. The mass spectrometer was operated in positive ion mode and the data was determined as described in the literature (Chen et al., 2020).

2.4.2 Protein identification, quantification and bioinformatic analysis

The MS raw data for each sample were combined and searched using the Max Quant 1.5.3.17 software for identification and quantitation analysis (Chen et al., 2020). The transcriptome of *D. asperoides* database was used for protein identification, and the database pattern was reversed. The protein sequences of the selected differentially expressed proteins were locally searched using the NCBI BLAST+ client software and InterProScan to find homologue sequences, then terms were mapped and sequences were annotated using Blast2GO. The GO annotation results were plotted by R scripts. Following annotation steps, proteins were blasted against KEGG database to retrieve orthology identifications and were subsequently mapped to pathways. Enrichment analysis was applied based on the Fisher' exact test, considering the whole quantified proteins as the background dataset. Benjamini-Hochberg correction for multiple testing was further applied to adjust derived *p*-values. And only functional categories and pathways with *p*-values under a threshold of 0.05 were considered significant. Data are available *via* ProteomeXchange with identifier PXD038580.

2.5 Gene co-expression network analysis

The WGCNA V1.41-1 R package was applied to conduct co-expression and module analyses (Langfelder and Horvath, 2008).

3 Results and discussion

3.1 Relative quantification assessment of differential compounds in *D. asperoides*

D. asperoides is a perennial plant commonly used as a traditional Chinese medicinal crop and mainly grows in the southern regions of China, such as Yunnan and Hunan Provinces (Yu et al., 2019). *D. asperoides* has been testified with pharmacological benefits for the treatments of a wide range of diseases, such as anti-inflammatory, anti-oxidatant, analgesic and anti-osteoporosis (Hung et al., 2006). To disclose the chemical compositions of *D. asperoides*, UPLC-Q-TOF-MS was used to investigate the distinct metabolites in aerial (leaves, flowers, and stems) and underground sections (roots and fibrous roots) (Figure 1A). As expected, eight components exhibit significant differences among these tissues (Figure 1B), including loganin, sweroside, loganic acid, hederagenin, alpha-Hederin, dipsacoside B, asperosaponin VI and hederacoside C (Figure 1C). Saponins presented great contents in the root of *D. asperoides*, including but not limited to hederacoside C and asperosaponin VI. Hederagenin was more abundant in fibrous roots than in other tissues, and asperosaponin VI was more abundant in roots. Hederacoside C was only detected in roots and leaves. Alpha-hederin and dipsacoside B were highly abundant in flowers and leaves, respectively. Notably, alpha-hederin, dipsacoside B and hederagenin have huge contents in flowers, leaves, fibrous roots, respectively. The study also discovered some non-saponins such as loganin and sweroside presented high contents in the root of *D. asperoides* compared to other tissues, but loganic acid was more abundant in stems. Through principal component analysis, it was found that the significant differences between root and stem tissues and other tissues (Figure 1D). However, only little differences were shown among fibrous root, flower and leave tissues. The analysis indicated that the relative content of compounds in the root was significantly different from that in the flower, which would be conducive to the further analysis of the relationship between the different compounds and differentially genes in the two tissues. The above results indicated the structure-specific and tissue-specific dependent patterns of saponins in *D. asperoides*.

3.2 Transcriptomic analysis and annotation

It is more accessible to explore metabolic processes of triterpenoid saponins in plants through the analysis of changes in compounds combined with functional genetics. NGS is capable of sequencing dozens or millions of DNA molecules synchronously and is used to analyze transcriptomes to get quantitative levels of gene expression. Nevertheless, the sequencing quality is relevant to the reading length and the synergy of gene cluster replication (Xu et al., 2020). SMRT sequencing can avoid the limitations of short-read sequences to obtain more long read length (Zhong et al., 2020). In this work, NGS and SMRT techniques were combinedly used to precisely assemble a comprehensive transcriptome of *D. asperoides*. The full-length transcriptome of *D. asperoides* was obtained using PacBio SMRT sequencing. The SMRT sequencing and next-generation

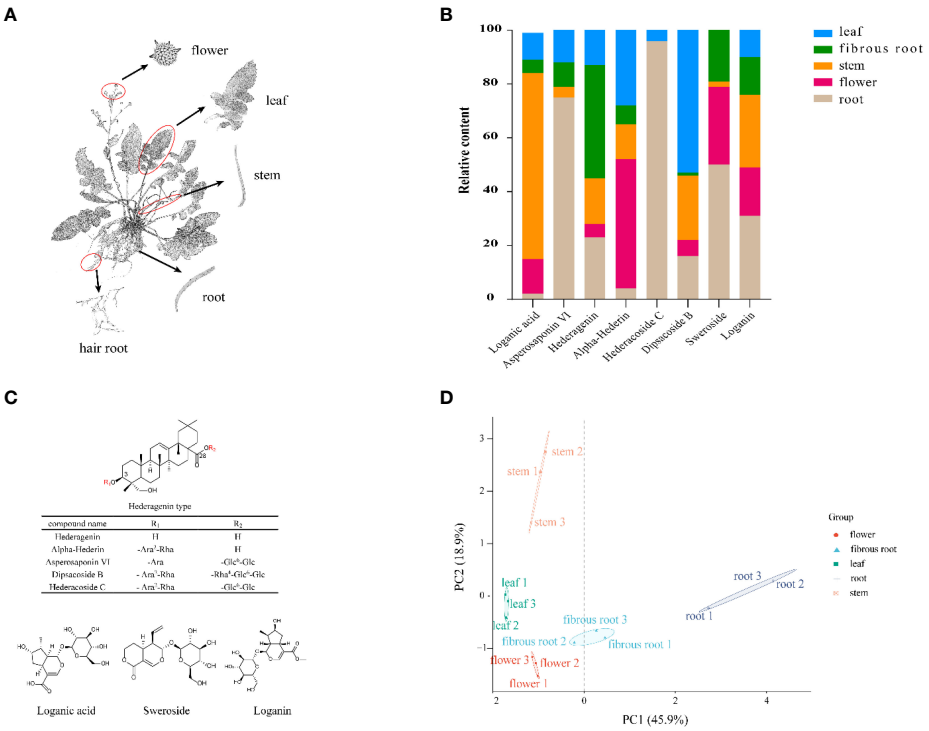


FIGURE 1
Structural, chemometric analyses and compositional variations in five tissues by UPLC-Q-TOF-MS in *D. asperoides*. **(A)** Five tissues of *D. asperoides* were analyzed in this study: root, leaf, flower, stem, and fibrous root. **(B)** Compositional variations in five tissues of *D. asperoides*. **(C)** Chemical structures of compounds isolated from *D. asperoides*. Glc, glucopyranosyl; Ara, arabinopyranosyl; Rha rhamnopyranosyl. **(D)** PCA analysis of relative quantification of differential compounds in five tissues.

sequencing (NGS) were concurrently combined to get more accurate transcriptomic database. First, all RNA specimens were sequenced by Illumina Novaseq 6000/MGISEQ-T7 instrument, generating 97.45 GB clean reads and Q30 up to 92.33% (Supplementary Table 1). Subsequently, 460,177 reads of insert were gained by SMRT sequencing, comprising a total of 420,803 full-length non-chimeric reads that incorporated 5'/3'-primers and a poly-(A) tail, along with 38,200 non-full length reads. To get high-quality isoforms with accuracy greater than 99%, iterative clustering for error correction was applied for predicting consensus isoforms, where the redundant sequences were clustered together to obtain a new consistency sequence, and then the non-full-length sequences are compared with the consistency sequence by quiver program. In total, 47,323 consensus isoforms were obtained, including 47246 high-quality (HQ) and 77 low-quality (LQ) transcripts. The clean Illumina reads were used to correct all SMRT reads to reduce high subread error rates, and the CD-HIT software was used to cluster the redundant sequences, obtaining 19526 unigene, with a mean length of 1961 bp, N50 of 2180 bp and GC content of 41%. To obtain a full-scale annotation of *D. asperoides* transcriptome, all full-length transcripts were annotated through NR, GO, SwissProt, KEGG, and Pfam databases (Supplementary Figure 1). Based on GO annotation, transcripts were sorted into the biological processes (BP), cellular component (CC), and molecular function (MF) (Supplementary Figure 2). A total number of 10383 transcripts were annotated in the KEGG database and classified into five main categories as follows: cellular processes (1069), environmental information processing (1023), genetic information processing (2104), metabolism (4436) and organismal Systems (1751) (Supplementary Figure 2). Notably, the "metabolic" pathways include the metabolism of terpenoids and polyketides (208), biosynthesis of other secondary metabolites (232) and carbohydrate metabolism (916). Furthermore, in the GO and KEGG enrichment analysis of differentially upregulated genes in roots and flowers (Supplementary Figure 3), it was shown that 486 transcripts were found in the "biosynthesis of secondary metabolites", which would contribute to revealing the biosynthesis pathways of saponins in *D. asperoides* in the future. The transcriptome data were deposited in NCBI with accession number PRJNA889678. The high-quality full-length transcriptome of *D. asperoides* offers much more information to reveal candidate genes involved triterpenoid saponins biosynthesis than other reports.

3.3 Tissue-specific dependent patterns of saponin-related genes in *D. asperoides*

To integrally analyze gene expression patterns in different *D. asperoides* tissue samples, PCA and Venn diagrams were established by processing transcriptome data. It was shown that there were significant differences among these five tissue samples (Figure 2B) with PC1, PC2, and PC3 interpretations varied by 20.17%, 15.28%, and 12.46%, respectively. In Venn diagram, 59881 transcripts were expressed in all five tissues, and 6245, 11575, 6559, 38744, and 22973 transcripts were particularly expressed in roots, flowers, stems, fibrous roots, and leaves, respectively (Figure 2A). Differentially expressed genes (DEGs) were further identified by comparing gene expression levels among samples, using coefficients calculated from log2 (fold change) and *p*-values for each transcript. Finally, 3575, 3696, and 4596 DEGs were found by comparing roots, stems and fibrous roots to flowers, respectively (Table 1). Different specific variations of triterpenoid saponins in *D. asperoides* are related to the expression of biosynthetic key genes. In MVA and MEP pathways, 2,3-oxidosqualene is the precursor to biosynthesis of triterpenoid saponins (Figure 3). SS and SE were responsible for the critical step of terpenoid carbocyclic skeleton compounds and intermediates biosynthesis (Xu et al., 2004). In addition, CYPs and UGTs were both significant to the diversification of triterpenoid saponin structures (Cheng et al., 2020). Beta-amyrin cyclase (β-AS) can cyclize 2,3-oxidosqualene to form β-amyrin. β-amyrin can be catalyzed to hederagenin through two CYPs, and hederagenin was further glycosylated to generate diverse saponins by UGTs. It was reported that CYP716A94 as a β-amyrin 28-oxidase could catalyze β-amyrin to oleanolic acid and CYP72A68 was essential to produce hederagenin through hydroxylation of C-23 in oleanolic acid (Han

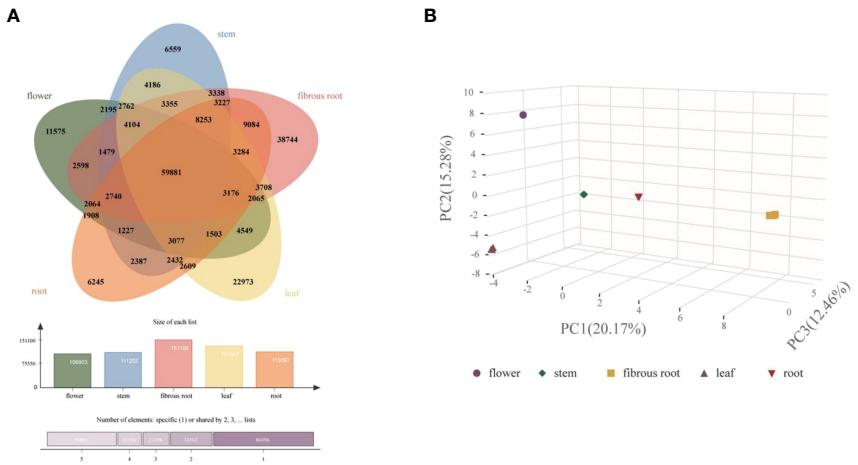


FIGURE 2
Clustering analysis of gene expression in five tissues of *D. asperoides*. **(A)** A Venn diagram comparing gene expression between different tissues. **(B)** PCA of gene expression in five tissues.

TABLE 1 The up-regulated and down-regulated DEGs in different tissues of *D. asperoides* (*p*-values ≤ 0.05 and fold change ≥ 2).

Item	Up-Regulated	Down-Regulated	Total
root -VS- flower.DEseq2	1365	2210	3575
root -VS- stem.DEseq2	548	831	1379
root -VS- fibrous root.DEseq2	395	1325	1720
root -VS- leaf.DEseq2	1117	1236	2353
stem -VS- flower.DEseq2	1547	2149	3696
stem -VS- fibrous root.DEseq2	511	964	1475
stem -VS- leaf.DEseq2	631	1660	2291
leaf -VS- flower.DEseq2	1319	2008	3327
leaf -VS- fibrous root.DEseq2	877	1393	2270
flower -VS- fibrous root.DEseq2	2297	2299	4596

et al., 2018; Tzin et al., 2019). UGT71G1 and UGT73K1 could catalyze the glycosylation of C-28 or C-3 hydroxyl group in hederagenin to produce hederagenin 3-*O*-glucoside or 28-*O*-glucoside (Achnine et al., 2005). Up to now, only a few CYPs and UGTs in *D. asperoides* have been functionally identified.

3.4 Identification of transcripts potentially involved in saponin biosynthesis

Transcripts related to upstream or downstream genes of saponin biosynthesis were found by analyzing the transcriptome of five tissue samples (roots, leaves, flowers, stems, fibrous roots). As shown in PCA (Figure 2B), the flower group was more differentiated from other groups, while the root group was clustered at the center. A total of 48 DEGs involved in MEP and MVA pathways were identified (Supplementary Table 2), and 4 (8.33%), 6 (12.50%), 7 (14.59%), 13 (27.08%), and 18 (37.50%) genes had prominent expression levels in flowers, roots, stems, fibrous roots, and leaves tissues, respectively. For instance, transcript-39509 and transcript-40019 (isopentenyl pyrophosphate isomerases, IDI) were expressed at higher levels in the leaf than in other tissues. Transcript-8753 (1-deoxy-D-xylulose-5-phosphate synthase, DXS) and transcript-23667 (GPS) showed higher expression levels in root and flower than other tissues, respectively. Transcript-9193 (hydroxymethylglutaryl-CoA reductase, HMGR), transcript-31058 (mevalonate kinase, MVK), transcript-31919 (4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, CMK) and TR4556_c1_g2 (2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, MCS) were supremely expressed in stem tissues. Transcript-18826 (hydroxymethylglutaryl-CoA synthase, HMGS), transcript-25812 (4-hydroxy-3-methylbut-2-enyl diphosphate reductase, HDR) and transcript-25516 (SS) showed high expression levels in fibrous root tissues. Transcript-14566, transcript-16847, transcript-5418 and transcript-10764 as SE also showed high expression levels in fibrous root tissues, as well as transcript-7961 and transcript-7223 (β-AS). In leaves tissues, two acetoacetyl-CoA thiolases (transcript-23660 and transcript-24461), three farnesyl diphosphate synthases (transcript-28558, transcript-24094 and transcript-28255), phosphomevalonate kinase (transcript-17266) and mevalonate pyrophosphate decarboxylase (transcript-25465) had remarkably expression levels (Figure 4 and Supplementary Table 2). Through hierarchical clustering analysis and gene expression modes in different tissues, upstream genes were identified that potentially participated in the triterpene saponin biosynthesis of *D. asperoides*.

Furthermore, CYPs and UGTs that related to the downstream biosynthetic pathway of triterpenoid saponin were screened in *D. asperoides* transcriptome. All 125 CYP transcripts were discovered, of which 13, 15, 23, 35 and 39 (10.4%, 12.0%, 18.4%, 28.0%, 31.2%) had the highest expression in fibrous root, stem, root, leaf, and flower (Supplementary Figure 4 and Supplementary Table 3). Meanwhile, 230 UGTs were identified, of which 29, 29, 33, 50 and 89 (12.61%, 12.61%, 14.35%, 21.74%, 38.69%) had the highest expression in flower, stem, leaf, fibrous root, and root (Supplementary Figure 4

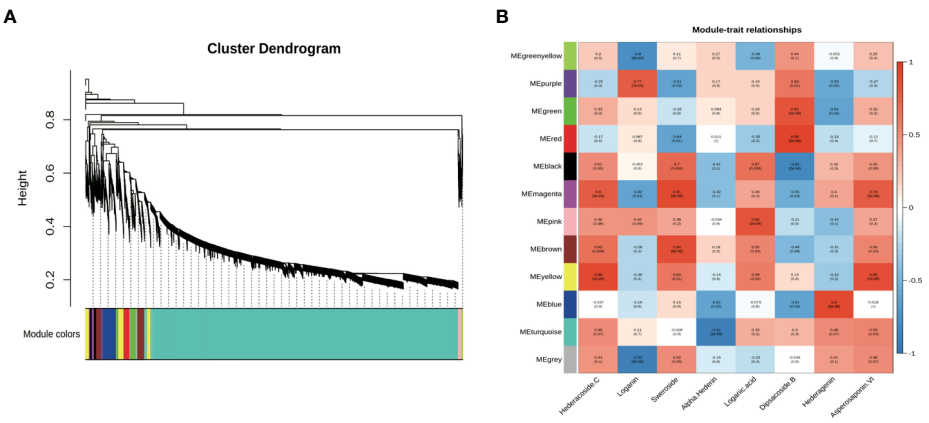


FIGURE 3
Co-expression network analysis of genes related to different saponins in five tissues of *D. asperoides* (root, leaf, flower, stem, fibrous root). **(A)** Gene dendrogram obtained by average linkage hierarchical clustering, demonstrating co-expression modules in *D. asperoides*. **(B)** The correlation coefficient between modules and saponins in *D. asperoides*.

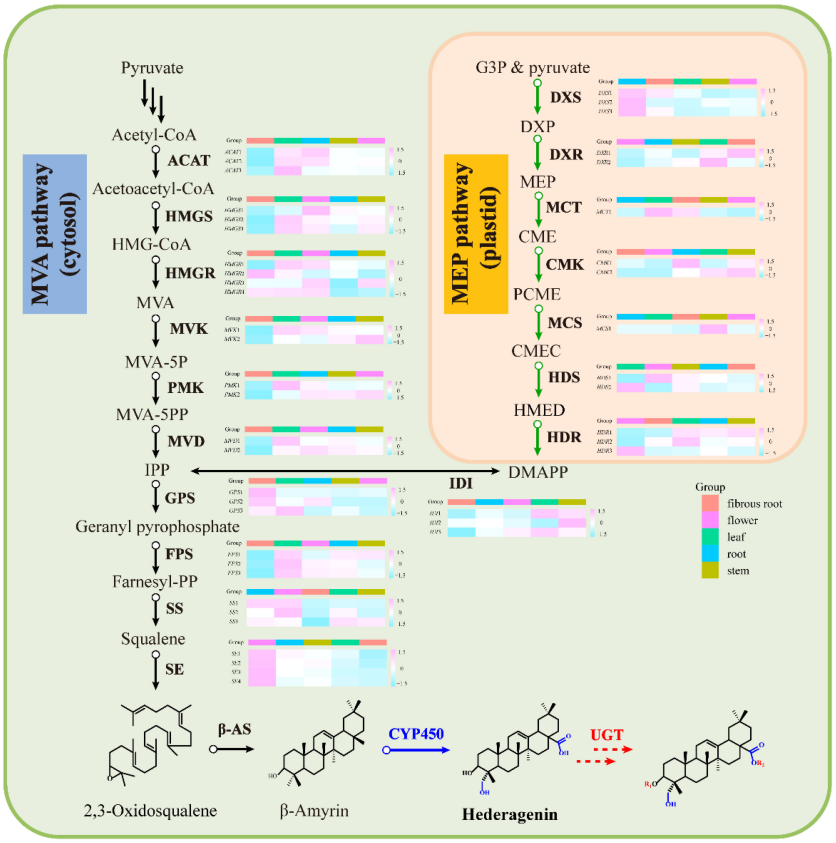


FIGURE 4
Gene expression in the MVA and MEP pathway for saponins in *D. asperoides*. ACAT, Acetyl Coenzyme a Acyltransferase; HMGS, 3-Hydroxy-3-Methylglutaryl Coenzyme a Synthase; HMGR, Hydroxymethylglutaryl-Coa Reductase; MVK, Mevalonate Kinase; PMK, Phosphomevalonate Kinase; MVD, Mevalonate Pyrophosphate Decarboxylase; IDI, Isopentenyl Diphosphate Delta-Isomerase; GPS, Geranyl Pyrophosphate Synthase; FPS, Farnesyl Pyrophosphate Synthase; SS, Squalene Synthase; SE, Squalene Epoxidase; DXS, 1-Deoxy-D-Xylulose-5-Phosphate Synthase; DXR, 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase; MCT, 2-C-Methyl-D-Erythritol 4-Phosphate Cytidylyltransferase; CMK, 4-(cytidine 50-diphospho)-2-C-methyl-D-erythritol kinase; MCS, 2-C-Methyl-D-Erythritol 2,4-Cyclodiphosphate Synthase; HDS, 4-Hydroxy-3-Methylbut-2-En-1-Yl Diphosphate Synthase; HDR, 1-Hydroxy-2-Methyl-2-(E)-Butenyl 4-Diphosphate Reductase; HMG-CoA, 3-Hydroxy-3-Methylglutaryl CoA; DXP, 1-Deoxy-D-Xylulose 5-Phosphate; CMEC, Carboxymethyl Ethyl Cellulose; DMAPP, Dimethylallyl Diphosphate; MVA-5P, Mevalonate-5-Pyrophosphate; CME, 4-(Cytidine 5′-Diphospho)-2-C-Methyl-D-Erythritol; PCME, 2-Phospho-4-(Cytidine 5′-Diphospho)-2-C-Methyl-D-Erythritol; HMED, 4-Hydroxy-3-Methylbut-2-Enyl-Diphosphate; β-AS, Beta-Amyrin Cyclase; CYP450, Cytochromep450; UGT, Glycosyltransferase; MVA, Mevalonic acid; IPP, Isopentenyl Diphosphate; MVA, Schematic of Mevalonate; MEP, 2-C-Methyl-D-Erythritol 4-Phosphate/1-Deoxy-D-Xylulose 5-Phosphate.

and Supplementary Table 3). According to the above results, a presumable conclusion could be drawn that the expression of CYPs and UGTs was different in five tissues, leading to differential contents of triterpenoid saponins.

3.5 Proteomics bioinformatics analysis

Transcriptomic analysis can only reveal triterpenoid saponin biosynthesis at the mRNA level, but cannot explain post-transcriptional processes such as translation and protein modification. Proteins are considered to have a greatly direct correlation with triterpenoid saponin. In this study, Label-free quantitative LC-MS/MS was used to obtain a full-scale proteomic profiles of three *D. asperoides* tissues. A total of 1,380,438 spectrums, 95,932 matched spectrums, 15,665 unique peptides and 3,774 identified proteins (Figure 5A) were collected. There were 2508, 1098, 143, and 28 proteins with molecular weights of 0-50 kDa, 50-100 kDa, 100-150 kDa, and over 150 kDa (Figure 5B), respectively. The above proteins with 1-5 peptides, 6-10 peptides, 11-14 peptides, and 15 or more peptides consisted of 2073, 993, 381 and 327 (Figure 5C), respectively. Protein sequences converging with 0-15%, 15-30%, 30-45%, 45-60% and 60-100% scope were accounted for 42.27%, 27.83%, 17.10%, 9.41%, and 3.39% (Figure 5D), respectively. As shown in Supplementary Figure 5, 643 out of 3,735 proteins were expressed in all three tissues, whereas 84, 103, and 475 were exclusively expressed in root, leaf, and flower, indicating that there were distinct proteins in different tissues. Therefore, significant differences in proteins were detected by comparing protein expression profiles between tissues using the fold change (FC) ≥ 2 and p-values < 0.05. Comparing Pleaf and Pflower samples with Proot samples, 102 and 132 proteins were discovered, respectively. Meanwhile, 740 differentially proteins were identified in Pleaf and Pflower samples (Supplementary Table 5). Proteomics analysis was further conducted to examine the changes in different tissues from protein levels as verified supplementary for transcriptome. In this study, it was found that there were significant differences between root and flower tissues by the analysis of compounds in five tissues (Figure 1D). Hence, the GO and KEGG enrichment analysis of different proteins in root and flower tissues were conducted. After

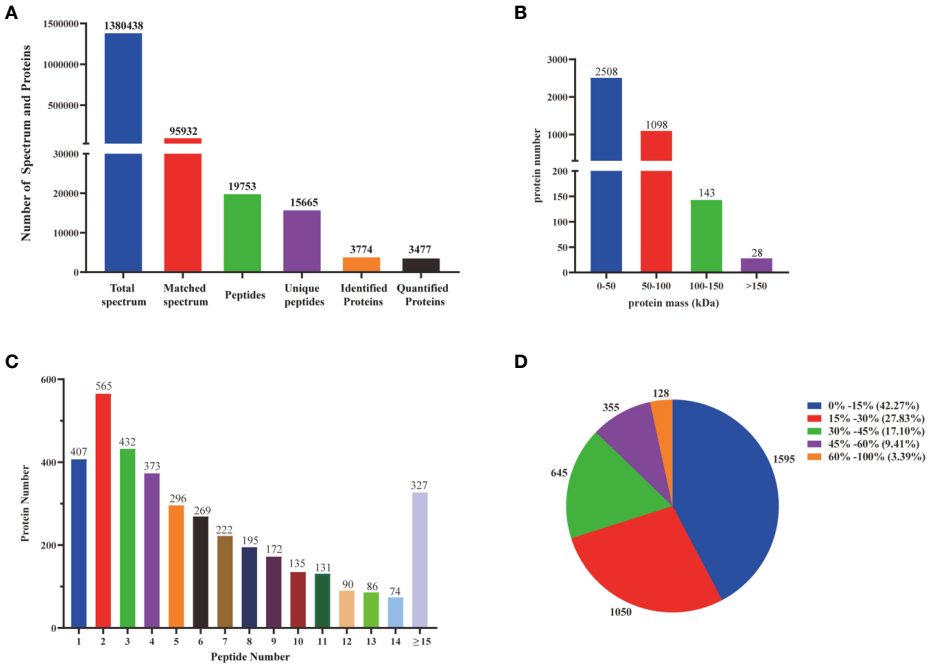


FIGURE 5
Identification and analysis of the proteome on *D. asperoides*. **(A)** Total spectrum, matched spectrum, peptides, unique peptides, identified proteins, and quantified proteins detected from Label- free proteomic analysis. **(B)** Identified proteins were grouped based on their protein mass. **(C)** The number of peptides matched to proteins was shown by Protein Pilot 5.0. **(D)** The identified proteins were classified into pie charts by protein sequence coverage.

go analysis, the above peptides were divided into BP (2985), MF (2266) and CC (3833). KEGG analysis showed that these different proteins were further assigned into 194 biological pathways, such as biosynthesis of cofactors, proteasome, glycerolipid metabolism, etc. (Supplementary Figure 6). In addition, heatmaps for all proteins were performed between three tissues. As shown in Supplementary Figure 7, the differentially proteins between various groups were diverse (Supplementary Table 6). Compared with the study on the proteomic analysis of *D. asperoides* roots from different habitats in China (Jin et al., 2020), our study identified some genes highly related to saponin biosynthesis through analyzing differentially proteins and binding transcriptome analysis in three tissues. In the analysis of transcriptome and proteomics, some genes were simultaneously identified, such as IDI (transcript-3950 and transcript-40019), HMGS (transcript-18826), ACAT (transcript-23660 and transcript-24461), and MVD (transcript-25465), etc. To some extent, this indicated that proteomics analysis was in keeping with the results in mRNA level.

3.6 Co-expression analysis of triterpenoid saponin contents and biosynthesis-associated transcripts

Co-expression analyses were generally used to exploit biological significance genes (Langfelder and Horvart, 2008). WGCNA is a system biology method for disclosing huge related gene clusters to different ingredients and figure out correlation coefficients between modules and target ingredients. It is convenient to seek out the modules related to triterpenoid saponins in tissues for further identifying critical genes involved in the biosynthesis of saponin. Genes involved in saponin biosynthesis of *D. asperoides* were identified through co-expression analysis and WGCNA. In this study, both saponin and non-saponin components were jointly analyzed to more accurately disclose genes related to triterpenoid saponin biosynthesis. As shown in Figure 3, a total of 18,940 transcripts were subdivided into twelve modular clusters based on transcripts expression levels and relative content of compounds, and all modules were inconsistently correlated with different saponins. This was conducted to disclosing the correlation of tissues and triterpenoid saponin contents. Genes with a positive correlation related to a certain saponin identified in modules can be selected for preferred candidate genes for further enzymatic function verification. Based on this, genes associated with a certain type of saponin biosynthesis were screened according to the coefficients ($R > 0.5$) and p-values ($p < 0.05$). For instance, in the MEmagenta module, 137 transcripts were remarkably associated with hederacoside C ($R = 0.8$, $p < 0.05$) and asperosaponin VI ($R = 0.79$, $p < 0.05$) composition in specimens, while dipsacoside B ($R = -0.55$, $p < 0.05$) showed negative correlation comparing with the above transcripts. Transcripts in the MEblue module displayed a highly positive correlation with hederagenin ($R = 0.9$, $p < 0.05$), while alpha-hederin exhibited a negative correlation. In MEgreen module, dipsacoside B ($R = 0.81$, $p < 0.05$) was significantly correlated with 344 transcripts. For non-saponin components, 131 transcripts in the MEpurple module were significantly correlated with loganin ($R = 0.77$, $p < 0.05$). In MEmagenta module, there was a positive relationship between 137 transcripts and sweroside ($R = 0.91$, $p < 0.05$), and 185 transcripts in the MEpink module were remarkably associated with loganic acid (Figure 3). Consequently, 1,256

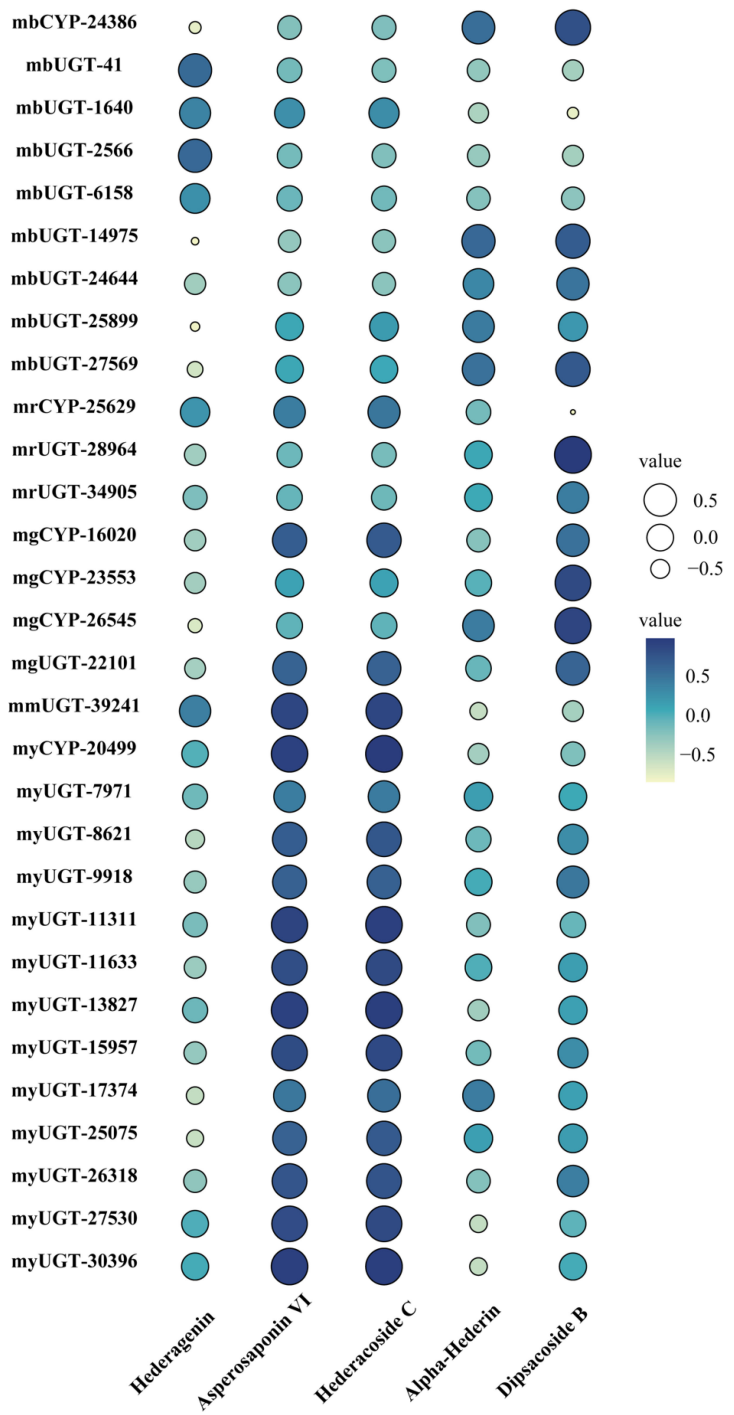


FIGURE 6
Pearson correlation bubble chart of gene expression patterns and saponin contents in five tissues of *D. asperoides*. mb, MEblue; mr, MEred; mg, MEgreen; mm, MEmagenta; my, MEyellow. The size of circles corresponds to correlation coefficient (R) values, and colors indicate whether a correlation is negative or positive.

transcripts were identified in seven modules correlated with target compositions. The red, yellow and bule modules contained saponins-type genes, but the purple, magenta, and pink modules contained non-saponin-type genes. More attention should be paid to red, yellow and bule modules for effectively screening essential genes participated in triterpenoid saponins of biosynthesis pathways in *D. asperoides*.

Different types of genes from modules positively correlated with triterpenoid saponin contents were obtained by WGCNA (Figure 3). In MEyellow and MEmagenta, 603 and 137 transcripts were strongly correlated with asperosaponin VI and hederacoside C, respectively. Furthermore, 284 and 343 transcripts were positively correlated with dipsacoside B in MEred and MEgreen modules, respectively. Moreover, 644 transcripts were highly associated with hederagenin in MEblue module. In total, 6 CYPs and 24 UGTs transcripts were identified, which were positively related to triterpenoid saponin contents. Four CYPs (transcript-25629, transcript-16020, transcript-

23553, transcript-26545) and three UGTs (transcript-28964, transcript-34905, transcript-22101) were highly associated with dipsacoside B. CYP (transcript-24386) and UGTs (transcript-41, transcript-1640, transcript-2566, transcript-6158, transcript-14975, transcript-24644, transcript-25899 and transcript-27569) were strongly associated with hederagenin. In addition, CYP (transcript-20499) and UGTs (transcript-7971, transcript-8621, transcript-9918, transcript-11311, transcript-11633, transcript-13827, transcript-15957, transcript-17374, transcript-25075, transcript-26318, transcript-27530, transcript-30396) were strongly correlated with asperosaponin VI and hederacoside C (Figure 3, Supplementary Tables 3 and 4).

CYPs and UGTs play important roles in saponins biosynthesis. It was found that 6 transcripts of CYPs and 24 transcripts of UGTs were highly expressed in five WGCNA modules (Figure 6). In Supplementary Table 4, it was summarized the correlation between saponins contents and the above genes examined. It is obvious that the significant correlation of seven UGTs (transcript-8621, transcript-11311, transcript-11633, transcript-13827, transcript-15957, transcript-25075 and transcript-26318) and one CYP (transcript-20499) was prominently positively correlated with hederacoside C and asperosaponin VI. Conversely, the above transcripts were inversely associated with hederagenin. Hederagenin can be catalyzed to hederacoside C and asperosaponin VI by UGTs, indicating that these genes could contribute to the biosynthesis of hederacoside C and asperosaponin VI. In addition, the expression of three CYPs (transcript-24386, transcript-23553 and transcript-26545) and two UGTs (transcript-27569 and transcript-28964) were notably correlated with dipsacoside B. As a result, those genes could be potentially related to the biosynthesis of dipsacoside B. Furthermore, there were one CYP (transcript-24386) and two UGTs (transcript-14975 and transcript-27569) were likely involved in the biosynthesis of alpha-hederin. It is worth mentioning that two CYPs (transcript-26545 and transcript-23553) were also identified in proteomics. These results will provide novel insights into understanding the biological functions of target genes in *D. asperoides*.

4 Conclusion

In summary, this is the first report on the full-length transcriptome of the medicinal plant *D. asperoides*. The distribution and contents of saponins exhibited tissue-specific dependent patterns in *D. asperoides*. Candidate CYPs, UGTs and other transcripts involved triterpenoid saponins biosynthesis were finally revealed through an integrated analysis strategy of the transcriptome, proteomics, and metabolites in five various tissues of *D. asperoides*, including root, leaf, flower, stem, and fibrous root. Together, these findings will offer novel insights into the molecular level for the control and regulation of saponin biosynthesis in *D. asperoides* and genetic elements for synthetic bioactive natural active compounds *de novo*.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number PRJNA889678, and ProteomeXchange, accession number PXD038580.

Author contributions

RW and XY were the leading investigators of this research program. RW designed the experiments. JP and CH performed most of the experiments and analyzed the data. WY, TN and XY assisted in experiments and discussed the results. JP and RW wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was financially sponsored by the Shanghai Rising-Star Program (20QA1408800) and the Natural Science Foundation of Shanghai (22ZR1461200, 20ZR1458200).

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/fpls.2023.1134352/full#supplementary-material

References

Achnine, L., Huhman, D. V., Farag, M. A., Sumner, L. W., Blount, J. W., and Dixon, R. A. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. *Plant J.* 41 (6), 875–887. doi: 10.1111/j.1365-313X.2005.02344.x

Chen, Q., Shi, J., Mu, B., Chen, Z., Dai, W., and Lin, Z. (2020). Metabolomics combined with proteomics provides a novel interpretation of the changes in nonvolatile compounds during white tea processing. *Food Chem.* 332, 127412. doi: 10.1016/j.foodchem.2020.127412

Cheng, Y., Liu, H., Tong, X., Liu, Z., Zhang, X., Li, D., et al. (2020). Identification and analysis of CYP450 and UGT supergene family members from the transcriptome of *Aralia elata* (Miq.) seem reveal candidate genes for triterpenoid saponin biosynthesis. *BMC Plant Biol.* 20 (1), 214. doi: 10.1186/s12870-020-02411-6

Choi, N. H., Jang, J. Y., Choi, G. J., Choi, Y. H., Jang, K. S., Nguyen, V. T., et al. (2017). Antifungal activity of sterols and dipsacus saponins isolated from *Dipsacus asper* roots against phytopathogenic fungi. *Pestic. Biochem. Physiol.* 141, 103–108. doi: 10.1016/j.pestbp.2016.12.006

Han, J. Y., Chun, J. H., Oh, S. A., Park, S. B., Hwang, H. S., Lee, H., et al. (2018). Transcriptomic analysis of *Kalopanax septemlobus* and characterization of ksbas, CYP716A94 and CYP72A397 genes involved in hederagenin saponin biosynthesis. *Plant Cell Physiol.* 59 (2), 319–330. doi: 10.1093/pcp/pcx188

Hung, T. M., Na, M., Thuong, P. T., Su, N. D., Sok, D., Song, K. S., et al. (2006). Antioxidant activity of caffeoyl quinic acid derivatives from the roots of *Dipsacus asper* wall. *J. Ethnopharmacol.* 108 (2), 188–192. doi: 10.1016/j.jep.2006.04.029

Jeong, S. I., Zhou, B., Bae, J. B., Kim, N. S., Kim, S. G., Kwon, J., et al. (2008). Apoptosis-inducing effect of akebia saponin d from the roots of *Dipsacus asper* wall in U937 cells. *Arch. Pharm. Res.* 31 (11), 1399–1404. doi: 10.1007/s12272-001-2123-0

Ji, D., Wu, Y., Zhang, B., Zhang, C. F., and Yang, Z. L. (2012). Triterpene saponins from the roots of *Dipsacus asper* and their protective effects against the $AB_{25\text{-}35}$ induced cytotoxicity in PC12 cells. *Fitoterapia* 83 (5), 843–848. doi: 10.1016/j.fitote.2012.03.004

Jia, X. H., Wang, C. Q., Liu, J. H., Li, X. W., Wang, X., Shang, M. Y., et al. (2013). Comparative studies of saponins in 1-3-Year-Old main roots, fibrous roots, and rhizomes of *Panax notoginseng*, and identification of different parts and growth-year samples. *J. Nat. Med.* 67 (2), 339–349. doi: 10.1007/s11418-012-0691-6

Jin, H., Yu, H., Wang, H., and Zhang, J. (2020). Comparative proteomic analysis of *Dipsacus asperoides* roots from different habitats in China. *Molecules* 25 (16), 3605. doi: 10.3390/molecules25163605

Jung, K. Y., Do, J. C., and Son, K. H. (1993). Triterpene glycosides from the roots of *Dipsacus asper*. *J. Nat. Prod.* 56 (11), 1912–1916. doi: 10.1021/np50101a007

Langfelder, P., and Horvath, S. (2008). WGCNA: An r package for weighted correlation network analysis. *BMC Bioinform.* 9, 559. doi: 10.1186/1471-2105-9-559

Li, F., Tanaka, K., Watanabe, S., Tezuka, Y., and Saiki, I. (2013). Dipasperoside a, a novel pyridine alkaloid-coupled iridoid glucoside from the roots of *Dipsacus asper*. *Chem. Pharm. Bull.* 61 (12), 1318–1322. doi: 10.1248/cpb.c13-00546

Liu, J. J., Wang, X. L., Guo, B. L., Huang, W. H., Xiao, P. G., Huang, C. Q., et al. (2011). Triterpenoid saponins from *Dipsacus asper* and their activities *in vitro*. *J. Asian Nat. Prod. Res.* 13 (9), 851–860. doi: 10.1080/10286020.2011.598858

Lu, C., Fan, G., and Wang, D. (2020). Akebia saponin d ameliorated kidney injury and exerted anti-inflammatory and anti-apoptotic effects in diabetic nephropathy by activation of NRF2/HO-1 and inhibition of NF-KB pathway. *Int. Immunopharmacol.* 84, 106467. doi: 10.1016/j.intimp.2020.106467

Seki, H., Tamura, K., and Muranaka, T. (2015). P450s and UGTs: Key players in the structural diversity of triterpenoid saponins. *Plant Cell Physiol.* 56 (8), 1463–1471. doi: 10.1093/pcp/pcv062

Sun, X., Ma, G., Zhang, D., Huang, W., Ding, G., Hu, H., et al. (2015). New lignans and iridoid glycosides from *Dipsacus asper* wall. *Molecules* 20 (2), 2165–2175. doi: 10.3390/molecules20022165

Tao, Y., Huang, S., Li, W., and Cai, B. (2019). Simultaneous determination of ten bioactive components in raw and processed radix *Dipsaci* by UPLC-Q-TOF-MS. *J. Chromatogr Sci.* 57 (2), 122–129. doi: 10.1093/chromsci/bmy093

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* 65, 225–257. doi: 10.1146/annurev-arplant-050312-120229

Tran, M. H., Phuong, T. T., UiJoung, Y., Zhang, X. F., Min, B. S., Mi, H. W., et al. (2008). Antioxidant activities of phenolic derivatives from *Dipsacus asper* wall. (II). *Nat. Prod. Res.* 14 (2), 107–112.

Tzin, V., Snyder, J. H., Yang, D. S., Huhman, D. V., Watson, B. S., Allen, S. N., et al. (2019). Integrated metabolomics identifies CYP72A67 and CYP72A68 oxidases in the biosynthesis of *Medicago truncatula* oleanate sapogenins. *Metabolomics* 15 (6), 85. doi: 10.1007/s11306-019-1542-1

Vranova, E., Coman, D., and Gruissem, W. (2013). Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.* 64, 665–700. doi: 10.1146/annurev-arplant-050312-120116

Wan, Z., Zhu, J., Tian, R., Yang, W., Chen, Z., Hu, Q., et al. (2021). Quality evaluation for *Dipacus asperoides* from enshi areas and optimization extraction of saponins and organic acids and its application. *Arab. J. Chem.* 14 (4), 103107. doi: 10.1016/j.arabjc.2021.103107

Wang, J., Deng, H., Zhang, J., Wu, D., Li, J., Ma, J., et al. (2020). α-hederin induces the apoptosis of gastric cancer cells accompanied by glutathione decrement and reactive oxygen species generation *via* activating mitochondrial dependent pathway. *Phytother. Res.* 34 (3), 601–611. doi: 10.1002/ptr.6548

Wang, Y., Li, Q., Wang, Q., Li, Y., Ling, J., Liu, L., et al. (2012). Simultaneous determination of seven bioactive components in oolong tea *Camellia sinensis*: Quality control by chemical composition and HPLC fingerprints. *J. Agric. Food Chem.* 60 (1), 256–260. doi: 10.1021/jf204312w

Wang, J. Y., Liang, Y. L., Hai, M. R., Chen, J. W., Gao, Z. J., Hu, Q. Q., et al. (2016). Genome-wide transcriptional excavation of *Dipsacus asperoides* unmasked both cryptic asperosaponin biosynthetic genes and SSR markers. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00339

Wang, Y., Shen, J., Yang, X., Jin, Y., Yang, Z., Wang, R., et al. (2018). Akebia saponin d reverses corticosterone hypersecretion in an alzheimer's disease rat model. *Biomed. Pharmacother.* 107, 219–225. doi: 10.1016/j.biopha.2018.07.149

Wang, Y., Zhang, H., Ri, H. C., An, Z., Wang, X., Zhou, J. N., et al. (2022). Deletion and tandem duplications of biosynthetic genes drive the diversity of triterpenoids in *Aralia elata*. *Nat. Commun.* 13 (1), 2224. doi: 10.1038/s41467-022-29908-y

Wang, Z. L., Zhou, J. J., Han, B. Y., Hasan, A., Zhang, Y. Q., Zhang, J. H., et al. (2022). GuRhaGT, a highly specific saponin 2''-O-Rhamnosyltransferase from *Glycyrrhiza uralensis*. *Chem. Commun.* 58 (34), 5277–5280. doi: 10.1039/D1CC07021E

Xu, R., Fazio, G. C., and Matsuda, S. P. (2004). On the origins of triterpenoid skeletal diversity. *Phytochemistry* 65 (3), 261–291. doi: 10.1016/j.phytochem.2003.11.014

Xu, R., Zhang, J., You, J., Gao, L., Li, Y., Zhang, S., et al. (2020). Full-length transcriptome sequencing and modular organization analysis of oleanolic acid- and dammarane-type saponins related gene expression patterns in *Panax japonicus*. *Genomics* 112 (6), 4137–4147. doi: 10.1016/j.ygeno.2020.06.045

Yu, X., Wang, L. N., Ma, L., You, R., Cui, R., Ji, D., et al. (2012). Akebia saponin d attenuates ibotenic acid-induced cognitive deficits and pro-apoptotic response in rats: Involvement of MAPK signal pathway. *Pharmacol. Biochem. Behav.* 101 (3), 479–486. doi: 10.1016/j.pbb.2012.02.014

Yu, J. H., Yu, Z. P., Wang, Y. Y., Bao, J., Zhu, K. K., Yuan, T., et al. (2019). Triterpenoids and triterpenoid saponins from *Dipsacus asper* and their cytotoxic and antibacterial activities. *Phytochemistry* 162, 241–249. doi: 10.1016/j.phytochem.2019.03.028

Zhong, F., Huang, L., Qi, L., Ma, Y., and Yan, Z. (2020). Full-length transcriptome analysis of *Coptis deltoidea* and identification of putative genes involved in benzylisoquinoline alkaloids biosynthesis based on combined sequencing platforms. *Plant Mol. Biol.* 102 (4-5), 477–499. doi: 10.1007/s11103-019-00959-y

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 March 2023
DOI 10.3389/fpls.2023.1102715

OPEN ACCESS

EDITED BY
Mark Chapman,
University of Southampton,
United Kingdom

REVIEWED BY
Liangsheng Zhang,
Zhejiang University, China
Wei Fan,
Agricultural Genomics Institute at Shenzhen (CAAS), China
Simone Scalabrin,
IGA Technology Services, Udine, Italy

*CORRESPONDENCE
Heng Zhang
hengzhang@psc.ac.cn
Hui Zhang
laohanzhang@hotmail.com

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 19 November 2022
ACCEPTED 06 March 2023
PUBLISHED 20 March 2023

CITATION
Li L, Song J, Zhang M, Iqbal S, Li Y, Zhang H and Zhang H (2023) A near complete genome assembly of chia assists in identification of key fatty acid desaturases in developing seeds. *Front. Plant Sci.* 14:1102715. doi: 10.3389/fpls.2023.1102715

A near complete genome assembly of chia assists in identification of key fatty acid desaturases in developing seeds

Leiting Li[1], Jingjing Song[1], Meiling Zhang[2], Shahid Iqbal[3], Yuanyuan Li[4], Heng Zhang[1]* and Hui Zhang[5]*

[1]National Key Laboratory of Molecular Plant Genetics, Shanghai Center for Plant Stress Biology, Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [2]Center for Excellence in Brain Science and Intelligence Technology, Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China, [3]Institute of Plant Breeding and Biotechnology, Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan, [4]Centre for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [5]Shandong Provincial Key Laboratory of Plant Stress Research, College of Life Science, Shandong Normal University, Jinan, Shandong, China

Chia is an annual crop whose seeds have the highest content of α-linolenic acid (ALA) of any plant known to date. We generated a high-quality assembly of the chia genome using circular consensus sequencing (CCS) of PacBio. The assembled six chromosomes are composed of 21 contigs and have a total length of 361.7 Mb. Genome annotation revealed a 53.5% repeat content and 35,850 protein-coding genes. Chia shared a common ancestor with *Salvia splendens* ~6.1 million years ago. Utilizing the reference genome and two transcriptome datasets, we identified candidate fatty acid desaturases responsible for ALA biosynthesis during chia seed development. Because the seed of *S. splendens* contains significantly lower proportion of ALA but similar total contents of unsaturated fatty acids, we suggest that strong expression of two *ShFAD3* genes are critical for the high ALA content of chia seeds. This genome assembly will serve as a valuable resource for breeding, comparative genomics, and functional genomics studies of chia.

KEYWORDS

chia, polyunsaturated fatty acids, transcriptome, FAD, HiFi

Introduction

Chia (*Salvia hispanica* L.) is an annual herbaceous crop belonging to the family of Lamiaceae, also commonly known as the mint family. Chia is native to central America and is believed to have served as a staple crop of the Aztec in pre-Columbian times (Valdivia-López and Tecante, 2015). Chia is currently cultivated for its seeds in Central and South America. Chia produces oily seeds with an oval shape and a diameter of ~2 mm. Thanks to its superior nutrient compositions, the chia seed is a trending functional food ingredient

(Muñoz et al., 2013; Cassiday, 2017). Chia seeds contain 30-40% total lipids, of which α-linolenic acid (ALA; C18:3, n-3), linoleic acid (LA; C18:2, n-6), and oleic acid (C18:1, n-9) account for ~60%, ~20%, and ~10% respectively (Ciftci et al., 2012; Kulczynski et al., 2019). ALA is an essential fatty acid (i.e., cannot be synthesized by human body) and up to 8-21% and 1-9% of ALA intake can be respectively converted to eicosapentaenoic acid (EPA; C20:5, n-3) and docosahexaenoic acid (DHA; C22:6, n-3) in the human body (Baker et al., 2016; Shahidi and Ambigaipalan, 2018). Studies indicate that these n-3 fatty acids are important for human development and growth (Li et al., 2019). The recommended Adequate Intake (AI) of ALA is 1.6 g/day for men and 1.1 g/day for women (Burns-Whitmore et al., 2019). In addition, a low n-6:n-3 ratio, as in the case of chia seeds, in the diet helps reduce inflammation (Simopoulos, 2002a; Simopoulos, 2002b; Lands, 2014). Chia seeds also have high contents of dietary fiber (up to 34.4%), proteins (16.5-24.2%), vitamin B3, multiple minerals (such as calcium, phosphorus, potassium, and ion), and antioxidants (Kulczynski et al., 2019). Because of these properties, chia seeds are increasingly used as an ingredient in food industry and restaurants.

In plants, fatty acid (FA) biosynthesis takes place within the plastid, where acetyl-coenzyme A (acetyl-CoA) is used as the main carbon donor for the initiation and elongation of acyl chains (Ohlrogge and Browse, 1995; Li-Beisson et al., 2013). During the elongation, fatty acids remain covalently attached to acyl carrier proteins (ACPs), which serve as a cofactor for FA biosynthesis. The fatty acids biosynthesis cycle is usually terminated when the acyl chain reaches 16 or 18 carbons in length, and two principal types of acyl-ACP thioesterases, FatA and FatB, hydrolyze acyl-ACP and release the corresponding FAs. Desaturation of common fatty acids (C16 and C18) begins at the C-9 position (Δ9) and progresses in the direction of the methyl carbon of the acyl chain. Thus, the conversion of stearic acid (C18:0) to α-linoleic acid (C18:3$^{\Delta9,12,15}$) involves the sequential action of three desaturases, including the stearoyl-ACP desaturase, the oleate desaturase, and the linoleate desaturase. In the model plant Arabidopsis, genetic analyses have identified the main enzymes with specific FA desaturase activities. While all the other FA desaturases are membrane-bound enzymes, the family of acyl-ACP desaturases (AADs) are stromal soluble enzymes that use stearoyl-ACP (C18:0) or palmitoyl-ACP (C16:0) as the substrate. The Arabidopsis genome encodes 7 AADs (Kachroo et al., 2007), named as FAB2 (FATTY ACID BIOSYNTHESIS 2) and AAD1-6. Genetic analyses indicate that FAB2, AAD1, ADD5, and AAD6 are redundant Δ9 stearoyl-ACP desaturases (SADs) (Kazaz et al., 2020), while AAD2 and AAD3 function as Δ9 palmitoyl-ACP desaturases (PADs) (Troncoso-Ponce et al., 2016). Further desaturation of oleic acids (C18:1$^{\Delta9}$) may take place within the plastid or the endoplasmic reticulum (ER). In the plastid, the oleic acids are incorporated into multiple types of glycerophospholipids and converted to C18:3 by FAD6 (FATTY ACID DESATURASE 6) and FAD7/8. Alternatively, the oleic acid may be exported and enters the acyl-CoA pool in the cytosol. The C18:1-CoA can be imported into ER, where it is incorporated into phosphatidylcholine (PC) and becomes sequentially desaturated by FAD2 and FAD3, which respectively prefer PC with C18:1 and C18:2 as the substrate. During seed development, the desaturated PCs are further converted to diacylglycerol (DAG) and triacylglycerol (TAG), the latter of which is the main form of storage lipids in the oil body of seeds.

In this study, we assembled a high-quality chia genome using accurate consensus long reads (PacBio HiFi reads) and genome-wide chromosome conformation capture (Hi-C). The chia genome is known to have 6 chromosomes ($2n$ = 12) (Estilai et al., 1990), which in our study are composed of 21 main contigs, with telomere repeats at 8 ends of the chromosomes. Utilizing this highly accurate and complete genome, we annotated transposable elements and protein-coding genes in the chia genome. Compared to a recently published chromosome-level assembly of chia (Wang et al., 2022), our assembly has better contiguity and ~15% more gene models (35,850 vs. 31,069) thanks to the highly accurate CCS reads. Alignment analyses also revealed multiple Mb-size structural variations between two assemblies, demonstrating the importance of multiple high-quality genomes for the same species. Finally, making use of a published seed development transcriptome, we identified the main ER-localized linoleate desaturases that underlie the extremely high ALA content in chia seeds.

Results

Genome assembly

We selected a chia cultivar with a Mexico origin (Supplemental Figure 1) for the assembly of the genome. About 24.7 Gb of circular consensus sequencing reads with an average read length of 16.1 kbp were generated from a single sequencing cell (Supplemental Figure 2). K-mer-based analyses of the HiFi reads estimated the nuclear genome to be ~352.7 Mb in size (Supplemental Figure 3).

We performed genome assembly using the hifiasm assembler (Cheng et al., 2021). The initial assembly was 388.0 Mb, consisting of 666 contigs with a N50 length of 21.8 Mb and an L50 number of 7, indicating a high contiguity of the assembly. The longest 21 contigs have a total length of 361.7 Mb and a minimum length of 1.7 Mb, while other contigs are significantly shorter, 636 of which have lengths shorter than 150 kbp (Figure 1A). The average HiFi read depth on the 21 long contigs varies between 43 and 58, which are around the 54-fold coverage of the nuclear genome calculated from the k-mer distribution (Figure 1A; Supplemental Figure 3). In contrast, the rest 645 contigs have a read coverage varying from 0 to 557, suggesting that they originate either from fragments of highly repetitive regions or from the high-copy organellar genomes.

We next analyzed the plastid and mitochondrion genomes. From the initial assembly, we identified a circular contig (ptg000033c) with a length of 313,444 bp and an average read coverage of 557 folds. Genome annotation identified 151 mitochondrion-encoded genes, including 21 transfer RNAs, 6 ribosomal RNAs (rRNAs), and 124 protein-coding genes (Supplemental Figure 4), indicating that this contig is the complete mitochondrion genome. We also identified 4 other contigs that show 100% sequence identity but structural variations to the mitochondrion genome (Supplemental Figure 5).

10.3389/fpls.2023.1102715

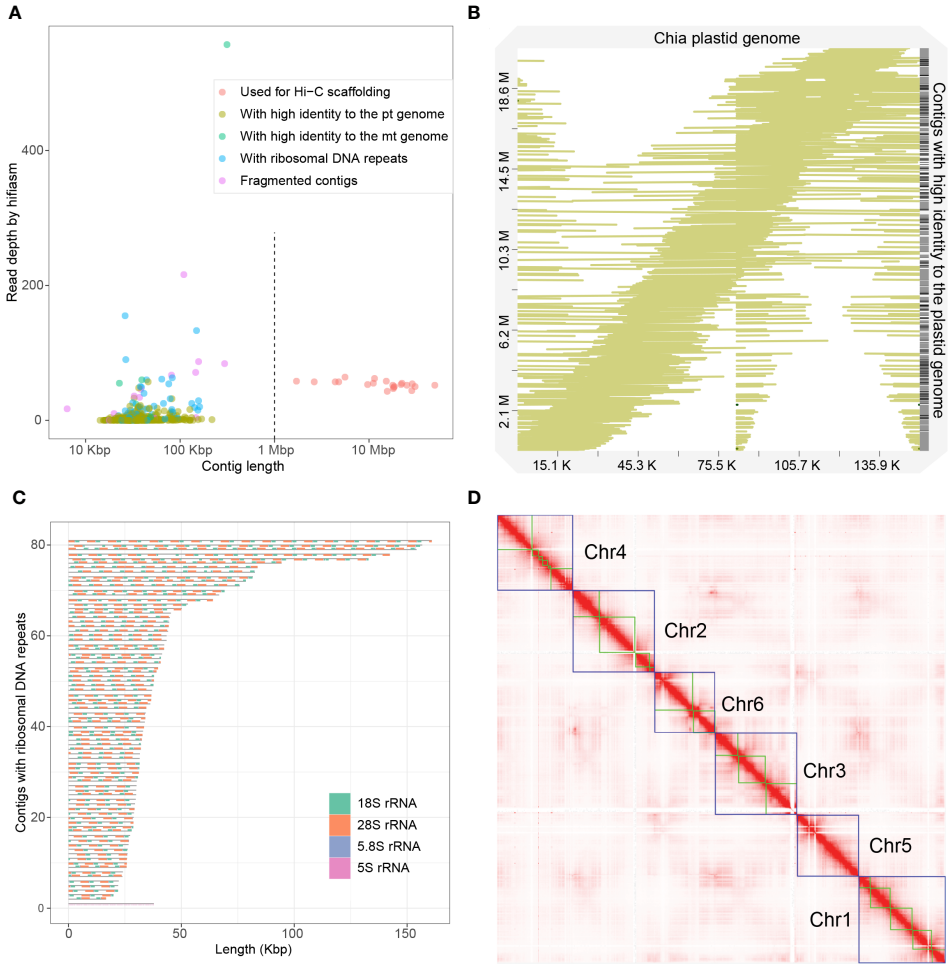


FIGURE 1
Assembly of the chia genome. **(A)** Dotplot showing the contig length and the read depth of the initial assembly. Contigs were classified into five categories based on the length, the read depth and their origins, as indicated in the legend. **(B)** Alignment of 538 initial contigs onto the chia plastid genome. The colored lines indicate the start and the end of the alignment relative to the plastid genome. Numbers in the X-axis indicate the length of the chia plastid genome while numbers in the Y-axis indicate the accumulated length of contigs that show high sequence identity to the plastid genome. **(C)** Structure of the 81 contigs containing ribosomal RNA repeats. **(D)** Hi-C contact map of the chia nuclear genome. Blue boxes indicate grouped pseudochromosomes, whereas green boxes indicate contigs.

Three of these contigs have a read depth similar to that of nuclear contigs (between 24 and 60) (Figure 1A). They might represent mitochondrial genome fragments recently transferred to the nuclear genome, or a minor population(s) of the heterozygous mitochondrial genome.

We could not identify a contig representing the complete plastid genome from the initial assembly. We thus assembled the plastid genome using Illumina short reads and the GetOrganelle software (Jin et al., 2020). The plastid genome has a length of 150,956 bp and 132 genes, including 87 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Supplemental Figure 6). Surprisingly, we found that 538 out of the 666 initial contigs could be mapped to the plastid genome with high coverage (>99%) and high identity rate (>99%) (Figure 1B). These contigs are short in length (14.2 to 217.6 kb) and most of them have low HiFi read coverage (with 530 contigs below 19-fold coverage) (Figure 1A). These plastid-originated contigs likely represent incompletely assembled plastid genome fragments and/or nuclear genome fragments with a plastid origin. The total length of these contigs was 20.7 Mb, accounting for most of the excessive part of the assembly compared to the predicted genome size.

Excluding the organellar-originated 543 contigs and the 21 high-confidence nuclear contigs, the remaining 102 contigs have a total length of 5.2 Mb. Ribosomal RNA (rRNA) repeats were identified in 81 of these contigs, indicating they were originated from genomic regions with high copy number of rRNA genes. Except for one contig mainly composed of 73 repeats of 5S rRNA, other contigs had a basic repeat unit of a “18S-5.8S-28S” structure with the copy number varied from 2 to 17 (Figure 1C). Considering the nuclear origin of most sequences, the 102 contigs were concatenated as Chr0.

We next used the 21 high-confidence nuclear contigs for Hi-C scaffolding. Based on ~180x (63.8 Gb) of Hi-C sequencing data, we clustered and ordered the 21 contigs into six pseudochromosomes,

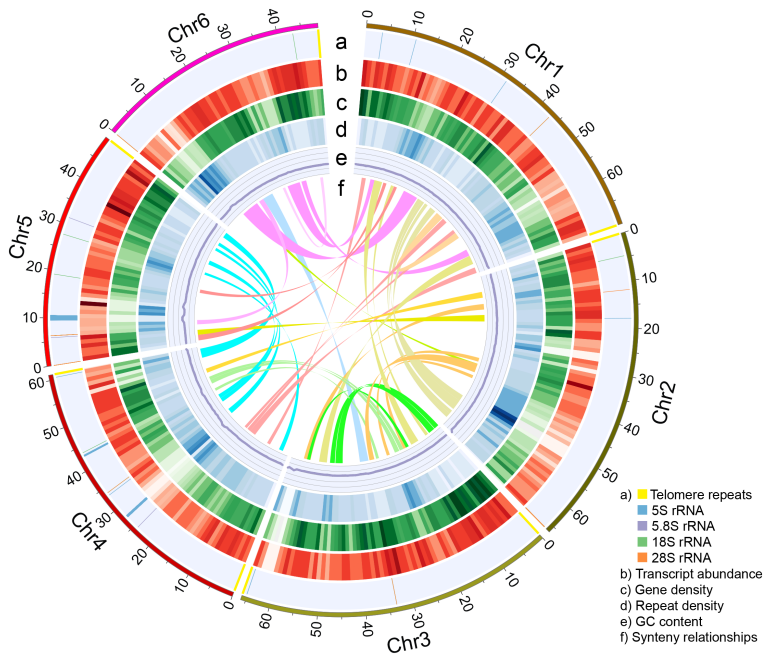


FIGURE 2
The nuclear genome of Shi_PSC_v1. Each ring indicates specific features of the nuclear genome. Data from non-overlapping 1-Mb windows were graphed: **(A)** Position of telomere repeats and ribosomal RNA genes; **(B)** Average transcript abundance; **(C)** Gene density; **(D)** LTR density; **(E)** GC content; **(F)** Synteny blocks >1 Mb in length.

whose sizes ranged from 47.8 Mb to 69.1 Mb (Figures 1D; 2, Table 1). The chromosome sequence names were decreasingly ordered based on sequence length. Chr5 was composed of a single contig while Chr4 contained the largest number (6) of contigs. The total length of the six pseudochromosomes was 361.7 Mb. The final v1 assembly (Shi_PSC_v1) of the chia genome composed of 9 sequences, seven of which (Chr0-Chr6) represent the nuclear genome, one for the mitochondrion genome, and one for the plastid genome.

Evaluation of genome assembly

We next evaluated the quality of the genome assembly using LTR Assembly Index (LAI) (Ou et al., 2018), Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni et al., 2021), Merqury (Rhie et al., 2020) and Illumina short reads. The whole genome had an LAI of 15.78, which was around the same level as the TAIR10 assembly of *Arabidopsis thaliana*, and could be considered as the reference level (Ou et al., 2018). The complete BUSCO of the chia genome assembly was 98.8%, indicating a high completeness of the gene space. Merqury compares k-mers from the assembly to those found in unassembled HiFi reads to estimate the completeness and accuracy. The completeness and quality value (QV) of Shi_PSC_v1 were 97.3 (out of 100) and 66.5 (>99.99% accuracy) respectively. Mapping of the Illumina short reads (Supplemental Table 1) against the chia genome assembly also revealed very high read mapping rate (99.9%) and a low apparent error rate (0.27%).

Genome annotation

For genome annotation, we first identified repetitive sequences in the Shi_PSC_v1 assembly. The analysis revealed that chia nuclear genome had a repeat content of 53.5% (Table 1). Similar to most plant genomes, retrotransposons accounts for the majority of the repetitive sequences of the genome. About half of the repeats were characterized as long terminal repeats (LTRs), with Gypsy (12.0% of the genome) and Copia (7.4% of the genome) being the main types. Besides, 65,851 simple repeats, 334 satellite sequences, 573 transfer RNAs (tRNAs) and 378 small nuclear RNAs (snRNAs) were also identified in the chia genome (Supplemental Table 2).

The repeat-masked assembly was then used for gene model prediction. Based on evidence from *ab initio* prediction, expressed sequence tags (ESTs) that assembled from the RNA-seq data by Gupta et al. (2021), and homologous protein sequences, a total of 35,850 protein-coding genes were annotated. Additionally, we also examined whether telomere signals were present at the end of each pseudochromosome. The results showed that all the six pseudochromosomes contain telomere repeats. Telomere repeats were detected at both ends of Chr3 and Chr4, and one end of Chr1, Chr2, Chr5, and Chr6 (Figure 2A). Comparing Shi_PSC_v1 to a recently published chia genome (Wang et al., 2022) revealed multiple Mb-size variations, including three inversions at the peri-telomeric region of Chr1 and the peri-centromeric regions of Chr2 and Chr3 (Supplemental Figure 7A). Further examination indicated that these regions are supported by raw reads in our assembly (Supplemental Figures 7B, C) but are composed of short

TABLE 1 Summary of chia genome assembly.

	Size	Number
Assembly features		
Estimated genome size	352,711,351 bp	
Total contigs	388,048,784 bp	666 contigs
Contig N50	21,830,104 bp	7 contigs
Longest contig	49,694,750 bp	
Chr1	69,924,378 bp	5 contigs/6997 genes
Chr2	66,361,501 bp	4 contigs/5756 genes
Chr3	66,031,358 bp	3 contigs/7894 genes
Chr4	61,126,009 bp	6 contigs/5365 genes
Chr5	49,694,750 bp	1 contig/4781 genes
Chr6	48,593,615 bp	2 contigs/4929 genes
Mitochondrial genome	313,444 bp	1 contig/151 genes
Plastid genome	150,956 bp	132 genes
GC content		37.00%
Annotation features		
Repetitive sequence		53.5%
Protein-coding genes		35,850

contigs concatenated together in the 2022 assembly (data not shown).

The complete BUSCO score of the protein sequences was 99.0%, close to the BUSCO score of the genome assembly (98.8%). Functional annotation showed that Gene Ontology (GO) terms (Gene Ontology, 2021), Pfam domains (Mistry et al., 2021), and InterPro families (Blum et al., 2021) were assigned to 58.9% (21,125), 72.0% (25,799), and 79.2% (28,405) of the protein-coding genes. In total, AHRD (Automated assignment of Human Readable Descriptions) function names were assigned to 89.5% (32,089) of the protein-coding genes (Boecker, 2021) (Supplemental Table 3). These metrics indicate high quality of the genome annotation.

Evolution of the chia genome

To understand the evolution of the chia genome, we selected five other species from the family of Lamiaceae, including three from the genus of *Salvia*, together with three species of Asterids and *Arabidopsis thaliana* for the orthology analysis (Figure 3A). A species tree constructed using orthologs shared in all analyzed species with STAG (Emms and Kelly, 2018) confirmed a close relationship between chia and *S. splendens*, as well as *S. bowleyana* and *S. miltiorrhiza* (Figure 3A). Using a reference divergence time of 115 million years ago (MYA) between Arabidopsis and other lineages (Hedges et al., 2015), chia was estimated to diverge with *S. splendens* ~6.2 million years ago (MYA) and the four *Salvia* species have a common ancestor ~21.8 MYA. The protein-coding genes of chia were assigned to 17,158 families. Relative to the common ancestor of chia and *S. splendens*, expansion in 528 families and reduction in 2,344 families were observed in chia (Figure 3A). In contrast, *S. splendens* had 8,777 expanded families and a large number of 2-copy gene families (Figure 3B). This is consistent with its recent tetraploidization event (Jia et al., 2021). Among the ten species analyzed, 8,812 families were shared while between 265 and 1,147 families were unique for each species (Figure 3C). Among the 720 gene families (2,529 genes) unique to chia, 72.6% of them were comprised of 2 or 3 members (Supplemental Figure 8) and the largest one contained 36 members. GO enrichment analysis was performed for genes in these chia-specific gene families. The results showed that the top enriched GO term in the category of biological process was "defense response" (GO:0006952) (Supplemental Figure 9), suggesting their potential roles in the environmental adaptation of chia. In addition, "acyl-[acyl-carrier-protein] desaturase activity" (GO:0045300) in the category of molecular function was enriched (Supplemental Figure 10). This expanded family mainly includes orthologous genes of *AtFAB2* (Supplemental Table 3; Supplemental Figure 11), the stearoyl-ACP (C18:0) or palmitoyl-ACP (C16:0) desaturases of Arabidopsis.

To investigate the whole-genome duplication events of chia, we performed intra-genome synteny analysis. In total, 323 synteny blocks with an average of 20.5 homologous gene pairs per block were identified (Figure 2F). The distribution of synonymous substitution rates (Ks) of these gene pairs revealed a single Ks peak at ~0.26 (Supplemental Figure 12), which was consistent with the whole genome duplication (WGD) event prior to the

10.3389/fpls.2023.1102715

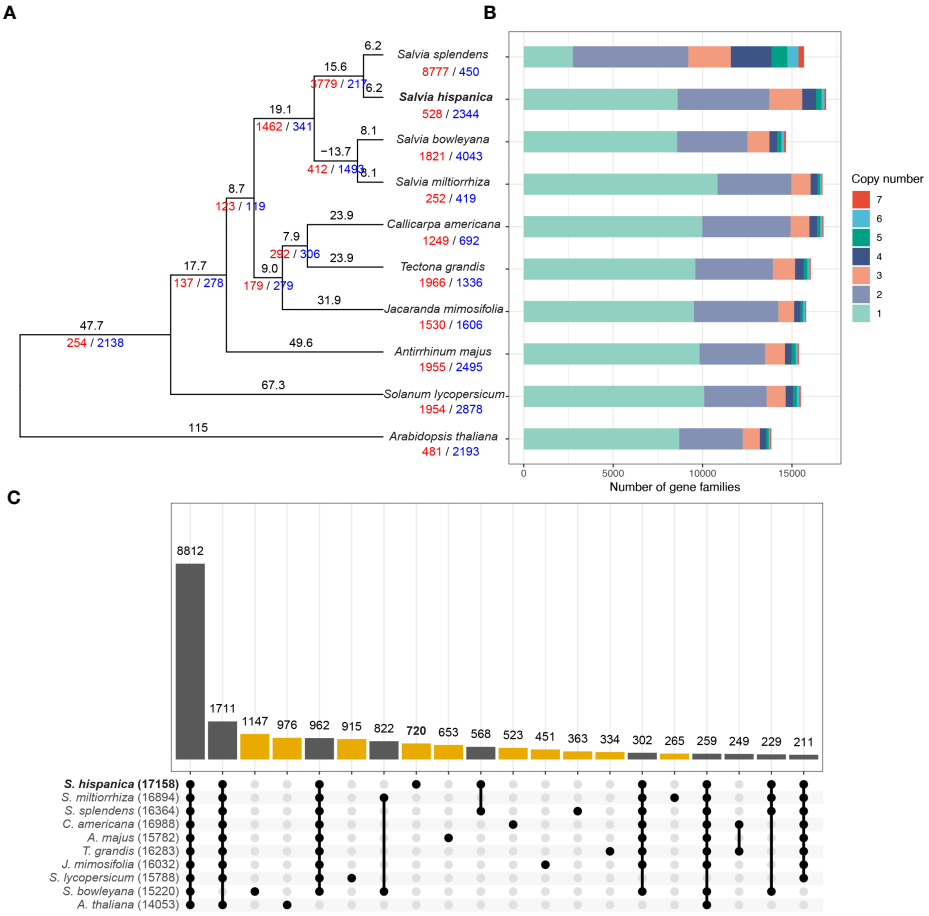


FIGURE 3
Evolution of the chia genome. **(A)** Phylogenetic tree for chia and 9 other plant species. Numbers of expanded and contracted gene families were indicated with red and blue numbers at each branch point. Branch length indicate the estimated divergence time in million years ago. **(B)** Numbers of gene families with different sizes (copy numbers) in each plant species. The X-axis indicate the number of gene families. **(C)** Upset plot indicating the number of gene families shared by different species. Yellow bars indicate numbers of species-specific gene families, whereas gene families that are commonly identified in two or more species are indicated by dark gray bars. Filled circles with a connected line below the x-axis indicate the species that share a specific number of gene families.

tetraploidization event of *S. splendens* (Jia et al., 2021). This indicates that this WGD event occurred before the divergence of chia and *S. splendens*.

Identification of genes involved in ALA biosynthesis

We next sought to identify genes underlying the high ALA content in chia seeds. We used kofamKOALA (Aramaki et al., 2020) to identify homologous genes of the lipid biosynthesis pathway (ko01004 of KEGG) in the chia genome (Supplemental Figure 13; Supplemental Table 4). We focused on genes encoding fatty acid desaturases. The analysis revealed 2 orthologs of *AtFatA* (K10782), 6 orthologs of *AtFatB* (K10781), 14 genes of the *AAD* family (K03921), 2 orthologs of *AtFAD2* (K10256), 2 orthologs of *AtFAD3* (K10257), and 2 orthologs of *AtFAD7/8* (K10257) among others (Figure 4A; Supplemental Figures 13, 14). Multiple sequence alignment (Supplemental Figure 15) indicated that *AtFAD7/8* and their orthologs in chia contain extra N-terminal sequences (plastid transit peptides) compared to the *AtFAD3* branch, consistent with their predicted localization in the plastid (Xue et al., 2018).

We utilized two published transcriptome dataset to help identify candidate ALA biosynthesis genes in the chia genome, one covering 13 different tissues or developmental stages of chia (Gupta et al., 2021) and one covering five different time points of chia seed development (3, 7, 14, 21, and 28 days after flower opening (DAF)) (Sreedhar et al., 2015). We reason that the ALA biosynthesis genes should be expressed at high levels during seed development. Indeed, we found that *Shi004382* (*ShFatA*), *Shi017381*, *Shi000260*, and *Shi006361* (*AtFAB2* orthologs), *Shi027338* and *Shi033531* (*AtFAD2* orthologs), and *Shi018884* and *Shi004328* (*AtFAD3* orthologs) are highly expressed in developing chia seeds, and their expression levels are decreased in the 28 DAF sample (Figure 4B). These genes are also expressed at significantly higher levels in developing seeds compared to other chia tissues/organs (Supplemental Figure 13). Although FAB2 homologs have either SAD or PAD activity, studies in Arabidopsis indicate that a single amino acid change (Tyr to Phe) is sufficient to confer PAD

10.3389/fpls.2023.1102715

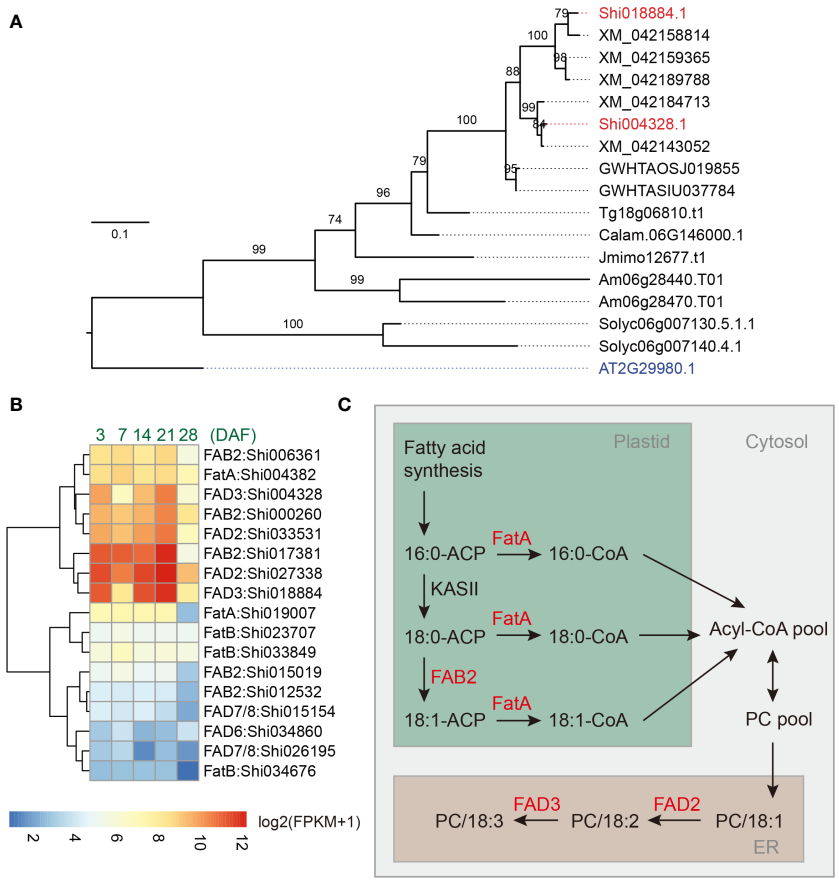


FIGURE 4
Identification of critical genes involved in fatty acid biosynthesis of chia seeds. **(A)** Phylogenetic tree of the *FAD3* genes. Shi: *Salvia hispanica*; AT: *Arabidopsis thaliana*; XM: *Salvia splendens*; GWHTAOSJ: *Salvia miltiorrhiza*; GWHTASIU: *Salvia bowleyana*; Tg: *Tectona grandis*; Jmimo: *Jacaranda mimosifolia*; Calam: *Callicarpa americana*; Am: *Antirrhinum majus*; Solyc: *Solanum lycopersicum*. **(B)** Expression pattern for *FatA*, *FatB*, *FAB2*, *FAD2*, *FAD3*, *FAD7/8*, and *FAD6* genes in developing chia seeds. DAF: Days after flower opening; FPKM: Fragments Per Kilobase of transcript per Million mapped reads. Only genes with maximum FPKM > 1 in seed samples were included in the plot. **(C)** A model for the biosynthesis of ALA in the chia genome. PC, phosphatidylcholine; ER, endoplasmic reticulum.

activity to AtFAB2 (SAD) (Troncoso-Ponce et al., 2016). The residue is predicted to locate at the bottom part of the substrate channel and the bulkier lateral chain of Phe may reduce the substrate binding pocket to better accommodate C16-ACP substrates. Multiple sequence alignment indicated that the highly expressed FAB2 homologs (*Shi017381*, *Shi000260*, and *Shi006361*) in chia seeds have a Tyr residue at the corresponding position, suggesting that they function as SADs (Supplemental Figure 16). In contrast, the two orthologs (*Shi015154* and *Shi026195*) of *AtFAD7/8*, the plastid localized omega-3 desaturase of Arabidopsis, were expressed at low to medium levels (FPKM values between 1.7 – 18.6) in developing seeds (Figure 4B; Supplemental Figure 13). In fact, multiple FA biosynthesis-related genes, such as genes encoding acyl carrier proteins (*Shi029800*, *Shi029801* and *Shi008432*), oil body-associated proteins (*Shi002948* and *Shi002148*), and lipid-transfer proteins (*Shi014949* and *Shi010250*), are also among the top 100 highly expressed genes in developing chia seeds (Supplemental Table 5). These results suggest a biosynthetic pathway involving plastid and ER localized enzymes, including *ShFAB2*, *ShFatA*, *ShFAD2* and *ShFAD3*, is responsible for the high ALA content in chia seeds (Figure 4C). Despite copy number variations were identified in some of these genes (Supplemental Table 4), we suggest that strong expression of fatty acid desaturase genes, particularly the ER localized FAD3s, are responsible for the high ALA content in chia seeds.

Discussion

De novo assembly of plant genomes has been greatly facilitated by the advancement of third-generation sequencing technologies that produce single-molecule long reads without the need of polymerase chain reactions. Commercially available 3rd-generation sequencing platforms suffer from high error rate of the raw reads (usually between 10-15%). The circular consensus sequencing (CCS) mode of PacBio significantly reduced consensus error rate by sequencing the same DNA insert multiple times. With carefully selected sizes of the DNA insert, a balance of sequencing length and accuracy can be achieved. In the current study, we performed CCS sequencing of the chia genomic DNA with a single SMRT cell, which produces 24.7 Gb of CCS data with median quality value of 31. The initial assembly included 666

contigs, while our analyses indicated that 623 of them originated from the organellar genomes or ribosome RNA repeats (Figure 1). The top 21 contigs have a total length of 361.7 Mb, which is slightly larger than the estimated genome size of 352.7 Mb based on k-mer analysis. Consistent with this high completeness of the nuclear genome, telomere repeats were identified at one or both ends of each of the six pseudochromosomes and rRNA repeats were identified in multiple chromosomes (Figure 2). Collapsing of repetitive regions was a common problem for *de novo* assembly of genomes with high repeat contents using longer but non-CCS PacBio reads. We did not observe similar phenomenon during the assembly of the chia genome. We reason that improved accuracy of the CCS mode helps resolving highly complex regions of the genome unless the repeat unit exceeds the read length, or the repeat sequences are highly similar.

Through phylogenetic and gene expression analyses, we identified candidate genes underlying high ALA contents of chia seeds. Two copies each of *ShFAB2*, *ShFAD2*, and *ShFAD3* exhibit very similar expression patterns (Figure 4B), suggesting these enzymes act together to promote the ALA content in chia seeds. This is consistent with the reported substrate channeling between FAD2 and FAD3 (Lou et al., 2014). Mature chia seeds have a lipid content of ~35%, of which up to 64% are ALA, the highest among all plant species (Muñoz et al., 2013; Kulczynski et al., 2019). Compared to its close relative, *S. splendens*, whose seeds were reported to have a ALA content of 34.5% and a LA content of 31.3% (Joh et al., 1988), the total content of ALA and LA of chia seeds are similar, suggesting that the elevated conversion rate from LA to ALA is the main event that drives high ALA content in chia seeds. In support of the idea that FAD3 is a rate limiting step in ALA biosynthesis, it was shown that overexpression of the rice *FAD3* gene is sufficient to increase the ALA content in seeds by ~28 fold (Liu et al., 2012). In addition to chia, seeds of flax (*Linum usitatissimum*) and perilla *(Perilla frutescens*) also have a relative ALA content around 60% (Ciftci et al., 2012). Although the genetic basis underlying their high ALA content remains to be determined, convergent high ALA contents in these species indicate that increasing omega-3 contents in seeds involve limited number of steps during evolution. This suggests a promising future for improving lipid composition in grains through transgenic or genome editing approaches.

Materials and methods

Library preparation and sequencing

Chia seeds were surface sterilized and grown in ½ MS medium supplemented with 0.7% agarose in a Percival growth chamber. Genomic DNA was extracted from two-week-old seedlings for genome survey sequencing and accurate consensus long-read sequencing (HiFi sequencing). The genome survey library was prepared and sequenced at the Genomics Core Facility of Shanghai Center for Plant Stress Biology following standard protocols. A 15-kb PacBio HiFi sequencing library were constructed and sequenced on a PacBio Sequel IIe platform at Berry Genomics (Beijing, China) following manufacturer's instructions. Etiolated 2-week-old seedlings were collected and used for crosslinking, proximity ligation, and library construction. The Hi-C library prepared by Biozeron (Shanghai, China) and sequenced at the Illumina NovaSeq platform with paired-end 150 bp sequencing mode.

Genome size estimation

To estimate the genome size of chia, 21 bp k-mer frequency of the PacBio HiFi reads was firstly counted with jellyfish (version 2.3.0) (Marcais and Kingsford, 2011). The k-mer frequency table was then used as input for GenomeScope2 (version 2.0) (Ranallo-Benavidez et al., 2020) to fit a diploid mathematical model to estimate the genome size, heterozygosity, and repetitiveness (Supplemental Figure 3).

Genome assembly

To assemble the nuclear genome using HiFi reads, three state-of-the-art genome assemblers were tested, including Flye (version 2.9) (Kolmogorov et al., 2019), HiCanu (version 2.2) (Nurk et al., 2020), and hifiasm (version 0.16.1) (Cheng et al., 2021). Flye applied a data structure of repeat graph (Kolmogorov et al., 2019). HiCanu was a modification of the Canu assembler (Koren et al., 2017) that was designed for HiFi reads with homopolymer compression, overlap-based error correction, and aggressive false overlap filtering (Nurk et al., 2020). Hifiasm is a genome assembler specifically designed for HiFi reads (Cheng et al., 2021). The previously estimated genome size was used as input parameter for Flye and HiCanu, while hifiasm does not require pre-estimated genome size. The results indicated that hifiasm with default parameters performed the best in terms of contiguity (Supplemental Table 6) and accuracy (Supplemental Figure 17).

To assemble the chia plastid genome, the GetOrganelle software (version 1.6.2) was used (Jin et al., 2020), which performs well in a comparison of chloroplast genome assembly tools (Freudenthal et al., 2020). GetOrganelle firstly extracted Illumina short reads that could be mapped to the embryophyte plastomes (a library composed of 101 plastid genomes) with bowtie2 (version 2.3.4.1) (Langmead and Salzberg, 2012) and then assembled them using SPAdes (version 3.13.0) (Bankevich et al., 2012). GetOrganelle produced three contigs representing the large single copy (LSC), small single copy (SSC) and inverted region (IR) of the chia plastid genome. Such three contigs were then aligned against the plastid genome of *Salvia miltiorrhiza* (accession number: NC_020431.1) (Qian et al., 2013), a close relative of chia. The alignment was performed with minimap2 (version 2.11) (Li, 2018) and visualized with D-Genies (version 1.3.1) (Cabanettes and Klopp, 2018). The three contigs were then ordered into a complete plastid genome using a customized Perl (version 5.34.0) script based on the BioPerl toolkit (version 1.7.4) (Stajich et al., 2002). Next, CHLOË (version

7c33699, https://chloe.plastid.org/) was used for the annotation of protein-coding genes, transfer RNAs, and ribosomal RNAs in the plastid genome.

To obtain the chia mitochondrial genome, we inspected contigs produced by hifiasm and found contig ptg000033c (length: 313,444 bp, read depth: 557) was circular and had the highest average read depth. Then we submitted this contig to the AGORA web tool (Jung et al., 2018) for genome annotation, with the protein-coding and rRNA genes of the *Salvia miltiorrhiza* mitochondrial genome (accession number: NC_023209.1) as a reference. The results of AGORA were then manually corrected by 1) removing protein-coding genes shorter than 30 amino acids, 2) removing protein-coding genes with pre-stop codons, 3) correcting mislabeled positions of ribosomal RNA genes. The chia mitochondrial genome was then visualized using OrganellarGenomeDRAW (OGDraw, version 1.3.1) (Greiner et al., 2019).

The "1-to-1" coverage and identity rate of contigs against the chia plastid and mitochondrial genomes were calculated using the dnadiff program of the MUMmer package (version 3.23) (Kurtz et al., 2004).

To obtain chia pseudochromosome sequences, the top 21 contigs in length and the Hi-C data was used for scaffolding. Illumina sequencing adapters and low-quality sequences of Hi-C data were trimmed by trim_galore (version 0.6.7, https://github.com/FelixKrueger/TrimGalore) with default parameters (quality score: 20; minimum length: 20 bp), which is a wrapper of cutadapt (version 3.4) (Martin, 2011). The clean Hi-C data were analyzed using Juicer (version 1.6) (Durand et al., 2016b), which produced high-quality DNA contact information. Then the 3D-DNA pipeline (version 180922) (Dudchenko et al., 2017) was used for ordering the contigs into pseudochromosomes. After visualizing the Hi-C contact map with Juicebox (version 1.9.1) (Durand et al., 2016a), we manually connect the contigs using "run-asm-pipeline-post-review.sh" of the 3D-DNA pipeline to avoid splitting the contigs.

Identification of rRNA repeats and telomere signatures

To predict the location of ribosomal RNA (rRNA) in the nuclear genome, Basic Rapid Ribosomal RNA Predictor (barrnap, version 0.9, https://github.com/tseemann/barrnap) was used, which using the nhmmer (version 3.1b1) (Wheeler and Eddy, 2013) to search the potential location of eukaryotes rRNA genes (5S, 5.8S, 28S, and 18S).

The telomere signature was examined using the program FindTelomeres (https://github.com/JanaSperschneider/FindTelomeres), which was a Python script for finding telomeric repeats (TTTAGGG/CCCTAAA). The results were further confirmed by TRF (version 4.09.1) (Benson, 1999) with parameters of "2 7 7 80 10 50 500 -m -d -h".

Genome circular plots were created in Circos (version 0.69.6) (Krzywinski et al., 2009). Dot plot of two genome assemblies was created using Assemblytics (Nattestad and Schatz, 2016). Visualization of the reads alignment file was performed using Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013).

Genome quality evaluation

The quality of the genome assembly was evaluated using three methods, including Benchmarking Universal Single-Copy Orthologs (BUSCO) (version 5.0.0) (Manni et al., 2021), LTR Assembly Index (LAI) (version 2.9.0) (Ou et al., 2018) and Merqury (version 1.3) (Rhie et al., 2020). Merqury is a tool for reference-free assembly evaluation. Additionally, Illumina short reads were mapped to chia genome assembly using bwa-mem (version 0.7.17) (Li, 2013). The mapping rate and error rate of the Illumina short reads were estimated by SAMtools (version 1.15.1) (Li et al., 2009).

Genome annotation

A combined method was used for chia gene prediction, including *ab initio* prediction, EST discovery and protein homology search. To predict gene models, we firstly masked the repeats using RepeatMasker (version 3.1.2-p1) (Tarailo-Graovac and Chen, 2009). A species-specific repeat library was constructed for RepeatMasker using Repeatmodeler2 (version 2.0.2) (Flynn et al., 2020) and LTR_retriever (version 2.9.0) (Ou and Jiang, 2018). The LTR candidates for LTR_retriever was identified by LTR_FINDER_parallel (version 1.1) (Ou and Jiang, 2019) and LTRharvest (version 1.6.0) (Ellinghaus et al., 2008). LTR_FINDER_parallel is a parallel wrapper of LTR_FINDER (version 1.07) (Xu and Wang, 2007). The chia transcriptome of 13 tissue types (involved seeds, cotyledon, shoots, leaves, internodes, racemes, and flowers) (Gupta et al., 2021) were retrieved from the NCBI SRA database (accession number: PRJEB19614) and *de novo* assembled using Trinity (version 2.11.0) (Grabherr et al., 2011). The assembled transcripts were used as expressed sequence tags (EST) evidence for further gene model prediction. Seven sets of protein sequences downloaded from public databases were used as protein homology evidences, including *Arabidopsis thaliana* (version Araport11) (Cheng et al., 2017), *Antirrhinum majus* (version IGDBV1) (Li et al., 2019), *Callicarpa americana* (Hamilton et al., 2020), *Salvia miltiorrhiza* (version 1.0) (Song et al., 2020), *Salvia splendens* (Dong et al., 2018), *Tectona grandis* (Zhao et al., 2019) and the UniprotKB/Swiss-Prot dataset (version release-2020_04) (Poux et al., 2017).

Maker (version 3.01.03) (Campbell et al., 2014) was run three rounds to train AUGUSTUS (version 3.4.0) (Stanke and Waack, 2003) and SNAP (version 2006-07-28) (Korf, 2004) gene prediction parameters. GeMoMa (version 1.8) (Keilwagen et al., 2019) and MetaEuk (release 5) (Levy Karin et al., 2020) were used with the above mentioned protein homology datasets to discover gene models. Finally, EVidenceModeler (EVM, version 1.1.1) (Haas et al., 2008) was used to combine all the above gene prediction evidences. The est2geome and protein2genome features produced

by Maker were used as transcript and protein evidence for EVM. The AUGUSTUS and SNAP gene models were used as *ab initio* prediction evidence for EVM. The GeMoMa and EetaEuk produced gene models were used as OTHER_PREDICTION evidence, which means they do not provide an indication of intergenic regions (Haas et al., 2008). As some of the gene models were overlapping with repetitive sequences, the final coding sequences and protein sequences were extracted from the unmasked genome assembly. Gene function annotation was performed by InterProScan (version 5.52-86.0) (Jones et al., 2014) and AHRD (version 3.3.3) (Boecker, 2021).

Genome evolution

Orthofinder (version 2.5.4) (Emms and Kelly, 2019) was used for the construction of orthologous groups. The STAG algorithm (Emms and Kelly, 2018) implemented in Orthofinder was used to estimate the species tree. Chia and other nine genomes were used for the construction of orthologous groups, including *Arabidopsis thaliana* (version Araport11) (Cheng et al., 2017), *Solanum lycopersicum* (version ITAG4.0) (Hosmani et al., 2019), *Antirrhinum majus* (version IGDBV1) (Li et al., 2019), *Tectona grandis* (Zhao et al., 2019), *Callicarpa americana* (Hamilton et al., 2020), *Jacaranda mimosifolia* (Wang et al., 2021), *Salvia bowleyana* (Zheng et al., 2021), *Salvia miltiorrhiza* (version 1.0) (Song et al., 2020), and *Salvia splendens* (version SspV2) (Jia et al., 2021). Gene family size expansion and contraction analysis was performed by CAFE5 (version 5.0.0) (Mendes et al., 2020). Synteny analysis was performed by the Python version of MCScan (version 1.1.17) (Tang et al., 2008). ParaAT (version 2.0) (Zhang et al., 2012) was used to prepare the alignment data for calculating Ks values, which was a wrapper of MUSCLE (version 3.8.1551) (Edgar, 2004) and PAL2NAL (version 13) (Suyama et al., 2006). KaKs_Calculator (version 2.0) (Wang et al., 2010) was used for calculating the Ks values using the YN model (Yang and Nielsen, 2000). The upset plot was created using the ggupset package (https://cran.r-project.org/package=ggupset) in R.

Gene expression analysis

Besides the chia transcriptome of 13 tissue types that retrieved from the NCBI SRA database (accession number: PRJEB19614) (Gupta et al., 2021), another set of transcriptome data for chia seed development was retrieved from the NCBI SRA database (accession number: PRJNA196477), which was sampled in 3, 7, 14, 21, and 28 DAF (Sreedhar et al., 2015). The raw RNA-seq data downloaded from the NCBI SRA database were firstly converted to FASTQ format using the fastq-dump command from the SRA Toolkit package (version 2.9.3, https://github.com/ncbi/sra-tools). Reads were then trimmed using trim_galore and then mapped to the chia reference genome by STAR (version 2.7.5c) (Dobin et al., 2013). Gene counts were summarized by featureCounts (version 2.0.1) (Liao et al., 2014). FPKM values were calculated using functions of the DESeq2 package (version 1.32.0) (Love et al., 2014) in the R platform (version 4.1.1) (R Core Team, 2021).

Multiple sequence alignment and phylogenetic tree construction

Visualization of multiple sequence alignment of the *FAD2, FAD3, FAD7, and FAD8* genes was performed using the Clustal Omega web tool (https://www.ebi.ac.uk/Tools/msa/clustalo/). Phylogenetic trees of the *FAB2*/*AAD*, *FAD2*, *FAD3*, *FAD7* and *FAD8* were constructed with the maximum likelihood method by IQ-TREE2 (Minh et al., 2020). The best-fitting amino acid substitution model was determined by ModelFinder (Kalyaanamoorthy et al., 2017).

Data availability statement

The datasets presented in this study can be found in online repositories. The genome assembly and corresponding sequencing data were deposited at NCBI (https://www.ncbi.nlm.nih.gov/) under accession number PRJNA864090 and at NGDC (https://ngdc.cncb.ac.cn/) under accession number PRJCA010915. The genome assembly and annotation data were deposited at CoGe (https://genomevolution.org/coge/) with genome ID 64745 for unmasked genome and genome ID 64746 for masked genome and figshare (https://doi.org/10.6084/m9.figshare.21976526).

Author contributions

LL performed data analyses; JS, MZ, and SI prepared plant materials; SI, YL, HeZ, and HuZ designed the project; LL and HeZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (31922008 to HeZ and 31900189 to LL), the Strategic Priority Research Program of CAS (XDB27040108 to HeZ), the Belt and Road Program of CAS (131965KYSB20190083-03 to HeZ), the Youth Innovation Promotion Association CAS (Y201844 to HeZ), and Central Guided Local Science and Technology Development Fund Project (YDZX2021079 to HuZ).

Acknowledgments

We thank the Core Facility for Genomics of Shanghai Center for Plant Stress Biology (PSC) for the construction of the Illumina sequencing library and the Core Facility for Bioinformatics of PSC for the maintenance of the high-performance computing (HPC) clusters.

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/fpls.2023.1102715/full#supplementary-material

References

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., et al. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36, 2251–2252. doi: 10.1093/bioinformatics/btz859

Baker, E. J., Miles, E. A., Burdge, G. C., Yaqoob, P., and Calder, P. C. (2016). Metabolism and functional effects of plant-derived omega-3 fatty acids in humans. *Prog. Lipid Res.* 64, 30–56. doi: 10.1016/j.plipres.2016.07.002

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Boecker, F. (2021). *AHRD: Automatically annotate proteins with human readable descriptions and gene ontology terms* (Germany: University of Bonn, Bonn).

Burns-Whitmore, B., Froyen, E., Heskey, C., Parker, T., and San Pablo, G. (2019). Alpha-linolenic and linoleic fatty acids in the vegan diet: Do they require dietary reference Intake/Adequate intake special consideration? *Nutrients* 11, 2365. doi: 10.3390/nu11102365

Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958. doi: 10.7717/peerj.4958

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-p. *Curr. Protoc. Bioinf.* 48, 4 11 11–39. doi: 10.1002/0471250953.bi0411s48

Cassiday, L. (2017). Chia: superfood or superfad? *Inform* 28, 6–13. doi: 10.21748/inform.01.2017.06

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 89, 789–804. doi: 10.1111/tpj.13415

Ciftci, O. N., Przybylski, R., and Rudzińska, M. (2012). Lipid components of flax, perilla, and chia seeds. *Eur. J. Lipid Sci. Technol.* 114, 794–800. doi: 10.1002/ejlt.201100207

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dong, A. X., Xin, H. B., Li, Z. J., Liu, H., Sun, Y. Q., Nie, S., et al. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience* 7, giy068. doi: 10.1093/gigascience/giy068

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016a). Juicebox provides a visualization system for Hi-c contact maps with unlimited zoom. *Cell Syst.* 3, 99–101. doi: 10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016b). Juicer provides a one-click system for analyzing loop-resolution Hi-c experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 18. doi: 10.1186/1471-2105-9-18

Emms, D., and Kelly, S. (2018). STAG: species tree inference from all genes. *BioRxiv*, 267914. doi: 10.1101/267914

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

Estilai, A., Hashemi, A., and Truman, K. (1990). Chromosome number and meiotic behavior of cultivated chia, salvia hispanica (Lamiaceae). *HortScience* 25, 1646–1647. doi: 10.21273/HORTSCI.25.12.1646

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

Freudenthal, J. A., Pfaff, S., Terhoeven, N., Korte, A., Ankenbrand, M. J., and Forster, F. (2020). A systematic comparison of chloroplast genome assembly tools. *Genome Biol.* 21, 254. doi: 10.1186/s13059-020-02153-6

Gene Ontology, C. (2021). The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 49, D325–D334. doi: 10.1093/nar/gkaa1113

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Gupta, P., Geniza, M., Naithani, S., Phillips, J. L., Haq, E., and Jaiswal, P. (2021). Chia (*Salvia hispanica*) gene expression atlas elucidates dynamic spatio-temporal changes associated with plant growth and development. *Front. Plant Sci.* 12, 667678. doi: 10.3389/fpls.2021.667678

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

Hamilton, J. P., Godden, G. T., Lanier, E., Bhat, W. W., Kinser, T. J., Vaillancourt, B., et al. (2020). Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing lamiaceae species, *Callicarpa americana*. *Gigascience* 9, giaa093. doi: 10.1093/gigascience/giaa093

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* 32, 835–845. doi: 10.1093/molbev/msv037

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., et al. (2019). An improved *de novo* assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-c proximity ligation and optical maps. *BioRxiv*, 767764. doi: 10.1101/767764

Jia, K. H., Liu, H., Zhang, R. G., Xu, J., Zhou, S. S., Jiao, S. Q., et al. (2021). Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome. *Hortic. Res.* 8, 177. doi: 10.1038/s41438-021-00614-y

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Joh, Y.-G., Lee, O.-K., and Lim, Y.-J. (1988). Studies on the composition of fatty acid in the lipid classes of seed oils of the labiatae family. *J. Korean Appl. Sci. Technol.* 5, 13–23. doi: 10.12925/jkocs.1988.5.1.2

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

Jung, J., Kim, J. I., Jeong, Y. S., and Yi, G. (2018). AGORA: organellar genome annotation from the amino acid and nucleotide references. *Bioinformatics* 34, 2661–2663. doi: 10.1093/bioinformatics/bty196

Kachroo, A., Shanklin, J., Whittle, E., Lapchyk, L., Hildebrand, D., and Kachroo, P. (2007). The *Arabidopsis* stearoyl-acyl carrier protein-desaturase family and the

contribution of leaf isoforms to oleic acid synthesis. *Plant Mol. Biol.* 63, 257–271. doi: 10.1007/s11103-006-9086-y

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kazaz, S., Barthole, G., Domergue, F., Ettaki, H., To, A., Vasselon, D., et al. (2020). Differential activation of partially redundant Delta9 stearoyl-ACP desaturase genes is critical for omega-9 monounsaturated fatty acid biosynthesis during seed development in arabidopsis. *Plant Cell* 32, 3613–3637. doi: 10.1105/tpc.20.00554

Keilwagen, J., Hartung, F., and Grau, J. (2019). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* 1962, 161–177. doi: 10.1007/978-1-4939-9173-0_9

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5, 59. doi: 10.1186/1471-2105-5-59

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kulczynski, B., Kobus-Cisowska, J., Taczanowski, M., Kmiecik, D., and Gramza-Michalowska, A. (2019). The chemical composition and nutritional value of chia seeds-current state of knowledge. *Nutrients* 11, 1242. doi: 10.3390/nu11061242

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi: 10.1186/gb-2004-5-2-r12

Lands, B. (2014). Historical perspectives on the impact of n-3 and n-6 nutrients on health. *Prog. Lipid Res.* 55, 17–29. doi: 10.1016/j.plipres.2014.04.002

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Levy Karin, E., Mirdita, M., and Soding, J. (2020). MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8, 48. doi: 10.1186/s40168-020-00808-x

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*. 1303.3997. doi: 10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

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

Li, D., Wahlqvist, M. L., and Sinclair, A. J. (2019). Advances in n-3 polyunsaturated fatty acid nutrition. *Asia Pac J. Clin. Nutr.* 28, 1–5. doi: 10.6133/apjcn.201903_28(1).0001

Li, M., Zhang, D., Gao, Q., Luo, Y., Zhang, H., Ma, B., et al. (2019). Genome structure and evolution of *Antirrhinum majus* l. *Nat. Plants* 5, 174–183. doi: 10.1038/s41477-018-0349-9

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., et al. (2013). Acyl-lipid metabolism. *Arabidopsis Book* 11, e0161. doi: 10.1199/tab.0161

Liu, H. L., Yin, Z. J., Xiao, L., Xu, Y. N., and Qu le, Q. (2012). Identification and evaluation of omega-3 fatty acid desaturase genes for hyperfortifying alpha-linolenic acid in transgenic rice seed. *J. Exp. Bot.* 63, 3279–3287. doi: 10.1093/jxb/ers051

Lou, Y., Schwender, J., and Shanklin, J. (2014). FAD2 and FAD3 desaturases form heterodimers that facilitate metabolic channeling *in vivo*. *J. Biol. Chem.* 289, 17996–18007. doi: 10.1074/jbc.M114.572883

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, 550. doi: 10.1186/s13059-014-0550-8

Manni, M., Berkeley, M. R., Seppey, M., Simao, 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, 4647–4654. doi: 10.1093/molbev/msab199

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*. 36, 5516–5518 doi: 10.1093/bioinformatics/btaa1022

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Muñoz, L. A., Cobos, A., Diaz, O., and Aguilera, J. M. (2013). Chia seed (*Salvia hispanica*): an ancient grain and a new functional food. *Food Rev. Int.* 29, 394–408. doi: 10.1080/87559129.2013.818014

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023. doi: 10.1093/bioinformatics/btw369

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. doi: 10.1101/gr.263566.120

Ohlrogge, J., and Browse, J. (1995). Lipid biosynthesis. *Plant Cell* 7, 957–970. doi: 10.1105/tpc.7.7.957

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* 10, 48. doi: 10.1186/s13100-019-0193-0

Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C. H., Lu, Z., et al. (2017). On expert curation and scalability: UniProtKB/Swiss-prot as a case study. *Bioinformatics* 33, 3454–3460. doi: 10.1093/bioinformatics/btx439

Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., et al. (2013). The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PloS One* 8, e57607. doi: 10.1371/journal.pone.0057607

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi: 10.1038/s41467-020-14998-3

R Core Team (2021). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. doi: 10.1186/s13059-020-02134-9

Shahidi, F., and Ambigaipalan, P. (2018). Omega-3 polyunsaturated fatty acids and their health benefits. *Annu. Rev. Food Sci. Technol.* 9, 345–381. doi: 10.1146/annurev-food-111317-095850

Simopoulos, A. P. (2002a). The importance of the ratio of omega-6/omega-3 essential fatty acids. *BioMed. Pharmacother.* 56, 365–379. doi: 10.1016/S0753-3322(02)00253-6

Simopoulos, A. P. (2002b). Omega-3 fatty acids in inflammation and autoimmune diseases. *J. Am. Coll. Nutr.* 21, 495–505. doi: 10.1080/07315724.2002.10719248

Song, Z., Lin, C., Xing, P., Fen, Y., Jin, H., Zhou, C., et al. (2020). A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* 13, e20041. doi: 10.1002/tpg2.20041

Sreedhar, R. V., Kumari, P., Rupwate, S. D., Rajasekharan, R., and Srinivasan, M. (2015). Exploring triacylglycerol biosynthetic pathway in developing seeds of chia (*Salvia hispanica* l.): a transcriptomic approach. *PloS One* 10, e0123580. doi: 10.1371/journal.pone.0123580

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2, ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 4, 10. doi: 10.1002/0471250953.bi0410s25

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Troncoso-Ponce, M. A., Barthole, G., Tremblais, G., To, A., Miquel, M., Lepiniec, L., et al. (2016). Transcriptional activation of two delta-9 palmitoyl-ACP desaturase genes by MYB115 and MYB118 is critical for biosynthesis of omega-7 monounsaturated fatty acids in the endosperm of arabidopsis seeds. *Plant Cell* 28, 2666–2682. doi: 10.1105/tpc.16.00612

Valdivia-López, MÁ, and Tecante, A. (2015). Chia (*Salvia hispanica*): A review of native Mexican seed and its nutritional and functional properties. *Adv. Food Nutr. Res.* 75, 53–75. doi: 10.1016/bs.afnr.2015.06.002

Wang, L., Lee, M., Sun, F., Song, Z., Yang, Z., and Yue, G. H. (2022). A chromosome-level genome assembly of chia provides insights into high omega-3 content and coat color variation of its seeds. *Plant Commun.* 3, 100326. doi: 10.1016/j.xplc.2022.100326

Wang, M., Zhang, L., and Wang, Z. (2021). Chromosomal-level reference genome of the Neotropical tree *Jacaranda mimosifolia* d. don. *Genome Biol. Evol.* 13, evab094. doi: 10.1093/gbe/evab094

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wheeler, T. J., and Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Xue, Y., Chen, B., Win, A. N., Fu, C., Lian, J., Liu, X., et al. (2018). Omega-3 fatty acid desaturase gene family from two omega-3 sources, salvia hispanica and perilla frutescens: Cloning, characterization and expression. *PloS One* 13, e0191432. doi: 10.1371/journal.pone.0191432

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43. doi: 10.1093/oxfordjournals.molbev.a026236

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhao, D., Hamilton, J. P., Bhat, W. W., Johnson, S. R., Godden, G. T., Kinser, T. J., et al. (2019). A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* 8, giz005. doi: 10.1093/gigascience/giz005

Zheng, X., Chen, D., Chen, B., Liang, L., Huang, Z., Fan, W., et al. (2021). Insights into salvianolic acid b biosynthesis from chromosome-scale assembly of the salvia bowleyana genome. *J. Integr. Plant Biol.* 63, 1309–1323. doi: 10.1111/jipb.13085

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 March 2023
DOI 10.3389/fpls.2023.1133616

OPEN ACCESS

EDITED BY
Mark Chapman,
University of Southampton,
United Kingdom

REVIEWED BY
Fangyuan Zhang,
Southwest University, China
Huasheng Peng,
China Academy of Chinese Medical Sciences, China
Daike Tian,
Shanghai Chenshan Plant Science Research Center (CAS), China

*CORRESPONDENCE
Ying Wang
yingwang@scib.ac.cn
Wei Sun
wsun@icmm.ac.cn

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 29 December 2022
ACCEPTED 07 March 2023
PUBLISHED 27 March 2023

CITATION
Mi Y, He R, Wan H, Meng X, Liu D, Huang W, Zhang Y, Yousaf Z, Huang H, Chen S, Wang Y and Sun W (2023) Genetic and molecular analysis of the anthocyanin pigmentation pathway in *Epimedium*. *Front. Plant Sci.* 14:1133616. doi: 10.3389/fpls.2023.1133616

Genetic and molecular analysis of the anthocyanin pigmentation pathway in *Epimedium*

Yaolei Mi[1], Ruikun He[2], Huihua Wan[1], Xiangxiao Meng[1], Di Liu[3], Wenjun Huang[4], Yanjun Zhang[4], Zubaida Yousaf[5], Hongwen Huang[6], Shilin Chen[1,7], Ying Wang[8]* and Wei Sun[1,7]*

[1]Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [2]By-Health Institute of Nutrition and health. By-health Co., Ltd. Guangzhou, Guangdong, China, [3]College of Pharmacy, Hubei University of Chinese Medicine, Wuhan, Hubei, China, [4]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, China, [5]Department of Botany, Lahore College for Women University, Lahore, Pakistan, [6]Lushan Botanical Garden, Chinese Academy of Sciences, Jiujiang, Jiangxi, China, [7]Institute of Herbgenomics, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China, [8]South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, Guangdong, China

Introduction: Flower color is an ideal trait for studying the molecular basis for phenotypic variations in natural populations of species. *Epimedium* (Berberidaceae) species exhibit a wide range of flower colors resulting from the varied accumulation of anthocyanins and other pigments in their spur-like petals and petaloid sepals.

Methods: In this work, the anthocyanidins of eight different *Epimedium* species with different floral pigmentation phenotypes were analyzed using HPLC. Twelve genes involved in anthocyanin biosynthesis were cloned and sequenced, and their expression was quantified.

Results: The expression levels of the catalytic enzyme genes DFR and ANS were significantly decreased in four species showing loss of floral pigmentation. Complementation of EsF3'H and EsDFR in corresponding *Arabidopsis* mutants together with overexpression of EsF3'5'H in wild type *Arabidopsis* analysis revealed that these genes were functional at the protein level, based on the accumulation of anthocyanin pigments.

Discussion: These results strongly suggest that transcriptional regulatory changes determine the loss of anthocyanins to be convergent in the floral tissue of *Epimedium* species.

KEYWORDS

gene expression, *Epimedium*, anthocyanin, spur, sepal

Introduction

Accumulation of the secondary metabolite anthocyanin is predominantly responsible for red, blue, and purple pigmentation in angiosperms. Pigmentation is a major determinant of a species' pollination syndrome, which refers to the selection of particular floral traits caused by the preference of their pollinators (Fenster et al., 2004). Flower color is intricately regulated by the specific combinations of certain pigment metabolites produced, and is subjected to ecological selection and convergent evolution. Therefore, flower color is an ideal trait for examining ecological and evolutionary selection processes. The anthocyanin biosynthetic pathway (ABP) has been well established in many model species, such as *Arabidopsis*, petunia (*Petunia hybrida* E. Vilm.), and snapdragon (*Antirrhinum majus* L.) (Buer et al., 2010; Pollastri & Tattini, 2011). Most of the knowledge of anthocyanin biosynthesis in *Arabidopsis* has been obtained from the analysis of transparent testa (*tt*) mutants, which show loss of seed pigmentation (Lepiniec et al., 2006). In the early steps of the pathway, the key enzymes chalcone synthase (CHS), chalcone isomerase (CHI), and flavanone 3-hydroxylase (F3H) condense and convert a phenylpropanoid precursor, *p*-coumaroyl-CoA, along with three molecules of malonyl CoA, to dihydrokaempferol (Lepiniec et al., 2006). Parallel catalyzation by flavonoid-3'-hydroxylase, flavonoid-3',5'-hydroxylase, dihydroflavonol-4-reductase (DFR), and anthocyanidin synthase (ANS) results in the production of various types of anthocyanidin (Figure 1) (Holton & Cornish, 1995; Boss et al., 1996). The transcriptional regulators controlling flavonoid biosynthetic enzymes have been extensively studied, and include the MYB, the bHLH, and the WD-repeat proteins. Yeast-three-hybrid protein interaction data suggested that a protein complex of the MYB-bHLH-WD40 transcription factors binds the regulatory promoter regions of the flavonoid pathway enzymatic, or structural, genes, to regulate anthocyanin biosynthesis (Gonzalez et al., 2008).

The evolutionary basis for the loss of anthocyanin pigments in floral tissue has been investigated by characterizing major floral pigmentation loci using controlled cross segregating populations (Schwinn, 2006; Whittall et al., 2006; Hoballah et al., 2007; Streisfeld and Rausher, 2009; Smith and Rausher, 2011). Evidence suggests that flower color transition is affected by the transcriptional regulation of several anthocyanin structural genes expression. For example, altered activity of specific transcriptional factors accounts for altered patterns of pigmentation in white *Petunia axillaris* and some *Antirrhinum* species (Schwinn, 2006; Hoballah et al., 2007). *Cis*-regulatory changes in the *F3'H* gene promoter cause down-regulation of *F3'H* transcription and altered flux in the anthocyanin pathway, resulting in increased production of the red pigment, pelargonidin, instead of blue, in *Ipomoea horsfalliae* Hook. (Des Marais & Rausher, 2010). Although it has been suggested that mutations in structural genes may incur higher deleterious pleiotropy than those in *cis*-regulatory elements or transcription

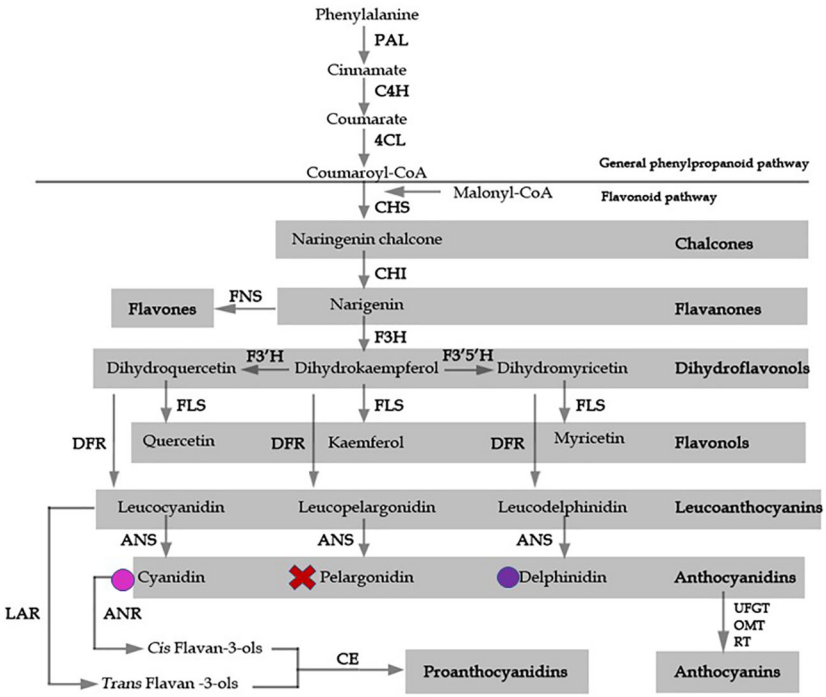


FIGURE 1
A model for flavonoid biosynthesis in *Epimedium* flowers based on classic investigation. Pathway enzymes are listed as an abbreviation beside arrows, and include 4CL, 4-coumarate: coenzyme A ligase; ANS, anthocyanidin synthase; C4H, cinnamate-4-hydroxylase; CHS, chalcone synthase, CHI, chalcone isomerase; DFR, dihydroflavonol 4-reductase; F3H, flavavone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; F3'5'H, flavonoid 3'5' hydroxylase; FLS, flavonol synthase; FNS, flavone synthase; LAR, leucoanthocyanidin reductase; UFGT, UDP flavonoid gulcosyl transferase; OMT, O-methyltransferase; PAL, phenylalanine ammonia lyase; RT, rhamnosyl transferase. The products of each enzymatic reaction are listed below the arrows. Colored circles indicate the presence of delphinidin and cyanidin anthocyanidins, X represents absence of pelargonidin.

factors, the possibility that enzyme coding sequence variation is involved in flower color transition cannot be excluded (Streisfeld and Rausher, 2009).

The *Epimedium* genus (Berberidaceae), known as "Yinyang Huo" by Chinese druggists, is one of the most popular traditional Chinese medicinal herb genera (Sun et al., 2014; Zhang et al., 2021). A monophyletic group of 50 species of *Epimedium* is found in western and central China (Huang et al., 2013a; Huang et al., 2013b; Huang et al., 2015; Huang et al., 2016). *Epimedium* species display a vast range of flower colors; from white and yellow to rose, crimson, and violet (Figure 2). These color pigments are distributed in petaloid sepals or petals or both. In this study, we studied the phenotypic variation of color in *Epimedium* species distributed in the Hubei province of China. The expression of genes involved in the anthocyanin biosynthetic pathway (ABP) was also analyzed for the association with the different flower color polymorphisms. Our study focused on answering two questions: (1) Has anthocyanin pigment loss, or variation, in different species resulted from the same mechanism? (2) Which candidate genes are involved in anthocyanin pigmentation in *E. sagittatum*?

Results

Analysis of pigments and flavonoid intermediates in different *Epimedium* species

Using HPLC, the major pigments from the floral tissues of anthocyanin species (A+) species were found to comprise delphinidin and cyanidin, whereas no detectable anthocyanins were found in the non-anthocyanin species (A-) flowers (Figure 3). To further characterize the mechanism responsible for the non-pigmentation of flowers in A- species, *E. sagittatum* was used as a model for the enzymatic function.

Expression of ABP genes in floral tissues of *Epimedium*

To determine whether changes in gene expression might be involved in the non-pigmentation phenotype of A- species, the transcript levels of putative anthocyanin biosynthetic enzymes were examined in petal tissue (Figure 4; Supplementary Figure 1A). Expression of *CHS1* not *CHS2* and *CHS3* was found to be significantly lower in *E. wushanese* (A-) than in other species. Similarly, down-regulation of *CHS2* was observed in *E. frachetii* (A-), suggesting that loss of anthocyanin may result from low levels of expression of different copies of *CHS* in *E. wushanese* and *E. franchetii*. For *CHI* and *F3H*, we found no significant correlation between expression level and the loss of anthocyanins in spur tissues of all A- species. Among the structural genes, *ANS* was the only ABP locus where all A- species had significantly lower expression levels than that of A+ species. This suggests that the lack of pigmentation production in all A- species could be caused primarily by lower *ANS* expression. The expression level of *DFR* was significantly lower in A- species than in A+ species, except for *E. lishihchenii*. It has been reported that substrate competition between *FLS* and *DFR* creates a metabolic flux of the flavonoid biosynthetic pathway in *Arabidopsis*. In this study, low expression of *DFR* in the A- species *E. franchetii* and *E. wushanese* was correlated with increased accumulation of *FLS* expression. On the other hand, up-regulation of *DFR* was positively correlated with *FLS* expression in *E. lishihchenii* but *E. sagittatum* showed no correlation with *DFR*. In summary, these results suggest that the loss of anthocyanin in *E. frachetii*, *E. wushanese* and *E. sagittatum*

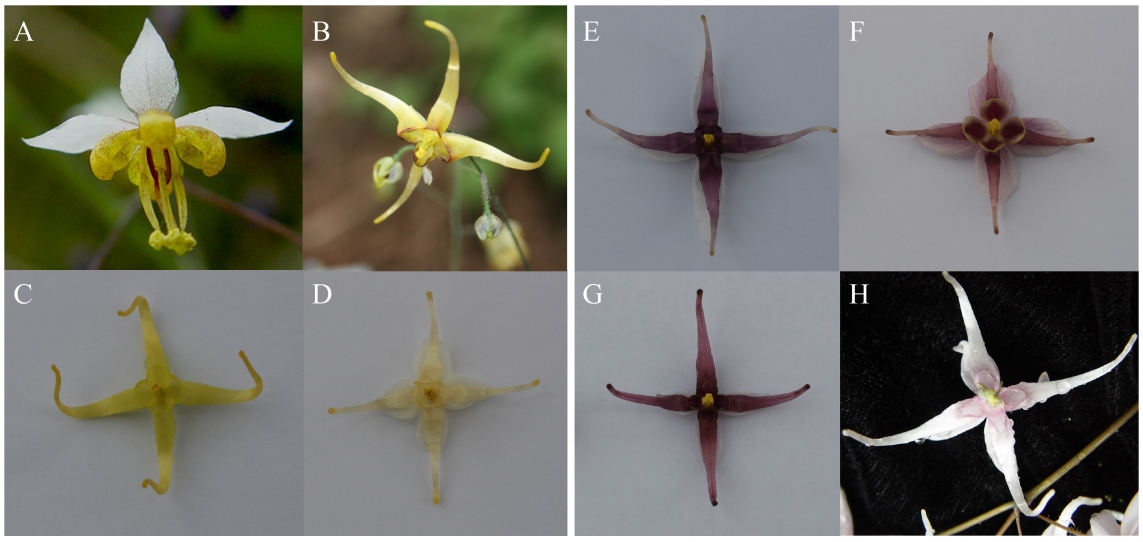


FIGURE 2
Floral phenotypes of accessions of different species within the genus *Epimedium*. **(A–D)** are non-pigmentation species (A-); **(E–H)** are classified as pigmentation species (A+). All photos were taken by W. S. **(A)** *E. sagittatum*, **(B)** *E. lishihchenii*, **(C)** *E. franchetii*, **(D)** *E. wushanense*, **(E)** *E. zhushanense*, **(F)** *E. epstenii*, **(G)** *E. acuminatum*, **(H)** *E. leptorrhizum*.

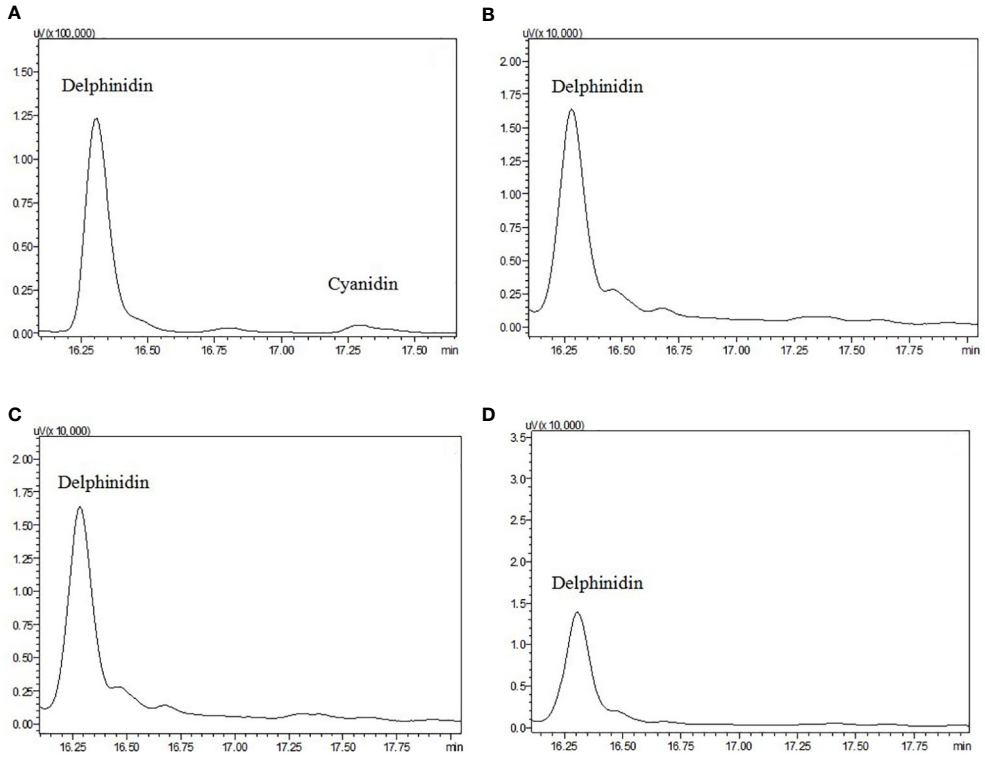


FIGURE 3
High-performance liquid chromatograms of extracts from petals and sepals of anthocyanidin pigments from **(A)** *E. acuminatum*, **(B)** *E. epstenii*, **(C)** *E. zhushanense* and **(D)** *E. leptorrhizum*. Peaks labeled represent the standards of each of the anthocyanin.

may be primarily related to alterations at the *ANS* locus, affecting gene expression.

To further analyze the loss of anthocyanin in sepals (Figure 5; Supplementary Figure 1B), gene expression was analyzed across five *Epimedium* species using the same primers. Expression of *DFR* was lowest in the three A- species and was correlated with *ANS* expression, suggesting the expression of *DFR* and *ANS* could be regulated by a common transcription factor. *CHS1* transcripts were not detected in the sepals of *E. wushanese*, which also had the lowest *CHS1* expression in petals. These observations suggest that negative regulation of the *DFR* and *ANS* genes together was also correlated with the lowest *CHS1* in sepals and petals in *E. wushanese*.

Complementation analyses

To study the catalytic activity of *E. sagittatum* ABP gene products, 35S::*EsF3'H* and 35S::*EsDFR* genes were individually transferred into their respective *Arabidopsis* mutants; *transparent testa 7* (*tt7)* lacking flavonoid 3'-hydroxylase, and *transparent testa3* (*tt3*) lacking dihydroflavonol reductase under the control of the cauliflower mosaic virus 35S promoter (Peer et al., 2001). Transgenic and mutant control seedlings were grown under nitrogen stress to determine if the *Epimedium* genes could rescue the *Arabidopsis* anthocyanin-null mutant phenotypes. Accumulation of anthocyanins was observed in transgenic seedlings ectopically expressing *EsF3'H* and *EsDFR* (Figure 6). However, the *tt7* and *tt3* mutant controls did not exhibit anthocyanin accumulation in cotyledons. Thus the *E. sagittatum* genes showed catalytic activity in *Arabidopsis*. Given the lack of an *Arabidopsis* mutant for *F3'5'H*, in order to determine whether *EsF3'5'H* can function *in vivo*, we overexpressed *35S::EsF3'5'H* in wild-type *Arabidopsis*. Under normal conditions on 1/2 MS medium, the seedlings overexpressing *EsF3'5'H* showed comparable anthocyanin production to wild-type controls (Figure 6).

Discussion

The four *Epimedium* A- species (*E. sagittatum*, *E. lishihchenii*, *E. franchetii* and *E. wushanense*) investigated in this study appeared to exhibit anthocyanin loss at the phenotypic level *via* reduced activity of the anthocyanin branch of the flavonoid pathway. In all species, this appears to involve reduced transcriptional activity of pathway genes, similar to studies in *Mimulus aurantiacus* (Streisfeld & Rausher, 2009). Interestingly, one A- species (*E. lishihchenii*) expressed all ABP loci except for *ANS* at a high level.

While the data linking conserved gene regulation changes to anthocyanin level changes are purely correlative, we found no evidence for the role of coding-region mutations in determining different anthocyanin levels. In A- specie *E. sagittatum*, the F3'H and DFR enzymes were shown to rescue anthocyanin production in their corresponding *Arabidopsis* mutants, suggestive of adequate catalytic function (Huang et al., 2012). Accumulation of

10.3389/fpls.2023.1133616

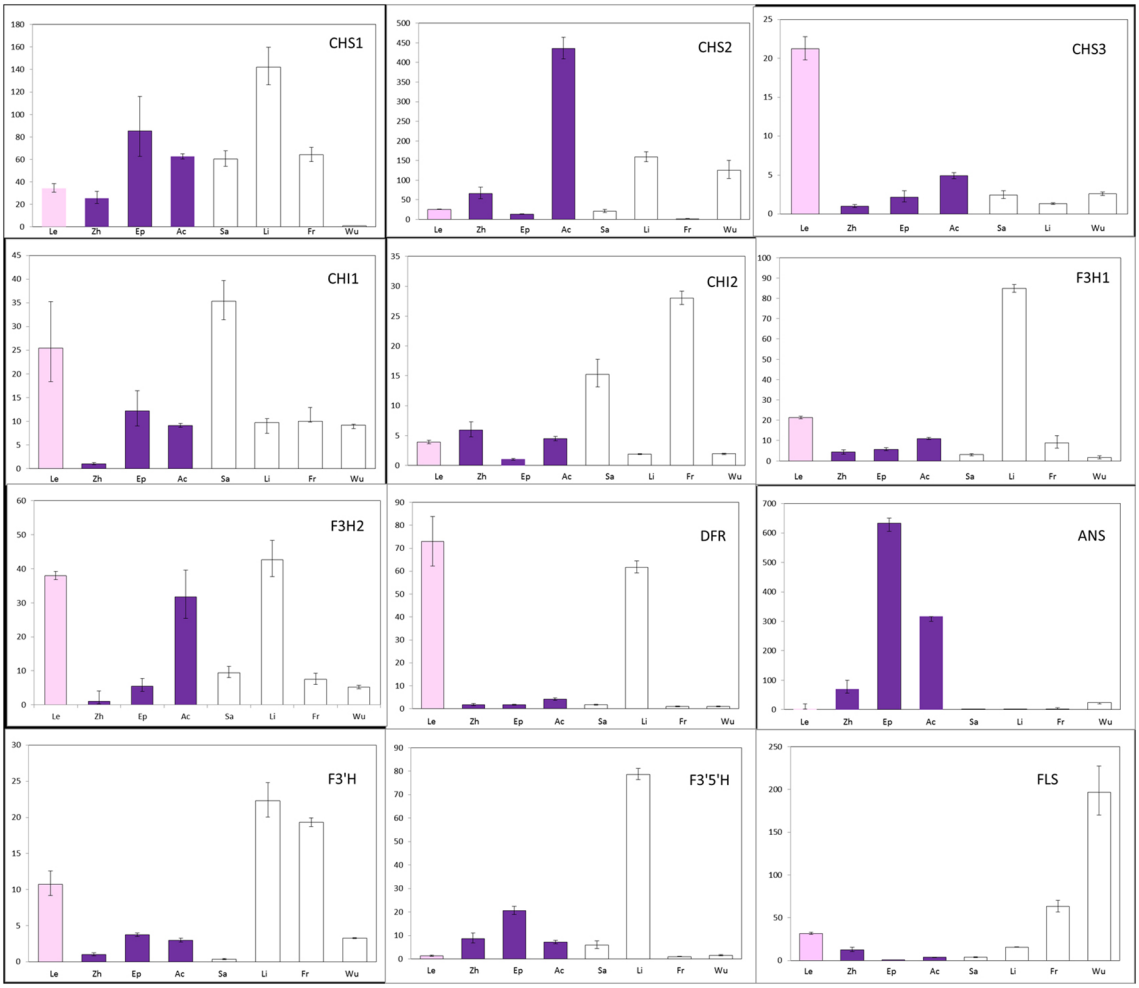


FIGURE 4
Quantitative expression pattern of ABP structural genes from petal tissue of eight *Epimedium* species. Colored and empty bars represent A+ and A- species, respectively. Le, Zh, Ep, Ac, Sa, Li, Fr and Wu represent *E. leptorrhizum*, *E. zhushanense*, *E. epstenii*, *E.acuminatum*, *E. sagittatum*, *E. lishihchenii*, *E. franchetii* and *E. wushanense*. Data presented here are the mean values of three replicates with error bars indicating SE.

anthocyanin in *35S::EsDFR* in this study and *35S::EsMYBA1* transformed *Arabidopsis* indicated functionality of the *EsDFR* and *EsMYBA1* coding region (Huang et al., 2013a). Thus, we concluded that the loss of flower color in *E. sagittatum* (A-) was due to a tissue-specific regulatory change affecting *EsDFR* and *EsANS* transcription and not coding-region mutations of *EsDFR*, *EsANS*, or *EsMYBA1*. We also suggested that the changes responsible for the loss of pigmentation in *E. franchetii* and *E. wushanense* flowers were shared with *E. sagittatum*, based on similar correlative gene expression patterns and anthocyanin production in leaves. Functional assays of the putative *cis*-elements and trans-regulators involved in *DFR* or *ANS* transcription were required to determine the precise regulatory mechanisms resulting in reduced *ANS* and *DFR* gene expression in A- species.

Downregulation of *CHS* was a major cause of white flowers in natural populations of *Aquilegia flavellata* and *Parrya nudicaulis* (Whittall et al., 2006; Dick et al., 2011). Although we found an association between the A- phenotype and downregulation of *CHS1* in yellow-flowered *E. wushanese*, *DFR*, and *ANS* were also downregulated, which may also have contributed to the A- phenotype. Therefore, the A- phenotype in four *Epimedium* species was also proposed to be due to alteration at the regulatory level, rather than functional mutations in ABP enzymes. The loci regulating anthocyanin in *Epimedium* were currently being fine-mapped and confirmed by transformation assays.

Materials and methods

Tissue harvest

Eight *Epimedium* species (*E. acuminatum*, *E. franchetii*, *E. leptorrhizum*, *E. epstenii*, *E. sagittatum*, *E. lishihchenii*, *E. wushanense*, *and E. zhushanense*) grown in the specialized *Epimedium* nurseries of Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China (Figure 2). All plants were transplanted from wild populations and growing under the same environmental conditions. Floral tissues including petaloid sepals and spur-like petals were collected in the spring of 2011. The eight species were separated into two groups corresponding to anthocyanin

10.3389/fpls.2023.1133616

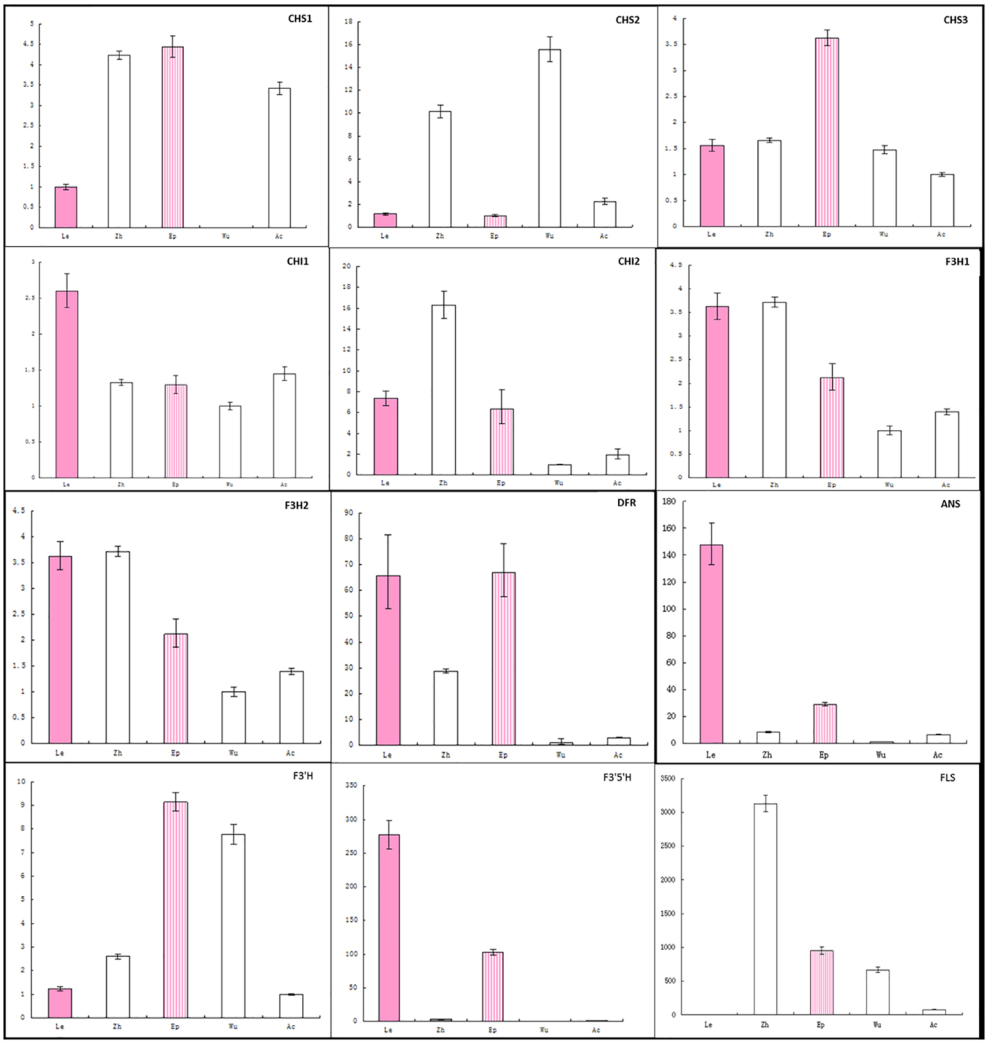


FIGURE 5
Analysis of expression profiles of anthocyanin genes in petaloid sepals of five *Epimedium* species using real-time PCR. The cDNA templates are listed as follows: Le, *E. leptorrhizum*; *Zh*, *E. zhushanense*; *Ep*, *E. epstenii*; Wu, *E. wushanense* and Ac, *E. acuminatum*. Data presented here are the mean values of three replicates with error bars indicating SE.

(A+) (*E. acuminatum*, *E. leptorrhizum, E. epstenii* and *E. zhushanense*) and non-anthocyanin (A-) (*E. franchetii*, *E. sagittatum*, *E. lishihchenii*, *E. wushanense*) based on visual observation of the floral tissues. The samples were weighed, packaged in aluminum foil, flash-frozen in liquid nitrogen, and then stored at -80°C.

HPLC analysis of flavonoid intermediates and anthocyanin

The profiles of anthocyanidins from the samples of A- species and A+ species were analyzed using HPLC. The precursors of anthocyanin pigments were extracted from 100 mg of fresh corolla tissue. For each sample, 20 μL of supernatant was injected into a Shimadzu LC-20 AT liquid chromatograph (Shimadzu Corporation, Japan) and a 250×4.6 mm reverse phase C18 column (Sigma-Aldrich, USA) at a flow rate of 1 ml min^{-1}. The organic solvent was composed of acetonitrile and 0.1% trifluoroacetic acid, and the polar solvent was 0.1% trifluoroacetic acid in HPLC-grade water. The anthocyanin was measured at 550 nm. The chemical compounds cyanidin, delphinidin, malvidin, pelargonidin, peonidin, and petunidin (Poypehenols Laboratories, Norway), were used as anthocyanidin standards.

Transferring ABP candidate genes into other *Epimedium* species

In total, 12 genes from *E. sagittatum* involved in the ABP were cloned following RT-PCR amplification using degenerate primers or specific primers based on our previous investigation(Zeng et al., 2010; Huang et al., 2013a; Huang et al., 2013b; Huang et al., 2015). These genes were *CHS1*, *CHS2*, *CHS3*, *CHI1*, *CHI2*, *F3H1*, *F3H2*, *F3'H*, *F3'5'H*, *FLS*, *DFR*, *ANS*. In this study, all pairs of primer from *E. sagittatum* were transferred to other *Epimedium* species.

FIGURE 6
Phenotypes of the *Arabidopsis* with overexpression of *EsF3'H*, *EsDFR and F3'5'H*(**A**) Image of anthocyanin in *tt7* mutant and rescuing line of *EsF3'H* in *tt7* background, (**B**) Phenotypes of wild-type, and transgenic *Arabidopsis* seedling with *EsF3'5'H*, (**C**) Phenotypes of *tt3* mutant, and transgenic *Arabidopsis* seedling with *EsDFR*.

Gene expression

Total RNA was extracted from inner sepals and petals at anthesis, at which time the biosynthesis of anthocyanin is completed. First-strand cDNA was synthesized using PrimeScript RT reagent Kit (Takara, Japan) following the manufacturer's instructions. In each 20 μL qRT-PCR reaction, 50 ng of cDNA was amplified using SYBR® Premix Ex TaqTM II (Takara, Japan) and 100 mM of primers in an ABI7500 Real-Time PCR machine (ABI, USA) as per the manual. Actin was amplified as the control gene. The samples from three tissues were used and three technical replicates were performed for each sample. Data were analyzed by ABI7500 software. In this study, all pairs of primer (*CHS1*, *CHS2*, *CHS3*, *CHI1*, *CHI2*, *F3H1*, *F3H2*, *F3'H*, *F3'5'H*, *FLS*, *DFR*, and *ANS*) from *E. sagittatum* were transferred in other *Epimedium* species. All primers used in this manuscript are listed in the supplementary database.

Complementation analysis

For functional analyses, the *E. sagittatum* (A-) genes *EsF3'H* and *EsDFR* were overexpressed in their respective *Arabidopsis thaliana* (ecotype Landsberg) mutants, each lacking anthocyanins at the seedling stage. *EsF3'5'H*, The coding regions of *EsF3'H*, *EsF3'5'H*, and *EsDFR* were cloned into pMD19-T (Takara, Japan). The SalI and SacI digested fragment of each gene was purified and ligated into the pMV plasmid (derived from pBI121) behind the cauliflower 35S promoter. The plasmids were then transformed into *Agrobacterium* strain EHA105. *Arabidopsis* wild-type and mutants (*tt3* and *tt7*) were transformed by the floral dip infiltration method (Zhang et al., 2006). Transformants were selected on 1/2 Murashige and Skoog medium supplemented with 50 μg/mL kanamycin. Resistant seedlings were then transferred into the soil to harvest seeds. T1 seedlings were screened on 1/2 MS medium minus nitrogen for observation of anthocyanin accumulation.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Reference.

Author contributions

WS, HH, and YW conceived and designed the experiments. WS and YM performed the experiments.WS, WH, ZY, XM, and HW analyzed the data. WS wrote the paper. XM and HW revised the paper. YZ provided Figure 2 and collected species. RH and SC supervised this investigation. All authors contributed to the article and approved the submitted version.

Funding

The authors are grateful for finical support from the Scientific and Technological Innovation Project of the China Academy of Chinese Medical Sciences (CI2021A04806 and CI2021A04008) and National Key Research and Development Program of China (2022YFC3501703).

Acknowledgments

We thank Ms. Xiaoping Pan for assistance with qRT-PCR analysis, Dr. Sen Lin and Guoxiang Dong for assistance with HPLC analysis, Ms. Xiaomin Hu for assistance with screening *Arabidopsis*.

Conflict of interest

Author RH was employed by company By-Health Institute of Nutrition and health. By-health Co., Ltd.

The reviewer HP declared a shared affiliation with the authors YM, SC, and WS to the handling editor at the time of review China Academy of Chinese Medical Sciences, Beijing, China.

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/fpls.2023.1133616/full#supplementary-material

References

Boss, P. K., Davies, C., and Robinson, S. P. (1996). Analysis of the expression of anthocyanin pathway genes. *Plant Physiol.* 111, 1059–1066. doi: 10.1104/pp.111.4.1059

Buer, C. S., Imin, N., and Djordjevic, M. A. (2010). Flavonoids: New roles for old molecules. *J. Integr. Plant Biol.* 52, 98–111. doi: 10.1111/j.1744-7909.2010.00905.x

Des Marais, D. L., and Rausher, M. D. (2010). Parallel evolution at multiple levels in the origin of hummingbird pollinated flowers in *Ipomoea*. *Evol. (N. Y).* 64, 2044–2054. doi: 10.1111/j.1558-5646.2010.00972.x

Dick, C. A., Buenrostro, J., Butler, T., Carlson, M. L., Kliebenstein, D. J., and Whittall, J. B. (2011). Arctic Mustard flower color polymorphism controlled by petal-specific downregulation at the threshold of the anthocyanin biosynthetic pathway. *PLoS One* 6. doi: 10.1371/journal.pone.0018230

Fenster, C. B., Armbruster, W. S., Wilson, P., Dudash, M. R., and Thomson, J. D. (2004). Pollination syndromes and floral specialization. *Annu. Rev. Ecol. Evol. Syst.* 35, 375–403. doi: 10.1146/annurev.ecolsys.34.011802.132347

Gonzalez, A., Zhao, M., Leavitt, J. M., and Lloyd, A. M. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* 53, 814–827. doi: 10.1111/j.1365-313X.2007.03373.x

Hoballah, M. E., Gubitz, T., Stuurman, J., Broger, L., Barone, M., Mandel, T., et al. (2007). Single gene-mediated shift in pollinator attraction in *Petunia*. *Plant Cell* 19, 779–790. doi: 10.1105/tpc.106.048694

Holton, T. A., and Cornish, E. C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* 7, 1071. doi: 10.2307/3870058

Huang, W., Khaldun, A. B. M., Lv, H., Du, L., Zhang, C., and Wang, Y. (2016). Isolation and functional characterization of a R2R3-MYB regulator of the anthocyanin biosynthetic pathway from *Epimedium sagittatum*. *Plant Cell Rep.* 35, 883–894. doi: 10.1007/s00299-015-1929-z

Huang, W., Sun, W., Lv, H., Luo, M., Zeng, S., Pattanaik, S., et al. (2013a). A R2R3-MYB transcription factor from *Epimedium sagittatum* regulates the flavonoid biosynthetic pathway. *PLoS One* 8. doi: 10.1371/journal.pone.0070778

Huang, W., Sun, W., Lv, H., Xiao, G., Zeng, S., and Wang, Y. (2013b). Isolation and molecular characterization of thirteen R2R3-MYB transcription factors from *Epimedium sagittatum*. *Int. J. Mol. Sci.* 14, 594–610. doi: 10.3390/ijms14010594

Huang, W., Sun, W., and Wang, Y. (2012). Isolation and molecular characterisation of flavonoid 3'-hydroxylase and flavonoid 3', 5'-hydroxylase genes from a traditional Chinese medicinal plant, *Epimedium sagittatum*. *Gene* 497, 125–130. doi: 10.1016/j.gene.2011.11.029

Huang, W., Zeng, S., Xiao, G., Wei, G., Liao, S., Chen, J., et al. (2015). Elucidating the biosynthetic and regulatory mechanisms of flavonoid-derived bioactive components in *Epimedium sagittatum*. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00689

Lepiniec, L., Debeaujon, I., Routaboul, J.-M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006). Genetics and biochemistry of deed flavonoids. *Annu. Rev. Plant Biol.* 57, 405–430. doi: 10.1146/annurev.arplant.57.032905.105252

Peer, W. A., Brown, D. E., Tague, B. W., Muday, G. K., Taiz, L., and Murphy, A. S. (2001). Flavonoid accumulation patterns of transparent testa mutants of *Arabidopsis*. *Plant Physiol.* 126, 536–548. doi: 10.1104/pp.126.2.536

Pollastri, S., and Tattini, M. (2011). Flavonols: Old compounds for old roles. *Ann. Bot.* 108, 1225–1233. doi: 10.1093/aob/mcr234

Schwinn, K. (2006). A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. *Plant Cell* 18, 831–851. doi: 10.1105/tpc.105.039255

Smith, S. D., and Rausher, M. D. (2011). Gene loss and parallel evolution contribute to species difference in flower color. *Mol. Biol. Evol.* 28, 2799–2810. doi: 10.1093/molbev/msr109

Streisfeld, M. A., and Rausher, M. D. (2009). Genetic changes contributing to the parallel evolution of red floral pigmentation among *Ipomoea* species. *New Phytol.* 183, 751–763. doi: 10.1111/j.1469-8137.2009.02929.x

Sun, W., Huang, W., Li, Z., Song, C., Liu, D., Liu, Y., et al. (2014). Functional and evolutionary analysis of the AP1/SEP/AGL6 superclade of MADS-box genes in the basal eudicot *Epimedium sagittatum*. *Ann. Bot.* 113, 653–668. doi: 10.1093/aob/mct301

Whittall, J. B., Voelckel, C., Kliebenstein, D. J., and Hodges, S. A. (2006). Convergence, constraint and the role of gene expression during adaptive radiation: Floral anthocyanins in *Aquilegia*. *Mol. Ecol.* 15, 4645–4657. doi: 10.1111/j.1365-294X.2006.03114.x

Zeng, S., Xiao, G., Guo, J., Fei, Z., Xu, Y., Roe, B. A., et al. (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. et zucc.) maxim. *BMC Genomics* 11. doi: 10.1186/1471-2164-11-94

Zhang, X., Henriques, R., Lin, S. S., Niu, Q. W., and Chua, N. H. (2006). Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nat. Protoc.* 1, 641–646. doi: 10.1038/nprot.2006.97

Zhang, Y., Li, J., Wang, Y., and Liang, Q. (2021). Taxonomy of Epimedium (Berberidaceae) with special reference to Chinese species. *Chin. Herb. Med.* 14 (1), 20–35. doi: 10.1016/j.chmed.2021.12.001

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 March 2023
DOI 10.3389/fpls.2023.1166710

OPEN ACCESS

EDITED BY
Wei Xu,
Kunming Institute of Botany (CAS), China

REVIEWED BY
Bo Feng,
Chengdu Institute of Biology (CAS), China
Cheng Liu,
Shandong Academy of Agricultural Sciences, China

*CORRESPONDENCE
Houyang Kang
houyang.kang@sicau.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 15 February 2023
ACCEPTED 09 March 2023
PUBLISHED 30 March 2023

CITATION
Zhang H, Zeng C, Li L, Zhu W, Xu L, Wang Y, Zeng J, Fan X, Sha L, Wu D, Cheng Y, Zhang H, Chen G, Zhou Y and Kang H (2023) RNA-seq analysis revealed considerable genetic diversity and enabled the development of specific KASP markers for *Psathyrostachys huashanica*. *Front. Plant Sci.* 14:1166710. doi: 10.3389/fpls.2023.1166710

RNA-seq analysis revealed considerable genetic diversity and enabled the development of specific KASP markers for *Psathyrostachys huashanica*

Hao Zhang[1,2†], Chunyan Zeng[1,2†], Liangxi Li[1,2], Wei Zhu[1,2], Lili Xu[2], Yi Wang[1,2], Jian Zeng[3], Xing Fan[1,2], Lina Sha[4], Dandan Wu[1,2], Yiran Cheng[1], Haiqin Zhang[4], Guoyue Chen[1,2], Yonghong Zhou[1,2] and Houyang Kang[1,2*]

[1]State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Sichuan Agricultural University, Chengdu, Sichuan, China, [2]Triticeae Research Institute, Sichuan Agricultural University, Chengdu, Sichuan, China, [3]College of Resources, Sichuan Agricultural University, Chengdu, Sichuan, China, [4]College of Grassland Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan, China

Psathyrostachys huashanica, which grows exclusively in Huashan, China, is an important wild relative of common wheat that has many desirable traits relevant for wheat breeding. However, the poorly characterized interspecific phylogeny and genomic variations and the relative lack of species-specific molecular markers have limited the utility of *P. huashanica* as a genetic resource for enhancing wheat germplasm. In this study, we sequenced the *P. huashanica* transcriptome, resulting in 50,337,570 clean reads that were assembled into 65,617 unigenes, of which 38,428 (58.56%) matched at least one sequence in public databases. The phylogenetic analysis of *P. huashanica*, Triticeae species, and Poaceae species was conducted using 68 putative orthologous gene clusters. The data revealed the distant evolutionary relationship between *P. huashanica* and common wheat as well as the substantial diversity between the *P. huashanica* genome and the wheat D genome. By comparing the transcriptomes of *P. huashanica* and Chinese Spring, 750,759 candidate SNPs between *P. huashanica* Ns genes and their common wheat orthologs were identified. Among the 90 SNPs in the exon regions with different functional annotations, 58 (64.4%) were validated as Ns genome-specific SNPs in the common wheat background by KASP genotyping assays. Marker validation analyses indicated that six specific markers can discriminate between *P. huashanica* and the other wheat-related species. In addition, five markers are unique to *P. huashanica*, *P. juncea*, and *Leymus* species, which carry the Ns genome. The Ns genome-specific markers in a wheat background were also validated regarding their specificity and stability for detecting *P. huashanica* chromosomes in four wheat–*P. huashanica* addition lines. Four and eight SNP markers were detected in wheat–*P. huashanica* 2Ns and 7Ns addition lines, respectively, and one marker was specific to both wheat–*P. huashanica* 3Ns, 4Ns, and 7Ns addition lines. These markers developed using transcriptome data

may be used to elucidate the genetic relationships among *Psathyrostachys*, *Leymus*, and other closely-related species. They may also facilitate precise introgressions and the high-throughput monitoring of *P. huashanica* exogenous chromosomes or segments in future crop breeding programs.

KEYWORDS

***Psathyrostachys huashanica*, transcriptome sequencing, phylogenetic relationship, interspecific variation, KASP markers**

1 Introduction

The genus *Psathyrostachys* Nevski, which belongs to the tribe Triticeae, comprises eight perennial diploid or tetraploid species that contain only the Ns genome (Yen and Yang, 2011). In China, *Psathyrostachys huashanica* Keng f. ex P. C. Kuo ($2n = 2x = 14$, NsNs) is a nationally protected rare plant that is native to the mountainous slopes of Huashan Pass in the Qinling Mountains of Shaanxi province (Sun et al., 1993). The superior characteristics of *P. huashanica* include early maturation, tolerance to drought and salinity, and resistance to stripe rust, take-all, powdery mildew, wheat scab, and yellow dwarf. Moreover, its genome includes genes associated with many beneficial yield-related traits (Jing et al., 1999). To transfer these desirable traits into wheat, *P. huashanica* was hybridized with common wheat starting in the 1990s (Chen et al., 1991). Some progeny lines harboring *P. huashanica* chromosomal segments incorporated into the wheat genome were developed as derivative lines (Kang et al., 2009) with chromosomal additions (Kishii et al., 2010; Du et al., 2014; Tan et al., 2021), substitutions (Bai et al., 2020; Qu et al., 2022), and translocations (Li et al., 2020; Liu et al., 2021a). These progeny lines outperformed their wheat parents in terms of abiotic and biotic stress resistance and agronomic traits. To date, there has been some progress in the mapping of important genes in the *P. huashanica* genome, including those conferring resistance to stripe rust and take-all (Ma et al., 2016; Bai et al., 2021; Sun et al., 2018). Therefore, *P. huashanica* is generally considered to be a potentially useful germplasm for the genetic improvement of wheat.

Developing species-specific molecular markers that facilitate the identification of alien chromosomes or segments associated with genes of interest is critical for wheat breeding programs (Liu et al., 2018). Scholars have reported some results of genome sequencing and assembly of *P. huashanica*, while the genome data was not available at present (Li, 2019). Unfortunately, there are currently relatively few genomic and molecular marker resources for *P. huashanica*. The reported markers specific to *P. huashanica* mainly consist of expressed sequence tag-simple sequence repeats (Kanwal, 2019), sequence characterized amplified region markers, random-amplified polymorphic DNAs (Du et al., 2014) and common PCR markers (Tan et al., 2021). However, the relatively low polymorphism and distribution densities of these markers have restricted their use in wheat breeding programs and investigations of the phylogenetic relationships among *P. huashanica* and related species. Thus, additional molecular markers will need to be developed on the basis of high-throughput genotyping. Rapid advances in next-generation sequencing technologies have facilitated the large-scale identification of single nucleotide polymorphisms (SNPs) in wheat and multiple wheat-related species (Zhang et al., 2017; Ma et al., 2019). In addition, RNA sequencing (RNA-seq) technology has been used for the high-throughput and cost-effective detection of SNPs and genes as well as for analyzing phylogenetic relationships, evaluating genetic diversity, and developing molecular markers for Triticeae species (Zhou et al., 2017). To date, RNA-seq approaches have been applied to develop novel SNP markers for several wild wheat relatives, such as *Agropyron cristatum* (Zhou et al., 2017) and *Thinopyrum elongatum* (Lou et al., 2017), as well as for *Aegilops* species, including *Aegilops umbellulata* (Okada et al., 2018) and *Aegilops tauschii* (Iehisa et al., 2014). These markers were widely used for the ongoing introgression of valuable alien genes into wheat, but they also clarified phylogenetic relationships and the genetic diversity among the various genomes in Triticeae species.

We previously reported the generation of hybrids from a cross between *P. huashanica* and common wheat that did not involve an embryo rescue step and the development of some wheat–*P. huashanica* lines with useful genes for enhancing wheat characteristics (Kang et al., 2009; Kang et al., 2016). Unfortunately, the phylogenetic relationships, genetic diversity, and SNPs in *P. huashanica* and wheat remain poorly investigated. In this study, we used Illumina RNA-seq technology to generate the basal transcriptome sequencing data for *P. huashanica* and revealed genetic polymorphisms as well as phylogenetic relationships between *P. huashanica* and other Triticeae species. Furthermore, on the basis of genome-specific SNPs, we developed Kompetitive allele-specific PCR (KASP) markers for *P. huashanica*. These markers were subsequently used to elucidate the genetic differences and phylogenetic relationships among Ns, H, R, P, V, and other closely-related genomes. They were also validated regarding their utility for detecting *P. huashanica* chromosomes in wheat–*P. huashanica* 2Ns, 3Ns, 4Ns, and 7Ns addition lines.

2 Materials and methods

2.1 Plant materials

Psathyrostachys huashanica ($2n = 2x = 14$, NsNs) accession ZY3157 was collected on Huashan Mountain (Shanxi, China) by Profs. C. Yen and J. L. Yang (Sichuan Agricultural University).

Wheat cultivar Chinese Spring (CS) and Chinese Spring *ph2b* (CS*ph2b*) were used as the positive controls for the molecular marker analysis and the source of blocking DNA for the Genomic *in situ* hybridization (GISH) experiments. The molecular markers were validated using the following wheat-related species: *Psathyrostachys juncea* (2*n* = 2*x* = 14, NsNs, PI314028), *Th. elongatum* (2*n* = 2*x* = 14, EE, PI531718), *Pseudoroegneria libanotica* (2*n* = 2*x* = 14, StSt, PI228391), *D. villosum* (2*n* = 2*x* = 14, VV, PI470279), *Hordeum vulgare* (2*n* = 2*x* = 14, HH, ZY11001), *A. cristatum* (2*n* = 2*x* = 14, PP, PI499389), *Secale cereale* (2*n* = 2*x* = 14, RR, QL), and *Leymus racemosus* (2*n* = 4*x* = 28, NsNsXmXm, ZY07023). Four previously identified wheat–*P. huashanica* addition lines with 2Ns, 3Ns, 4Ns, and 7Ns chromosomes were used for a KASP genotyping assay that was conducted to verify the utility of the Ns genome-specific SNPs (Zhang et al., 2022). Voucher specimens were deposited in the herbarium of the Triticeae Research Institute, Sichuan Agricultural University, China.

2.2 RNA–seq, transcriptome assembly and annotation

Total RNA was extracted from the flag leaves and young roots collected from five *P. huashanica* ZY3157 plants at the jointing stage using TRIzol reagent (Thermo Fisher Scientific Inc., Shanghai, China) according to the manufacturer's instructions. The purity and concentration of the extracted RNA were then determined using the NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc.). The RNA integrity and quantity were determined using the Agilent 4200 system and the Agilent High Sensitivity DNA Kit (Agilent Technologies Inc., USA). The mRNA purified from the total RNA was used to construct cDNA libraries, after which the library concentrations were determined using the Qubit 3.0 fluorometer (Thermo Fisher Scientific Inc.) and by performing a qPCR assay. After preparing the sequencing libraries and pooling the libraries for different tissues, the samples were sent to the BerryGenomics Corporation (Beijing, China) for an Illumina paired-end sequencing analysis using the NovaSeq 6000 platform, which generated 150-bp paired-end reads. Reads containing adapters or more than three Ns and low-quality reads (more than 20% nucleotides with Phred quality score ≤5) were removed using an in-house Perl script to produce clean reads. The Q20, Q30, and GC content of the clean reads were calculated. The clean reads were mapped using the SILVA database and Bowtie2 (version 2.4.5) to eliminate the rRNA (Langmead and Salzberg, 2012). The Trinity program (version 2.13.2) was used for the transcriptome *de novo* assembly; the default parameters were applied, but the minimum K-mer coverage was set to 2 (Grabherr et al., 2011). The first transcript sequence generated by Trinity was selected if there were multiple isoforms. All of the unigenes identified on the basis of the Trinity assembly results were used as queries to screen the following databases using BLAST (2.11.0) (Altschul et al., 1990), with an E-value cut-off of 1e-5: NR (non-redundant protein sequences), Pfam (protein families), Swiss-Prot (manually annotated and reviewed protein sequences), eggNOG (Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups), GO (Gene Ontology), and KEGG (Kyoto Encyclopedia of Genes and Genomes).

2.3 Phylogenetic analysis

A BLASTn search was conducted using Orthofinder (version 2.5.4) to identify the single-copy orthologous pairs between the *P. huashanica* unigenes and the coding sequences (CDSs) from *Triticum aestivum* (A, B, and D genomes were separated), *Triticum turgidum* (A and B genomes were separated), *Ae. tauschii*, *Triticum urartu*, *S. cereale*, *H. vulgare*, *Brachypodium distachyon*, *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Setaria italica*, and *Arabidopsis thaliana* (Emms and Kelly, 2015). The CDSs were downloaded from the EnsemblPlants (https://www.plants.ensembl.org) database. The orthologous pairs that were single-copy genes in one genome and conserved in all genomes were aligned using MAFFT (version 7.508) (Katoh and Standley, 2013). The aligned genes were merged in a series for the construction of a phylogenetic tree using MEGA7 according to the neighbor-joining method (Kumar et al., 2016). The bootstrap support values were calculated with 1,000 replications and presented at each node. The maximum likelihood method was used to calculate the evolutionary distance in terms of the number of base substitutions per site, with all positions containing gaps and missing data eliminated (Tamura et al., 2004). The phylogenetic tree was drawn to scale, with branch lengths representing the evolutionary distances.

2.4 SNP discovery and development of KASP markers

The default parameters of the HISAT 2 software (version 2.2.1) were used to map the *P. huashanica* clean reads to the CS reference genome sequence (IWGSC RefSeqv1.0) and generate bam files (Kim et al., 2019). Picard-tools (version 1.92) and Samtools (version 1.15.1) were used to sort and mark duplicated reads as well as reorder the bam alignment results (Li et al., 2009). The SNPs and insertions/deletions (indels) were identified using Samtools, with the minimum mapping quality set to 1 and the filtering flag set to 0.002 (Li et al., 2009). Variations were annotated using Annovar on the basis of their genomic locations and filtered (total coverage >4 and quality >30) to obtain high-confidence variations (Yang and Wang, 2018). The distribution of SNPs and indels on *P. huashanica* and common wheat chromosomes was visualized using the Circos software (0.69.8) (Krzywinski et al., 2009). Homozygous SNPs and contextual sequences that did not overlap the intron in the wheat genome were analyzed using the online platform Polymarker (https://www.polymarker.info) with 100-bp flanking sequences on each side to generate KASP primers. All primers were synthesized by Sangon Biological Engineering Co., Ltd. (Shanghai, China). Details regarding the primers are presented in Supplementary Table S1.

2.5 KASP marker validation and sequential FISH and GISH analyses

The specificity, stability, and universality of the KASP markers were assessed using five individual plants of *P. huashanica* ZY3157,

CS, CS*ph2b*, wheat–*P. huashanica* addition lines (2Ns, 3Ns, 4Ns, and 7Ns), and 14 wheat-related species, including *P. juncea*, *L. racemosus*, *Th. elongatum*, *Pse. libanotica*, *D. villosum*, *H. vulgare*, *A. cristatum*, and *S. cereale*. The KASP genotyping and fluorescence data analysis were performed as previously described (Ma et al., 2019). The slides for the mitotic metaphase chromosomes were prepared according to a published method (Han et al., 2006). The FISH assay involving oligonucleotide probes and the GISH assay involving *P. huashanica* DNA labeled with fluorescent tags (nick translation method of the Atto550 NT labeling kit; Jena Bioscience, Jena, Germany) were completed using the mitotic chromosomes of the wheat–*P. huashanica* addition lines to identify the added *P. huashanica* chromosomes. For the GISH assay, CS or CS*ph2b* genomic DNA was used for blocking. The FISH and GISH protocols were described in our previous report (Zhang et al., 2022). The samples on slides were counterstained with a 4,6-diamino-2-phenylindole solution (Vector Laboratories, Burlingame, CA, USA) and then examined using the BX-63 microscope (Olympus, Tokyo, Japan). Images were captured using the DP80 CCD camera (Olympus) installed on the microscope.

3 Results

3.1 Analysis of the *P. huashanica* transcriptome data and annotation of unigenes

The sequencing of the *P. huashanica* transcriptome generated 53,288,309 raw reads, of which 50,337,570 were clean reads, with a GC content of 55.75% as well as Q20 and Q30 scores of 97.01% and 92.83%, respectively (Table 1). The *de novo* assembly using the high-quality filtered reads produced 264,519 transcripts and 65,617 unigenes. The unigene lengths ranged from 306 to 11,879 bp, with a mean length of 1,208 bp and an N50 of 1,879 bp (Table 2; Supplementary Figure S1).

The 65,617 *P. huashanica* unigenes were annotated on the basis of six publicly available databases. More specifically, 57.25%, 28.77%, 31.43%, 54.02%, 40.66%, and 10.34% of the unigenes were annotated according to the NR, Pfam, Swissprot, eggNOG, GO, and KEGG databases, respectively (Figure 1; Supplementary Table S2). During the BLAST search, 38,428 unigenes (58.56%) had a match in at least one database (Supplementary Table S2). The NR database had more matches with the *P. huashanica* unigenes than the other databases. In addition, 20.94% of the unigenes were highly similar to sequences in the NR database (95%–100% sequence identity), but most of the annotated unigenes had sequence identities ranging from 80% to 95% (Supplementary Figure S2).

3.2 Identification of single-copy orthologous genes and analysis of phylogenetic relationships

The comparison between the *P. huashanica* unigenes and the CDSs in other Poaceae and Triticeae species revealed a total of 68 putative orthologous gene clusters comprising single-copy genes that were conversed in all genomes. These putative orthologous genes were used to construct a phylogenetic tree. The phylogenetic analysis indicated that all Triticeae members were clustered into a sister clade and were closely related to *B. distachyon*, whereas the gramineous species, including *O. sativa*, *S. italica*, *S. bicolor*, and *Z. mays*, were divided into other clades. Notably, the species with the Ns genome had a more distant evolutionary relationship with *T. aestivum* and their ancestral species, including *T. urartu*, *Ae. tauschii*, and *T. turgidum*, than with *H. vulgare* and *S. cereale* (Figure 2). These results were consistent with the difficulties associated with the hybridization between wheat and *P. huashanica* and also suggested that the Ns genome likely contains desirable genetic variations that may be beneficial for wheat genetic studies and breeding. Moreover, the Ns genome was more distantly related to the wheat D genome than to the other wheat genomes (Figure 2).

3.3 Comparison of *P. huashanica* and *T. aestivum* sequences

To identify the variations between *P. huashanica* and *T. aestivum*, the transcript sequences were compared with the published sequences in the wheat CDS database. According to the BLASTn search, the average transcript sequence identity between wheat and *P. huashanica* was 95.13%, with 97.50% revealed as the most common sequence identity (Figure 3). Accordingly, the transcript sequences were relatively conserved between *P. huashanica* and wheat.

3.4 Identification of the variants between *P. huashanica* and *T. aestivum* and analysis of their effects

To identify genomic variants, the clean *P. huashanica* transcriptome sequencing reads were mapped to the CS reference genome sequence (IWGSC RefSeqv1.0) (Figure 4A). The average gene density was calculated for the wheat chromosomes. Additionally, the average gene density and variant density increased from the centromeres to the telomeres of the chromosomes (Figures 4C–F). Moreover, 750,759 SNPs and 3,883 indels were identified between the *P. huashanica* and CS transcripts (Figures 4E, F; 5A). The variants were spread across the wheat

TABLE 1 Summary of the RNA_Seq data.

Raw reads	Clean reads	Clean reads/raw reads (%)	Q20 (%)	Q30 (%)	Average GC content (%)
53,288,309	50,337,570	94.46	97.01	92.83	55.75

10.3389/fpls.2023.1166710

TABLE 2 Summary of the *P. huashanica* transcriptome assembly.

Assembly parameters	
Number of transcripts assemed	264,519
Number of unigene	65,617
Maximum unigene length (bp)	11,879
Minimum unigene length (bp)	306
Average unigene length (bp)	1208
N50 length of unigene (bp)	1879

genome, but they were not directly proportional to the chromosome length and gene number (Figures 5A, B). An average of 55.67 SNPs/Mb were detected in the wheat genome (Figure 5B). The SNP density was highest in the D genome (69.71/Mb), followed by the A genome (51.17/Mb) and the B genome (49.24/Mb) (Figure 5B). Furthermore, the highest and lowest SNP densities were observed for the homologous group 5 chromosomes (64.59/Mb) and the homologous group 7 chromosomes (50.11/Mb), respectively (Figure 5B).

The mapping of all SNPs to the wheat genes indicated 54,277 of the 107,891 annotated genes (50.31%) contained one or more SNPs, of which 668 were related to disease resistance (Figures 4B; 5C; Supplementary Table S3). There was an average of 13.83 SNPs per gene and 25.16% of the genes had more than 20 SNPs (Figure 5C). Of these genes, 1,416 genes had more than 50 substitutions in the identified SNPs, whereas 9,743 genes had fewer than five substitutions (Figure 5C). These 1,416 genes may be highly diverse, making them potentially useful for exploring the genetic diversity among Triticeae species and for breeding novel varieties.

The distribution of the SNPs in various wheat genomic regions was determined (Figure 5D). A total of 12,530 (1.58%), 36,690 (4.64%), 2,811 (0.36%), and 4,108 (0.52%) variants were detected in the intronic, intergenic, upstream, and downstream regions, respectively (Figure 5D). Approximately 30,387 SNPs (3.84%) were located in the 5′- or 3′-UTRs, whereas 88.83% of the variants were present in the coding regions, in which the non-synonymous-to-synonymous variants ratio was 48.61% (Figure 5D). These results implied that the transcribed regions were likely under purifying selection. In terms of the substitution types, transitions (62.80%; A/G: 31.45% and T/C: 31.35%) were more common than transversions (37.19%; A/C: 7.87%, A/T: 5.96%, C/G: 15.5%, and T/G: 7.86%) (Figure 5E). The proportions of the A/G and T/C transitions were similar (i.e., 31.45% and 31.35%, respectively) (Figure 5E). Of the transversions, C/G was the most common (15.5%), followed by A/C and T/G (7.87%) and A/T (5.96%) (Figure 5E). The transitions-to-transversions ratio was 1.69:1, which reflected the genetic conservation during evolution.

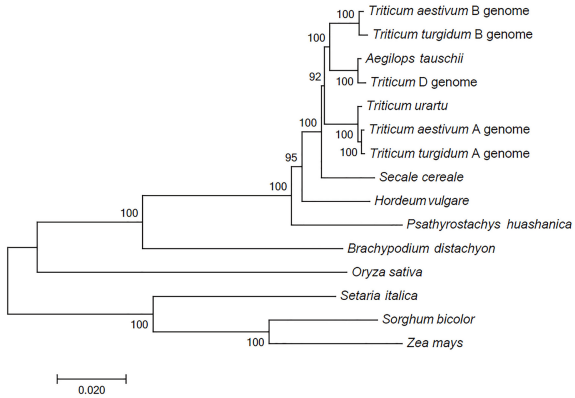


FIGURE 2
Phylogenetic relationships between *P. huashanica* and other genomes. Numbers at the node are the bootstrap values, shown as percentages. The branch lengths are the same as the evolutionary distances.

3.5 Identification of SNPs on the basis of KASP genotyping results

The KASP genotyping assays validated 90 candidate SNPs between the *P. huashanica* and CS genomes (Supplementary Table S1). The allele-specific primers uncovered single nucleotide substitutions between *P. huashanica* and the other accessions. In the KASP genotyping assays, 58 primers identified obvious clusters and detected allele 2 in *P. huashanica*, but allele 1 in CS and CS*ph2b*. Hence, these SNPs, which represented 64.4% of the developed markers, were validated as specific for *P. huashanica* in a common wheat background (Supplementary Table

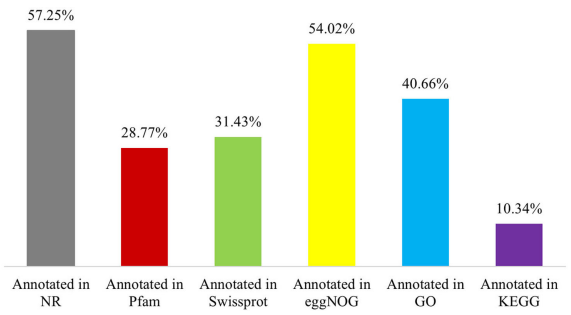


FIGURE 1
Function annotation of *P. huashanica* unigenes.

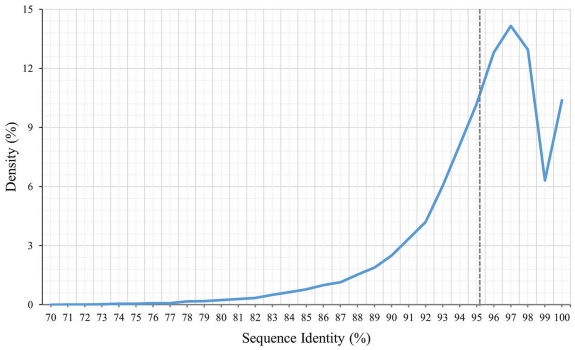


FIGURE 3
Density distribution of *P. huashanica*. Unigene sequence identity compared to the wheat.

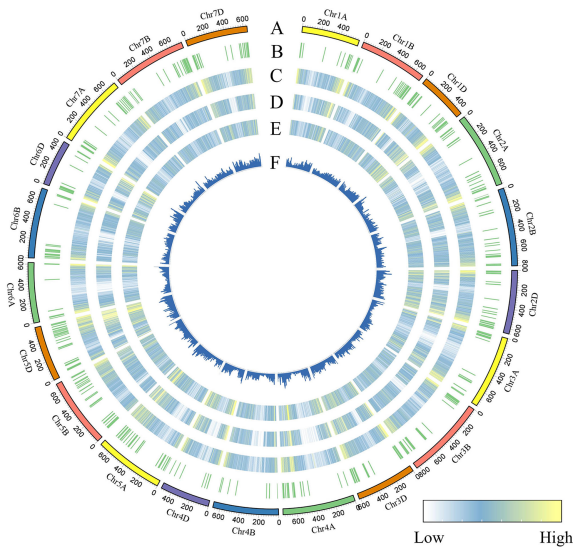


FIGURE 4
CIRCOS visualization of different data at the wheat genome-wide level. **(A)** Karyotype of the wheat genome. **(B)** Location of variant genes related to disease resistance. **(C)** Gene content density distribution; Gene density was calculated in a 5-Mb window. **(D)** Variant transcriptional density of *P.huashanica*. Variant density was calculated in transcript regions at 7-Mb window intervals. (EF) Variant density in transcript regions of *P.huashanica*. Variant density in transcript regions was visualized by heatmap **(E)** and bar **(F)** and calculated at 7-Mb window intervals.

S1). Among these specific markers, four and eight markers detected heterozygous alleles (allele 1/allele 2) in the wheat–*P. huashanica* 2Ns and 7Ns addition lines, respectively, whereas one marker simultaneously detected heterozygosity (allele 1/allele 2) in both wheat–*P. huashanica* 3Ns, 4Ns, and 7Ns addition lines (Figures 6A, B; Supplementary Table S1). The chromosomal compositions in these addition lines were confirmed by conducting FISH and GISH assays (Supplementary Figure S3). Thus, these markers were useful for detecting the corresponding *P. huashanica* chromosomes in a wheat background.

To evaluate the specificity and stability of the 58 P. *huashanica*-specific markers, the KASP genotyping assay was performed using eight wheat-related species that differed regarding their basic genome. Six of these markers detected allele 2 in *P. huashanica* while detected allele 1 or no amplification in all other wheat-related species (Figure 6C; Supplementary Table S1). In contrast, three markers detected allele 2 in both *P. huashanica* and *P. juncea*, but detected allele 1 or no amplification in all other analyzed species. Therefore, they were specific to the Ns genome in the genus *Psathyrostachys*. Five markers were specific to the Ns genome, of which two detected allele 2 in Ns genome-containing species, but allele 1 in the other species. In contrast, three markers detected allele 2 in *P. huashanica* and *P. juncea*, but allele 1/allele 2 (heterozygosity) in *L. racemosus* (Figure 6D; Supplementary Table S1).

4 Discussion

4.1 Powerful method for exploring *P. huashanica* genomic polymorphisms and developing markers

Psathyrostachys huashanica genes are potentially useful for increasing wheat resistance to biotic and abiotic stresses. Hence,

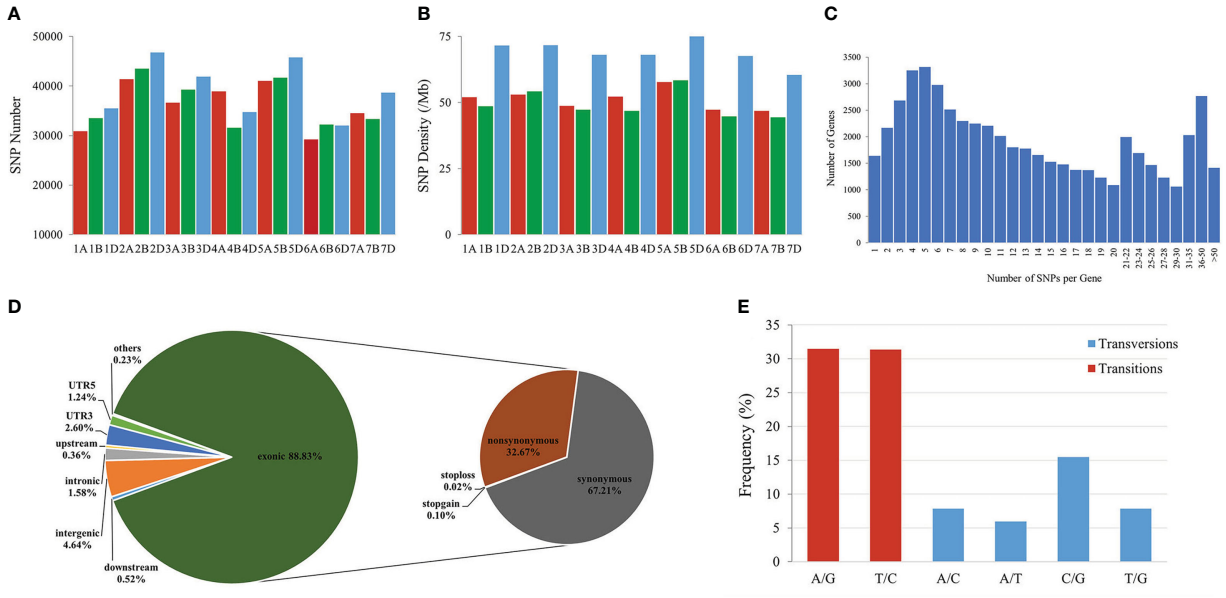


FIGURE 5
Features of SNPs between *P. huashanica* and wheat. **(A)** Distribution of SNPs in the wheat chromosomes. **(B)** Density distribution of SNPs in the wheat chromosomes. **(C)** Distribution of the number of SNPs per gene. **(D)** Distribution of SNPs in different genomic regions. **(E)** Frequency of different substitution types in the identified SNPs.

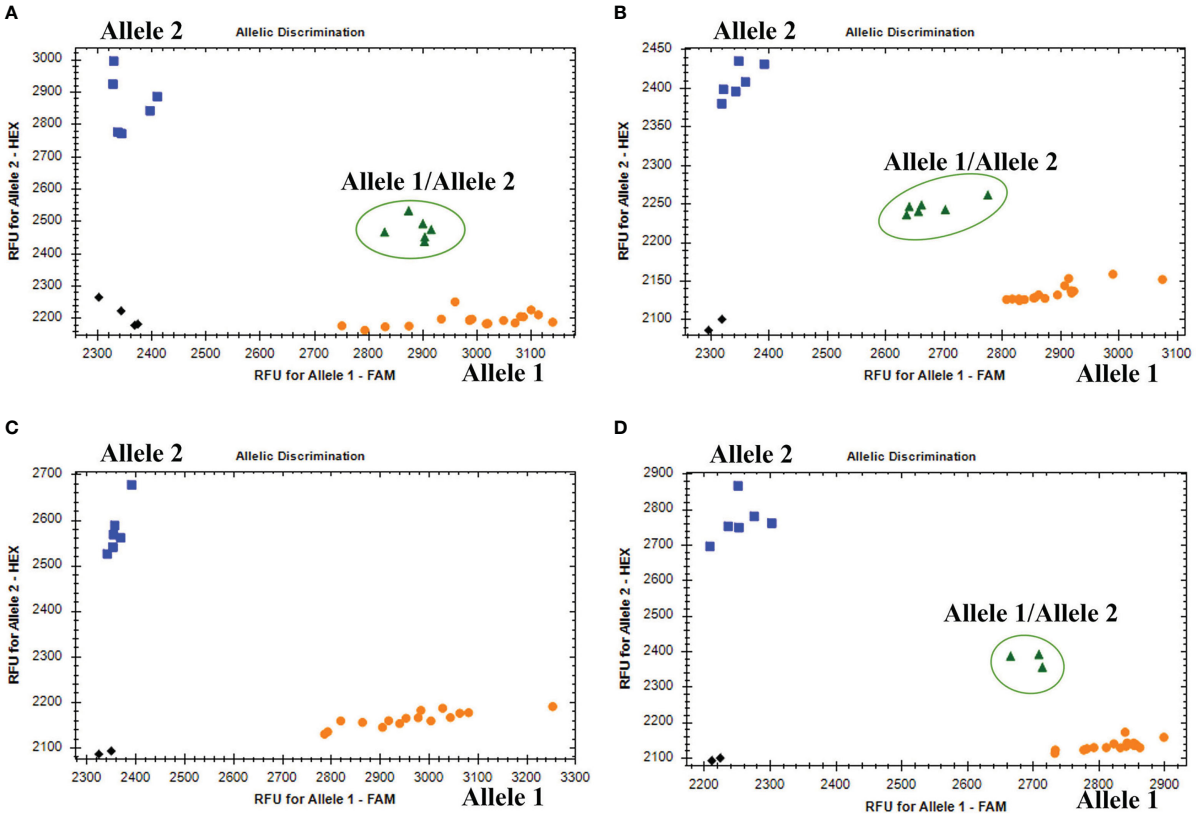


FIGURE 6
(A) Amplification results of marker KP2D-180993989 in CS, CS*ph2b*, wheat- *P. huashania* 3Ns, 4Ns, 7Ns addition lines (Allele 1); *P. huashanica* (Allele 2); wheat-*P. huasahnica* 2Ns addition lines (Allele 1/Allele 2). **(B)** Amplification results of marker KP7D-437429145 in CS, CS*ph2b*, wheat- *P. huashania* 2Ns, 3Ns, 4Ns addition lines (Allele 1); *P. huashanica* (Allele 2); wheat-*P. huasahnica* 7Ns addition lines (Allele 1/Allele 2). **(C)** Amplification results of marker KP2A-719278641 in *P. juncea*, *Th. elongatum*, *A cristatum*, *D villosum*, *H vulgare*, *S. cereale*, *Pse. libanotica*, *Ley. racemosus* (Allele 1); *P. haushancia* (Allele 2). **(D)** Amplification results of marker KP1A-440448839 in *Th. elongatum*, *A cristatum*, *D villosum*, *H vulgare*, *S. cereale*, *Pse. libanotica* (Allele 1); *P. haushancia*, *P. juncea* (Allele 2); *Ley. racemosus* (Allele 1/Allele 2).

P. huashanica has been used in wheat breeding programs for a long time. However, *P. huashanica* has primarily been used for molecular cytogenetic research, which resulted in the production and identification of wheat–*P. huashanica* introgression lines as well as the development of genome-specific markers, while the available genetic diversity and polymorphisms in the transcribed regions of the genome remained largely underexploited (Bai et al., 2020; Li et al., 2020; Liu et al., 2021a; Qu et al., 2022). To further explore the utility of *P. huashanica* for improving wheat, its genetic diversity should be elucidated at the molecular level and additional genetic markers should be developed.

Reference genomes are not widely available for most wild relatives of wheat because of their substantial abundance of repetitive sequences and their complexity. Although some statistics of the *P. huashanica* genome assembly have been reported, the genome resource has not been released (Li, 2019). Advances in next-generation sequencing technologies (e.g., Illumina RNA-seq) have provided researchers with alternative approaches for studying global transcriptome profiles for species lacking reference genomes. Moreover, RNA-seq-based methods are generally unaffected by the repetitiveness of non-transcribed regions and genome complexity for analyzing the transcripts of wheat relatives, evaluating genetic diversity, and identifying novel molecular markers (Zeng et al., 2017; Zhou et al., 2017). For example, Okada et al. (2018) performed an RNA-seq analysis of 12 representative *Ae. umbellulata* accessions. Many SNPs and indels were called and anchored to the pseudomolecules of *Ae. tauschii* and barley, revealing the greater genetic diversity in *Ae. umbellulata* than in *Ae. tauschii*. Zhou et al. (2017) conducted a transcriptome sequencing study to explore the genetic relationships between *A. cristatum* and wheat and wheat relatives as well as to identify the variations between *A. cristatum* and wheat. Potential SNPs were detected, of which 53 were validated according to a KASP genotyping assay. In another study, transcriptome data were used to develop 134 *Ae. longissima*-specific PCR markers, which may enable the transfer of desirable *Ae. longissima* genes into wheat *via* marker-assisted selection (Wang et al., 2018). Kanwal (2019) characterized 11 polymorphic EST-SSR primers to reveal the population genetic diversity among 12 P. *huahsancia* accessions by transcriptome analysis.

To detect genome-wide polymorphisms and develop novel genome-specific SNP markers, we conducted an RNA-seq analysis of *P. huashanica* leaf and root tissues using the Illumina NovaSeq 6000 platform, which generated 150-bp paired-end reads.

The comparative analysis of *P. huashanica* and wheat transcripts indicated the average sequence identity was 95.13%, whereas the peak sequence identity was 97.6%. These findings reflect the relatively close genetic relationship between *P. huashanica* and wheat, implying molecular markers may be developed on the basis of the wheat reference genome. By mapping the transcriptome sequencing data to the wheat reference genome, many high-quality SNPs on 21 wheat chromosomes were called, with 50.31% of the predicted wheat genes containing one or more variants, indicative of a great genetic diversity in *P. huashanica*. Using the called SNPs enabled the development of genus-specific markers. Compared with the other sequencing approaches used in earlier molecular investigations, such as genotyping-by-sequencing and specific-locus amplified fragment sequencing (Elshire et al., 2011; Poland et al., 2012; Kantarski et al., 2016; Song et al., 2020), RNA-seq is a relatively cost-effective method for producing transcription data-based markers closely linked to genes associated with useful agronomic traits.

4.2 Phylogenetic relationships between *P. huashanica* and wheat as well as wheat relatives

Phylogenetic relationships are useful for further characterizing crops and for selecting varieties in wheat breeding programs. Although wheat–*P. huashanica* progeny lines have long been used as sources of genes related to value-added traits, the genetic relationships between *P. huashanica* and other Triticeae species remain unclear. Our phylogenetic analysis showed that *P. huashancia* has closer genetic relationship with barley and rye than with wheat. These findings are consistent with the results of similar phylogenetic analyses involving trnL-F sequences and the chloroplast genome (Chen et al., 2020) as well as the SNP validation results in this study. According to the analysis of the Ns chromosome markers in wheat-related species, 28 of the 58 specific SNP markers (48.28%) occupied similar positions in *P. huashanica* (Ns) and *A. cristatum*. The next highest percentages were observed for the comparisons with *S. cereale* (R, 46.55%), *Pse. libanotica* (St, 43.10%) and *H. vulgare* (H, 31.03%). In contrast, only 27.59% of the SNP markers occupied similar positions in *P. huashanica* (Ns), *D. villosum* and *Th. elongatum*. Thus, *P. huashanica* has a closer genetic relationship with *A. cristatum*, *S. cereale*, *Pse. libanotica* and *H. vulgare*, than with *D. villosum* and *Th. elongatum*. Notably, 47 and 43 of the specific SNP markers detected allele 2 in *P. juncea* and allele 2 or heterozygosity in *L. racemosus*, respectively. These observations provide new evidence that the genera *Leymus* and *Psathyrostachys* are genetically closely related and that *Leymus* species contain the Ns genome from *Psathyrostachys*, which is in accordance with the results of previous research (Wang and Lu, 2014; Sha et al., 2017). In addition, the SNP density was obviously higher in the wheat D genome than in the A and B genomes, suggesting that *P. huashanica* is more distantly related to wheat relatives with the D genome than to wheat relatives containing the A and B genomes. This is also supported by the phylogenetic analysis involving Ns and the wheat A, B, and D genomes. Moreover, the average gene density and variant transcriptional density increased from the centromeres to the telomeres of the *P. huashanica* chromosomes, indicating that the frequency of allelic variations was greater for the telomeres than for the centromeres. Earlier studies demonstrated that the likelihood of genetic changes (e.g., exchange, recombination, and elimination) in chromosomal regions increases as the distance from the centromere increases (Jiang et al., 1994; Fan et al., 2020), which explains the rarity of intercalary translocation lines during interspecific or intergeneric hybridizations.

4.3 SNP marker application

Molecular markers have been extensively used to detect and trace alien chromosomes or chromosomal segments carrying desirable genes in a wheat background to increase the selection efficiency during breeding and shorten the breeding cycle (Fedak, 1999). An increasing number of elite genes have been identified in *P. huashanica*, but the previously developed molecular tools for *P. huashanica* were relatively imprecise and inefficient (Wang et al., 2014; Kanwal, 2019). For instance, Kanwal (2019) developed a series of EST-SSR markers of *P. huashanica* based on transcriptome data. In one of our earlier studies, we developed specific PCR markers for *P. huashanica* 7Ns chromosomes, but they were not co-dominant and could not be used to simultaneously trace alien chromosomes and their homoeologous groups (Tan et al., 2021). However, KASP markers for SNP genotyping are viable alternatives that are increasingly becoming the preferred markers because of their efficiency, accuracy, and ease-of-use as well as the fact they are not dependent on gel electrophoresis (Khera et al., 2013; Semagn et al., 2014). Therefore, KASP markers have been commonly used for mapping genes, identifying alien introgression lines, and marker-assisted selection-based wheat breeding. For example, Ma et al. (2019) quickly and reliably characterized two wheat–*A. cristatum* introgression lines with increased grain numbers per spike and resistance to powdery mildew by completing a KASP genotyping assay involving 6P-specific SNP markers. Diagnostic KASP markers have been developed to trace functional genes relevant for breeding, including the leaf rust resistance gene *Lr42* (Liu et al., 2021b), the Fusarium head blight resistance gene *Fhb1* (Su et al., 2018), and putative pre-harvest sprouting resistance genes (Liu et al., 2022). In the current study, we successfully developed the specific KASP markers for *P. huashanica* and Ns genome-containing species by RNA-seq. The markers described herein may be exploited to identify *P. huashanica* chromosomes in a wheat background. Moreover, these KASP markers and SNPs potentially useful for designing KASP markers may be applicable for monitoring wheat–*P. huashanica* cryptic small alien segment introgressions, while also facilitating the marker-assisted transfer of desirable traits from *P. huashanica* into adapted wheat cultivars in breeding programs. They can also further clarify the phylogenetic and functional relationships among *Psathyrostachys* and *Leymus* species.

Data availability statement

Raw data were deposited in NCBI SRA database. (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA937391).

Author contributions

HZ, CZ, LL and HK conducted the experiment, analyzed the data, and drafted the manuscript. WZ, LX, YW and JZ characterized addition lines. XF, LS, HQZ, DW, YC and GC provided technique guidance for bioinformatics analysis. YZ and HK designed the experiment and formulated the questions. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 31971883), and the Science and Technology Bureau of Sichuan Province (2023NSFSC1995, 2022YFH0069, 2022ZDZX0014, 2022NSFSC1671), and the Science and Technology Bureau of Chengdu City (2021-YF05-00681-SN, 2022-YF05-00449-SN).

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/fpls.2023.1166710/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bai, S. S., Yuan, F. P., Zhang, H. B., Zhang, Z. Y., Zhao, J. X., Yang, Q. H., et al. (2020). Characterization of the wheat-*Psathyrostachys huashanica* keng 2Ns/2D substitution line H139: A novel germplasm with enhanced resistance to wheat take-all. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00233

Bai, S. S., Zhang, H. B., Han, J., Wu, J. H., Li, J. C., Geng, X. X., et al. (2021). Identification of genetic locus with resistance to take-all in the wheat-*Psathyrostachys huashanica* keng introgression line H148. *J. Integr. Agric.* 12, 3101–3113. doi: 10.1016/S2095-3119(20)63340-8

Chen, N., Chen, W. J., Yan, H., Wang, Y., Kang, H. Y., Zhang, H. Q., et al. (2020). Evolutionary patterns of plastome uncover diploid-polyploid maternal relationships in triticeae. *Mol. Phylogenet. Evol.* 149, 106838. doi: 10.1016/j.ympev.2020.106838

Chen, S. Y., Zhang, A. J., and Fu, J. (1991). The hybridization between *Triticum aestivum* and *Psathyrostachys huashanica*. *Acta Genet. Sin.* 18 (6), 508–512.

Du, W. L., Wang, J., Pang, Y. H., Wu, J., Zhao, J., Liu, S. H., et al. (2014). Development and application of PCR markers specific to the 1Ns chromosome of *Psathyrostachys huashanica* keng with leaf rust resistance. *Euphytica* 200 (2), 207–220. doi: 10.1007/s10681-014-1145-x

Elshire, R. J., Glaubitz, J. C., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Emms, D. M., and Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16 (1), 157. doi: 10.1186/s13059-015-0721-2

Fan, C. L., Hao, M., Jia, Z. Y., Neri, C., Chen, X., Chen, W. S., et al. (2020). Some characteristics of crossing over in induced recombination between chromosomes of wheat and rye. *Plant J.* 105, 1665–1676. doi: 10.1111/tpj.15140

Fedak, G. (1999). Molecular aids for integration of alien chromatin through wide crosses. *Genome* 42 (4), 584–591. doi: 10.1139/g99-046

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29 (7), 644–652. doi: 10.1038/nbt.1883

Han, F. P., Lamb, J. C., and Birchler, J. A. (2006). High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Proc. Natl. Acad. Sci. U. S. A.* 103 (9), 3238–3243. doi: 10.1073/pnas.0509650103

Iehisa, J. C. M., Shimizu, A., Sato, K., Nishijima, R., Sakaguchi, K., Matsuda, R., et al. (2014). Genome-wide marker development for the wheat d genome based on single nucleotide polymorphisms identified from transcripts in the wild wheat progenitor *Aegilops tauschii*. *Theor. Appl. Genet.* 127 (2), 261–271. doi: 10.1007/s00122-013-2215-5

Jiang, J. M., Friebe, B., and Gill, B. S. (1994). Recent advances in alien gene transfer in wheat. *Euphytica* 73 (3), 199–212. doi: 10.1007/BF00036700

Jing, J. X., Fu, J., Yuan, H. X., Wang, M. N., and Li, Z. Q. (1999). A preliminary study on heredity of the resistance to stripe rust in three wild relatives of wheat. *Acta Phytopathol. Sin.* 29, 147–150.

Kang, H. Y., Wang, Y., Sun, G. L., Zhang, H. Q., Fan, X., and Zhou, Y. H. (2009). Production and characterization of an amphiploid between common wheat and *Psathyrostachys huashanica* keng ex kuo. *Plant Breed.* 128 (1), 36–40. doi: 10.1111/j.1439-0523.2008.01542.x

Kang, H. Y., Zhang, Z. J., Xu, L. L., Qi, W. L., Tang, Y., Wang, H., et al. (2016). Characterization of wheat-*Psathyrostachys huashanica* small segment translocation line with enhanced kernels per spike and stripe rust resistance. *Genome* 59 (4), 221–229. doi: 10.1139/gen-2015-0138

Kantarski, T., Larson, S., Zhang, X. F., Dehaan, L., Borevitz, J., Anderson, J., et al. (2016). Development of the first consensus genetic map of intermediate wheatgrass (*Thinopyrum intermedium*) using genotyping-by-sequencing. *Theor. Appl. Genet.* 130 (1), 137–150. doi: 10.1007/s00122-016-2799-7

Kanwal, N. (2019). Transcriptome analyses and population genetics of *Psathyrostachys huashancia*. *Northwest Univ.* doi: 10.27405/d.cnki.gxbdu.2019.000350

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

Khera, P., Upadhyaya, H. D., Pandey, M. K., Roorkiwal, M., Sriswathi, M., Janila, P., et al. (2013). Single nucleotide polymorphism–based genetic diversity in the reference set of peanut (*Arachis* spp.) by developing and applying cost-effective kompetitive allele specific polymerase chain reaction genotyping assays. *Plant Genome* 6 (3). doi: 10.3835/plantgenome2013.06.0019

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi: 10.1038/s41587-019-0201-4

Kishii, M., Dou, Q. W., Garg, M., Ito, M., Tanaka, H., and Tsujimoto, H. (2010). Production of wheat-*Psathyrostachys huashanica* chromosome addition lines. *Jpn. J. Genet.* 85 (4), 281–286. doi: 10.1266/ggs.85.281

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi: 10.1093/molbev/msw054

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Li, Z. H. (2019). *Advances in genome sequencing of p. huashanica. abstract from the 10th national congress of wheat genomics and molecular breeding*. (Yantai, China).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). 1000 genome project data processing subgroup. the sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, J. C., Zhao, L., Cheng, X. N., Bai, G. H., Li, M., Wu, J., et al. (2020). Molecular cytogenetic characterization of a novel wheat-*Psathyrostachys huashanica* keng T3DS-5NsL•5NsS and T5DL-3DS•3DL dual translocation line with powdery mildew resistance. *BMC Plant Biol.* 20, 163. doi: 10.1186/s12870-020-02366-8

Liu, Y., Chen, H., Li, C. X., Zhang, L. R., Shao, M. Q., Pang, Y. H., et al. (2021b). Development of diagnostic markers for a wheat leaf rust resistance gene *Lr42* using RNA-sequencing. *Crop J.* 9 (6), 1357–1366. doi: 10.1016/j.cj.2021.02.012

Liu, Y. X., Huang, S. H., Han, J., Hou, C. C., Zheng, D. S., Zhang, Z. M., et al. (2021a). Development and molecular cytogenetic identification of a new wheat-*Psathyrostachys huashanica* keng translocation line resistant to powdery mildew. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.689502

Liu, L. Q., Luo, Q. L., Teng, W., Li, B., Li, H. W., Li, Y. W., et al. (2018). Development of *Thinopyrum ponticum*- specific molecular markers and FISH probes based on SLAF-seq technology. *Planta* 247, 1099–1108. doi: 10.1007/s00425-018-2845-6

Liu, G. N., Mullan, D., Zhang, A. M., Liu, H., Liu, D. C., and Yan, G. J. (2022). Identification of KASP markers and putative genes for pre-harvest sprouting resistance in common wheat (*Triticum aestivum* l.). *Crop J.* doi: 10.1016/j.cj.2022.09.002

Lou, H., Dong, L., Zhang, K., Wang, D. W., Zhao, M., Li, Y., et al. (2017). High-throughput mining of e-genome-specific SNPs for characterizing *Thinopyrum elongatum* introgressions in common wheat. *Mol. Ecol. Resour.* 17, 1318–1329. doi: 10.1111/1755-0998.12659

Ma, D. F., Fang, Z. W., Yin, J. L., Chao, K. X., Jing, J. X., Li, Q., et al. (2016). Molecular mapping of stripe rust resistance gene *YrHu* derived from *Psathyrostachys huashanica*. *Mol. Breed.* 36, 64. doi: 10.1007/s11032-016-0487-6

Ma, H. H., Zhang, J. P., Zhang, J., Zhou, S. H., Han, H. M., Liu, W. H., et al. (2019). Development of p genome-specific SNPs and their application in tracing *Agropyron cristatum* introgressions in common wheat. *Crop J.* 37, 151–162. doi: 10.1016/j.cj.2018.07.003

Okada, M., Yoshida, K., Nishijima, R., Michikawa, A., Motoi, Y., Sato, K., et al. (2018). RNA-Seq analysis reveals considerable genetic diversity and provides genetic markers saturating all chromosomes in the diploid wild wheat relative *Aegilops umbellulata*. *BMC Plant Biol.* 18, 271. doi: 10.1186/s12870-018-1498-8

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS One* 7 (2), e32253. doi: 10.1371/journal.pone.0032253

Qu, X. J., Zhang, D. S., Zhang, X. Y., Wang, S. W., Wang, C. Y., Wang, Y. Z., et al. (2022). Cytogenetic and marker assisted identification of a wheat–*Psathyrostachys huashanica* keng f. ex P.C.Kuo alien substitution line conferring processing quality and resistance to stripe rust. *Genet. Resour. Crop Evol.* 69, 687–698. doi: 10.1007/s10722-021-01253-x

Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2014). Single nucleotide polymorphism genotyping using kompetitive allele specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol. Breed.* 33, 1–14. doi: 10.1007/s11032-013-9917-x

Sha, L. N., Fan, X., Li, J., Liao, J. Q., Zeng, J., Wang, Y., et al. (2017). Contrasting evolutionary patterns of multiple loci uncover new aspects in the genome origin and evolutionary history of *Leymus* (Triticeae; poaceae). *Mol. Phylogenet. Evol.* 114, 175–188. doi: 10.1016/j.ympev.2017.05.015

Song, L. Q., Zhao, H., Zhang, Z., Zhang, S., Liu, J. J., Zhang, W., et al. (2020). Molecular cytogenetic identification of wheat-*Aegilops biuncialis* 5Mb disomic addition line with tenacious and black glumes. *Int. J. Mol. Sci.* 21 (11), 4053. doi: 10.3390/ijms21114053

Su, Z. Q., Jin, S. J., Zhang, D. F., and Bai, G. H. (2018). Development and validation of diagnostic markers for Fhb1 region, a major QTL for fusarium head blight resistance in wheat. *Theor. Appl. Genet.* 131, 2371–2380. doi: 10.1007/s00122-018-3159-6

Sun, C., Liu, Y. K., Chao, K. X., Fang, Z. W., Wang, S. P., Tang, Q., et al. (2018). Characterization and molecular mapping of stripe rust resistance in wheat – *Psathyrostachys huashanica* introgression line H9015-17-1-9-6. *Can. J. Plant Pathol.* 41 (1), 65–75. doi: 10.1080/07060661.2018.1523230

Sun, G. L., Yang, J. L., and Yen, C. (1993). Endangering reason and reproductive strategy of *Psathyrostachys huashanica* population. *J. Syst. Evol.* 25 (5), 393–398. doi: 10.1088/0256-307X/18/11/313

Tamura, K., Nei, M., and Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.* 101 (30), 11030–11035. doi: 10.1073/pnas.0404206101

Tan, B. W., Zhao, L., Li, L. Y., Zhang, H., Zhu, W., Xu, L. L., et al. (2021). Identification of a wheat-*Psathyrostachys huashanica* 7Ns ditelosomic addition line conferring early maturation by cytological analysis and newly developed molecular and FISH markers. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.784001

Wang, K. Y., Lin, Z. S., Wang, L., Wang, K., and Ye, X. G. (2018). Development of a set of PCR markers specific to *Aegilops longissima* chromosome arms and application in breeding a translocation line. *Theor. Appl. Genet.* 131 (1), 13–25. doi: 10.1007/s00122-017-2982-5

Wang, R. R. C., and Lu, B. (2014). Biosystematics and evolutionary relationships of perennial triticeae species revealed by genomic analyses. *J. Syst. Evol.* 52 (6), 697–705. doi: 10.1111/jse.12084

Wang, J., Wang, L. M., Du, W. L., Chen, L. G., Liu, S. H., Wu, J., et al. (2014). Development of 5Ns chromosome-specifc SCAR markers for utilization in future wheat breeding programs. *Genetika* 50 (6), 692–699. doi: 10.1134/S1022795414060131

Yang, H., and Wang, K. (2018). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10 (10), 1556–1566. doi: 10.1038/nprot.2015.105

Yen, C., and Yang, J. L. (2011). *Biosystematics of triticeae. 1st ed* Vol. Volume IV (Beijing: China Agricultural Press).

Zeng, F. Q., Biligetu, B., Coulman, B., Schellenberg, M. P., and Fu, Y. B. (2017). RNA-Seq analysis of gene expression for floral development in crested wheatgrass (*Agropyron cristatum* l.). *PloS One* 12 (5), e0177417. doi: 10.1371/journal.pone.0177417

Zhang, J. P., Liu, W. H., Lu, Y. Q., Liu, Q. X., Yang, X. M., Li, X. Q., et al. (2017). A resource of large-scale molecular markers for monitoring *Agropyron cristatum* chromatin introgression in wheat background based on transcriptome sequences. *Sci. Rep.* 7 (1), 11942. doi: 10.1038/s41598-017-12219-4

Zhang, H., Wang, F., Zeng, C. Y., Zhu, W., Xu, L. L., Wang, Y., et al. (2022). Development and application of specific FISH probes for karyotyping *Psathyrostachys huashanica* chromosomes. *BMC Genomics* 23, 309. doi: 10.1186/s12864-022-08516-6

Zhou, S. H., Yan, B. Q., Li, F., Zhang, J. P., Zhang, J., Ma, H. H., et al. (2017). RNA-Seq analysis provides the first insights into the phylogenetic relationship and interspecific variation between *Agropyron cristatum* and wheat. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01644

frontiers | Frontiers in Plant Science

TYPE Data Report
PUBLISHED 04 July 2023
DOI 10.3389/fpls.2023.1220925

OPEN ACCESS

EDITED BY
Mark Chapman,
University of Southampton,
United Kingdom

REVIEWED BY
Daniel B. Marchant,
Stanford University, United States
Xiao Chun Wan,
Anhui Agricultural University, China
Hong Chen,
Jiangsu Province and Chinese Academy of Sciences, China

*CORRESPONDENCE
Ke-Wang Xu
xukw10@njfu.edu.cn
Xiaozeng Yang
yangxz@sRNAworld.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 11 May 2023
ACCEPTED 16 June 2023
PUBLISHED 04 July 2023

CITATION
Guo Z, Wei J, Xu Z, Lin C, Peng Y, Wang Q, Wang D, Yang X and Xu K-W (2023) HollyGTD: an integrated database for holly (Aquifoliaceae) genome and taxonomy. *Front. Plant Sci.* 14:1220925. doi: 10.3389/fpls.2023.1220925

HollyGTD: an integrated database for holly (Aquifoliaceae) genome and taxonomy

Zhonglong Guo[1†], Junrong Wei[1†], Zhenxiu Xu[1], Chenxue Lin[1], Ye Peng[1], Qi Wang[1], Dong Wang[2,3], Xiaozeng Yang[2*] and Ke-Wang Xu[1*]

[1]Co–Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing, China, [2]Institute of Biotechnology, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [3]WeiRan Biotech, Beijing, China

KEYWORDS

holly, Aquifoliaceae, genome, taxonomy, database

Introduction

Aquifoliaceae, also known as the holly family, comprising the single species-rich genus *Ilex* L. and more than 600 species (Loizeau et al., 2016). Species in this family are dioecious shrubs or trees. It is sub-cosmopolitan, but is best represented in mountainous areas of the tropics, especially in Asia, Central and South America. Many holly species possess great economic value and folk cultural significance. Some of them are commonly used as ornamental plants in parks and gardens for their foliage and decorative berries, such as the common holly *I. aquifolium*, the American holly *I. opaca*, the horned holly *I. cornuta*, and the Japanese holly *I. crenata*. The fruiting branches are also popularly applied to decorate temple courts in China and Christmas trees in the West. Some hollies can also be made into beverages, including *I. paraguariensis* (the "Yerba Mate" or Paraguay Tea in South America), *I. vomitoria* (the "Cassena" or Black Drink in North America and Mexico), *I. latifolia* (Kudingcha in East Asia).

In recent years, genome sequencing has become an important step to decipher the genetic structure and to understand the biological principles controlling the various traits of these plants (Boutanaev et al., 2015; Bredeson et al., 2022; Shen et al., 2023). In order to better store, inquire, mine, integrate, and disseminate the abundant datasets, more and more special comprehensive databases have been launched during the past several years (Harper et al., 2016; Jung et al., 2019; Guo et al., 2023). As a group with important economic value, the genomic and genetic data have been rapidly accumulated for hollies (Kong et al., 2022; Xu et al., 2022a; Yao et al., 2022). However, there is still no integrative database for comparative genomics and transcriptomics of hollies to study gene function and genome evolution. The research community for holly has gathered a significant amount of taxonomic information over the last few decades, including type locality, type specimens, and herbarium code (Manen et al., 2010; Xu et al., 2022b; Yang et al., 2023). But

the lack of a standardized platform for data processing and visualization limits the accessibility of such data.

Herein, we developed the Holly Genome and Taxonomy Database (HollyGTD) (https://hollygdb.com/), which integrates the holly data from public databases with the data produced by our group. The HollyGTD combines a variety of multi-omics data (genome, re-sequencing, and transcriptome) and taxonomic resources with a wealth of phenotypic images. HollyGTD offers a couple of easy-to-use access functions/interfaces and eight built-in tools for data analysis, for instance, Blast, JBrowse, Search Gene, Tissue Expression, Gene Annotation, Phylogenetic Tree, Primer Design, and Literature. Therefore, we believe that HollyGTD, a comprehensive database with useful data on genome, genotype, and taxonomy, may represent a valuable resource for the entire holly research community.

Materials and methods

Hardware and software

On a Linux server powered by Alibaba Cloud technology, the HollyGTD website is hosted. Technical assistance and web application development have both used the PHP language. The back-end servers were developed by MySQL. HollyGTD's website interfaces were created using HTML, CSS, and JavaScript. To produce interactive data visualizations, Highcharts (https://www.highcharts.com) was integrated with histograms and heatmaps.

Resources of genome references and annotations

Two chromosome level genomes in HollyGTD, *Ilex asprella* and *I. polyneura*, were retrieved from NGDC (CNCB-NGDC Members and Partners, 2022) and NCBI (Barrett et al., 2013), respectively. The assembly and annotation of the *Ilex latifolia* genome were done by our group. Genome resources were available in Supplementary Table S1.

Genotyping of re-sequencing data

The raw re-sequencing data of 114 *Ilex* species were produced using Illumina Hiseq X Ten platform by our group (Supplementary Table S1). After removing the adapter using trim_galore v0.5.0 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), clean reads were mapped to the *I. latifolia* genome using bwa v0.7.17 (Li, 2013). The variants were then invoked using the standard GATK v4.1.2.0 pipeline (Van der Auwera et al., 2013). SNPs and allele frequency (more than 0.05) were further analyzed. SnpEff v5.1 (Cingolani et al., 2012) was performed to identify SNPs in exons, introns, intergenic regions, 5' UTRs and 3' UTRs according the GFF3 file of *I. latifolia*.

Gene annotation *via* InterProScan

Using InterProScan (5.30), functional domains of protein-coding genes were discovered (Jones et al., 2014). A detailed page with information on homologous, families, domains, repeats, and GO terms was assigned to each gene.

Taxonomy and phylogenetic tree

Nomenclature of 808 scientific names of Aquifoliaceae were retrieved from Tropicos (https://www.tropicos.org/home) and Jstor (https://www.jstor.org/). Photos of leaves, flowers, pollens, whole plants, and so on were collected from our group. The phylogenetic tree was obtained from Yang's research (Yang et al., 2023).

Literature collection

Using the Python Entrez library, automated searches for the terms "*Ilex* AND Aquifoliaceae" were created. Then, 709 holly-related literatures were kept after manual filtration.

Content of HollyGTD

HollyGTD is made up of three parts: modules, data, and tools (Figure 1). These three parts work together to better organize all of the current data stored in bulk on HollyGTD and to provide users with user-friendly interfaces and easy-to-use tools.

HollyGTD harbors three major modules or interfaces to present the genome, genotype, and taxonomy datasets (Figure 1). Through these modules, users can easily access the underlying data. 1) Genome, which offers comprehensive details on three reference genomes and associated annotations; 2) Genotype, which provides variations produced from re-sequencing data of 114 species *via* visual and searchable access ports; 3) Taxonomy, which houses taxonomic data on every Aquifoliaceae species and arranges all of the manually collected phenotypic images by our group.

Data in HollyGTD include three genomes and associated annotations, 114 re-sequencing data from distinct species of holly, 21 RNA-Seq datasets with different developmental stages, taxonomic information of 808 scientific names, more than 700 research papers published in the last decades, and batched phenotypic photos.

The third part of HollyGTD is designed to create and integrate eight related tools with various functions or data in order to make it easier for users to use and download these data (Figure 1). Blast, JBrowse, Primer Design, Search Gene, and Gene Annotation are tools related to various genomics data. Tissue Expression tool interactively displays transcriptomic datasets among distinct developmental stages of fruits and leaves. Phylogenetic Tree enables users to search against the most recent taxonomic relationship of Aquifoliaceae according to Yang's study (Yang et al., 2023). Literature is used to fast retrieval and access

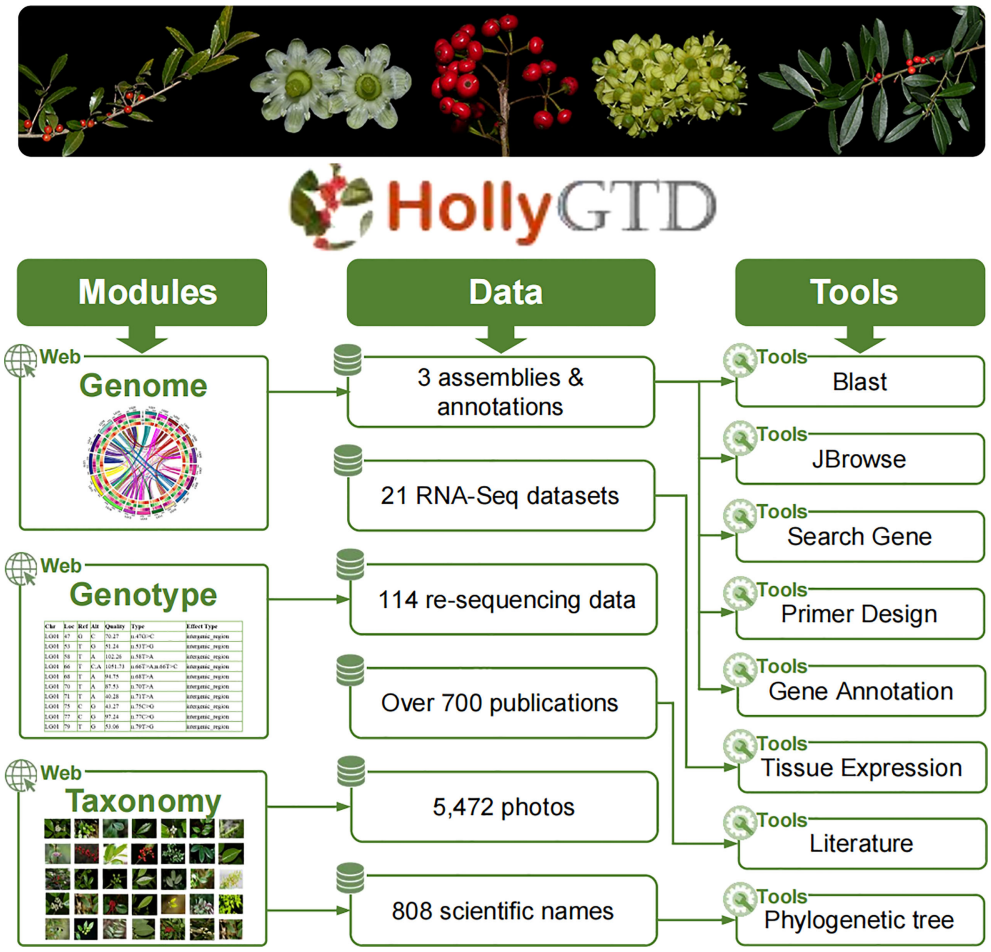


FIGURE 1
Framework of three parts at Holly Genome Database.

published researches on holly. In addition to these tools, browsers, search engines, filters, and other tools are available to make HollyGTD use easier.

Tools of HollyGTD

Blast

Blast allows users to search the homologous sequences of interest against three holly genomes (Figure 2A), either through filling a sequence in the text box or uploading a fasta file. Users can customize their query with advanced options and choose one of the five Blast options (blastn, blastp, blastx, tblastn, or tblastx) that are available. The output results of Blast hits are shown as collapsible fields in a standard table with the following columns: Query name, Target name, Score, Identities, Percentage, and Expect.

JBrowse

JBrowse is an open-source, extensible and comprehensive computational platform used to visualize and integrate genomic and multi-omics data (Buels et al., 2016). The integrated data of three genomes and annotated genomic datasets are displayed in HollyGTD using JBrowse2 (Figure 2B). HollyGTD currently provides three genome data, and users can easily browse and explore the information they need or are interested in, like the level of expression of particular genes.

Search gene

Users can search all annotated holly genes using the Search Gene tool, download the genomics, CDS, and protein of a particular gene, and view the gene structure and sequence using a graphic panel. This tool was developed to make it easier for users to use and download each gene's information (Figure 2C).

Tissue expression

Using *I. latifolia* as the reference genome, RNA-Seq datasets were used to determine each gene's expression level (Figure 2D). The Tissue Expression tool can find out the expression level of a given gene in green fruits, red fruits, and different developmental stages of leaves. To visualize the expression data, Highcharts

10.3389/fpls.2023.1220925

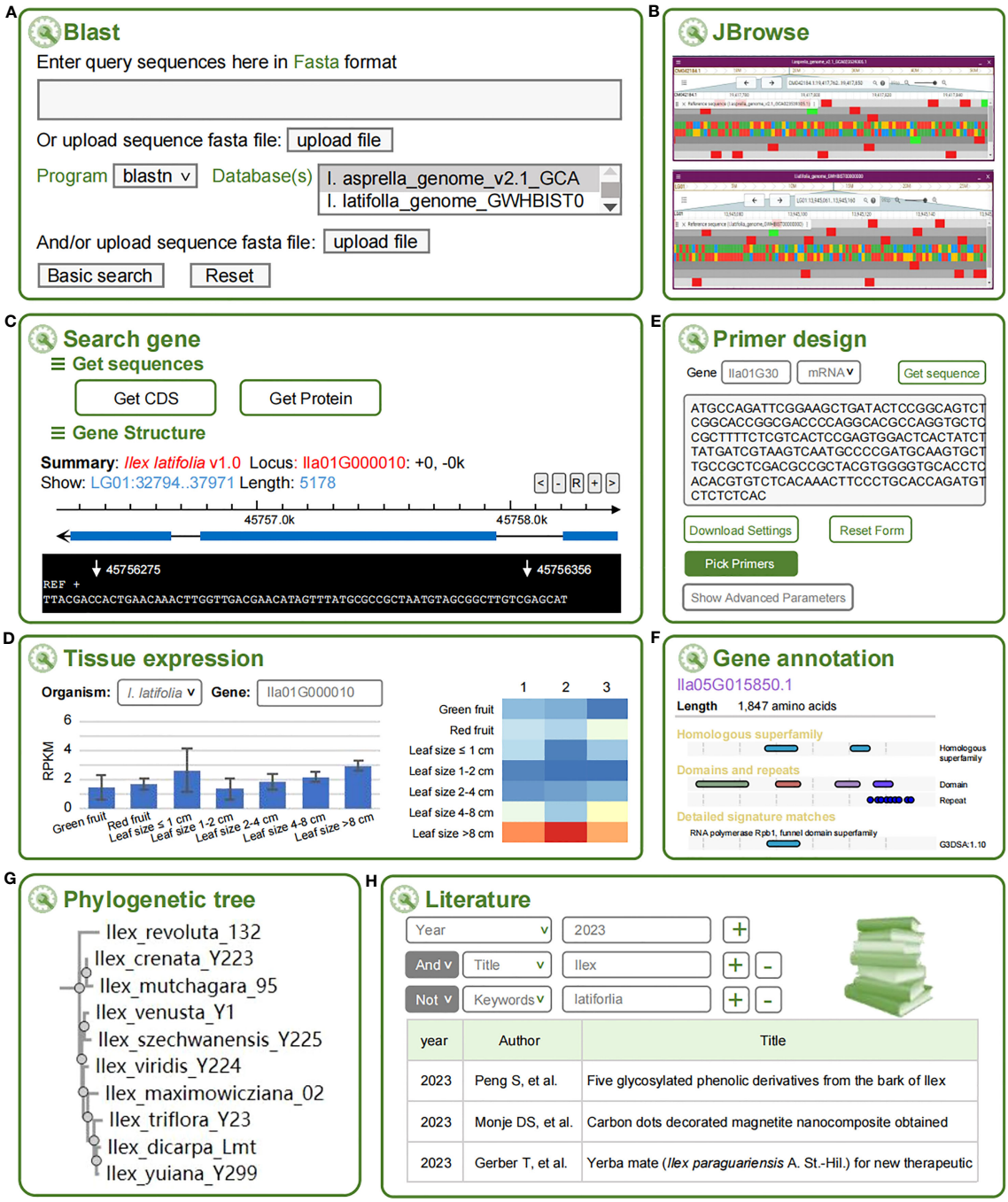


FIGURE 2
Eight tools at HollyGTD. **(A)** Blast. **(B)** JBrowse. **(C)** Search gene. **(D)** Tissue expression. **(E)** Primer design. **(F)** Gene annotation. **(G)** Phylogenetic tree. **(H)** Literature.

(https://www.highcharts.com) was performed to generate an interactive and dynamic histogram and heatmap. When the cursor is placed over a point on the heatmap, the gene ID, SRR ID, FPKM, and other pertinent data are displayed.

Primer design

A web-based PCR primer design tool, Primer-Design, is created with primer3 (Untergasser et al., 2012) as the core program to facilitate the users' molecular experiment (Figure 2E). In addition to the standard primer design function, some novel features for genetic experiment design are available. For instance, by entering the gene ID, the genomic sequences can be automatically loaded into the input field. Additionally, users have a variety of parameters for primer design.

Gene annotation

Gene Annotation tool gathers additional functional annotations for each gene, such as detailed information on a specific gene family, homologous superfamily, domains, repeats and GO (Gene Ontology) terms obtained through the InterPro database (Blum et al., 2021) (Figure 2F).

Phylogenetic tree

Based on the newly generated phylogenetic tree using rigorously identified 202 species and closely authenticated gene sequences of three nuclear genes (ITS, ETS, and *nepGS*), Phylogenetic Tree tool serves users with a convenient web search to retrieve the systematic status of the queried species (Figure 2G).

Literature

HollyGTD offers a specialized literature retrieval tool for holly scientific research, consisting of more than 700 papers published in the past few decades, to facilitate efficient literature triage and curation (Figure 2H). The literature search tool supports keyword searches for years, authors, titles, and journals, while the hyperlinks to full-texts publications are provided in the list of research result.

Data availability statement

The sources of omics data in HollyGTD are available at Supplementary Table S1. The original contributions presented in the study are publicly available. This data can be found here: https://ngdc.cncb.ac.cn/gwh, GWHBIST00000000.

Author contributions

K-WX, XY and ZG designed the project. ZG and JW designed and developed the HollyGTD website. JW and DW improved the web interface. CL and YP collected and collated the data. ZG and JW performed the bioinformatic analyses. K-WX, ZG and JW wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Natural Science Foundation of Jiangsu Province (#BK20210612), the National Natural Science Foundation of China (#32100167), the Nanjing Forestry University project funding (#163108093) and Beijing Academy of Agriculture and Forestry Sciences (#JKZX2022201).

Conflict of interest

Author DW was employed by company WeiRan Biotech.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1220925/full#supplementary-material

References

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J., et al. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E81–E88. doi: 10.1073/pnas.1419547112

Bredeson, J. V., Lyons, J. B., Oniyinde, I. O., Okereke, N. R., Kolade, O., Nnabue, I., et al. (2022). Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat.Commund* 13, 2001. doi: 10.1038/s41467-022-29114-w

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17, 66. doi: 10.1186/s13059-016-0924-1

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

CNCB-NGDC Members and Partners (2022). Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res.* 50, D27–d38. doi: 10.1093/nar/gkab951

Guo, Z., Li, B., Du, J., Shen, F., Zhao, Y., Deng, Y., et al. (2023). LettuceGDB: the community database for lettuce genetics and omics. *Plant Commun.* 4, 100425. doi: 10.1016/j.xplc.2022.100425

Harper, L., Gardiner, J., Andorf, C., and Lawrence, C. J. (2016). MaizeGDB: the maize genetics and genomics database. *Methods Mol. Biol.* 1374, 187–202. doi: 10.1007/978-1-4939-3167-5_9

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

Jung, S., Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., et al. (2019). 15 years of GDR: new data and functionality in the genome database for rosaceae. *Nucleic Acids Res.* 47, D1137–D1145. doi: 10.1093/nar/gky1000

Kong, B. L., Nong, W., Wong, K. H., Law, S. T., So, W. L., Chan, J. J., et al. (2022). Chromosomal level genome of *Ilex asprella* and insight into antiviral triterpenoid pathway. *Genomics* 114, 110366. doi: 10.1016/j.ygeno.2022.110366

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2*. doi: 10.48550/arXiv.1303.3997

Loizeau, P. A., Savolainen, V., Andrews, S., and Spichiger, R. (2016). "Aquifoliaceae," in *Flowering plants. eudicots, the families and genera of vascular plants.* Ed. K. Kubitzki (Berlin: Springer), 31–36.

Manen, J. F., Barriera, G., Loizeau, P. A., and Naciri, Y. (2010). The history of extant *Ilex* species (Aquifoliaceae): evidence of hybridization within a Miocene radiation. *Mol. Phylogenet. Evol.* 57, 961–977. doi: 10.1016/j.ympev.2010.09.006

Shen, F., He, H., Huang, X., Deng, Y., and Yang, X. (2023). Insights into the convergent evolution of fructan biosynthesis in angiosperms from the highly characteristic chicory genome. *New Phytol.* 238, 1245–1262. doi: 10.1111/nph.18796

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115. doi: 10.1093/nar/gks596

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf.* 43, 11.10.11–11.10.33. doi: 10.1002/0471250953

Xu, K., Lin, C., Lee, S. Y., Mao, L., and Meng, K. (2022b). Comparative analysis of complete *Ilex* (Aquifoliaceae) chloroplast genomes: insights into evolutionary dynamics and phylogenetic relationships. *BMC Genom.* 23, 203. doi: 10.1186/s12864-022-08397-9

Xu, K. W., Wei, X. F., Lin, C. X., Zhang, M., Zhang, Q., Zhou, P., et al. (2022a). The chromosome-level holly (*Ilex latifolia*) genome reveals key enzymes in triterpenoid saponin biosynthesis and fruit color change. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.982323

Yang, Y., Jiang, L., Liu, E.-D., Liu, W.-L., Chen, L., Kou, Y.-X., et al. (2023). Time to update the sectional classification of *Ilex* (Aquifoliaceae): new insights from *Ilex* phylogeny, morphology, and distribution. *J. Syst. Evol.* doi: 10.1111/jse.12935

Yao, X., Lu, Z., Song, Y., Hu, X., and Corlett, R. T. (2022). A chromosome-scale genome assembly for the holly (*Ilex polyneura*) provides insights into genomic adaptations to elevation in southwest China. *Hortic. Res.* 9, uhab049. doi: 10.1093/hr/uhab049

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 July 2023
DOI 10.3389/fpls.2023.1205051

OPEN ACCESS

EDITED BY
Gao Jihai,
Chengdu University of Traditional Chinese Medicine, China

REVIEWED BY
Huasheng Peng,
China Academy of Chinese Medical Sciences, China
Yifei Liu,
Hubei University of Chinese Medicine, China

*CORRESPONDENCE
Peiyao Xin
xpytgyx@163.com
Yu Song
songyu@gxnu.edu.cn

RECEIVED 20 April 2023
ACCEPTED 29 May 2023
PUBLISHED 07 July 2023

CITATION
Cao Z, Yang L, Xin Y, Xu W, Li Q, Zhang H, Tu Y, Song Y and Xin P (2023) Comparative and phylogenetic analysis of complete chloroplast genomes from seven *Neocinnamomum* taxa (Lauraceae). *Front. Plant Sci.* 14:1205051. doi: 10.3389/fpls.2023.1205051

Comparative and phylogenetic analysis of complete chloroplast genomes from seven *Neocinnamomum* taxa (Lauraceae)

Zhengying Cao[1,2], Linyi Yang[3], Yaxuan Xin[1,2], Wenbin Xu[4], Qishao Li[1,2], Haorong Zhang[1,2], Yuxiang Tu[1,2], Yu Song[5]* and Peiyao Xin[1,2]*

[1]Southwest Research Center for Landscape Architecture Engineering, National Forestry and Grassland Administration, Southwest Forestry University, Kunming, China, [2]Key Laboratory of Forest Resources Conservation and Utilization in the Southwest Mountains of China Ministry of Education, Southwest Forestry University, Kunming, China, [3]Yunnan Forestry Vocational and Technical College, Kunming, Yunnan, China, [4]Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, China, [5]Key Laboratory of Ecology of Rare and Endangered Species and Environmental Protection (Ministry of Education) & Guangxi Key Laboratory of Landscape Resources Conservation and Sustainable Utilization in Lijiang River Basin, Guangxi Normal University, Guilin, Guangxi, China

The genus *Neocinnamomum* is considered to be one of the most enigmatic groups in Lauraceae, mainly distributed in tropical and subtropical regions of Southeast Asia. The genus contains valuable oilseed and medicinal tree species. However, there are few studies on the genus *Neocinnamomum* at present, and its interspecific relationship is still unclear. In order to explore the genetic structure and evolutionary characteristics of the *Neocinnamomum* chloroplast genome and to resolve the species relationships within the genus, comparative genomic and phylogenetic analyses were performed on the whole chloroplast genome sequences of 51 samples representing seven *Neocinnamomum* taxa. The whole *Neocinnamomum* chloroplast genome size ranged from 150,753-150,956 bp, with a GC content of 38.8%-38.9%. A total of 128 genes were annotated within the *Neocinnamomum* chloroplast genome, including 84 protein coding genes, 8 rRNA genes, and 36 tRNA genes. Between 71-82 SSRs were detected, among which A/T base repeats were the most common. The chloroplast genome contained a total of 31 preferred codons. Three highly variable regions, *trnN-GUU-ndhF*, *petA-psbJ*, and *ccsA-ndhD*, were identified with Pi values > 0.004. Based on the whole chloroplast genome phylogenetic tree, the phylogenetic relationships among the seven *Neocinnamomum* taxa were determined. *N. delavayi* and *N. fargesii* were the most closely related species, and *N. lecomtei* was identified as the most basal taxon. In this study, the characteristics and sequence variation of the chloroplast genomes of seven *Neocinnamomum* taxa were revealed, and the genetic relationship among the species was clarified. The results of this study will provide a reference for subsequent molecular marker development and phylogenetic research of *Neocinnamomum*.

KEYWORDS

***Neocinnamomum*, chloroplast genome, genome comparison, sequence characteristic, phylogenetic analysis**

Introduction

Neocinnamomum (Neocinnamomeae, Lauraceae) is a genus of evergreen shrubs and small trees distributed across south-central, western, and northwestern China, as well as Nepal, Myanmar, and Vietnam, among other areas (Flora of China, 1982). To date, there are 7 known species of *Neocinnamomum* worldwide. Five of these species (*N. caudatum*, *N. fargesii*, *N. lecomtei*, *N. mekongense*, and *N. delavayi*) are located in China, and are primarily distributed across Hainan, Guangxi, Yunnan, and Sichuan, as well as Tibet (Flora of China, 1982). In 2017, Xu et al. (2017) discovered a variant of *N. caudatum* (*N. caudatum* var. *macrocarpum*) in Baise, Guangxi, China. Outside of China, *N. atjehense* and *N. parvifolium* (revised as *N. delavayi*) are also recorded in the world flora (WFO, 2023). *Neocinnamomum* includes important oil-bearing tree species as well as valuable medicinal resources. In China, *N. caudatum* is known as "Baigui" and *N. delavayi* is known as "Sangujin" (Wu, 1983). The bark and leaves of these species have a long history of use in traditional Chinese medicine (TCM) for dispelling wind and cold, promoting blood circulation, and relieving blood stasis (Jiangsu New Medical College, 1977). In addition, the seeds of *Neocinnamomum* species are rich in fatty acids (Gan et al., 2018). For example, the seeds of *N. caudatum* contain approximately 54% oil (dry weight), and the seeds of *N. delavayi* contain approximately 57% oil. These oils have the potential to be developed and promoted as raw materials for biodiesel production (Wang et al., 2013), which is important for sustainable energy production. Genomic research on the genus *Neocinnamomum* currently lags far behind that of other oil plants. In addition, the degree of exploitation and utilization of wild *Neocinnamomum* resources is also extremely low.

The genus *Neocinnamomum* was originally established by Liu in 1934, and was distinguished from other genera of Lauraceae by the presence of four-locular anthers with collateral pollen sacs (Liu, 1934). However, this feature appears to only exist in *N. delavayi* and a few individuals of *N. caudatum*. Because of this, Kostermans (Kostermans, 1974) suggested that these anther characteristics were insufficient to distinguish the genus *Neocinnamomum*. Instead, he suggested that the presence of compound cymes; shallow, fleshy fruiting receptacles; persistently enlarged tepals; and dichotomous leaves were better distinguishing characteristics. Despite these traits, Kostermans did not deny the morphological similarity between *Neocinnamomum* and *Cinnamomum*. Traditional classification systems have long considered the two genera to be closely related. However, more recent molecular phylogenetic studies have called this relationship into question, obscuring the phylogenetic position of the genus *Neocinnamomum* within Lauraceae. The development of molecular systematics techniques has allowed researchers to utilize chloroplast gene fragments, whole chloroplast genomes, and nrDNA to study *Neocinnamomum* phylogenetics (Chanderbali et al., 2001; Rohwer and Rudolph, 2005; Wang et al., 2010; Song et al., 2020). In contrast to the traditional understanding that *Neocinnamomum* and *Cinnamomum* are closely related, these recent studies indicate that *Neocinnamomum* is a monophyletic group with *Cassytha* and *Caryodaphnopsis* as its closest relatives.

Although the monophyly of the genus *Neocinnamomum* has been confirmed, the relationships between species within the genus are less clear. Kostermans (1974) divided the genus *Neocinnamomum* into six species, four of which are easily distinguished. However, *N. mekongense* and *N. delavayi* are extremely similar morphologically, and can be distinguished only based on the presence of pubescent twigs. Recently, Wang et al. (2010) used three molecular fragments of *psbA-trnH*, the *trnK* cpDNA region, and ITS nrDNA to draw a phylogenetic tree of the genus *Neocinnamomum* and found that *N. mekongense* and *N. delavayi* were located on the same branch. Despite this, molecular fragment analysis has not satisfactorily resolved the relationship between the two species. Compare with the use of chloroplast (cp) gene fragments, the whole cp genome can provide much more robust variation information (Wariss et al., 2017). For example, Ren et al. (2019) constructed a maximum likelihood (ML) phylogenetic tree using the whole chloroplast genomes of *N. lecomtei*, *N. mekongense*, and *N. delavayi*. This highly-supported analysis indicated that *N. lecomtei* and *N. mekongense* were located on different branches. Unfortunately, only 5 samples from 3 species of *Neocinnamomum* were used for the study, limiting the power of the results.

Chloroplasts are the central hub for photosynthetic and certain other metabolic reactions in plants (Tao et al., 2017). Like mitochondria, cp are maternally inherited semi-autonomously and contain a relatively independent genetic system. In addition, cp have been central drivers of evolutionary processes (Neuhaus and Emes, 2000; Xing and Liu, 2008). As sequencing technology has developed, the publication of an ever-increasing number of plant cp genomes has given us a deeper understanding of the structure and variation of cp genomes. For the vast majority of angiosperms, the cp genome is characterized by a closed, double-stranded circular structure, and consists of two inverted repeat (IRs) regions, a large single-copy (LSC) region, and a small single-copy (SSC) region (Jansen et al., 2011). The structural stability of the cp genome is largely maintained by the conserved IR region, which is characterized by a base substitution rate of only 1/4 of that of the SC region (Drouin et al., 2008). The cp genome offers several advantages over the nuclear and mitochondrial genomes, including structural conservation, low molecular weight, simple structure, genetic stability, and moderate evolution rate (slower than the nuclear genome but higher than the mitochondrial genome) (Goremykin et al., 2003). Owing to these unique characteristics, cp genomes are widely used to explore the phylogeny and genetic relationship among plant clades (Yang et al., 2018; Cao et al., 2022; Lan et al., 2022; Wang et al., 2023; Xin et al., 2023). The cp genome is particularly advantageous for studies of species identification, molecular geography, and speciation processes (Xu et al., 2001).

To date, only a few *Neocinnamomum* genomes have been sequenced, and detailed cp genomic comparisons and phylogenetic analyses are lacking. In order to further clarify the phylogenetic relationships among species of the genus *Neocinnamomum*, and to obtain useful genetic resources, we sequenced and assembled the cp genomes of 50 samples representing 7 *Neocinnamomum* taxa. Included in the analysis

was one unique taxa recently collected from Wenshan, Yunnan, China, which shares some traits with *N. complanifructum* (merged into *N. lecomtei*). Based on 51 cp genome sequences (including 50 sequencing sequences and one sequence from NCBI), we analyzed the cp genome structure, gene content, codon usage frequencies, simple sequence repeats (SSRs), highly variable regions, and IR expansion and contraction. Finally, we reconstructed the phylogenetic relationships of 7 *Neocinnamomum* taxa. These results will be of great significance for studies of the population genetics, species identification, and conservation biology of the genus *Neocinnamomum*.

Materials and methods

Plant material sampling

Fifty samples of either fresh leaves or silica gel-dried material representing 7 *Neocinnamomum* species were collected, including 5 samples of *N. delavayi*, 6 samples of *N. mekongense*, 4 samples of *N. fargesii*, 2 samples of *N. caudatum* var. *macrocarpum*, 8 samples of *N. lecomtei*, 23 samples of *N. caudatum*, and 2 samples of *N. sp* (Supplementary Table S1). Plant samples were primarily collected in the Yunnan, Sichuan, Guangxi, and Hainan provinces of China, at an altitude of 1100-2300 m. Only fresh, tender, healthy leaves with no visible pest or disease damage were collected. Fresh samples were dried in self-sealed bags of silica gel prior to DNA extraction and sequencing. At the same time, the cp genome sequence data of *N. caudatum* (RL01) was downloaded from NCBI database for subsequent analysis.

DNA extraction and sequencing

Total genomic DNA was extracted from fresh and silica gel-dried leaves using a modified cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987). DNA integrity was evaluated using 1% agarose gel electrophoresis. DNA purity and concentration were evaluated by Nanodrop. Using qualified DNA samples, 150 bp double-end libraries were constructed and sequenced on the Illumina sequencing platform at Novogene Bioinformatics Technology Co. Ltd (Beijing, China). More than 4.0 Gb of reads was generated per sample.

Genome assembly and annotation and codon preference analysis

After the sequencing data were filtered and screened, the cp genome was automatically assembled using the GetOrganelle v1.7.1 (Jin et al., 2020). Bandage v0.8.1 (Wick et al., 2015) was used to visualize the assembly, remove redundant contigs, and edit the sequences into loops. The *N. mekongense* whole cp genome sequence (GenBank accession number NC_039718) was used as a reference, automatically annotated using CPGAVAS2 (http://47.96.249.172:16019/analyzer/home) (Liu et al., 2012), and adjusted using Geneious v9.1.7 (Kearse et al., 2012). The circular cp genome diagram was created using the OGDRAW online tool (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) (Lohse et al., 2013). The final assembled and annotated cp genome sequences were uploaded to the NCBI database to obtain GenBank accession numbers.

Eleven individuals of *Neocinnamomum* plants were selected for analysis (one sample was randomly selected from each taxon, and RL01 *N. caudatum*, 7751 N. *caudatum*, 6068 N. *delavayi* and 7683 N. *mekongense*).The relative synonymous codon usage (RSCU) of 11 cp genomes was statistically analyzed using CodonW1.4.2 software (https://galaxy.pasteur.fr/?form=codonw) (Liu and Xiu, 2005). Frequent codon usage (i.e., codon usage preference) is indicated when RSCU >1, while RSCU = 1 indicates no usage preference.

Simple sequence repeat analysis

The SSRs present in 11 cp genomes were detected by MISA (https://webblast.ipk-gatersleben.de/misa/) (Beier et al., 2017), using the following parameters: the threshold values for the number mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide repeats were set to 10, 5, 4, 3, 3, and 3, respectively (Wu et al., 2021).

Comparative chloroplast genome analysis

By examining the genes bordering IR regions, we can analyze IR expansion and contraction within the cp genomes of *Neocinnamomum* species. We manually examined the IR junctions of 11 cp genomes in Geneious software. To identify differences among the 11 cp genomes of the *Neocinnamomum* taxa, a comparative analysis of the full-length cp genome sequences was performed using mVISTA (http://genome.lbl.gov/vista/mvista/submit.shtml) (Frazer et al., 2004), with the cp genome of *N. delavayi* (6068) used as a reference. MAFFT v7.455 software (Katoh and Standley, 2013) was used to align the 51 cp genomes. The nucleotide polymorphism (Pi) between genomes was calculated using DnaSP v6.12.03 (Rozas et al., 2017), with a window length of 600 and a step size of 200 (Shen et al., 2022). The high-variance regions were screened by combining the mVISTA and Pi results.

Phylogenetic analysis

Phylogenetic relationships were reconstructed based on the 51 cp genomes, using *Caryodaphnopsis henryi*, *C. tonkinensis* and *C. malipoensis* from NCBI database as outgroup. Both the ML and Bayesian Inference (BI) methods were used to construct the phylogenetic tree. Each of the complete cp genomes were aligned using MAFFT v7.455 (Katoh and Standley, 2013), and then manually

10.3389/fpls.2023.1205051

edited using BioEdit v 7.2.5 (Hall, 1999). IQ-TREE v1.6.7 (Nguyen et al., 2015) was used to construct the ML phylogenetic tree, and the "GTR+F+R4" model was used for nucleic acid substitution. The support of each branch was verified by bootstrapping, with 1,000 iterations. BI analyses were performed using the Mrbayes v3.2.6 module of the CIPRES website (http://www.phylo.org/) (Huelsenbeck and Ronquist, 2001). Briefly, the processed data was uploaded to the CIPRES platform and jModeltest v2.1.10 (Posada, 2008) was used to determine the best nucleotide substitution model for phylogenetic reconstruction. Then, the Markov chain Monte Carlo (MCMC) algorithm was run for 1,000,000 generations. The results were sampled every 100 generations, and the first 25% of the generated trees were discarded. The "TPM1uf+I+G" model (freqA=0.3027, freqC=0.1972, freqG=0.1894, freqT=0.3107, R(a)[AC]=1.0000, R(b)[AG]=2.8265, R(c)[AT]=0.3346, R(d)[CG]=0.3346, R(e)[CT]=2.8265, R(f)[GT]=1.000, p-inv=0.6960, gamma shape=1.0440) was used to construct the BI phylogenetic tree. The phylogenetic trees were visualized and adjusted using FigTree v1.4.3 (Dong et al., 2022b).

Results

Features of the chloroplast genome

The structure of the *Neocinnamomum* cp genome is a typical circular double-stranded tetrad (Figure 1). The cp genome size of 51 *Neocinnamomum* plant samples ranged from 150,753 to 150,956 bp (Supplementary Table S2). The GC content was between 38.8%-38.9%, with *N. lecomtei* and *N. mekongense* having a higher GC content than the other studied taxa. We observed some variation in the length and GC content of the LSC, SSC, and IR regions of the cp genomes. The LSC region ranged in length between 91,850-92,006 bp, and had a GC content of 37.4%-37.5%. The SSC region ranged in length between 18,096-18,457 bp, and had a GC content of 33.2%-33.4%. The IR region ranged in length between 20,257-20,425 bp, and had a GC content of 44.5%-44.6%. Overall, the GC content of the IR region was significantly higher than in the SSC and LSC regions.

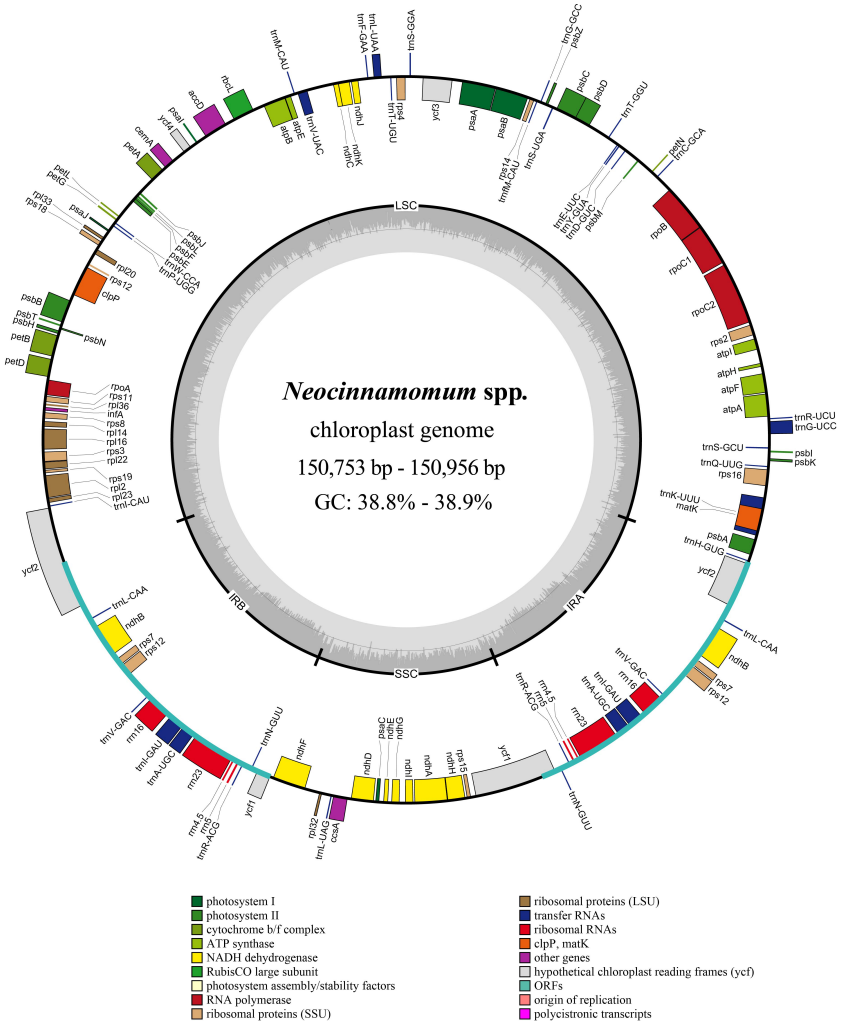


FIGURE 1
Circular map of the *Neocinnamomum* chloroplast genome. Genes belonging to different functional groups are color-coded. In the inner circle, the GC and AT content are denoted by the dark gray and light gray, respectively.

10.3389/fpls.2023.1205051

TABLE 1 Genes present in the *Neocinnamomum* chloroplast genome.

Category	Group of Genes	Genes Names
Photosynthesis gene	Photosystems I	*psaA*, *psaB*, *psaC*, *psaJ*, *psaI*
	Photosystems II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Cytochrome b/f complex	*petA*, *petB* [A], *petD* [A], *petG*, *petL*, *petN*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF* [A], *atpH*, *atpI*
	NADH dehydrogenase	*ndhA* [A], *ndhB* [A,C], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Rubisco Large subunit	*rbcL*
Self-replication gene	RNA polymerase	*rpoA*, *rpoB*, *rpoC1* [A], *rpoC2*
	Ribosomal proteins (SSU)	*rps2*, *rps3*, *rps4*, *rps7* [C], *rps8*, *rps11*,*rps12* [B,C], *rps14*, *rps15*, *rps16* [A], *rps18*, *rps19*
	Ribosomal proteins (LSU)	*rpl2* [A], *rpl14*, *rpl16* [A], *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*
	Transfer RNAs	*trnA-UGC* [A,C], *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-GCC*, *trnG-UCC* [A], *trnH-GUG*, *trnI-CAU* [C], *trnI-GAU* [A], *trnK-UUU* [A], *trnL-CAA* [C], *trnL-UAA* [A], *trnL-UAG*, *trnM-CAU*, *trnN-GUU* [C], *trnP-UGG*, *trnQ-UUG*, *trnR-ACG* [C], *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC* [C], *trnV-UAC* [A], *trnW-CCA*, *trnY-GUA*
	Ribosomal RNAs	*rrn4.5* [C], *rrn5* [C], *rrn16* [C], *rrn23* [C]
Other genes	Maturase	*matK*
	Envelop membrane protein	*cemA*
	Subunit of acetyl-CoA-carboxylase	*accD*
	Translational initiation factor	*infA*
	C-type cytochrome synthesis	*ccsA*
	Proteolysis	*clpP* [B]
Functions unknown	Hypothetical chloroplast reading frames (ycf)	*ycf1* [C], *ycf2* [C], *ycf3* [B], *ycf4*

A and B indicate one intron and two introns, respectively. C indicates two copies of the gene.

A total of 128 genes were annotated within the *Neocinnamomum* cp genome, including 84 protein coding genes, 8 rRNA genes, and 36 tRNA genes (Table 1). The coding genes were primarily composed of self-replication genes, photosynthesis genes, ycf genes, and "other" genes. Two copies each of 15 genes were located in the IR region, including 5 protein coding genes (*ndhB*, *rps7*, *rps12*, *ycf1*, and *ycf2*), 4 rRNA genes (*rrn4.5*, *rrn5*, *rrn16*, and *rrn23*) and 6 tRNA genes (*trnA-UGC*, *trnI-CAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*). In addition, 15 genes (9 protein coding genes and 6 tRNA genes) contained one intron and 3 genes (*rps12*, *clpP*, ycf3) contained two introns.

Codon bias analysis

Using MEGAX64 to determine the codon usage bias of the cp genomes, 64 codons, encoding 20 amino acids, were detected (Figure 2). Among these, three codons had RSCU = 1 and 31 codons had RSCU > 1, indicating a preference for these codons within the *Neocinnamomum* cp genome. Among 11 cp genomes of the 7 *Neocinnamomum* taxa, the AGA codon encoding arginine (R) was the most frequently used among the synonymous codons, with RSCU values ranging from 1.82-1.83. Least frequently used was the CGC codon encoding arginine (R), with RSCU values ranging from

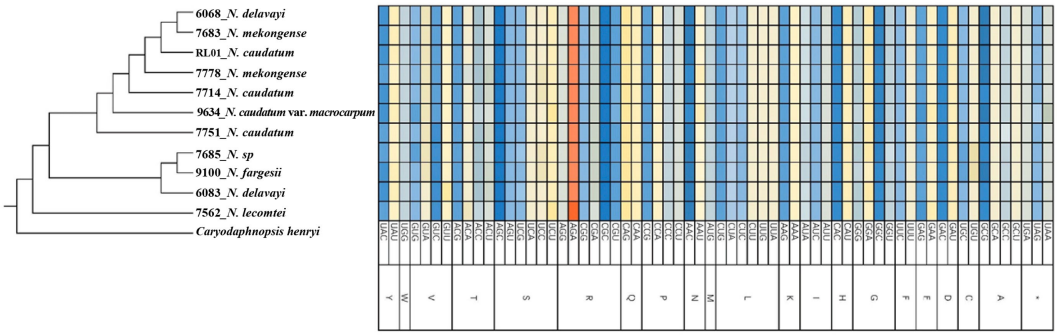


FIGURE 2
Relative synonymous codon usage (RSCU) of 20 amino acids and the stop codon for 11 chloroplast genomes. 20 uppercase letters such as A, C and D are abbreviations of 20 amino acids, and the symbol "*" is the stop codon.

0.41-0.42. Among the 20 amino acids, the most common were arginine (R), leucine (L), and serine (S), using 6 codons. The next most common were alanine (A), glycine (G), proline (P), threonine (T), and valine (V), using 4 codons. The least common were tryptophan (W) and methionine (M), using 1 codon. Except for methionine (M) and tryptophan (W), the preferred codons for all amino acids ended in A/U, indicating a preference for codons ending in these bases across the *Neocinnamomum* cp genome.

Simple sequence repeats analysis

SSRs are widely distributed in chloroplast genomes. We detected between 71-82 SSRs across each of 11 cp genomes representing 7 *Neocinnamomum* taxa, for a total of 828 SSRs. The SSRs ranged from single to hexanucleotide repeats, although not all six types of SSR were present in each taxa (Figure 3). Single nucleotide SSRs were the most common (52-63), and were mainly comprised of A/T repeats, with only 1-3 C/G repeats. There were 5-7 dinucleotide SSRs, all of which were AT/AT repeats. Only one trinucleotide SSR, an AAT/ATT repeat, was identified, with only one repeat per sample. There were 11-12 tetranucleotide SSRs, including four repeat types. However, the AAAC/GTTT repeat was present only in *N. delavayi* (6068), *N. mekongense* (7683), and *N. caudatum* (RL01). There were 0-2 pentanucleotide SSRs, including three repeat types (AATTC/AATTG, AAAAC/GTTTT, and ACGAT/ATCGT), which were present only in *N. delavayi* (6083) and *N. caudatum* (7714). The hexanucleotide AATTAG/AATTCT SSR was identified only in *N. caudatum* var. *macrocarpum* and *N. caudatum* (7751), with one repeat per sample.

Contraction and expansion of the IR regions

IR junctions analysis of 11 cp genomes representing 7 *Neocinnamomum* taxa indicated the presence of a contracted expansion of the tetrad structure SC/IR boundary (Figure 4). The LSC/IRb (JLB), IRb/SSC (JSB), SSC/IRa (JSA), and IRa/LSC (JLA) linkage boundaries were primarily associated with four genes: *ycf1*, *ycf2*, *ndhF*, and *ndhH*. The LSC/IRb (JLB) boundary and the SSC/IRa (JSA) boundary were located within the coding regions of *ycf2*

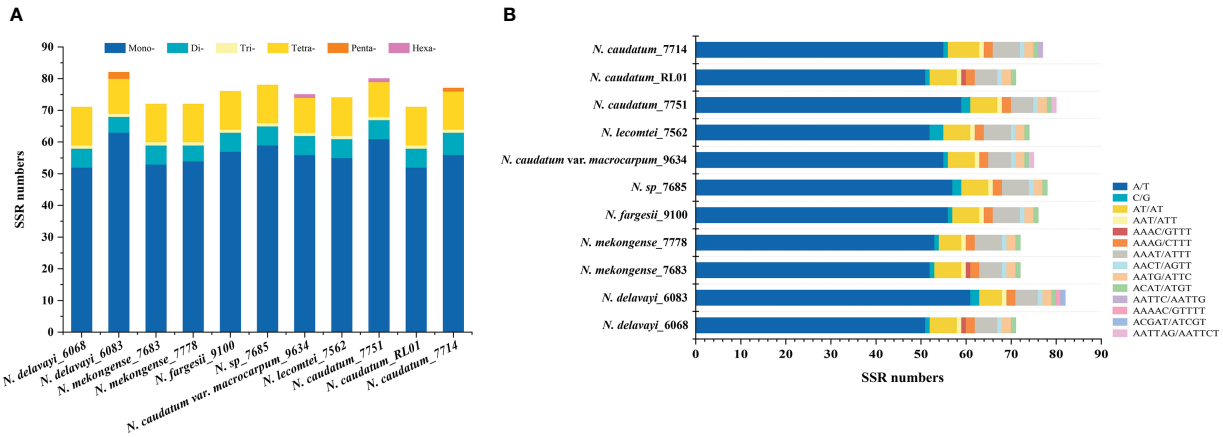


FIGURE 3
Comparison of SSRs across 11 *Neocinnamomum* cp genomes. **(A)** The type frequency of different SSR types. **(B)** The type frequency of SSR motifs in different repeat class types.

10.3389/fpls.2023.1205051

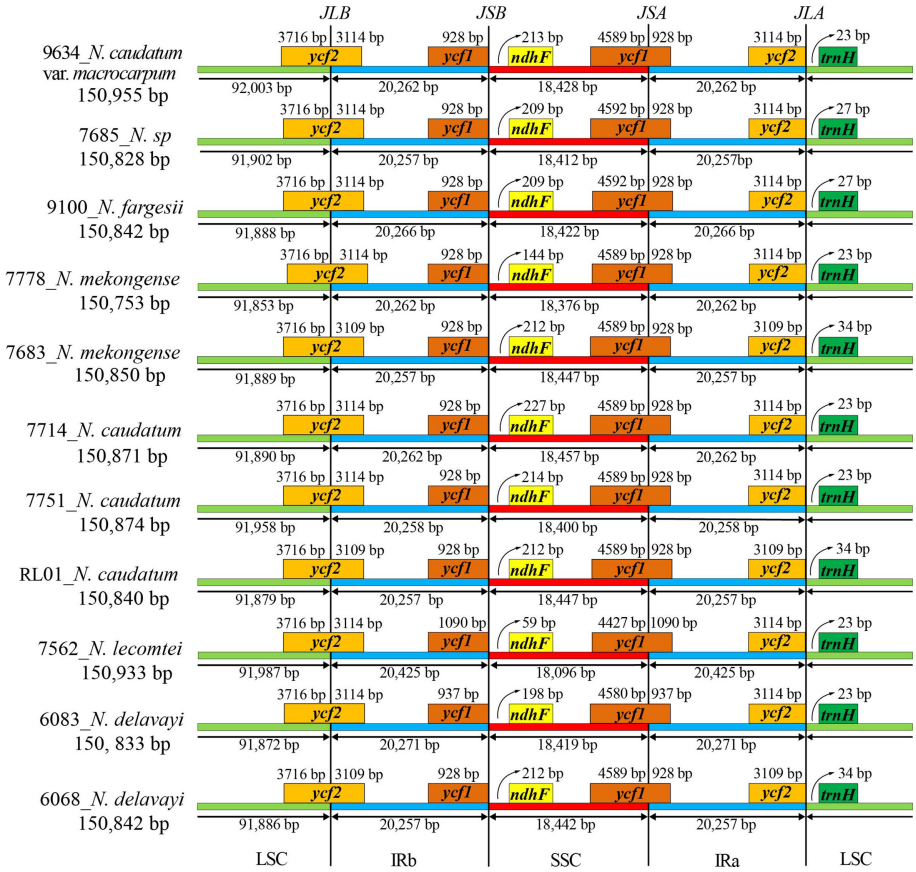


FIGURE 4
Expansion and contraction of the IR/SC boundary of the *Neocinnamomum* chloroplast genome.

and *ycf1*, respectively. In three samples (7683, 6068, and RL01), the *ycf2* gene expanded 3109 bp into the IRb region, which was 5 bp less than in other samples. The *ycf1* gene is different in both *N. delavayi* and *N. lecomtei* compared to the other taxa, with an additional expansion of 9 bp and 162 bp into the IRa region, respectively. The IRb/SSC (JSB) boundary was located in the noncoding region between *ycf1* and *ndhF*, with *trnF* of *N. lecomtei* closest to the boundary at 59 bp. The IRa/LSC (JLA) boundary was located in the noncoding region between *ycf2* and *ndhH*, with *trnH* of three samples (7683, 6068, and PL01) farthest from the boundary at 34 bp. Interestingly, the IR boundaries of *N. lecomtei* differed significantly from the other taxa. We hypothesize that this taxa either has an increased rate of evolution or diverged earlier. The distribution of genes at the SC/IR boundary was similar for samples 7683, 6068, and PL01, but differed somewhat from other samples of their respective taxa. Collectively, although the SC/IR boundary of the *Neocinnamomum* cp genome was relatively conservative, it also exhibited significant diversity.

Identification of variability hotspots

Using mVISTA software, the *N. delavayi* (6068) cp genome was used as the reference and compared with the full-length cp genomes of *N. delavayi* (6083), *N. mekongense* (7683 and 7778), *N. fargesii* (9100), *N. caudatum* (7714, 7751 and RL01), *N. caudatum* var. *macrocarpum* (9634), *N. sp* (7685), and *N. lecomtei* (7562) (Figure 5). Overall, both gene composition and order was relatively conserved across all 7 *Neocinnamomum* taxa. The IR region was more conserved than the LSC and SSC regions. The coding region was more conserved than the non-coding region, and variants primarily occurred in the spacer regions of adjacent genes, such as *trnN-GUU-ndhF*, *petA-psbJ*, *rbcL-accD*, and *psbE-petL*.

DnaSP software was used to compare the nucleotide variation values (Pi) between all genes and intergenic regions of the cp genomes. The Pi values of the 51 *Neocinnamomum* cp genomes varied from 0 to 0.01571, with a mean of 0.00098 (Figure 6). Although the *Neocinnamomum* cp genome was highly conserved, we identified three divergent hotspot regions (Pi > 0.004): *trnN-GUU-ndhF*, *petA-psbJ*, and *ccsA-ndhD*. Among them, *petA-psbJ* was located in the LSC region and had a Pi value of 0.00696. Both *trnN-GUU-ndhF* and *ccsA-ndhD* were located in the SSC region and had Pi values of 0.01571 and 0.00522, respectively.

Phylogenetic analysis

The phylogenetic relationships among 7 taxa of *Neocinnamomum* were reconstructed based on the cp genomes, with *C. henryi*, *C. tonkinensis* and *C. malipoensis* used as outgroup

10.3389/fpls.2023.1205051

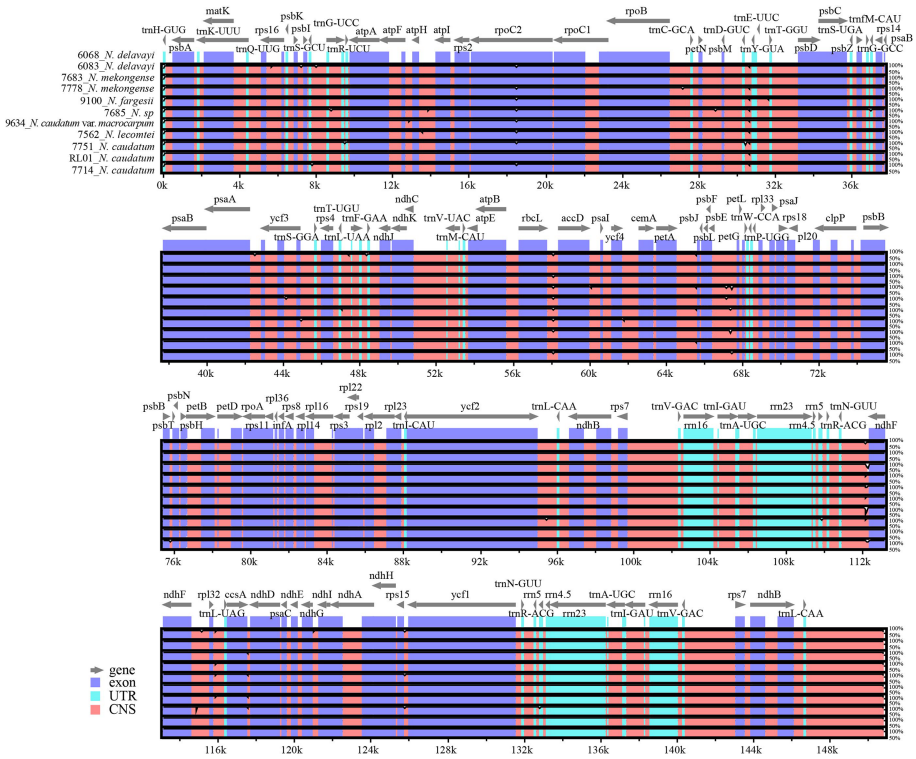


FIGURE 5
Complete chloroplast genome alignments for seven species of *Neocinnamomum*. mVISTA software was used to perform the alignment, with the *N. delavayi* (6068) chloroplast genome used as the reference sequence. The gray arrows indicate genes and their directions.

(Figure 7). In the ML phylogenetic tree, 40.91% of the branches had 100% support and 93.18% of the branches had ≥ 75% support. In the BI phylogenetic tree, only a small branch had a support rate of 0.5, and the other branches had a support of 1. The topology of the phylogenetic tree constructed based on both ML and BI methods was nearly identical. Overall, the 51 samples representing 7 *Neocinnamomum* taxa were divided into six branches. Clade I included only *N. caudatum*. Clade II included *N. mekongense*, as well as *N. delavayi* (6068) and *N. caudatum* (RL01). Clade III included only *N. caudatum* var. *macrocarpum*. Clade IV included only *N. caudatum* (7751). Clade V included *N. delavayi*, *N. fargesii*, and *N.* sp. Clade VI included only *N. lecomtei*, which was first differentiated as a basal taxon.

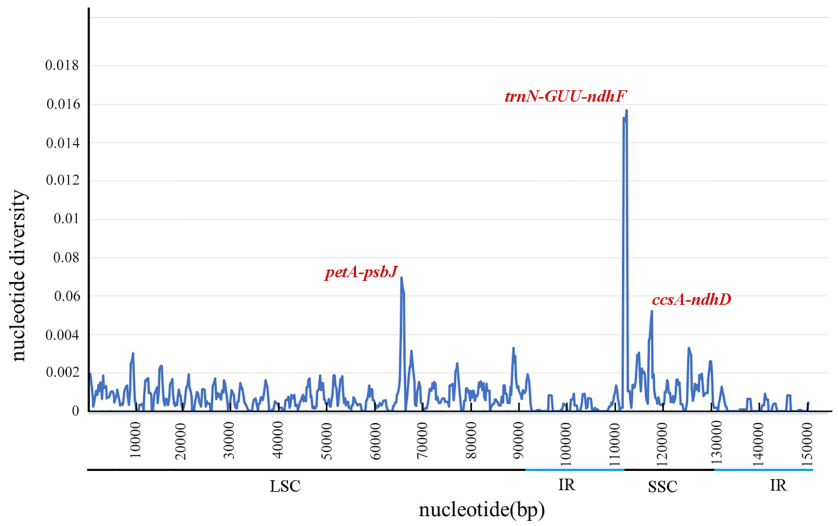


FIGURE 6
Nucleotide diversity (Pi) in matching areas of *Neocinnamomum* chloroplast genomes.

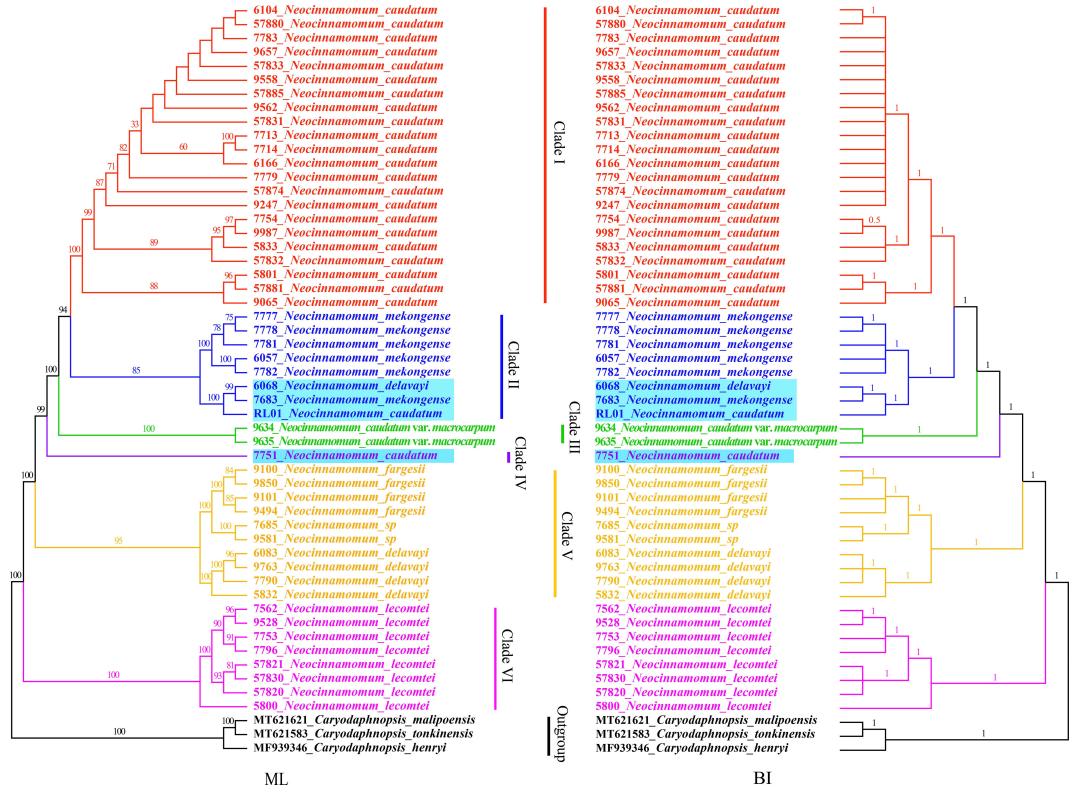


FIGURE 7
Phylogenetic tree constructed with *Neocinnamomum* chloroplast genomes.

Discussion

The structure, size, and gene content of the cp genomes of the 7 *Neocinnamomum* taxa were relatively conserved. The 51 *Neocinnamomum* cp genomes exhibited a typical tetrad structure, with a length between 150,753 bp-150,956 bp and a GC content of 38.8%-38.9%. Notably, the *Neocinnamomum* chloroplast genome is smaller than that of most other Lauraceae genera (except for *Cassytha*). Song et al. (2017) found that the cp genomes of the core Lauraceae group ranged from 150,749bp to 152,739bp in length, whereas the cp genomes of the basal Lauraceae group ranged from 157,577bp to 158,530bp in length. Their analysis indicated that the core Lauraceae group lost *trnI-CAU*, *rpl23*, *rpl2*, and a fragment of *ycf2*, as well as the intergenic regions of these genes in IRb region. To a great extent, the loss of fragments in the IR region will lead to plastid contraction, which may account for the smaller *Neocinnamomum* cp genome. A total of 128 genes were annotated in the *Neocinnamomum* cp genome, including 84 protein-coding genes, 8 rRNA genes, and 36 tRNA genes. After excluding duplicates, a total of 113 unique genes were identified. These results are consistent with previous studies on *Neocinnamomum* cp genomes (Song et al., 2017; Tian, 2021). Overall, the cp genomes were very similar between all 7 taxa of *Neocinnamomum*, with no significant differences in either gene sequence or gene content.

SSR markers are widely employed in plant germplasm research because they are rich in polymorphism, strongly co-dominant, highly reproducible, stable, and reliable (Echt et al., 1996). We detected a total of 828 SSRs across 11 *Neocinnamomum* cp genomes, with 71-82 detected per sample. Single nucleotide repeats accounted for 74.64% of all SSRs, followed by tetranucleotide repeats (15.58%). These results differ from studies of the Lauraceae genera *Litsea* (Liu et al., 2021) and *Ocotea* (Trofimov et al., 2022), the cp genomes of which contain more dinucleotide repeats than tetranucleotide repeats. According to the principle of base complementarity, the 6 repeat types include 14 repeat motifs, among which A/T accounted for 72.71% of all SSRs, followed by AT/AT and AAAT/ATTT, which accounted for 7.85% and 7.25% of all SSRs, respectively. These results indicate that A/T are the preferred bases across the cp genomes of *Neocinnamomum* taxa. Studies suggest that energy consumption affects base preference. Because the nitrogen content of A/T bases is lower than that of G/C bases, A/T enrichment can make base mutations consume less energy (Niu et al., 2017). In addition to affecting the SSR types and codon bias, base bias can also affect the stability of the four partitions of the cp genome (Mukhopadhyay et al., 2007). To date, there are few published studies of SSR markers in the *Neocinnamomum* cp genome. The identification of SSRs in the *Neocinnamomum* cp genome can provide a reference for the development of molecular markers and the analysis of genetic variation within the genus *Neocinnamomum*.

During cp genome evolution, the boundaries of the IR region may undergo contraction and expansion. This process is central to cp genome evolution and is the primary driver of cp genome

diversity among species (Park et al., 2018). Song et al. (2017) detected a double intact copy of the *ycf2* gene, as well as one intact copy and one fragment of *ycf1*, in the plastids of basal Lauraceae group. However, only one intact copy and one fragment of both *ycf2* and *ycf1* were detected in the plastids of *Neocinnamomum*. Consistent with Song's results, we identified one complete copy and one fragment of *ycf1* and *ycf2* located at the boundary between the IR region and the SC region, respectively. The *ycf1* fragment at the IRb/SSC boundary and the *ycf2* fragment at the IRa/LSC boundary were considered pseudogenes. Zhao et al. (2018) found that variation in the lengths of *ycf1*, *ycf2*, *ψycf1*, and *ndhF-ψycf1* drives the contraction and expansion of the IR region in the *Lindera* cp genome. This is similar to *Neocinnamomum*, wherein the contraction and expansion of the IR region was driven by variation in the lengths *ycf1*, *ycf2*, *ndhF*-ψycf1, and *trnH-ψycf2*.

Although the cp genome is relatively conserved and exhibits a slow rate of evolution, it also contains several mutation sites such as t*rnH-psbA* and *trnQ-rps16*, among other fragments (Nithaniyal and Parani, 2016; Linh et al., 2022). We performed a comparative analysis of the cp genomes of 7 *Neocinnamomum* taxa by combining the use of the DnaSP 6 software with the mVISTA online program. Three hypervariable regions, *trnN-GUU-ndhF*, *petA-psbJ*, and *ccsA-ndhD*, were identified. These hypervariable regions can be used as candidate molecular markers for the development of specific DNA barcodes to aid in the taxonomic identification of *Neocinnamomum* samples. Among these, the highly-variable *petA-psbJ* was previously identified as a variation hotspot in the Lauraceae species *Litsea glutinosa*, *Persea americana*, and *Machilus chuanchienensis* (Hinsinger et al., 2017; Song et al., 2016; Bai et al., 2022). The variability of the IR region was significantly lower than that of the SSC and LSC regions, and the variability of the coding regions was much lower than that of the noncoding regions, in the *Neocinnamomum* cp genome. This result is consistent with studies of cp genomes in other higher plants, including *Hernandia nymphaeifolia* (Li et al., 2020) and *Saposhnikovia divaricata* (Yi et al., 2022). These results also suggest that the noncoding regions evolve faster than the coding regions in the genus *Neocinnamomum*. Therefore, priority should be given to noncoding regions when screening DNA barcodes for *Neocinnamomum* species.

Neocinnamomum has been confirmed as a monophyletic by several studies. However, the phylogenetic relationships between species within the genus remain controversial. *N. delavayi* and *N. mekongense* are virtually indistinguishable in terms of morphological and molecular characteristics, and *N. mekongense* was once considered to be a variety of *N. delavayi* (Handel-Mazzetti, 1925). Wang et al. (2010), based on the phylogenetic study of *psbA-trnH*, *trnK*, and ITS, reported that *N. delavayi* and *N. mekongense* are closely related sister taxa. Li et al. (2016) used four species of *Neocinnamomum* in a phylogenetic study of *Caryodaphnopsis* based on RPB2, LEAFY, and ITS sequences, and the results supported the sister relationship between *N. delavayi* and *N. mekongense*. However, our phylogenetic analysis yielded different results. In our whole cp genome phylogenetic tree, *N. delavayi* was most closely related to and formed a sister group with *N. fargesii* and *N. sp*, while it was distant from *N. mekongense*. We speculate that the selection and combination of different gene fragments led to divergent phylogenetic trees. In addition, differences in biparental and maternal inheritance could have caused the divergent results of the ITS and cpDNA evolutionary trees (Wu, 2018). Of course, sampling bias may also lead to phylogenetic errors, and increasing the sampling size can effectively improve the overall phylogenetic accuracy (Murphy et al., 2001; Zwickl and Hillis, 2002; Dong et al., 2022a). Although the results are different, it has been recognized that the cp genome can well analyze the phylogeny of Lauraceae, and the whole cp genome is more effective than the fragment (Tian et al., 2021). *N. sp*, a special taxa recently collected from Wenshan, Yunnan, China, is morphologically similar to *N. complanifructum*, which has been merged into *N. lecomtei*. Our results indicated that *N. sp* is a sister taxon of *N. fargesii*. In addition, nesting was observed for 4 samples, including *N. caudatum* (RL01), *N. caudatum* (7751), *N. delavayi* (6068), and *N. mekongense* (7683), which may have been due to partial gene exchange between closely distributed species (Wu et al., 2022). The geographic distributions of *N. delavayi*, *N. mekongense*, and *N. caudatum* are similar, and often overlap, suggesting that hybridization may have occurred among these species. However, whether these four samples represent hybrids will require further morphological and molecular studies.

Conclusions

In this study, we sequenced and assembled 50 cp genomes representing 7 *Neocinnamomum* taxa, and obtained the cp genome data of *N. caudatum* (RL01) from NCBI. A total of 51 cp genomes were analyzed. Similar to most angiosperms, the *Neocinnamomum* cp genome exhibited a relatively conserved gene structure and gene content. Sequence variations were primarily concentrated in the LSC and SSC regions. Three hotspot regions and 71-82 SSRs were identified, which may be used as resources for DNA barcoding and molecular marker development for further genetic diversity analyses and molecular marker-assisted breeding. The whole cp genome phylogenetic tree revealed that the 51 samples representing 7 *Neocinnamomum* taxa were divided into six branches. Among them, *N. lecomtei* was the most primitive, basal taxon, and *N. delavayi* was found to be most closely related to *N. fargesii*. These results deepen our understanding of the *Neocinnamomum* cp genome and provide the basis for subsequent taxonomic identification, phylogenetic evolution, and population genetics studies of *Neocinnamomum* species.

Data availability statement

The 50 new sequencing data presented in the study are deposited in the NCBI (https://www.ncbi.nlm.nih.gov/nuccore), accession numbers OR085909- OR085920, and OR095860- OR095897.

Author contributions

PX and YS designed the research study. YS and WX contributed materials. ZC annotated and analyzed the genomes, wrote the manuscript. LY isolated DNA, assembled the genomes. YX, QL, HZ and YT helped ZC to analyze the data. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by Yunnan Science and Technology Talents and Platform Program (No. 202205AF150022), the National Natural Science Foundation of China (No. 32260060, 32060710), and the Special Program for Technology Bases and Talents of Guangxi (Grant No. 2022AC20002).

Acknowledgments

We thank TopEdit (www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript. We also thank Xishuangbanna Tropical Botanical Garden, CAS. Genomic data processing and analyses were conducted at the High Performance Computing Cluster from the Institutional Center for Shared Technologies and Facilities of Xishuangbanna Tropical Botanical Garden.

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/fpls.2023.1205051/full#supplementary-material

References

Bai, X., Peng, J., Yang, Y., and Xiong, B. (2022). The complete chloroplast genome sequence of *Machilus chuanchienensis* (Lauraceae): genome structure and phylogenetic analysis. *Genes (Basel)* 13 (12), 2402. doi: 10.3390/genes13122402

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Cao, Z., Zhao, W., Xin, Y., Shen, W., Wang, F., Li, Q., et al. (2022). Characteristics of the complete chloroplast genome of *Pourthiaea* (Rosaceae) and its comparative analysis. *Horticulturae* 8 (12), 1144. doi: 10.3390/horticulturae8121144

Chanderbali, A. S., Werff, H., and Renner, S. S. (2001). Phylogeny and historical biogeography of lauraceae: evidence from the chloroplast and nuclear genomes. *Ann. Missouri Bot. Gard.* 88, 104–134. doi: 10.2307/2666133

Dong, Z., Qu, S., Landrein, S., Yu, W. B., Xin, J., Zhao, W., et al. (2022a). Increasing taxa sampling provides new insights on the phylogenetic relationship between *Eriobotrya* and *Rhaphiolepis*. *Front. Genet.* 13. doi: 10.3389/fgene.2022.831206

Dong, Z., Zhang, R., Shi, M., Song, Y., Xin, Y., Li, F., et al. (2022b). The complete plastid genome of the endangered shrub *Brassaiopsis angustifolia* (Araliaceae): comparative genetic and phylogenetic analysis. *PloS One* 17 (6), e0269819. doi: 10.1371/journal.pone.0269819

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19 (1), 11–15.

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49 (3), 827–831. doi: 10.1016/j.ympev.2008.09.009

Echt, C. S., May-Marquardt, P., Hseih, M., and Zahorchak, R. (1996). Characterization of microsatellite markers in eastern white pine. *Genome* 39 (6), 1102–1108. doi: 10.1139/g96-138

Editorial Committee of Chinese Flora of China Academy of Sciences (1982). *Flora of China* Vol. 31 (Beijing: Science Press), 229.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic. Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Gan, Y., Song, Y., Chen, Y., Liu, H., Yang, D., Xu, Q., et al. (2018). Transcriptome analysis reveals a composite molecular map linked to unique seed oil profile of *Neocinnamomum caudatum* (Nees) merr. *BMC Plant Biol.* 18 (1), 303. doi: 10.1186/s12870-018-1525-9

Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S., and Hellwig, F. H. (2003). Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* 20 (9), 1499–1505. doi: 10.1093/molbev/msg159

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic. Acids Symp. Ser.* 41, 95–98. doi: 10.1021/bk-1999-0734.ch008

Handel-Mazzetti, H. (1925). *Cinnamomum delavayi* lecomte var. *mekongense*. *Anz Akad Wiss Wien Math-Naturwiss Kl* 62, 218.

Hinsinger, D. D., and Strijk, J. S. (2017). Toward phylogenomics of lauraceae: the complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and pacific ocean islands. *Plant Gene* 9, 71–79. doi: 10.1016/j.plgene.2016.08.002

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17 (8), 754–755. doi: 10.1093/bioinformatics/17.8.754

Jansen, R. K., Saski, C., Lee, S. B., Hansen, A. K., and Daniell, H. (2011). Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. *Mol. Biol. Evol.* 28 (1), 835–847. doi: 10.1093/molbev/msq261

Jiangsu New Medical College (1977). *Dictionary of traditional Chinese medicine* (Shanghai: Shanghai People Press).

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21 (1), 241. doi: 10.1186/s13059-020-02154-5

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28 (12), 1647–1649. doi: 10.1093/bioinformatics/bts199

Kostermans, A. J. G. H. (1974). A monograph of the genus *Neocinnamomum* liou ho. *Reinwardtia* 9, 85–96.

Lan, Z., Shi, Y., Yin, Q., Gao, R., Liu, C., Wang, W., et al. (2022). Comparative and phylogenetic analysis of complete chloroplast genomes from five *Artemisia* species. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1049209

Li, J., Liu, Q., Zhang, J., Yang, Y., Zhang, S., and Zhang, Y. (2020). Chloroplast genome of endangered mangrove plants *Hernandia nymphaeifolia* and its phylogenetic evolution. *J. Northwest Forest. Univ.* 35 (5), 54–61. doi: 10.3969/j.issn.1001-7461.2020.05.09

Li, L., Madriñán, S., and Li, J. (2016). Phylogeny and biogeography of *Caryodaphnopsis* (Lauraceae) inferred from low-copy nuclear gene and ITS sequences. *Taxon* 65 (3), 433–443. doi: 10.12705/653.1

Linh, N. N., Hang, P. L. B., Hue, H. T. T., Ha, N. H., Hanh, H. H., Ton, N. D., et al. (2022). Species discrimination of novel chloroplast DNA barcodes and their application for identification of *Panax* (Aralioideae, araliaceae). *PhytoKeys* 188, 1–18. doi: 10.3897/phytokeys.188.75937

Liu, H. (1934). *Lauracées de chine et d'Indochine: contribution à l'étude systématique et phytogéographique* (Paris: Hermann et Cie).

Liu, C., Chen, H. H., Tang, L. Z., Khine, P. K., Han, L. H., Song, Y., et al. (2021). Plastid genome evolution of a monophyletic group in the subtribe lauriineae (Laureae, lauraceae). *Plant Diversity* 44 (4), 377–388. doi: 10.1016/j.pld.2021.11.009

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., et al. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* 13, 715. doi: 10.1186/1471-2164-13-715

Liu, Q., and Xue, Q. (2005). Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J. Genet.* 84 (1), 55–62. doi: 10.1007/BF02715890

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic. Acids Res.* 41, W575–W581. doi: 10.1093/nar/gkt289

Mukhopadhyay, P., Basak, S., and Ghosh, T. C. (2007). Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. *Gene* 400 (1-2), 71–81. doi: 10.1016/j.gene.2007.05.027

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* 409 (6820), 614–618. doi: 10.1038/35054550

Neuhaus, H. E., and Emes, M. J. (2000). Nonphotosynthetic metabolism in plastids. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 51 (51), 111–140. doi: 10.1146/annurev.arplant.51.1.111

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 (1), 268–274. doi: 10.1093/molbev/msu300

Nithaniyal, S., and Parani, M. (2016). Evaluation of chloroplast and nuclear DNA barcodes for species identification in *Terminalia* l. *Biochem. Syst. Ecol.* 68, 223–229. doi: 10.1016/j.bse.2016.08.001

Niu, Z., Xue, Q., Wang, H., Xie, X., Zhu, S., Liu, W., et al. (2017). Mutational biases and GC-biased gene conversion affect GC content in the plastomes of dendrobium genus. *Int. J. Mol. Sci.* 18 (11), 2307. doi: 10.3390/ijms18112307

Park, S., An, B., and Park, S. (2018). Reconfiguration of the plastid genome in *Lamprocapnos spectabilis*: IR boundary shifting, inversion, and intraspecific variation. *Sci. Rep.* 8 (1), 13568. doi: 10.1038/s41598-018-31938-w

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* 25, 1253–1256. doi: 10.1093/molbev/msn083

Ren, S., Song, Y., Zhao, M., and Xu, W. (2019). The plastid genome sequence of *Neocinnamomum delavayi* (Lec.) liou. *Mitochondrial. DNA B Resour* 4 (2), 3711–3712. doi: 10.1080/23802359.2019.1679051

Rohwer, J. G., and Rudolph, B. (2005). Jumping genera: the phylogenetic positions of *Cassytha*, *Hypodaphnis*, and *Neocinnamomum* (Lauraceae) based on different analyses of *trnK* intron sequences. *Ann. Mo. Bot. Gard.* 92, 153–178. doi: 10.3417/2014033

Rozas, J., Ferrer-Mata, A., Sánchez-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 (12), 3299–3302. doi: 10.1093/molbev/msx248

Shen, W., Dong, Z., Zhao, W., Ma, L., Wang, F., Li, W., et al. (2022). Complete chloroplast genome sequence of *Rosa lucieae* and its characteristics. *Horticulturae* 8 (9), 788. doi: 10.3390/horticulturae8090788

Song, Y., Yao, X., Tan, Y., Gan, Y., and Corlett, R. T. (2016). Complete chloroplast genome sequence of the avocado: gene organization, comparative analysis, and phylogenetic relationships with other lauraceae. *C. Can. J. For. Res.* 46, 1293–1301. doi: 10.1139/cjfr-2016-0199

Song, Y., Yu, W., Tan, Y., Jin, J., Wang, B., Yang, J., et al. (2020). Plastid phylogenomics improve phylogenetic resolution in the lauraceae. *J. Syst. Evol.* 58 (4), 423–439. doi: 10.1111/jse.12536

Song, Y., Yu, W. B., Tan, Y., Liu, B., Yao, X., Jin, J., et al. (2017). Evolutionary comparisons of the chloroplast genome in lauraceae and insights into loss events in the magnoliids. *Genome Biol. Evol.* 9 (9), 2354–2364. doi: 10.1093/gbe/evx180

Tao, X., Ma, L., Nie, B., Wang, Y., and Liu, Z. (2017). The draft and characterization of the complete chloroplast genome of *Vicia sativa* cv. langjian No.3. *Pratacult. Sci.* 34 (2), 321–330. doi: 10.11829/j.issn.1001-0629.2016-0054

Tian, Y. (2021). The complete chloroplast genomes of lauraceae species: comparative genomic and phylogenetic analyses (Nanjing University). doi: 10.27235/d.cnki.gnjiu.2021.001893

Tian, Y., Zhou, J., Zhang, Y., Wang, S., Wang, Y., Liu, H., et al. (2021). Research progress in plant molecular systematics of lauraceae. *Biol. (Basel)* 10 (5), 391. doi: 10.3390/biology10050391

Trofimov, D., Cadar, D., Schmidt-Chanasit, J., Rodrigues de Moraes, P. L., and Rohwer, J. G. (2022). A comparative analysis of complete chloroplast genomes of seven *Ocotea* species (Lauraceae) confirms low sequence divergence within the *Ocotea* complex. *Sci. Rep.* 12 (1), 1120. doi: 10.1038/s41598-021-04635-4

Wang, Z., Li, J., Conran, J. G., and Li, H. (2010). Phylogeny of the southeast Asian endemic genus *Neocinnamomum* h. liu (Lauraceae). *Plant Syst. Evol.* 290 (1), 173–184. doi: 10.1007/s00606-010-0359-1

Wang, Z., Yang, J., Tan, Y., Hu, G., and Long, C. (2013). Comprehensive evaluation of woody oil-bearing plants in yunnan as sources for biodiesel. *Plant Diversity Res.* 35 (5), 630–640. doi: 10.7677/ynzwyj201312168

Wang, F., Zhao, W., Dong, Z., Ma, L., Li, W., Li, Z., et al. (2023). Analysis of the chloroplast genome characteristics of 6 species of *Yucca*. *Bull. Bot. Res.* 43 (1), 109–119. doi: 10.7525/j.issn.1673-5102.2023.01.012

Wariss, H. M., Yi, T., Wang, H., and Zhang, R. (2017). The chloroplast genome of a rare and an endangered species *Salweenia bouffordiana* (Leguminosae) in China. *Conserv. Genet. Resour.* 10 (3), 405–407. doi: 10.1007/s12686-017-0836-8

WFO (2023) *Neocinnamomum atjehense kosterm*. Available at: http://www.worldfloraonline.org/taxon/wfo-0000381722 (Accessed February 23, 2023).

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* 31 (20), 3350–3352. doi: 10.1093/bioinformatics/btv383

Wu, Z. (1983). *Flora yunnanica* Vol. 3 (Beijing: Science Press).

Wu, Y. (2018). The molecular phylogenetic analysis of mulberry based on the nucleus ITS and chloroplast DNA sequences (Southwest University).

Wu, L., Cui, Y., Wang, Q., Xu, Z., Wang, Y., Lin, Y., et al. (2021). Identification and phylogenetic analysis of five *Crataegus* species (Rosaceae) based on complete chloroplast genomes. *Planta* 254 (1), 14. doi: 10.1007/s00425-021-03667-4

Wu, Z. Y., Milne, R. I., Liu, J., Slik, F., Yu, Y., Luo, Y. H., et al. (2022). Phylogenomics and evolutionary history of *Oreocnide* (Urticaceae) shed light on recent geological and climatic events in SE Asia. *Mol. Phylogenet. Evol.* 175, 107555. doi: 10.1016/j.ympev.2022.107555

Xin, Y., Yu, W. B., Eiadthong, W., Cao, Z., Li, Q., Yang, Z., et al. (2023). Comparative analyses of 18 complete chloroplast genomes from eleven *Mangifera* species (Anacardiaceae): sequence characteristics and phylogenomics. *Horticulturae* 9 (1), 86. doi: 10.3390/horticulturae9010086

Xing, S. C., and Liu, C. J. (2008). Progress in chloroplast genome analysis. *Prog. Biochem. Biophys.* 35, 21–28. doi: 10.3321/j.issn:1000-3282.2008.01.004

Xu, D. H., Abe, J., Kanazawa, A., Gai, J. Y., and Shimamoto, Y. (2001). Identification of sequence variations by PCR-RFLP and its application to the evaluation of cpDNA diversity in wild and cultivated soybeans. *Theor. Appl. Genet.* 102 (5), 683–688. doi: 10.1007/s001220051697

Xu, W., Xia, B., Zhang, S., and Chen, Z. (2017). A new variety of *Neocinnamomum* h. liou from guangxi, China. *Guihaia* 37 (7), 855–858. doi: 10.11931/guihaia.gxzw201610026

Yang, Z., Zhao, T., Ma, Q., Liang, L., and Wang, G. (2018). Comparative genomics and phylogenetic analysis revealed the chloroplast genome variation and interspecific relationships of *Corylus* (Betulaceae) species. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00927

Yi, S., Lu, H., Wang, W., Wang, G., Xu, T., Li, M., et al. (2022). The chloroplast genome of wild *Saposhnikovia divaricata*: genomic features, comparative analysis, and phylogenetic relationships. *Genes* 13 (5), 931. doi: 10.3390/genes13050931

Zhao, M. L., Song, Y., Ni, J., Yao, X., Tan, Y. H., and Xu, Z. F. (2018). Comparative chloroplast genomics and phylogenetics of nine *Lindera* species (Lauraceae). *Sci. Rep.* 8 (1), 8844. doi: 10.1038/s41598-018-27090-0

Zwickl, D. J., and Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51 (4), 588–598. doi: 10.1080/10635150290102339

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 18 August 2023
DOI 10.3389/fpls.2023.1211247

OPEN ACCESS

EDITED BY
Shuangyang Wu,
Gregor Mendel Institute of Molecular Plant Biology (GMI), Austria

REVIEWED BY
Rufeng Wang,
Shanghai University of Traditional Chinese Medicine, China
Hengchang Wang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Inom Juramurodov
ijuramurodov@mail.ru
Komiljon Tojibaev
ktojibaev@mail.ru

RECEIVED 24 April 2023
ACCEPTED 20 July 2023
PUBLISHED 18 August 2023

CITATION
Juramurodov I, Makhmudjanov D, Yusupov Z and Tojibaev K (2023) First comparative analysis of complete chloroplast genomes among six *Hedysarum* (Fabaceae) species. *Front. Plant Sci.* 14:1211247. doi: 10.3389/fpls.2023.1211247

First comparative analysis of complete chloroplast genomes among six *Hedysarum* (Fabaceae) species

Inom Juramurodov[1,2,3,4]*, Dilmurod Makhmudjanov[1,2,3,4], Ziyoviddin Yusupov[5] and Komiljon Tojibaev[2,3]*

[1]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China, [2]Yunnan International Joint Laboratory for Biodiversity of Central Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China, [3]Flora of Uzbekistan Laboratory, Institute of Botany of the Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan, [4]University of Chinese Academy of Sciences, Beijing, China, [5]International Joint Lab for Molecular Phylogeny and Biogeography, Institute of Botany, Academy Sciences of Uzbekistan, Tashkent, Uzbekistan

Hedysarum is one of the largest genera in the Fabaceae family, mainly distributed in the Northern Hemisphere. Despite numerous molecular studies on the genus *Hedysarum*, there is still a lack of research aimed at defining the specific characteristics of the chloroplast genome (cp genome) of the genus. Furthermore, the interrelationships between sections in the genus based on the cp genome have not yet been studied. In this study, comprehensive analyses of the complete cp genomes of six *Hedysarum* species, corresponding to sections *Multicaulia*, *Hedysarum*, and *Stracheya* were conducted. The complete cp genomes of *H. drobovii*, *H. flavescens*, and *H. lehmannianum* were sequenced for this study. The cp genomes of six *Hedysarum* species showed high similarity with regard to genome size (except for *H. taipeicum*), gene sequences, and gene classes, as well as the lacking IR region. The whole cp genomes of the six species were found to contain 110 genes ranging from 121,176 bp to 126,738 bp in length, including 76 protein-coding genes, 4 rRNA genes, and 30 tRNA genes. In addition, chloroplast SSRs and repetitive sequence regions were reported for each species. The six *Hedysarum* species shared 7 common SSRs and exhibited 14 unique SSRs. As well, three highly variable genes (*clpP, accD*, and *atpF*) with high Pi values were detected among protein-coding genes. Furthermore, we conducted phylogenetic analyses using the complete cp genomes and 76 protein-coding genes of 14 legume species, including the seven *Hedysarum* species. The results showed that the *Hedysarum* species form a monophyletic clade closely related to the genera *Onobrychis* and *Alhagi*. Furthermore, both of our phylogenetic reconstructions showed that section *Stracheya* is more closely related to section *Hedysarum* than to section *Multicaulia*. This study is the first comprehensive work to investigate the genome characteristics of the genus *Hedysarum*, which provides useful genetic information for further research on the genus, including evolutionary studies, phylogenetic relationships, population genetics, and species identification.

KEYWORDS

***Hedysarum*, chloroplast genome, comparative analysis, phylogeny, protein-coding genes**

Introduction

Hedysarum L. is one of the large genera in the Fabaceae family, containing more than 160 species (Lock, 2005). Plants belonging to the genus *Hedysarum* are distributed in Eurasia, North Africa, and North America (Lock, 2005; Liu et al., 2017). The species occur in meadows, clayey and stony places, deserts, steppes, forests, tundra, river valleys, and mountain slopes (Choi and Ohashi, 2003). The genus *Hedysarum* includes perennial herbs and rarely semi-shrubs, which differ from closely related genera in pod and pollen morphology (Fedtschenko, 1948; Choi and Ohashi, 1996). Previous studies have shown that many species of the genus *Hedysarum* have been employed in traditional Chinese medicine to strengthen the immune system and improve the energy of the body (Dong et al., 2013).

Recent molecular studies have proposed to divide this genus into three main clades, largely corresponding to the sections *Hedysarum, Stracheya*, and *Multicaulia* (Amirahmadi et al., 2014; Nafisi et al., 2019). In these molecular phylogenetic studies, species were divided into three sections that were well supported, but intersection relationships remained unresolved, especially in section *Multicaulia*. This could be due to the selection of regions with low variability in the cp genome. Therefore, it is necessary to identify regions with high nucleotide diversity in the cp genome as molecular markers for future molecular phylogenetic studies of *Hedysarum*. Furthermore, previous studies have reported conflicting results regarding the phylogenetic relationship of section *Stracheya* with the other two sections within the *Hedysarum* genus. Duan et al. (2015) found that section *Stracheya* was closely related to section *Multicaulia* based on both nrDNA ITS and some plastid markers. However, Liu et al. (2017) reported that section *Stracheya* was placed together with section *Hedysarum* based on nrDNA ITS and plastid markers. Conversely, Nafisi et al. (2019) supported a closer relationship between section *Stracheya* and section *Multicaulia*. Therefore, determining the phylogenetic position of section *Stracheya* in the genus *Hedysarum* based on the cp genome is necessary.

Chloroplasts are important intracellular organelles, having an independent genome with several genes responsible for the process of photosynthesis in green algae and plants (Nazareno et al., 2015; Smith and Keeling, 2015; Yin et al., 2017). Most complete cp genomes harbor a typical quadripartite structure including a long single copy (LSC) region, a small single copy (SSC) region, and two copies of inverted repeat (IR) regions (Bock, 2007). *Hedysarum* belonging to the IRLC (Inverted Repeat Lacking Clade) clade is described by the lack of one copy of the inverted repeat (IR) region in the whole cp genome (Wojciechowski et al., 2004; She et al., 2019). Species belonging to the IRLC clade are characterized by having a cp genome size of around 121,000–133,000 bp (Moghaddam et al., 2022; Yuan et al., 2022; Tian et al., 2021). To date, the size of the cp genomes of only four *Hedysarum* species are known, including *H. petrovii*, *H. semenovii*, *H. polybotrys*, and *H. taipeicum*, with genome sizes of 122,571 bp, 123,407 bp, 122,232 bp, and 126,699 bp, respectively.

Comparative genomics can be used to identify important structural sequences and detect evolutionary changes across genomes since the comprehensive analysis of the cp genome of genera belonging to the IRLC clade such as *Astragalus*, *Onobrychis*, *Caragana*, and *Glycyrrhiza* have been reported (Kang et al., 2018; Moghaddam et al., 2022; Yuan et al., 2022; Tian et al., 2021). A detailed characterization of these species' cp genome, including size, gene content, structure repeats, and GC content, as well as information about highly variable nucleotide regions, was provided. However, comprehensive studies on the genome structure of the genus *Hedysarum* have not been conducted so far.

In the present investigation, we detailed an overview of the complete sequence of the six *Hedysarum* species cp genome. We sequenced the complete cp genome of *H. drobovii*, *H. flavescens*, and *H. lehmannianum* to explore the relationships among *Hedysarum* species. We obtained the other three species (*H. petrovii*, *H. semenovii*, and *H. taipeicum*) from the National Center for Biotechnology Information (NCBI). The following questions were addressed: (1) what are the features of the cp genome of selected *Hedysarum* species? (2) How many potential microsatellite markers can the cp genome provide? (3) Which regions in the cp genome can be used as candidate molecular markers for future molecular phylogenetic studies? (4) What is the phylogenetic placement of section *Stracheya* within the genus *Hedysarum* based on the cp genome data?

Materials and methods

Plant materials

For the comparative genome analysis, species from each section of *Hedysarum* were selected in this study: *Hedysarum drobovii* and *H. petrovii* from the *H.* sect. *Multicaulia*; *H. flavescens, H. semenovii*, and *H. taipeicum* from the *H.* sect. *Hedysarum*; *H. lehmaniannum* from the *H.* sect. *Stracheya*. Fresh material for *H. drobovii*, *H. flavescens*, and *H. lehmaniannum* was collected from Uzbekistan (*H. drobovii*: Western Tien Shan, Chatkal Range, E. 70.1045, N. 41.560301, altitude: 970 m a.s.l., 06 June 2020, *Dekhkonov, Ortiqov, Turdiev, Juramurodov 19062020117*; *H. flavescens*: Western Tien Shan, Chatkal Range, E. 70.019246 N. 41.508587, altitude: 2290 m a.s.l., 06 June 2020, *Dekhkonov, Ortiqov, Turdiev, Juramurodov 19062020089*; *H. lehmaniannum*: Hisar Range, Boysun district, E. 67.163574, N. 38.337148, altitude: 2390 m a.s.l., 13 June 2021, *Turginov, Rahmatov 13062021007*), and their complete chloroplast (cp) genome sequences were generated (Figure 1). Herbarium materials of these three species were stored in the National Herbarium of Uzbekistan (TASH).

Sequencing, assembly, and annotation

Total genomic DNA was extracted from leaf material using DP305 Plant Genomic DNA kits (Tiangen, Beijing, China) following the manufacturer's protocol. The sequencing library was generated using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA, Catalog: E7370L) following the manufacturer's recommendations, and index codes were added to each sample. Briefly, the genomic DNA sample was fragmented by sonication to a size of 350 bp. Then DNA fragments were end

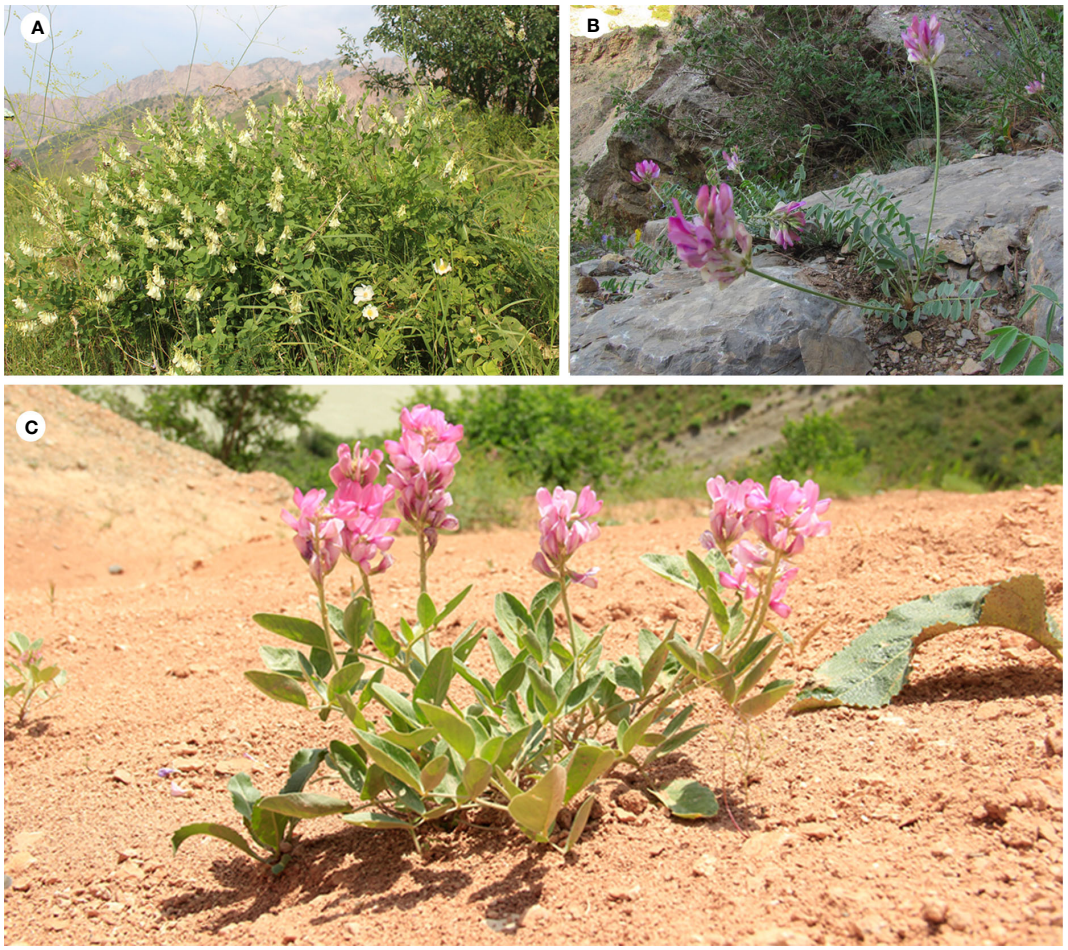


FIGURE 1
Three sequenced samples in this study. **(A)** *H. flavescens*, **(B)** *H. lehmaniannum*, and **(C)** *H. drobovii*. The photo of *H. lehmaniannum* was taken by O.Turginov.

polished, A-tailed, and ligated with the full-length adapter for Illumina sequencing, followed by further PCR amplification. After, PCR products were purified by the AMPure XP system (Beverly, USA). Subsequently, the library quality was assessed on the Agilent 5400 system(Agilent, USA) and quantified by QPCR (1.5 nM). The qualified libraries were pooled and sequenced on Illumina platforms with PE150 strategy in Novogene Bioinformatics Technology Co., Ltd (Beijing, China), according to effective library concentrations and data amount required.

The resulting clean reads were assembled using the GetOrganelle pipeline (Jin et al., 2020) with the optimized parameters "-F plant_cp -w 0.6 -o -R 20 -t 8 -k 75,95,115,127 and". Gene annotation was performed in Geneious v.10.0.2 and *H. polybotrys* (unpublished, accession number: MZ322397) was set as the reference. Start and stop codons and intron/exon boundaries for protein-coding genes were checked manually (Kearse et al., 2012).

Simple sequence repeats

The chloroplast simple sequence repeats (SSRs) were identified using the MIcroSAtellite (MISA) web tool (Beier et al., 2017). The search conditions for SSRs were set to isolate perfect mono-, di-, tri-, tetra-, penta-, and hexa nucleotide motifs with a minimum of 10, 5, 4, 3, 3, and 3 repeats, respectively. The REPuter program (Kurtz et al., 2001) was used to identify repeats: forward, reverse, palindrome, and complement sequences in cp genomes. The following settings for repeat identification were used: (1) a hamming distance equal to three, (2) minimal repeat size set to 30 bp, and (3) maximum computed repeats set to 90 bp.

Comparative analysis of chloroplast genomes

The cp genome was drawn using OGDRAWv1.1 (Lohse et al., 2007). Nucleotide variability (Pi) was calculated for the whole cp genome and protein-coding genes separately using DnaSP v. 6.12.03 software (Rozas et al., 2018). The window length was set to 800 bp and the step size was 200 bp. Furthermore, pairwise chloroplast genomic alignment among six species was compared by mVISTA in Shuffle-LAGAN mode (Frazer et al., 2004), and *H. polybotrys* (MZ322397) was used as a reference.

Phylogenetic analysis

The three sequenced cp genomes of *Hedysarum* and 11 genomes from other species (including *Onobrychis gaubae, O. viciifolia, Caragana jubata, C. kozlowii, Oxytropis aciphylla*, and *O. glabra* as outgroups) retrieved from NCBI (Supplementary Table 1) were used to construct a phylogenetic tree. Phylogenetic tree reconstruction was performed using complete cp genomes and protein-coding sequences that were first aligned multiple times using MAFFT software (Katoh and Standley, 2013).

We reconstructed phylogenetic trees using Bayesian inference (BI), Maximum Parsimony (MP), and Maximum Likelihood (ML) methods. Nucleotide substitution models were selected statistically with the help of jModelTest2 on XSEDE (www.phylo.org) by considering the Akaike Information Criterion (AIC). The GTR+G model for the protein-coding sequences and the TVM+G model for the complete cp genomes were selected as the best model. For BI, we used MrBayes v. 3.2.7a (Ronquist et al., 2012) with 10 million generations with random trees sampled every 1000 generations. In the latter analysis, after discarding the first 25% of the trees as burn-in, a 50% majority-rule consensus tree was constructed from the remaining trees to estimate posterior probabilities (PP). The ML phylogeny was reconstructed using IQ-TREE 2.1.2 software (Minh et al., 2021) with 1000 bootstrap (BS) replicates to assess clade support (Nguyen et al., 2015). For MP analysis, we used PAUP* 4.0a169 (Swofford, 2002). The MP bootstrap analysis was performed with heuristic search, TBR branch-swapping, 1000 bootstrap replicates, random addition sequence with 10 replicates, and a maximum of 1000 trees saved per round.

Results

Chloroplast genome features of *Hedysarum* species

The complete cp genomes of *H. drobovii*, *H. flavescens*, and *H. lehmannianum* were sequenced for this study. The sizes of the three newly sequenced species were 121,176 bp, 123,127 bp, and 123,586 bp, respectively. The *H. petrovii, H. semenovii*, and *H. taipeicum* species that were obtained from NCBI and the three newly sequenced species were without the typical quadripartite structure that contains a pair of IRs separated by LSC (large single-copy) and SSC (small single-copy) regions (Figure 2). The GC (guanine +cytosine) contents of the genomes of *H. drobovii*, *H. petrovii, H. flavescens*, *H. semenovii*, *H. lehmannianum*, and

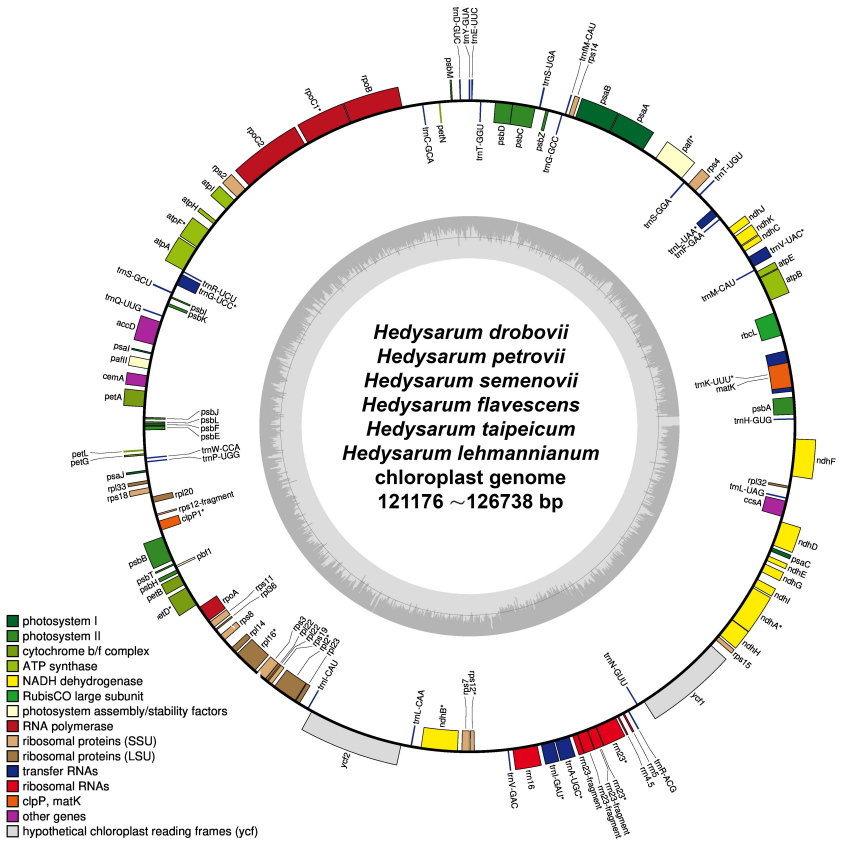


FIGURE 2
The chloroplast genome structure of six *Hedysarum* species. Genes shown outside the circles are transcribed clockwise, while those drawn inside are transcribed counterclockwise. Genes are color-coded according to their functional group.

H. taipeicum was 34.6%, 34.6%, 34.8%, 34.9%, 34.6%, and 35.1%, respectively. All six species' genomes formed 110 genes including 76 protein-coding genes, 4 rRNA genes, and 30 tRNA genes (Table 1). A total of 16 genes in the cp genomes of six *Hedysarum* species consisted of introns, among which the genes *trnK-UUU, trnC-ACA, trnL-UAA, rpoC1, atpF, trnG-UCC, clpP, petB, petD, rpl16, rpl2, ndhB, trnE-UUC, trnA-UGC*, and *ndhA* each contained one intron, and only *ycf3* gene contained two introns (Supplementary Table 2). The *trnK-UUU* gene contained the largest intron, from 2407 (*H. petrovii*) to 2503 (*H. taipeicum*). Additionally, the *rps12* protein-coding gene is a trans-splicing gene that does not have introns in the 3'-end.

Repeat sequences and SSRs analysis

A total of 188 SSRs were detected using the MISA web tool in the cp genome of each *H. drobovii*, *H. petrovii*, and *H. lehmannianum* species, while *H. flavescens*, *H. semenovii*, and *H. taipeicum* had different SSRs of 184, 190, and 172, respectively.

TABLE 1 List of genes annotated in the chloroplast genomes of six *Hedysarum* species in this study.

Category of genes	Group of genes	Genes
Genes for photosynthesis (44)	Subunits of photosystem I Subunits of photosystem II	*psaA, psaB, psaC, psaI, psaJ* *psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbH, psbK, psbL, psbM, psbN, psbT, psbZ,*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of NADH-dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	RubisCO large subunit	*rbcL*
Self-replication (57)	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl23, rpl32, rpl33, rpl36*
	Small subunit of ribosome	*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps18, rps19*
	RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Ribosomal RNAs	*rrn5S, rrn4.5S, rrn16S, rrn23S*
	tRNA genes	*trnH-GUG, trnK-UUU, trnM-CAU, trnV-UAC, trnF-GAA, trnL-UAA, trnS-GGA, trnT-UGU, trnG-GCC, trnT-GGU, trnC-GCA, trnfM-CAU, trnS-UGA, trnE-UUC* (×2 or ×1), *trnY-GUA, trnD-GUC, trnR-UCU, trnG-UCC, trnS-GCU, trnQ-UUG, trnW-CCA, trnP-UGG, trnI-CAU, trnL-CAA, trnN-GUU, trnL-UAG, trnV-GAC, trnA-UGC, trnI-GAU* (or *trnE-UUC), trnR-ACG*
Other genes (5)	Subunit of acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Protease	*clpP*
	Maturase	*matK*
Genes with unknown functions (4)	hypothetical chloroplast reading frames (ycf)	*ycf1, ycf2, ycf3, ycf4*

10.3389/fpls.2023.1211247

Among six *Hedysarum* cp genomes, the most abundant repeats were the mononucleotides from 145 (*H. taipeicum*) to 156 (*H. lehmannianum*), and the most dominant SSR was A/T (Figures 3A, B). Di-nucleotides (especially AT) were the second most predominant, varying from eight (*H. taipeicum*) to 21 (*H. lehmannianum*). A high number of trinucleotides was detected in *H. semenovii* (12), whereas a low number of trinucleotides was in *H. lehmannianum* (3). A total of 57 repeats of tetranucleotides, varying from seven (*H. drobovii*) to 12 (*H. petrovii*) were identified among the six *Hedysarum* cp genomes. Our analysis identified five pentanucleotide repeats in three *Hedysarum* species: *H. drobovii* (AAAAT, AAAGG, and AAGAC), *H. petrovii* (TTTCC), and *H. taipeicum* (AACCG), while the remaining three species did not exhibit any pentanucleotide repeats. Additionally, hexanucleotide repeats were detected only in *H. flavescens* (ATCAGT), *H. semenovii* (AAGACG, ATAGCT, and ATATTT), and *H. taipeicum* (AAGACG(× 2) and ATTCTT).

In our study, we examined common and unique SSRs in six *Hedysarum* species (Supplementary Tables 3, 4), and we found that the majority of repeat units were composed of A and T, with rare occurrences of C or G, indicating that the SSRs of different species had an obvious bias in the base types of repeat units. Common SSRs

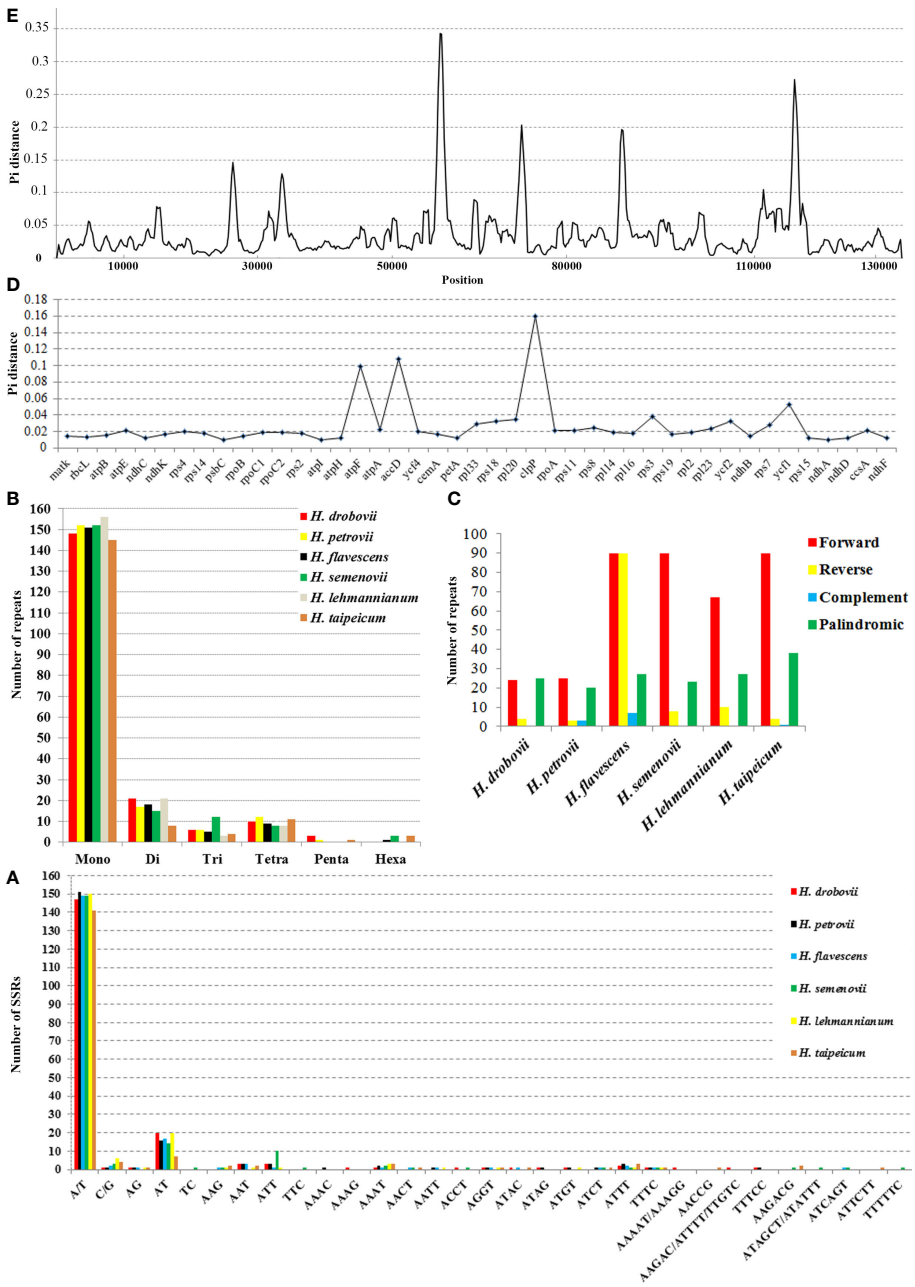


FIGURE 3
Chloroplast genome features of six *Hedysarum* species. Type of SSRs **(A)**; long repetitive sequences **(B)**; SSR distribution **(C)**. Nucleotide diversity (Pi) in protein-coding genes **(D)** and whole chloroplast genomes **(E)**. Among protein-coding genes, genes with nucleotide diversity < 0.01 are not shown.

10.3389/fpls.2023.1211247

included A, G, T, AT, AAAT, ATTT, and TTTC, which were present in all six species. We also identified 14 unique SSRs, including AAAG, AAAGG, AAAAT, AAGAC, ATTTT, and TTGTC in *H. drobovii*; TC, TTC, ATAGCT, and TTTTTC in *H. semenovii*; and AACCG and ATTCTT in *H. taipeicum*. Only one AAAC SSR was detected in *H. petrovii*, while no unique SSRs were identified in *H. flavescens* and *H. lehmannianum*.

In this study, we found many repeat regions including forward, reverse, palindromic, and complementary repeats (Figure 3C). Among the six studied *Hedysarum* species, the longest repetitive sequences were detected in the *H. flavescens* cp genome, which had 214 repetitive sequences with lengths of no more than 48 bp. On the contrary, the smallest repetitive sequences were found in *H. drobovii* and *H. petrovii* cp genomes, of which 53 and 51 scattered repetitive sequences with lengths of no more than 18 bp and 13 bp, respectively. The length of the largest forward and palindromic repeats were 62 bp and 36 bp in the *H. taipeicum* cp genome, respectively, whereas the largest reverse and complement repeat lengths were 48 bp and 7 bp in the *H. flavescens* cp genome, respectively. Equal numbers of forward repeats (90) were detected in *H. flavescens*, *H. semenovii*, and *H. taipeicum*. Additionally, the complement repeat was not found in the cp genomes of *H. drobovii, H. semenovii*, and *H. lehmannianum*.

Comparative genomic divergence and hotspot regions

We calculated the nucleotide diversity (Pi) values to estimate the divergence levels of the whole cp genome and protein-coding genes of the six *Hedysarum* species (Figures 3D, E). The most high-variation regions (Pi=0.3425) of the whole *Hedysarum* cp genomes were mainly concentrated between 55000 bp and 60000 bp. According to the Pi value results of protein-coding genes of six *Hedysarum* species, *clpP* (0.16), *accD* (0.108), and *atpF* (0.099) genes had the highest variability, while the *psaC* gene had a low nucleotide diversity (0.00136). In addition, the Pi values were less than 0.01 in 43.4% of the total protein-coding genes, while in 39.5% were 0.01–0.02. Only 26.3% of total protein-coding genes had Pi > 0.02 (Supplementary Table 5).

The cp genome sequences of six *Hedysarum* species were compared using the mVISTA software, and their alignments were visualized with annotation data (Figure 4). According to this visualization analysis, differences among sequences occurred in *clpP*, *accD*, and *ycf1* genes from the coding regions and mainly in non-coding intergenic regions. Encoded gene classes and alignments of the main part of the coding regions among the six *Hedysarum* were highly congruous.

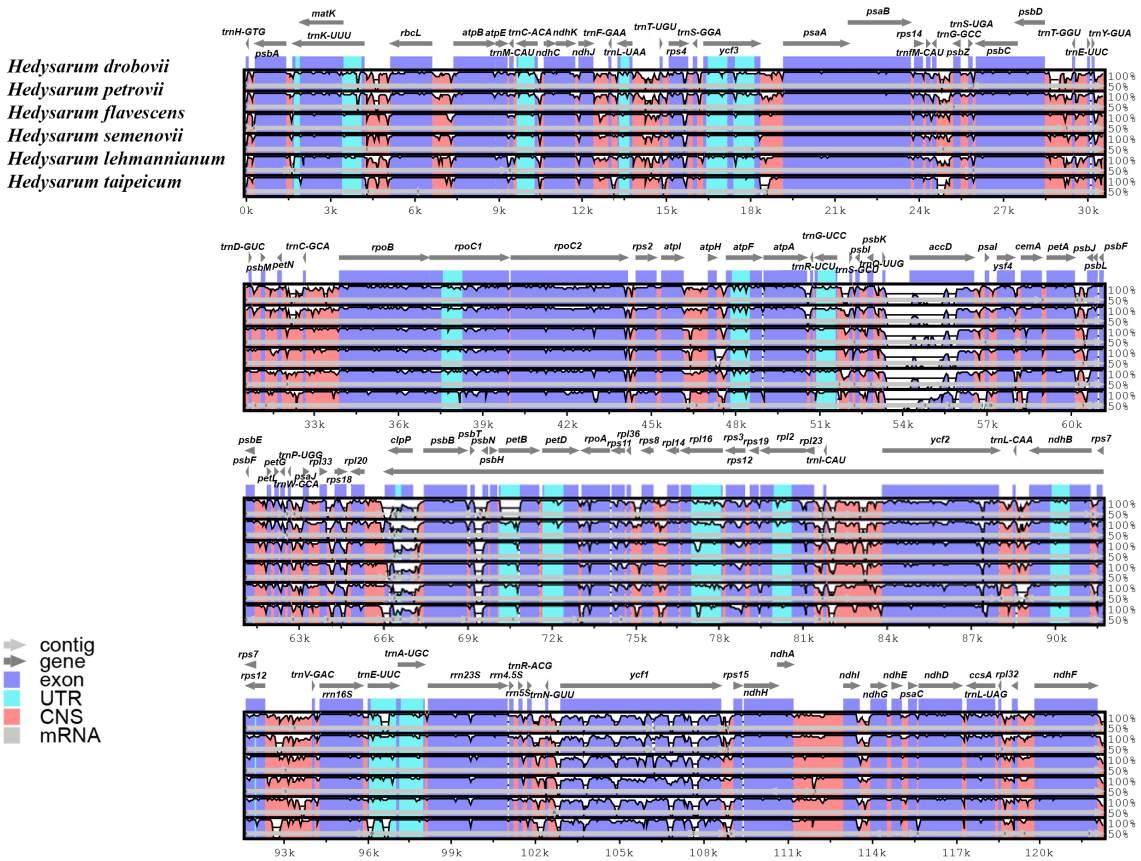


FIGURE 4
Alignment visualization of the chloroplast genome sequences of six *Hedysarum* species using mVISTA. Annotated genes are shown along the top. Genomic regions are color-coded to indicate protein-coding regions, exons, UTRs, and CNS. The similarity among the chloroplast genomes is shown on a vertical scale ranging from 50 to 100%.

Phylogenetic analysis

Seven *Hedysarum* and seven related cp genome data were analyzed phylogenetically. Phylogenetic reconstructions based on the complete cp genome and protein-coding genes yielded similar results (Figures 5A, B). All clades in both trees were strongly supported by BI, ML, and MP analyses, with 1.00, 100%, and 100% bootstrap values, respectively. Additionally, the results of the phylogenetic analysis based on the complete cp genomes and 76 protein-coding genes showed that *Hedysarum* was monophyletic. The clade including species of the genus *Onobrychis* was sister to *Hedysarum*. The *Hedysarum* species used in this study were formed into two clades. One is a clade containing *H. drobovii* and *H. petrovii* species corresponding to section *Multicaulia*. The second clade was formed by five *Hedysarum* species. *H. flavesens, H. semenovii, H. taipeicum*, and *H. polybotrys* species were placed into the section *Hedysarum*, and *H. lehmannianum* belonged to the section *Stracheya*, which was sister to the section *Hedysarum*.

Discussion

This study is the first to comprehensively examine the features of cp genomes in *Hedysarum* species. We compared the cp genomes of six species belonging to three sections that were distributed in different regions. We sequenced the cp genome of *H. drobovii, H. flavescens*, and *H. lehmannianum* for this study. The sizes of the six cp genomes ranged from 121,176 bp (*H. drobovii*) to 126,738 bp (*H. taipeicum*). It is worth noting that many related genera with similar cp genome sizes to *Hedysarum* have been reported in recent years (Tian et al., 2021; Bei et al., 2022; Moghaddam et al., 2022).

The cp genomes of *Hedysarum* species have 110 genes, including 76 protein-coding genes, 30 transfer RNA genes, and 4 ribosomal RNA genes. The structural composition of *Hedysarum* cp genomes revealed similarity with other IRLC clade species (Su et al., 2019; Bei et al., 2022; Moghaddam et al., 2022). The cp genomes of six *Hedysarum* species showed high similarity with regard to genome size (except for *H. taipeicum* which was 126,738 bp), gene sequences, gene classes, and the lacking IR region. All selected *Hedysarum* species were found to have lost one copy of the IR region, which was first identified in *H. taipeicum* by She et al. (2019). This loss of the IR region is common in most species belonging to the subfamily Papilionoideae in the family Fabaceae, forming a clade named the IR-lacking clade (IRLC) (Wojciechowski et al., 2004). The GC content of the six *Hedysarum* species in this study was highly similar, which is an important indicator of species affinity according to Tamura et al. (2011).

Introns are recognized as being central to the regulation of gene expression in plants and animals (Callis et al., 1987; Emami et al., 2013: Choi et al., 1991). In the present study, 15 genes with one intron and one gene (*ycf3*) with two introns were identified in each of the cp genomes of the six studied *Hedysarum* species. Most of the 16 identified genes have a high similarity in the structure of introns. However, a structural change was detected in the intron of the *petB* and *clpP* genes of *H. drobovii* and *H. lehmannianum*, respectively. The intron of the *petB* gene in the cp genome of *H. drobovii* is very short (9 bp); whereas, in the other five species, it ranges from 806 bp (*H. flavescens*) to 864 bp (*H. taipeicum*). Similarly, the intron of the *clpP* gene in the cp genome of *H. lehmannianum* is 6 bp long, while in other species it is from 159 bp (*H flavescens*) to 613 bp (*H. taipeicum*). However, the implications or link between gene expression and short or long introns for *Hedysarum* have not been studied. Further experimental work on the roles of introns in *Hedysarum* is therefore essential and should prove interesting. In consonance with previous studies, the *trnK-UUU* gene in the *Hedysarum* cp genome was observed to be harboring the largest intron (2407-2503 bp) which includes the *matK* gene.

The chloroplast SSRs were used in evolutionary studies, phylogenetic relationships, and plant population genetics and species identification as molecular markers (Olmstead and Palmer, 1994; Saski et al., 2005). A total of 172 SSRs (*H. taipeicum*) to 190 SSRs (*H. semenovii*) were found in the cp genome of six *Hedysarum* species. Several studies found that the mononucleotide repeats were dominant among SSRs in the cp genome, where A/T bases account for the majority (Ellegren, 2004; George et al., 2015; Ren et al., 2021). Likewise, A/T mononucleotide repeats were dominant among SSRs in the six *Hedysarum* cp genomes, ranging from 78.7% to 84.3%.

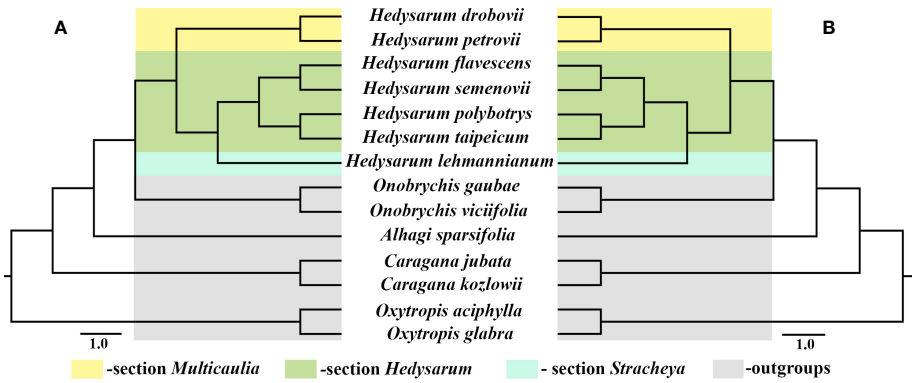


FIGURE 5
Phylogenetic tree of 14 species including seven *Hedysarum* species using BI, ML, and MP analyses based on complete cp genomes **(A)** and their 76 protein-coding genes **(B)**. All branches were maximally supported by BI (1.00), ML (100%), and MP (100%) methods.

Furthermore, the identified common and unique SSRs might play an important role in the analysis of the genetic diversity of the genus *Hedysarum*. In particular, the unique pentanucleotide SSRs identified in *H. drobovii* (AAAGG, AAAAT, AAGAC, ATTTT, and TTGTC) and *H. taipeicum* (AACCG), as well as the unique hexanucleotide SSRs identified in *H. semenovii* (ATAGCT and TTTTTC) and *H. taipeicum* (ATTCTT) may be effectively utilized in the future for species identification and assessment of genetic diversity in their populations. In addition, repeat sequences are known to play an important role in cp genome rearrangement, recombination, gene duplication, deletion, and gene expression (Gemayel et al., 2010; Do et al., 2014; Vieira et al., 2014; Li and Zheng, 2018),. They also have been reported to be responsible for substitutions and indels in the cp genome (Yi et al., 2013). We identified 51 (*H. petrovii*) to 214 (*H. flavescens*) repeat sequences among the six *Hedysarum* cp genomes analyzed, with forward repeats being the most common in *H. petrovii*, *H. flavescens*, *H. semenovii*, *H. lehmannianum*, and *H. taipeicum*, whereas palindromic and forward repeats were the most abundant in *H. drobovii*. Notably, *H. drobovii* and *H. petrovii*, both belonging to section *Multicaulia*, had significantly fewer repeat regions (53 and 51, respectively) compared to the other species (104–214). Further investigation into repeat sequences in section *Multicaulia* is necessary.

DNA barcodes with high variability are crucial for species identification, resource conservation, and phylogenetic analyses (Gregory, 2005; Bringloe and Saunders, 2019; Chen et al., 2019; Liu et al., 2019). Our comparison of *Hedysarum* species' cp genomes revealed high similarity in gene content and gene order, with genome lengths ranging from 121,176 to 126,738 bp. However, mVISTA analyses indicated that sequence variation was higher in non-coding regions than in other regions. Nucleotide diversity analysis identified six highly variable regions in the whole cp genome of *Hedysarum* species, mainly located in non-coding regions. Three protein-coding genes, *clpP*, *accD*, and *atpF*, exhibited higher Pi values and were found to be highly variable regions. The variability of the *clpP* gene can be attributed to the large variation in its Exon I length between species, ranging from 3 bp (*H. drobovii*) to 219 bp (*H. taipeicum* and *H. lehmannianum*) (Supplementary Table 2). The *clpP* and *accD* genes have been reported to play an important role in counteracting biotic and abiotic stress (Singh et al., 2015; Sinha et al., 2018; Ali and Baek, 2020), while the *atpF* gene is involved in the synthesis of ATF during photosynthesis (Ghulam et al., 2012), which is greatly influenced by altitude conditions (Wang et al., 2017). The high Pi values observed in these genes may reflect adaptation to different environmental conditions. Moreover, these highly variable regions can serve as candidate molecular markers and a reference for identifying future *Hedysarum* species. The *clpP, accD*, and *atpF* gene exon regions have similarly been identified as some of the most highly variable hotspot regions in cp genomes of some species (Mo et al., 2020; Mascarello et al., 2021; Moghaddam et al., 2022; Long et al., 2023).

Our phylogenetic analysis based on complete cp genomes and protein-coding genes confirmed previous studies on cp genome data of IRLC clade species, determining the phylogenetic position of *Hedysarum* as a sister to *Onobrychis* (She et al., 2019; Jin et al., 2021; Moghaddam et al., 2022; Tian et al., 2021). Our study also confirms the monophyly of *Hedysarum* based on plastid DNA genes, which is consistent with previous studies by Duan et al. (2015); Liu et al. (2017), and Nafisi et al. (2019). Although a limited number of species were used in our study, the phylogenetic relationships among the three sections of *Hedysarum* were analyzed using complete cp genomes and 76 protein-coding genes for the first time. Our results suggested that sections of *Hedysarum* could be monophyletic based on both cp genome data. However, further studies with more species, particularly from section *Stracheya*, are necessary to confirm this outcome. Furthermore, both our phylogenetic reconstructions revealed a close relationship between section *Stracheya* and *Hedysarum*, which is consistent with previous findings by Liu et al. (2017), but incongruous with the outcomes of Duan et al. (2015) and Nafisi et al. (2019). Additionally, this relationship is supported by the shared morphological characteristics between species of section *Stracheya* and *Hedysarum*, including leaves with numerous leaflets (4-15 paired), wings longer than half of the keel, and pods lacking ribs, bristles, or spines.

Conclusion

Our study is the first research work to investigate the genome characteristics of the genus *Hedysarum*. We sequenced, assembled, and annotated the cp genome of *H. drobovii*, *H. flavescens* va *H. lehmannianum* using high-throughput technology. Our study is based on cp genome data from a total of six *Hedysarum* species, including three previously published species. The cp genomes of all six *Hedysarum* species analyzed contained 110 genes, including 76 protein-coding genes, 4 rRNA genes, and 30 tRNA genes. We identified between 172 and 190 microsatellites and 51 to 214 pairs of repeat sequences among the six *Hedysarum* species cp genomes. In addition, we identified seven common SSRs and 14 unique SSRs in the studied *Hedysarum* species. Furthermore, we detected highly variable regions in the *clpP*, *accD*, and *atpF* protein-coding genes. These repeat motifs and highly variable genes could be used for evolutionary studies, phylogenetic relationships, plant population genetics, and species identification. Our phylogenetic reconstructions using the complete cp genome and protein-coding genes confirmed the monophyly of *Hedysarum*. Additionally, we supported the close relationship between section *Stracheya* and section *Hedysarum* using all three BI, ML, and MP methods. However, future studies using more species will provide a better understanding of the relationships among *Hedysarum* sections.

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

IJ: Conceptualization, methodology, data analysis, identification, visualization, writing, original draft preparation, reviewing, editing, and discussing. DM: methodology and data analysis. ZY: methodology and collection. KT: Supervision, investigation, identification, reviewing, editing, and discussing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the state research project "Taxonomic revision of polymorphic plant families of the flora of Uzbekistan" (FZ-20200929321) and the State Programs for 2021–2025 years "Grid mapping of the flora of Uzbekistan" and the "Tree of life: monocots of Uzbekistan" of the Institute of Botany of the Academy of Sciences of the Republic of Uzbekistan, the National Natural Science Foundation of China (32170215), the International Partnership Program of Chinese Academy of Sciences (151853KYSB20180009), Yunnan Young and Elite Talents Project (YNWRQNBJ-2019-033), the Ten Thousand Talents Program of Yunnan Province (202005AB160005), and the Chinese Academy of Sciences "Light of West China" Program.

Acknowledgments

The authors thank two reviewers for their comments and suggestions, which greatly improved the article.

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/fpls.2023.1211247/full#supplementary-material

References

Ali, M. S., and Baek, K. H. (2020). Protective roles of cytosolic and plastidal proteasomes on abiotic stress and pathogen invasion. *Plants* 9 (7), 832. doi: 10.3390/plants9070832

Amirahmadi, A., Osaloo, S. K., Moein, F., Kaveh, A., and Maassoumi, A. A. (2014). Molecular systematics of the tribe *Hedysareae* (Fabaceae) based on nrDNA ITS and plastid *trnL-F* and *matK* sequences. *Plant System. Evol.* 300 (4), 729–747. doi: 10.1007/s00606-013-0916-5

Bei, Z., Zhang, L., and Tian, X. (2022). Characterization of the complete chloroplast genome of *Oxytropis aciphylla* Ledeb.(Leguminosae). *Mitochondrial DNA Part B.* 7 (9), 1756–1757. doi: 10.1080/23802359.2022.2124822

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198

Bock, R. (2007). "Structure, function, and inheritance of plastid genomes," in *Cell and molecular biology of plastids* (Berlin, Heidelberg: Springer Berlin Heidelberg). doi: 10.1007/4735_2007_0223

Bringloe, T. T., and Saunders, G. W. (2019). DNA barcoding of the marine macroalgae from Nome, Alaska (Northern Bering Sea) reveals many trans-Arctic. *Polar Biol.* 42, 851–864. doi: 10.1007/s00300-019-02478-4

Callis, J., Fromm, M., and Walbot, V. (1987). Introns increase gene expression in cultured maize cells. *Genes Dev.* 1, 1183–1200. doi: 10.1101/gad.1.10.1183

Chen, K. C., Zakaria, D., Altarawneh, H., Andrews, G. N., Ganesan, G. S., John, K. M., et al. (2019). DNA barcoding of fish species reveals low rate of package mislabeling in Qatar. *Genome* 62, 69–76. doi: 10.1139/gen-2018-0101

Choi, T., Huang, M., Gorman, C., and Jaenisch, R. A. (1991). Generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.* 11, 3070–3074. doi: 10.1128/mcb.11.6.3070-3074.1991

Choi, B. H., and Ohashi, H. (1996). Pollen morphology and taxonomy of *Hedysarum* and its related genera of the tribe *Hedysareae* (Legiminosae-Papilionoideae). *J. Japanese Bot.* 71, 191–213.

Choi, B. H., and Ohashi, H. (2003). Generic criteria and an infrageneric system for *Hedysarum* and related genera (Papilionoideae-Leguminosae). *Taxon* 52, 567–576. doi: 10.2307/3647455

Do, H. D., Kim, J. S., and Kim, J. H. (2014). A trnI_CAU triplication event in the complete chloroplast genome of *Paris verticillata* M. Bieb.(Melanthiaceae, Liliales). *Genome Biol. Evol.* 6 (7), 1699–1706. doi: 10.1093/gbe/evu138

Dong, Y., Tang, D., Zhang, N., Li, Y., Zhang, C., Li, L., et al. (2013). Phytochemicals and biological studies of plants in genus *Hedysarum*. *Chem. Cent. J.* 7 (1), 1–3. doi: 10.1186/1752-153X-7-124

Duan, L., Wen, J., Yang, X., Liu, P. L., Arslan, E., Ertuğrul, K., et al. (2015). Phylogeny of *Hedysarum* and tribe *Hedysareae* (Leguminosae: Papilionoideae) inferred from sequence data of ITS, *matK*, *trnL-F* and *psbA-trnH*. *Taxon* 64 (1), 49–64. doi: 10.12705/641.26

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* 5 (6), 435–445. doi: 10.1038/nrg1348

Emami, S., Arumainayagam, D., Korf, I., and Rose, A. B. (2013). The efects of a stimulating intron on the expression of heterologous genes in *Arabidopsis thaliana*. *Plant Biotechnol. J.* 11, 555–563. doi: 10.1111/pbi.12043

Fedtschenko, B. A. (1948). *Flora of URSS* Vol. 13 (St. Petersburg: Academiae Scientiarum URSS, Mosqua-Leningrad Press), 259–319.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* 32, 273–279. doi: 10.1093/nar/gkh458

Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 1 (44), 445–477. doi: 10.1146/annurev-genet-072610-155046

George, B., Bhatt, B. S., Awasthi, M., George, B., and Singh, A. K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr. Genet.* 61 (4), 665–677. doi: 10.1007/s00294-015-0495-9

Ghulam, M. M., Zghidi-Abouzid, O., Lambert, E., Lerbs-Mache, S., and Merendino, L. (2012). Transcriptional organization of the large and the small ATP synthase operons, *atpI/H/F/A* and *atpB/E*, in *Arabidopsis thaliana* chloroplasts. *Plant Mol. Biol.* 79 (3), 259–272. doi: 10.1007/s11103-012-9910-5

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature* 434, 1067. doi: 10.1038/4341067b

Jin, Z., Jiang, W., Yi, D., and Pang, Y. (2021). The complete chloroplast genome sequence of Sainfoin (*Onobrychis viciifolia*). *Mitochondrial DNA Part B.* 6 (2), 496–498. doi: 10.1080/23802359.2020.1871439

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 1–31. doi: 10.1186/s13059-020-02154-5

Kang, S. H., Lee, J. H., Lee, H. O., Ahn, B. O., Won, S. Y., Sohn, S. H., et al. (2018). Complete chloroplast genome and 45S nrDNA sequences of the medicinal plant species *Glycyrrhiza glabra* and *Glycyrrhiza uralensis*. *Genes Genet. sys.* 93 (3), 83–89. doi: 10.1266/ggs.17-00002

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29 (22), 4633–4642. doi: 10.1093/nar/29.22.4633

Li, B., and Zheng, Y. (2018). Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. *Sci. Rep.* 8, 9285. doi: 10.1038/s41598-018-27453-7

Liu, X., Chang, E. M., Liu, J. F., Huang, Y. N., Wang, Y., and Yao, N. (2019). Complete chloroplast genome sequence and phylogenetic analysis of *Quercus bawanglingensis* Huang, Li et Xing, a vulnerable oak tree in China. *Forests* 10 (7), 587. doi: 10.3390/f10070587

Liu, P. L., Wen, J., Duan, L., Arslan, E., Ertuðrul, K., and Chang, Z. Y. (2017). *Hedysarum* L. (Fabaceae: Hedysareae) is not monophyletic – evidence from phylogenetic analyses based on five nuclear and five plastid sequences. *PloS One* 12, e0170596. doi: 10.1371/journal.pone.0170596

Lock, J. M. (2005). *Legumes of the World* (Richmond: Royal Botanic Gardens Press).

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. doi: 10.1007/s00294-007-0161-y

Long, L., Li, Y., Wang, S., Liu, Z., Wang, J., and Yang, M. (2023). Complete chloroplast genomes and comparative analysis of *Ligustrum* species. *Sci. Rep.* 13, 212. doi: 10.1038/s41598-022-26884-7

Mascarello, M., Amalfi, M., Asselman, P., Smets, E., Hardy, O. J., Beeckman, H., et al. (2021). Genome skimming reveals novel plastid markers for the molecular identification of illegally logged African timber species. *PloS One* 16 (6), e0251655. doi: 10.1371/journal.pone.0251655

Minh, B. Q., Lanfear, R., Trifinopoulos, J., Schrempf, D., and Schmidt, H. A. (2021) *IQ-TREE version 2.1.2: Tutorials and manual phylogenomic software by Maximum Likelihood*. Available at: http://www.iqtree.org/doc/iqtree-doc.pdf.

Mo, Z., Lou, W., Chen, Y., Jia, X., Zhai, M., Guo, Z., et al. (2020). The chloroplast genome of *Carya illinoinensis*: genome structure, adaptive evolution, and phylogenetic analysis. *Forests* 11 (2), 207. doi: 10.3390/f11020207

Moghaddam, M., Ohta, A., Shimizu, M., Terauchi, R., and Kazempour-Osaloo, S. (2022). The complete chloroplast genome of *Onobrychis gaubae* (Fabaceae-Papilionoideae): comparative analysis with related IR-lacking clade species. *BMC Plant Biol.* 22 (1), 75. doi: 10.1186/s12870-022-03465-4

Nafisi, H., Kazempour-Osaloo, Sh., Mozaffarian, V., and Schneeweiss, G. M. (2019). Molecular phylogeny and divergence times of *Hedysarum* (Fabaceae) with special reference to section *Multicaulia* in Southwest Asia. *Plant System. Evol.* 305 (10), 1001–1017. doi: 10.1007/s00606-019-01620-3

Nazareno, A. G., Carlsen, M., and Lohmann, L. G. (2015). Complete chloroplast genome of *tanaecium tetragonolobum*: the first bignoniaceae plastome. *PloS One* 10, e0129930. doi: 10.1371/journal.pone.0129930

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 (1), 268–274. doi: 10.1093/molbev/msu300

Olmstead, R. G., and Palmer, J. D. (1994). Chloroplast DNA systematics: a review of methods and data analysis. *Am. J. bot.* 81 (9), 1205–1224. doi: 10.1002/j.1537-2197.1994.tb15615.x

Ren, F., Wang, L., Li, Y., Zhuo, W., Xu, Z., Guo, H., et al. (2021). Highly variable chloroplast genome from two endangered Papaveraceae lithophytes *Corydalis tomentella* and *Corydalis saxicola*. *Ecol. Evol.* 11 (9), 4158–4171. doi: 10.1002/ece3.7312

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *System. Biol.* 61 (3), 539–542. doi: 10.1093/sysbio/sys029

Rozas, J., Ferrer-Mata, J. C., Sanchez-DelBarrio, P., Librado, P., and Guirao-Rico, S. E. (2018) *DnaSP version 6.12.03: A software for comprehensive analysis of DNA polymorphism data*. Available at: http://www.ub.es/dnasp.

Saski, C., Lee, S. B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., et al. (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Mol. Biol.* 59, 309–322. doi: 10.1007/s11103-005-8882-0

She, R. X., Li, W. Q., Xie, X. M., Gao, X. X., Wang, L., Liu, P. L., et al. (2019). The complete chloroplast genome sequence of a threatened perennial herb species *Taibai sweetvetch* (*Hedysarum taipeicum* KT Fu). *Mitochondrial DNA Part B.* 4 (1), 1439–1440. doi: 10.1080/23802359.2019.1598817

Singh, R. P., Shelke, G. M., Kumar, A., and Jha, P. N. (2015). Biochemistry and genetics of ACC deaminase: a weapon to "stress ethylene" produced in plants. *Front. Microbiol.* 6. doi: 10.3389/fmicb.2015.00937

Sinha, R., Pal, A. K., and Singh, A. K. (2018). Physiological, biochemical and molecular responses of lentil (*Lens culinaris* Medik.) genotypes under drought stress. *Indian J. Plant Physiol.* 23, 772–784. doi: 10.1007/s40502-018-0411-7

Smith, D. R., and Keeling, P. J. (2015). Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. U.S.A.* 112 (33), 10177–10184. doi: 10.1073/pnas.1422049112

Su, C., Liu, P. L., Chang, Z. Y., and Wen, J. (2019). The complete chloroplast genome sequence of *Oxytropis bicolor* Bunge (Fabaceae). *Mitochondrial DNA Part B.* 4, 3762–3763. doi: 10.1080/23802359.2019.1682479

Swofford, D. L. (2002). *PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version. 4* (Sunderland: Sinauer Associates). doi: 10.1111/j.0014-3820.2002.tb00191.x

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. evol.* 28 (10), 2731–2739. doi: 10.1093/molbev/msr121

Tian, C., Li, X., Wu, Z., Li, Z., Hou, X., and Li, F. Y. (2021). Characterization and comparative analysis of complete chloroplast genomes of three species from the genus *Astragalus* (Leguminosae). *Front. Genet.* 12. doi: 10.3389/fgene.2021.705482

Vieira, L., Faoro, H., Rogalski, M., Fraga, H., Cardoso, R. L. A., de Souza, E. M., et al. (2014). The complete chloroplast genome sequence of *Podocarpus lambertii*: genome structure, evolutionary aspects, gene content and SSR detection. *PloS One* 9 (3), e90618. doi: 10.1371/journal.pone.0090618

Wang, H., Prentice, I. C., Davis, T. W., Keenan, T. F., Wright, I. J., and Peng, C. (2017). Photosynthetic responses to altitude: an explanation based on optiMality principles. *New Phytol.* 213 (3), 976–982. doi: 10.1111/nph.14332

Wojciechowski, M. F., Lavin, M., and Sanderson, M. J. (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid gene resolves many well-supported subclades within the family. *Amer J. Bot.* 91, 1846–1862. doi: 10.3732/ajb.91.11.1846

Yi, X., Gao, L., Wang, B., Su, Y. J., and Wang, T. (2013). The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of Cephalotaxus chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. *Genome Biol. evol.* 5 (4), 688–698. doi: 10.1093/gbe/evt042

Yin, D., Wang, Y., Zhang, X., Ma, X., He, X., and Zhang, J. (2017). Development of chloroplast genome resources for peanut (Arachis hypogaea L.) and other species of Arachis. *Sci. Rep.* 7, 11649. doi: 10.1038/s41598-017-12026-x

Yuan, M., Yin, X., Gao, B., Gu, R., and Jiang, G. (2022). The chloroplasts genomic analyses of four specific Caragana species. *PLoSONE* 17 (9), e0272990. doi: 10.1371/journal.pone.0272990

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 September 2023
DOI 10.3389/fpls.2023.1210078

OPEN ACCESS

EDITED BY
Mark Chapman,
University of Southampton,
United Kingdom

REVIEWED BY
Linchun Shi,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Adriana Sacco,
National Research Council (CNR), Italy

*CORRESPONDENCE
Vineet K. Sharma
vineetks@iiserb.ac.in

RECEIVED 21 April 2023
ACCEPTED 15 August 2023
PUBLISHED 01 September 2023

CITATION
Mahajan S, Bisht MS, Chakraborty A and Sharma VK (2023) Genome of *Phyllanthus emblica*: the medicinal plant Amla with super antioxidant properties.
Front. Plant Sci. 14:1210078.
doi: 10.3389/fpls.2023.1210078

Genome of *Phyllanthus emblica*: the medicinal plant Amla with super antioxidant properties

Shruti Mahajan, Manohar S. Bisht, Abhisek Chakraborty and Vineet K. Sharma*

MetaBioSys Group, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India

Phyllanthus emblica or Indian gooseberry, commonly known as amla, is an important medicinal horticultural plant used in traditional and modern medicines. It bears stone fruits with immense antioxidant properties due to being one of the richest natural sources of vitamin C and numerous flavonoids. This study presents the first genome sequencing of this species performed using 10x Genomics and Oxford Nanopore Technology. The draft genome assembly was 519 Mbp in size and consisted of 4,384 contigs, N50 of 597 Kbp, 98.4% BUSCO score, and 37,858 coding sequences. This study also reports the genome-wide phylogeny of this species with 26 other plant species that resolved the phylogenetic position of *P. emblica*. The presence of three ascorbate biosynthesis pathways including L-galactose, galacturonate, and myo-inositol pathways was confirmed in this genome. A comprehensive comparative evolutionary genomic analysis including gene family expansion/contraction and identification of multiple signatures of adaptive evolution provided evolutionary insights into ascorbate and flavonoid biosynthesis pathways and stone fruit formation through lignin biosynthesis. The availability of this genome will be beneficial for its horticultural, medicinal, dietary, and cosmetic applications and will also help in comparative genomics analysis studies.

KEYWORDS

***Phyllanthus emblica*, amla, medicinal plant, genome sequencing, antioxidant, vitamin C biosynthesis**

Introduction

Vitamin C, also known as ascorbic acid, is a vital vitamin due to its multifaceted roles in animals as well as plants, and is an essential component of the human diet (Gallie, 2013; Carr and Maggini, 2017). The prolonged deficiency of this vitamin causes scurvy which was infamously responsible for killing thousands of sailors in the medieval period, since humans and primates cannot synthesise vitamin C due to the absence of an enzyme gulono-lactone oxidase (GULO), which is responsible for the final conversion to ascorbic

acid (Martini, 2003; Wheeler et al., 2015). Thus, they depend majorly on plants that are the dominant sources of vitamin C for animals (Wheeler et al., 2015).

Phyllanthus emblica is one of the richest sources of natural vitamin C, and is, also known as Indian gooseberry or amla. It is an economically important medicinal horticultural plant that belongs to the family Phyllanthaceae in order Malpighiales, and is widely used in pharmaceuticals, nutraceuticals, food industry, and cosmetics sectors with an estimated market of USD 49.34 billion by 2025 (Muzaffar et al., 2022). Genus *Phyllanthus* is the largest genus of its family with approximately 1,000 species of which several are used as ethnomedicinal herbs due to the presence of medicinal phytochemicals (Sarin et al., 2014; Mao et al., 2016; Geethangili and Ding, 2018). The morphological characteristics of *P. emblica* include a light grey stem with thin flaky bark, simple leaves, and greenish-yellow unisexual flowers that are arranged like female flowers on the top and male flowers on the lower side. The fruits are typical drupes of about 2 cm in diameter, also known as stone fruits, with seeds encased in a hard lignified endocarp known as stone that helps in seed protection and dispersal (Dardick and Callahan, 2014; Dasaroju and Gottumukkala, 2014).

The geographical distribution of this prominent ethnomedicinal herbal species spreads from tropical to warm temperate regions like India, China, Sri Lanka, Bangladesh, Indonesia, Thailand, etc., among which India is the top producers of amla with annual production of 1,275,660 metric tonnes (Mao et al., 2016; Department of Agriculture & Farmers Welfare et al., 2021). This plant has also been used in many traditional medicine systems like Indian Ayurveda, Traditional Chinese Medicine System, etc. and is now widely used in modern medicines (Mao et al., 2016; Gul et al., 2022). Extracts of almost all parts of this plant such as leaf, bark, seed, root and fruit show medicinal properties like anti-microbial, anti-viral, anti-inflammatory, anti-oxidant, anti-aging, anti-diabetic, hypolipidemic, hypoglycaemic, neuroprotective, anti-cancer, immunomodulatory and hepatoprotective, etc. due to the presence of various secondary metabolites (phytochemicals) like alkaloids, phenolic acids, hydrolysable tannins, flavonoids, etc. with significance to human health and diseases (Gantait et al., 2021; Gul et al., 2022; Saini et al., 2022; Yan et al., 2022). The clinical effectiveness of *P. emblica* has been confirmed in diseases like dyslipidaemia, type 2 diabetes, chronic periodontitis, symptomatic knee osteoarthritis, etc. (Gantait et al., 2021). Amla is used in treating COVID-19 patients where its consumption shortened the recovery time (Varnasseri et al., 2022). Additionally, its phytochemicals are reported as potential protease inhibitors of SARS-CoV-2 virus through *in-silico* evidences (Murugesan et al., 2021; Pandey et al., 2021). Its extracts are proven to have protective effects by maintaining gut microbiome homeostasis *in vivo* (Li X. et al., 2022; Luo et al., 2022). Along with its benefits in human health, it is also effective in aquaculture, dairy and poultry as feed additives (Nguse et al., 2022; Van Doan et al., 2022; Abo Ghanima et al., 2023).

Among the vitamin C-rich fruits, *P. emblica* is known to contain the highest content of vitamin C (up to 720mg/100g fruit) along with other phytochemicals, minerals and amino acids (Kubola et al., 2011; Chavhan, 2017; Abeysuriya et al., 2020; Gul et al., 2022). Plants produce this vitamin to protect them against biotic (pathogens) and abiotic stresses (heat or light), and is also needed for the biosynthesis of plant hormones, and plant pigments, and acts as a cofactor in the cell cycle and metabolism, etc. (Gallie, 2013). The ascorbate biosynthesis occurs in plants through four proposed pathways i.e., L-galactose (also known as Smirnoff-Wheeler pathway), galacturonate (uronic acid pathway), L-gulose and myo-inositol pathways (Fenech et al., 2019; Paciolla et al., 2019). Among these pathways, the Smirnoff-Wheeler (SW) pathway is considered as the most common pathway of ascorbate biosynthesis (Gómez-García and Ochoa-Alejo, 2016; Sodeyama et al., 2021). Various genome-wide studies revealed ascorbic acid biosynthesis pathways in *Psidium guajava*, *Citrus sinensis*, kiwifruits, etc., however the ascorbate biosynthesis pathways have not been examined in *P. emblica* (Xu et al., 2013; Feng et al., 2021; Liao et al., 2021; Han et al., 2022).

Despite being a pharmaceutically and nutritionally important plant, the genome sequence of *P. emblica* still remains unknown. However, transcriptome studies were carried out previously to explore a few biosynthesis pathways in *P. emblica* (Kumar et al., 2016; Xiong-fang et al., 2018). The number of chromosomes in *P. emblica* was first reported as 28 in 1943 (Perry, 1943). Several following studies reported the chromosome numbers ranging from 52 to 104, and the most recent study has reported the presence of 100 chromosomes in *P. emblica* (Ammal and Raghavan, 1958; Soontornchainaksaeng and Chaiyasut, 1999; Rahman et al., 2021). Thus, to gain genomic insights into the medicinal properties of *P. emblica*, we performed its genome sequencing and assembly using a hybrid approach that includes 10x Genomics and Oxford Nanopore Technology (ONT) long-read sequencing technologies along with transcriptomic sequencing using the Illumina technology. Further, we analysed the genes involved in vitamin C, lignin and flavonoid biosynthesis pathways. We also constructed a genome-wide phylogenetic tree of *P. emblica* with 26 plant species, which were further analysed for gene family expansion and contraction. Furthermore, this study performed a comprehensive comparative evolutionary genomic analysis across 19 plant species to uncover the genes with multiple signatures of adaptive evolution in *P. emblica*.

Materials and methods

DNA-RNA extraction, species identification and sequencing

The leaves sample from an individual plant located at the campus of Indian Institute of Science Education and Research Bhopal, India (23.2858° N, 77.2755° E) were used in this study (Supplementary Figure 1). The DNA was extracted from the leaves sample using Carlson lysis buffer except for the precipitation step that was carried out with 0.5X volume of NaCl ad 0.7X volume of isopropanol (Jaiswal et al., 2021). The extracted DNA was quantified, and quality was checked on Qubit 2.0 fluorometer and Nanodrop 8000 spectrophotometer, respectively. Species identification assay was performed using marker genes: Internal

Transcribed Spacer (*ITS*) and Maturase K (*MatK*). The extracted DNA was utilised to prepare libraries for 10x Genomics and nanopore sequencing that were sequenced on Illumina NovaSeq 6000 and MinION Mk1C sequencers, respectively. The RNA extraction from leaf tissue was performed as per the protocol described by Kumar and Singh (2012) with a few modifications (Kumar and Singh, 2012). The RNA was used for preparing the library using TruSeq Stranded Total RNA Library Preparation kit (Illumina Inc., CA, USA) with Ribo-zero Plant workflow and sequenced on Illumina NovaSeq 6000 instrument for generating 150 bp paired-end reads. The detailed method of DNA and RNA extraction with library preparation and sequencing is discussed in Supplementary Text 1.

Genome assembly

The proc10xG set of python scripts (https://github.com/ucdavis-bioinformatics/proc10xG) was used to pre-process the 10x Genomics raw reads by removing the barcode sequences. The obtained reads were processed by SGA-preqc (paired-end mode) for genome size estimation which works on a k-mer distribution-based approach (Simpson and Durbin, 2012). For genome complexity assessment, these pre-processed reads were used by Jellyfish v2.2.10 and GenomeScope v2.0 for generating k-mer count histograms and calculating heterozygosity, respectively (Marçais and Kingsford, 2011; Ranallo-Benavidez et al., 2020).

Guppy v3.2.1 (Oxford Nanopore Technologies) was used to carry out base calling of nanopore raw reads. Adaptor removal was performed on this base called raw data using Porechop v0.2.4 (Oxford Nanopore Technologies). The pre-processed reads were utilised for genome assembly using three different assemblers: wtdbg v2.0.0, SMARTdenovo (https://github.com/ruanjue/smartdenovo), and Flye v2.9 (Kolmogorov et al., 2019; Ruan and Li, 2020). wtdbg v2.0.0 and Flye v2.9 were used with default settings whereas SMARTdenovo was used with zero as minimum read length (Kolmogorov et al., 2019; Ruan and Li, 2020). Quast v5.0.2 was employed to assess the genome assembly statistics (Gurevich et al., 2013). The genome assembly resulting from Flye v2.9 was considered for further analysis due to its better assembly statistics and assembled genome size (Kolmogorov et al., 2019). The assembly was polished three times by Pilon v1.24 using filtered reads. ARCS v1.2.2 and LINKS v2.0.0 (default settings) were employed for the first round of scaffolding using Longranger basic v2.2.2 (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/installation) barcode filtered 10x Genomics linked reads and adaptor-removed nanopore reads, respectively (Walker et al., 2014; Warren et al., 2015; Yeo et al., 2018). After scaffolding, the quality-filtering of RNA-Seq paired-end raw reads was performed using Trimmomatic v0.39 with parameters- "LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:50" which were subsequently utilised by AGOUTI v0.3.3 for the second round of scaffolding (Bolger et al., 2014; Zhang et al., 2016).

Supernova v2.1.1 was used to perform *de novo* assembly of *P. emblica* with maxreads=all options with other default parameters (Weisenfeld et al., 2017). The obtained genome assembly was corrected by Tigmint v1.2.6 using Longranger basic v2.2.2 processed linked reads (Jackman et al., 2018). Further, the first round of scaffolding was performed with ARCS v1.2.2 and LINKS v2.0.0 using Longranger basic v2.2.2 processed linked reads and adaptor removed nanopore reads, respectively (Warren et al., 2015; Yeo et al., 2018). To increase the assembly contiguity, AGOUTI v0.3.3 was used with the pre-processed transcriptomic paired-end reads (quality-filtered) (Zhang et al., 2016).

RagTag v2.1.0 was used to merge both the assemblies obtained from Supernova and Flye assemblers using the patch command line utility (Alonge et al., 2021). RagTag uses the main assembly as reference and query assembly to fill the gap in the reference assembly. LR_Gapcloser was used to perform gap-closing of the assembly using pre-processed nanopore reads (Xu et al., 2019). Sealer v2.3.5 was used for gap-closing of the assembly using barcode-removed linked reads with 30-120 k-mer value and 10 bp interval (Paulino et al., 2015). The fixation of small indels, base errors, and local misassemblies was performed by Pilon v1.24 using pre-processed linked reads to provide the draft genome assembly of *P. emblica* (Walker et al., 2014). Obtained draft genome assembly was further length base filtered and scaffolds having length ≥5 kbp were retained. BUSCO v5.2.2 with embryophyta_odb10 single-copy orthologs dataset was employed to assess the completeness of genome assembly (Simão et al., 2015). For further assessment of the assembly quality, the barcode-filtered 10x Genomics reads, nanopore long reads and the quality-filtered transcriptomic reads were mapped onto the genome assembly using BWA-MEM v0.7.17, MiniMap2 v2.17 and HISAT2 v2.2.1, respectively, and SAMtools v1.13 "flagstat" was used to calculate the percentage of mapped reads (Li et al., 2009; Li, 2013; Kim et al., 2015; Li, 2018).

Annotation of genome and construction of gene set

For annotation of repeats in the final genome assembly, RepeatModeler v2.0.3 was used to generate a *de novo* repeat library (Flynn et al., 2020). The clustering of repeats was performed using CD-HIT-EST v4.8.1 with parameters - 8 bp seed size and 90% sequence identity to eliminate redundant sequences (Fu et al., 2012). The resultant repeat library was utilised by RepeatMasker v4.1.2 (http://www.repeatmasker.org) to soft-mask the final genome assembly of *P. emblica*.

The coding gene set was constructed on the resultant repeat-masked genome assembly using MAKER v3.01.04 pipeline which deploys approaches such as *ab initio* and evidence alignment for prediction (Campbell et al., 2014). The construction of *de novo* transcriptome assembly was performed by Trinity v2.14.0 (default parameters) using quality-filtered transcriptomic reads of *P. emblica* from this study and previously reported studies (Haas et al., 2013; Liu et al., 2018). The gene set was constructed with transcriptome assembly and protein sequences of species belonging to the same order Malpighiales (*Populus trichocarpa* and *Manihot esculenta*) that were used as EST and protein evidence, respectively. In the MAKER pipeline, AUGUSTUS v3.2.3 was used for *ab initio* gene prediction while empirical evidence alignments and alignment polishing were performed using BLAST and Exonerate v2.2.0

(https://github.com/nathanweeks/exonerate), respectively (Altschul et al., 1990; Stanke et al., 2006). Based on the length and Annotation Edit Distance (AED) of gene models, the final gene set was constructed by selecting genes with length (≥150bp) and AED values <0.5. The completeness of this final gene set (also termed a high-confidence gene set) was checked using BUSCO v5.2.2 with embryophyta_odb10 dataset (Simão et al., 2015).

Additionally, Barrnap v0.9 (https://github.com/tseemann/barrnap) and tRNAscan-SE v2.0.9 were used to perform *de novo* prediction of rRNA and tRNA, respectively (Chan and Lowe, 2019). Based on homology, miRNA gene sequences in the *P. emblica* genome were identified using miRbase database with e-value 10^{-9} and 80% identity (Griffiths-Jones et al., 2007).

Phylogenetic tree construction

The 26 plant species were selected from Ensembl plant release 54 for phylogenetic analysis considering the representation of each plant family among the selected species (Bolser et al., 2016). Besides the protein sequences of these selected 26 plant species, MAKER-derived protein sequences of *P. emblica* were used for phylogenetic tree construction. Among all the protein files, the longest isoform for each protein was selected and provided to OrthoFinder v2.5.4 to construct the set of orthologous genes (Emms and Kelly, 2019). KinFin v1.0 was used to extract fuzzy one-to-one orthologs protein sequences that were present in all 27 species (Laetsch and Blaxter, 2017). MAFFT v7.310 was used to individually align all the obtained fuzzy one-to-one orthologs which were filtered and concatenated using BeforePhylo v0.9.0 (https://github.com/qiyunzhu/BeforePhylo) (Katoh and Standley, 2013). These obtained protein sequences were used to construct a phylogenetic tree based on maximum likelihood using RAxML with the 'PROTGAMMAAUTO' amino acid substitution model and 100 bootstrap values (Stamatakis, 2014).

Amino acid sequences of *MatK* gene from 49 *Phyllanthus* species (top 49 species except *P. emblica* based on sequence length) and *Zea mays* (outgroup) obtained from UniProt database along with MatK protein sequence of *P. emblica* were used for the phylogenetic analysis. MAFFT v.7.310 was used to align these protein sequences, and RAxML v8.2.12 was used with 1000 bootstrap value and 'PROTGAMMAAUTO' amino acid substitution model to construct the *MatK*-based phylogenetic tree (Katoh and Standley, 2013; Stamatakis, 2014).

Gene family expansion and contraction analysis

The proteome files containing the longest isoform for every protein from selected 27 species along with generated species phylogenetic tree were provided to CAFE v5 to assess the evolution of gene families (Mendes et al., 2020). The species phylogenetic tree was adjusted to an ultrametric tree based on the calibration point of 120 million years between *P. emblica* and *Beta vulgaris* obtained from TimeTree database v5.0 (Kumar et al., 2022). BLASTP was performed on protein sequences of all 27 species in All-versus-All mode (Altschul et al., 1990). The BLASTP results were clustered using MCL v14-137 and gene families containing clade-specific genes and >100 gene copies for minimum of one species were eliminated. These resultant gene families and ultrametric species tree were used to analyse the evolution (expansion/contraction) of gene families using a two-lambda (λ) model where λ signifies a random birth-death parameter. Among the obtained contracted/expanded gene families, gene families with >10 genes were considered as highly contracted/expanded gene families.

Identification of signatures of adaptive evolution

The 19 plant species were selected for identification of genes with evolutionary signatures that included five species of order Malpighiales i.e., *Linum usitatissimum, Manihot esculenta, Phyllanthus emblica*, *Populus trichocarpa* and *Ricinus communis* along with *Actinidia chinensis* (order Ericales), *Arabidopsis thaliana* (order Brassicales), *Coffea canephora* (order Gentianales), *Cucumis sativus* (order Cucurbitales), *Daucus carota* (order Apiales), *Eucalyptus grandis* (order Myrtales), *Ficus carica* (order Rosales), *Gossypium raimondii* (order Malvales), *Helianthus annuus* (order Asterales), *Olea europaea* (order Lamiales), *Pistacia vera* (order Sapindales), *Quercus lobata* (order Fagales), *Solanum lycopersicum* (order Solanales) and *Vitis vinifera* (order Vitales). Orthologous gene sets were constructed by OrthoFinder v2.5.4 using the proteome files from 19 selected plant species. Orthogroups that contained protein sequences from all these selected species were retrieved and in case multiple protein sequences were present for a species, the longest isoform of that protein was selected and retained for further analysis.

Identification of genes with higher rate of evolution

The resulting orthogroups across 19 plant species were aligned individually using MAFFT v7.310 (Katoh and Standley, 2013). These obtained alignments were used to construct a phylogenetic tree for individual orthogroups using RAxML v8.2.12 with 'PROTGAMMAAUTO' amino acid substitution model and a bootstrap value of 100 (Stamatakis, 2014). R package "adephylo" was used to calculate root-to-tip branch length distance for genes of all species in the phylogenetic trees (Jombart et al., 2010). The genes of *P. emblica* with comparatively higher root-to-tip branch length distance values were extracted and listed as the genes with higher nucleotide divergence or rate of evolution.

Identification of *P. emblica* genes with unique amino acid substitutions

Using the multiple sequence alignments obtained from MAFFT v7.310, which were used for the identification of genes with a high rate of evolution, amino acid positions alike in all the species except *P. emblica* were extracted and labelled as genes with unique amino

acid substitution. Ten amino acids around any gap were not included in this analysis. The functional impact of obtained genes showing amino acid substitution was evaluated using Sorting Intolerant From Tolerant (SIFT) with UniProt database (Ng and Henikoff, 2003).

Identification of positively selected genes

MAFFT v7.310 was used for individual alignment of nucleotide sequence of all orthologous gene sets across selected 19 species (Katoh and Standley, 2013). PAML v4.9a with "codeml" program based on a branch-site model used nucleotide alignments in PHYLIP format and a species phylogenetic tree of these 19 species (constructed using RAxML) to identify genes with positive selection (Yang, 2007). These obtained genes with their log likelihood values were further processed through likelihood-ratio tests and genes with FDR-corrected p-values of <0.05 were labelled as positively selected genes. Positively selected codon sites were identified using Bayes Empirical Bayes (BEB) analysis with criteria of >95% probability for the foreground lineage.

Genes with multiple signatures of adaptive evolution (MSA)

The high rate of evolution, unique amino acid substitution with functional impact and positive selection are the evolutionary signatures of adaptive evolution. *P. emblica* genes that showed at least two of these evolutionary signatures were considered as the genes with multiple signatures of adaptive evolution (MSA).

Functional annotation

The annotation of high-confidence gene sets of *P. emblica* was performed using NCBI non-redundant (nr) database, SWISS-PROT database and Pfam-A v32.0 database using BLASTP (10^{-5} e-value), BLASTP (10^{-5} e-value), and HMMER v3.3, respectively (Bairoch and Apweiler, 2000; Bateman et al., 2004; Finn et al., 2011). The coding genes including the genes with evolutionary signatures were functionally annotated using KAAS and eggNOG mapper (Moriya et al., 2007; Huerta-Cepas et al., 2017). Further, the considered contracted and expanded gene families of *P. emblica* were extracted and provided to KAAS v2.1 and eggNOG mapper v2.1.9 for functional annotation, respectively (Moriya et al., 2007; Huerta-Cepas et al., 2017). The functional annotation of contracted and expanded gene families was also checked manually on Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

Analysis of vitamin C biosynthesis genes

The protein sequences of all the available enzymes of all four proposed pathways of vitamin C biosynthesis for *A. thaliana* were downloaded from Swiss-Prot or NCBI database. The gene D-galactose reductase (*GalUR*) was not available for *A. thaliana*, thus, sequence from strawberry plant species was used. These protein sequences were matched against the protein sequences of *P. emblica* using BLASTP with e-value 10^{-9} (Altschul et al., 1990). The enzymes involved in galactose pathway from vitamin C-rich plants i.e., *Actinidia chinensis* (kiwi), *Capsicum annuum* (chilli pepper), *Carica papaya* (papaya), *Citrus sinensis* (sweet orange), *Malpighia glabra* (acerola), *Myrciaria dubia* (camu-camu), *Solanum lycopersicum* (tomato) and *Vitis vinifera* (grapes) along with *Arabidopsis thaliana* as an outgroup were also obtained from UniProt or NCBI databases. Six genes i.e., Mannose 1-phosphate guanylyl transferase (*GMPP*), GDP-D-Mannose 3',5'-epimerase (*GME*), GDP-L-galactose phosphorylase (*GGP/VTC2/VTC5*) and L-galactose-1-phosphate phosphatase (*GPP/VTC4*), L-galactose dehydrogenase (*GalDH*) and L-galactono-1,4-lactone dehydrogenase (*GLDH*) were selected for phylogeny due to their sequence availability for all nine species. The phylogenetic tree was constructed using these six genes each from 10 selected species (including *P. emblica*) using RAxML (Stamatakis, 2014). Among the above expanded and contracted genes, we checked for the copy number of all the gene families involved in ascorbate biosynthesis and regeneration pathways in the *P. emblica* genome.

The distant orthologs of 20 genes involved in ascorbate biosynthesis and regeneration pathways were identified to elucidate their origin by HHblits web server with default parameters using UniRef30 database (Remmert et al., 2012). The top 10 hits considering unique genus for each gene were extracted and aligned using MAFFT v.7.310 (Katoh and Standley, 2013). These alignments were used to construct the phylogenetic trees for each of the genes using RAxML v8.2.12 (1000 bootstrap value and 'PROTGAMMAAUTO' amino acid substitution model) (Stamatakis, 2014).

Analysis of flavonoid biosynthesis pathway

The protein sequences of genes involved in flavonoid biosynthesis pathway for *Manihot esculenta* were downloaded from UniProt and NCBI databases and matched against the protein sequences of *P. emblica* using BLASTP (e-value 10^{-9}) (Altschul et al., 1990).

Analysis of lignin biosynthesis pathway

The protein sequences of genes involved in lignin biosynthesis pathway for *Arabidopsis thaliana* were downloaded from UniProt and NCBI databases and matched against the protein sequences of *P. emblica* using BLASTP (e-value 10^{-9}) (Altschul et al., 1990).

Analysis of gene structure

Exonerate v2.2.0 was used to examine the exon-intron structure of genes involved in ascorbate biosynthesis and regeneration pathways, flavonoid biosynthesis, and lignin biosynthesis.

Results

Species identification and sequencing

The species was confirmed using the sequencing of two DNA markers, ITS and *MatK*, that were aligned to *P. emblica* sequences available at NCBI-nt database with the highest identity of 100% and 99.89%, respectively. A total of 136 Gbp (~237x coverage) and 18.3 Gbp (~32x coverage) of genome sequence data were generated using third-generation sequencing technologies i.e., 10x Genomics and Oxford Nanopore Technology (ONT), respectively (Supplementary Table 1). Further, ~85 million transcriptomic reads from leaf tissue were used for analysis in this study. (Supplementary Table 2).

Genome assembly and annotation

We computationally estimated the genome size of *P. emblica* to be 579 Mbp. The final genome assembly had a size of 519 Mbp and consisted of 4,384 contigs with GC content of 33.49%, largest contig of 3.3 Mbp, and N50 of 597 Kbp (Supplementary Table 3). The heterozygosity was estimated to be 1.37%, which appears to be high given its small genome size. The 98.4% complete and 0.4% fragmented BUSCOs of this genome assembly indicated its completeness (Supplementary Table 4). Further, 96.8% of linked reads and 93.45% of nanopore long reads could be mapped on the final genome assembly. The repeats constituted 53.39% of the genome with 2,051 *de novo* repeat family sequences that were clustered into 1,803 repeat families. Among the interspersed repeats, 10.96% and 10.13% were predicted as Ty1/Copia and Gypsy/DIRS1 elements, respectively (Supplementary Table 5). A total of 815 transfer RNA (tRNA) and 141 ribosomal RNA (rRNA) genes were identified in the genome. The detailed information on 216 microRNAs (miRNA) of *P. emblica* genome is mentioned in Supplementary Table 6.

The *de novo* transcriptome assembly comprised of a total of 238,454 transcripts and these transcripts were used as EST (empirical evidence) in the MAKER pipeline. The high-confidence gene set constituted of 37,858 genes and had an 89.9% complete BUSCO score (Supplementary Table 4). Overall, ~ 96% (36,296 out of 37,858) high-confidence coding genes of *P. emblica* could be annotated using the three reference databases; NCBI-nr, Swiss-Prot, and Pfam-A (Supplementary Table 7). The functional annotations of coding genes of *P. emblica* are mentioned in Supplementary Tables 8–10.

Phylogenetic tree construction

We identified 145,194 orthogroups, of which 123 one-to-one fuzzy orthogroups were predicted from the selected 27 plant species. These selected concatenated protein sequence alignments of these fuzzy one-to-one orthogroups contained 104,108 alignment positions were used to construct a phylogenetic tree based on maximum likelihood using 26 eudicot species and *Zea mays* as an outgroup. The phylogenetic tree showed *Populus trichocarpa* and *Manihot esculenta* as the closest species to *P. emblica* as they belong to the same order Malpighiales. As per the phylogenetic tree, *P. emblica* diverged earlier (88 million years ago) than the other considered species of the order Malpighiales (Figure 1; Supplementary Text 2). Similarly, the phylogeny constructed with vitamin C biosynthesis genes followed the genome-wide phylogeny of *P. emblica* where the species like *P. emblica* and *Malpighia glabra*, *Arabidopsis thaliana* and *Carica papaya* and, *Capsicum annuum* and *Solanum lycopersicum* belonging to the same orders were sharing their common ancestral node (Figure 2). The phylogenetic tree of 50 *Phyllanthus* species using *MatK* indicated that *P. emblica* is evolutionarily closer to *P. urinaria* (Supplementary Figure 2), which has also been supported by other studies (Kathriarachchi et al., 2006; Bouman et al., 2021).

Gene family expansion and contraction analysis

Gene family expansion and contraction analysis helps to identify the gene families that have increased or decreased in

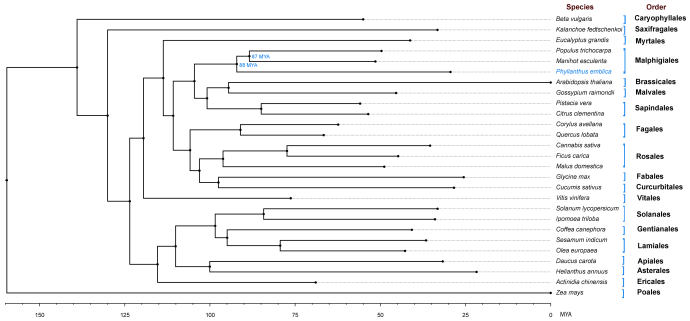


FIGURE 1
Genome-wide phylogeny of *P. emblica* with 26 other plant species Genome-wide phylogeny of *P. emblica* with 25 other eudicot species and a monocot species, *Zea mays* that was used as an outgroup. The indicated adjusted divergence time for Malpighiales order species were obtained from TimeTree database v5.0 (Kumar et al., 2022). The schematic representation method of the evolutionary time scale is similar to the previous studies (Teh et al., 2017; Xia et al., 2021).

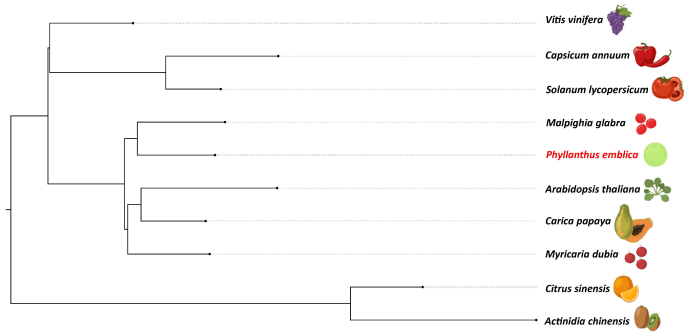


FIGURE 2
Vitamin C biosynthesis genes phylogeny of *P. emblica* with other vitamin C-rich fruits The phylogeny constructed for *P. emblica*, *Actinidia chinensis*, *Capsicum annuum*, *Carica papaya*, *Citrus sinensis*, *Myrciaria dubia*, *Malpighia glabra*, *Solanum lycopersicum*, *Vitis vinifera* and *Arabidopsis thaliana* using six genes of Smirnoff Wheeler pathway of ascorbate biosynthesis. The genes used were GDP-mannose pyrophosphorylase (*GMPP*), GDP-D-Mannose 3',5'-epimerase (*GME*), GDP-L-galactose phosphorylase (*GPP/VTC2*), L-galactose-1-phosphate phosphatase (GPP/*VTC4*), L-galactose dehydrogenase (*GalDH*) and L-galactono-1,4-lactone dehydrogenase (*GLDH*).

number in a given species. The analysis of adaptive evolution provides the clues of natural selection of specific phenotypic traits in a given species to cope with diverse environmental conditions that help in its survival. The gene expansion and contraction analysis showed a contraction of 1,048, and an expansion of 3,520 gene families in this species. Among these families, five and 42 gene families were found to be highly contracted and highly expanded, respectively. Among the 42 expanded gene families, 38 could be functionally annotated (Supplementary Table 11). The expanded gene families were majorly involved in lignin biosynthesis pathway, MAPK signaling pathway, transcription (as transcription factors), phenylpropanoid pathway, brassinosteroid biosynthesis, terpenoid biosynthesis, transportation (as transporters), plant hormone signal transduction, etc.

Identification of signatures of adaptive evolution

For evolutionary analysis, 7,864 orthogroups were obtained across 19 selected species. Among these orthogroups, 46 genes showed higher nucleotide divergence and 488 genes showed positive selection in *P. emblica*. The genes of *P. emblica* were present in ~35% (2,791 of 7,864) of the orthogroups showed unique amino acid substitutions. A total of 236 genes were identified as MSA genes in *P. emblica*. The MSA genes were found to be involved in physiological processes like plant growth, ROS regulation and detoxification, DNA damage response, immune signaling, abiotic stress response, pathogen resistance, response to hormones like ethylene, abscisic acid, gibberellin and cytokinin, and cell wall modification. The list and functional details of the MSA genes of *P. emblica* are mentioned in Supplementary Tables 12, 13.

Vitamin C biosynthesis pathway

Ascorbate, a non-enzymatic antioxidant, plays an important role in ROS detoxification and is a part of the ascorbate-glutathione pathway. *P. emblica* contains all the genes of the SW pathway similar to the other vitamin C-rich plant species like guava, kiwi, chilli pepper, etc., (Gómez-García and Ochoa-Alejo, 2016; Wang et al., 2018; Feng et al., 2021) (Figure 3). The gene structures of these genes are mentioned in Supplementary Table 14. The evolutionary analysis of six genes of SW pathway showed that *P. emblica* genes were phylogenetically closer to genes of *Malpighia glabra*, which also lies in the same order Malphigiales (Figure 2). Further, the genomic clues for presence of another pathway of ascorbate biosynthesis, i.e., galacturonate pathway were apparent from the presence of *PME*, *PL*, *PG*, *GalUR* and *GLDH* genes in *P. emblica* genome. This pathway is also proposed in tomatoes, strawberries, oranges, and grapes due to the presence of gene *GalUR*, which is the key gene of this pathway and was also present in *P. emblica* (Agius et al., 2003; Cruz-Rus et al., 2010; Cruz-Rus et al., 2011; Badejo et al., 2012; Xu et al., 2013). In addition, the genes *MIOX* and *GulLDH* involved in myo-inositol pathway were found, which supports the presence of the third pathway of ascorbate biosynthesis in *P. emblica*. However, the presence of the fourth pathway (L-gulose pathway) could not be confirmed due to lack of sufficient identification of genes involved in this pathway in *P. emblica* genome.

Among the genes of all proposed ascorbate biosynthesis pathways, six genes *HK*, *GPI*, *GMPP*, *PME*, *PL* and *PG* were identified with unique amino acid substitutions. Gene family of pectin methylesterase (*PME*) involved in galacturonate pathway of ascorbate biosynthesis was highly expanded. Other genes of the galacturonate pathway i.e., polygalacturonase (*PG*) was found with MSA, and pectin lyase (*PL*) gene had unique amino acid substitutions. Along with these biosynthesis genes, *MATE* (Multidrug And Toxic compound Extrusion) gene family, which is a vacuolar ascorbate transporter, was also highly expanded in *P. emblica* (Hoang et al., 2021). The details of ascorbate biosynthesis and regeneration pathways are described in Figure 3 and Supplementary Table 15.

The phylogenetic relationships of 16 genes involved in ascorbate biosynthesis pathway with their distant orthologs are shown in Supplementary Figures 3–18. 10 out of 16 genes of

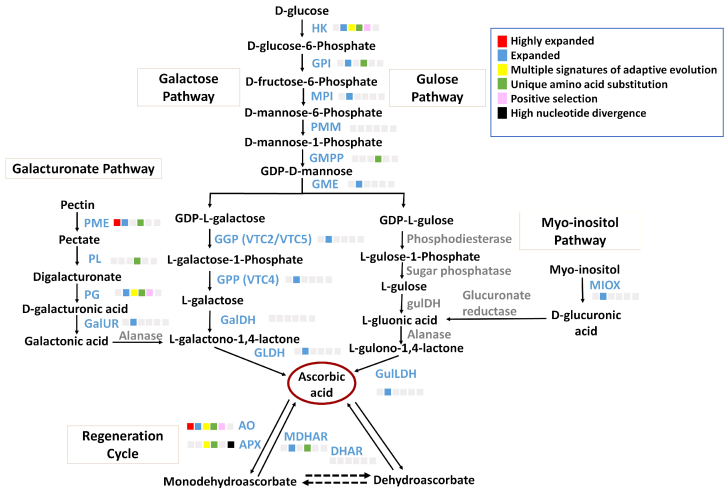


FIGURE 3
Ascorbate biosynthesis pathways The figure represents four proposed Ascorbate biosynthesis pathways i.e., Galactose pathway, Gulose pathway, Galacturonic acid pathway and Myo-inositol pathway and Ascorbate regeneration cycle. The enzymes of galactose pathway are HK, Hexokinase; GPI, Glucose 6-phosphate isomerase; MPI/PMI, Mannose 6-phosphate isomerase; PMM, Phosphomannomutase; GMPP, GDP-mannose pyrophosphorylase; GME, GDP-D-Mannose 3',5'-epimerase; GGP/VTC2, GDP-L-galactose phosphorylase; GPP/VTC4, L-galactose-1-phosphate phosphatase; GalDH, L-galactose dehydrogenase; GLDH, L-galactono-1,4-lactone dehydrogenase. The Gulose pathway includes HK, GPI, MPI, PMM, GMPP, GME, Phosphodiesterase, Sugar phosphatase, L-gulose dehydrogenase, Aldonolactonase and Gulono-1,4-lactone dehydrogenase. The Galacturonic acid pathway includes PME, Pectin methylesterase; PL, Pectin lyase; PG, Polygalacturonse; GalUR, D-galacturonate reductase; Alanase, Aldono-lactonase; and GLDH. The Myo-inositol pathway includes MIOX, Myo-inositol oxidase; Glucuronate reductase; and GulLDH, Gulono-1,4-lactone dehydrogenase. The regeneration cycle of ascorbate includes AO, Ascorbate oxidase; APX, Ascorbate peroxidase; DHAR, Dehydroascorbate; and MDHAR, Monodehydroascorbate. The enzymes in blue colour and grey colour indicate presence and absence of their genes in *P.* emblica, respectively. The colour panel in front of each enzyme indicates the evolutionary signatures like highly expanded gene family, expanded gene family, MSA, unique amino acid substitution, positive selection and high nucleotide divergence shown by its gene in *P. emblica*.

ascorbate biosynthesis pathways had distant orthologs from other Malpighiales order members and among these genes, seven were phylogenetically closer to Malpighiales members. 11 genes had distant orthologs from monocot species. *GME* and *GPI* involved in the L-galactose pathway of ascorbate biosynthesis had algal and fungal orthologs, respectively. *GMPP* had orthology with algal and animal genes. *MPI* also had fungal and animal orthologs. *GalDH* and *PMM* had protozoan and animal orthologs along with a bacterial ortholog for *GalDH*. *GalUR* involved in the L-galacturonate pathway of ascorbate biosynthesis showed a fungal ortholog.

Glutathione metabolism and ascorbate-glutathione pathway

Glutathione is another non-enzymatic antioxidant that plays a key role in different environmental stresses mainly oxidative stress and is a part of the ascorbate-glutathione pathway (Hasanuzzaman et al., 2019; Dorion et al., 2021). A total of four genes (*GPX*, *G6PD*, *gshA*, and *APX*) involved in glutathione metabolism showed multiple signatures of adaptive evolution and along with these genes, *GST* showed positive selection in *P. emblica*. The functional details of these genes are mentioned in Supplementary Table 16.

The ascorbate-glutathione pathway, also known as ascorbate regeneration cycle, plays an important role in oxidative stress response by converting the oxidised ascorbate forms to ascorbate and vice versa *via* four enzymes i.e., ascorbate oxidase (AO), ascorbate peroxidase (APX), dehydroascorbate reductase (DHAR) and monodehydroascorbate reductase (MDHAR) (Chen et al., 2003; Li et al., 2017). All these genes of ascorbate regeneration pathway were found in *P. emblica*. Among the four genes, *AO* and *APX* showed multiple signatures of adaptive evolution, and *MDHAR* showed unique amino acid substitutions. The functional details of these genes are mentioned in Supplementary Table 16. The phylogenetic relationships of these genes with their distant orthologs are shown in Supplementary Figures 19–22. Three genes involved in the ascorbate regeneration pathway had orthologs from bryophyte species (*APX* and *DHAR*) and algal species (*MDHAR*). Also, *DHAR* had a phylogenetically closer ortholog from Malpighiales order.

Flavonoid biosynthesis pathway

All 15 key genes of flavonoid biosynthesis pathway were found in *P. emblica* genome (Figure 4), and their gene structures are mentioned in Supplementary Table 14. Seven of these genes contained unique amino acid substitutions. Flavonoid 3'5'-hydroxylase (*F3'5'H*) was among the genes with MSA, and flavonol synthase (*FLS*), flavone synthase (*FS*), isoflavone synthase (*IFS*), flavavone 3-hydroxylase (*F3H*), leucoanthocyanidin reductase (*LDOX*) and *F3'5'H* gene families were highly expanded. The detailed pathway of flavonoid biosynthesis is showed in Figure 4 and Supplementary Table 17.

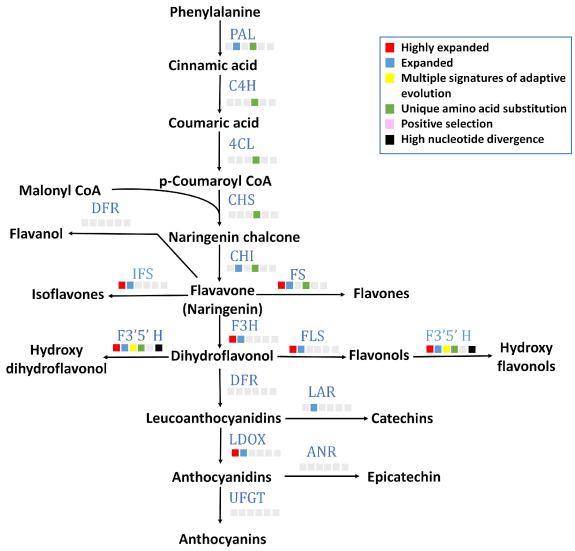


FIGURE 4
Flavonoid biosynthesis pathway The figure indicates the flavonoid biosynthesis pathway in *P. emblica*. The pathway includes enzymes PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase;4CL, 4-coumaroyl CoA ligase; CHS, chalcone synthase; CHI, chalcone isomerase; FS, flavone synthase; DFR, dihydroflavonol 4-reductase; IFS, isoflavone synthase; F3'5'H, flavonol 3'5' -hydroxylase; FLS, flavonol synthase; F3H, Flavanone 3-hydroxylase; LAR, leucoanthocyanidin reductase; LDOX, leucoanthocyanidin dioxygenase; ANR, anthocyanidin reductase; and UFGT, uridine diphosphate (UDP)-glucose flavonoid-3-O-glycosyltransferase. The blue colour of enzymes indicates presence of their genes in *P. emblica*. The colour panel in front of each enzyme indicates the evolutionary signatures like highly expanded gene family, expanded gene family, MSA, unique amino acid substitution, positive selection and high nucleotide divergence shown by its gene in *P. emblica*.

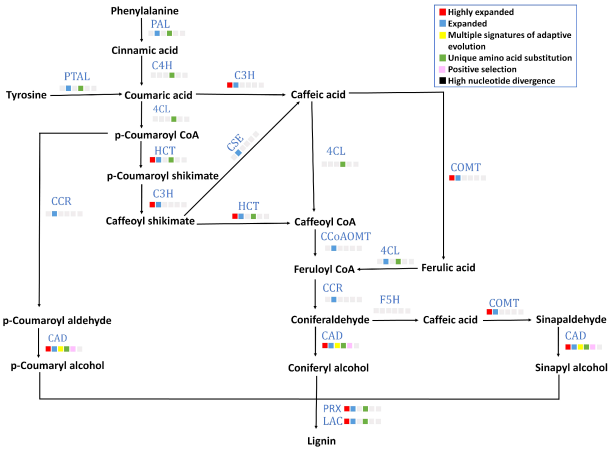


FIGURE 5
Lignin biosynthesis pathway The figure denotes lignin biosynthesis pathway in *P. emblica*. The pathway includes enzymes PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase;4CL, 4-coumaroyl CoA ligase; PTAL, phenylalanine/tyrosine ammonia-lyase; C3H, 5-O-(4-coumaroyl)-D-quinate 3'-monooxygenase; HCT, shikimate O-hydroxycinnamoyl transferase; CSE, caffeoyl shikimate esterase; COMT, caffeic acid 3-O-methyltransferase; CCoAOMT, caffeoyl-CoA O-methyltransferase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl-alcohol dehydrogenase; F5H, ferulate-5-hydroxylase; PRX, peroxidase; and LAC, laccase. The blue colour of enzymes indicates presence of their genes in *P. emblica*. The colour panel in front of each enzyme indicates the evolutionary signatures like highly expanded gene family, expanded gene family, MSA, unique amino acid substitution, positive selection and high nucleotide divergence shown by its gene in *P. emblica*.

Lignified endocarp in this stone fruit

A lignified endocarp is a trait found in stone fruits like *P. emblica*. Lignin is important in stone cell formation in a drupe fruit that provides rigidity for seed protection and dispersal. Eight out of 13 genes (gene structures are mentioned in Supplementary Table 14) of lignin biosynthesis pathway including *PAL*, *4CL*, *C4H*, *PTAL*, *HCT*, *CAD*, *POD* and *LAC* contained unique amino acid substitutions (Figure 5). Among these eight genes, *CAD* showed multiple signatures of adaptive evolution. Furthermore, six gene families *C3H*, *HCT*, *COMT*, *CAD*, *POD* and *LAC* were highly expanded. Moreover, the gene families of transcription factors *MYB* and *LBD18* that are involved in lignin biosynthesis pathway were also found to be highly expanded (Supplementary Table 18).

Plant growth, hormone and stress response

Among the 236 MSA genes, 36 genes were found to be involved in plant growth and development. These 36 genes are involved in cell division, flower development, seed development, seed germination, shoot development, cell elongation, sugar metabolism, cell wall biosynthesis, and root development, etc. (Supplementary Table 19).

Phytohormones such as auxin, cytokinin, gibberellic acid, abscisic acid (ABA), ethylene, etc. help in plant growth, development, stress tolerance, etc. throughout the plant life. 20 MSA genes were responsible for plant hormone biosynthesis, signalling and response in plants. These include genes associated with plant hormone responses related to auxin, abscisic acid, ethylene, jasmonic acid, and cytokinin (Supplementary Table 19).

Plants have mechanisms for stress tolerance against abiotic and biotic stresses. 38 out of 236 MSA genes were associated with various responses against these stresses. Among these, 30 genes were associated with responses to abiotic stresses like salt, cold, heat, drought, etc., whereas 19 genes were associated with biotic stress tolerance. These genes are involved in stress responses like stress signal transduction, secondary metabolite biosynthesis, abscisic acid biosynthesis, ROS detoxification, stress-specific gene regulation, degradation of misfolded and damaged proteins, DNA damage response, cell wall modification, disease resistance, etc. (Supplementary Table 19). Also, six MSA genes were found to be involved in DNA damage repair mechanism in plants against environmental stresses (Supplementary Table 19).

ROS regulation and detoxification

Reactive oxygen species (ROS) are metabolic by-products produced in mitochondria, plastids and peroxisomes, which can cause irreversible DNA damage resulting into cell death. In plants,

ROS not only cause harmful effects but also act as signalling molecules for plant growth and stress responses. 18 out of 236 MSA genes were associated with ROS response, regulation and detoxification. These genes are involved in porphyrin biosynthesis, ROS-induced responses, ROS scavenging, biosynthesis and protecting antioxidant enzymes, maintaining homeostasis, accumulation, and biosynthesis of antioxidants, and activation of Fe-S cluster (Supplementary Table 20).

Discussion

P. emblica or amla is a widely used medicinal plant with enormous antioxidant properties (Gul et al., 2022). To understand the genomic basis of these properties, we successfully constructed the first draft genome assembly of *P. emblica* using a hybrid sequencing approach using 10x Genomics, Illumina and ONT technologies. Despite a repetitive and highly heterozygous nature of this genome, implementation of a hybrid approach helped in constructing a high-quality genome assembly with N50 of ~0.6 Mbp and high BUSCO completeness (98.4%).

Further, this study is the first to resolve the genome-wide phylogenetic position of *P. emblica* with respect to 26 other plant species and found its early divergence from *Manihot esculenta* and *Populus trichocarpa* species of order Malpighiales, which was also confirmed by the adjusted time obtained from TimeTree database. Our phylogeny is also supported by the revised classification of order Malpighiales where Phyllanthaceae was separated as individual family from Euphorbiaceae in the Angiosperm Phylogeny Group Classification (APG III) (Group, 2009; Kawakita and Kato, 2017). The phylogeny of vitamin C biosynthesis genes also followed the genome-wide phylogeny. Phylogenetic analysis of genes involved in the L-galactose pathway of ascorbate biosynthesis showed that six out of 10 genes in *P. emblica* were closer to orthologs from other Malpighiales members. Further, key genomic insights were gained from the results of gene family expansion and contraction and from the genes with multiple signatures of adaptive evolution in *P. emblica*. The genes related to the biosynthesis of ascorbic acid, lignin and flavonoid were found to be evolutionarily selected in *P. emblica*.

Ascorbic acid is the major antioxidant in *P. emblica* and its fruit "amla" is one of the richest natural sources. A transcriptomic study of oranges had shown the attribution of different pathways of ascorbate biosynthesis in a tissue-specific as well as fruit developmental stage specific manner (Caruso et al., 2021). This could also be possible in *P. emblica* where the presence of three pathways of ascorbate biosynthesis is traced, and they could have roles in different stages and tissues. It was noted that the genes of one of the ascorbic acid biosynthesis pathways i.e., galacturonate pathway were found with MSA (*PG*), amino acid substitutions (*PME*, *PL* and *PG*) and highly expanded gene family (*PME*) in the *P. emblica* genome. The involvement of enzymes PME and PG in increased ascorbate production in tomatoes along with the role of PME in regulating ascorbate content through galacturonate pathway is shown in previous studies (Di Matteo et al., 2010; Badejo et al., 2012; Ruggieri et al., 2015; Rigano et al., 2018). Thus, it is tempting to speculate that the evolution of genes of galacturonate pathway could be associated with the high ascorbate production in *P. emblica*.

P. emblica is also rich in flavonoids that are synthesised in response to plant stress and contribute to its antioxidant property. The genes from *PAL* to *CHI* involved in the initial part of flavonoid biosynthesis pathway were found with unique amino acid substitutions, and the *F3'5'H* gene, which is previously reported to increase flavonoid accumulation, showed multiple signatures of adaptation (Wang et al., 2014; Nguyen et al., 2021). In addition to these, *FLS*, *F3'5'H*, *FS*, *LDOX*, *F3H* and *IFS* gene families were highly expanded which collectively indicates evolution of flavonoid biosynthesis genes in *P. emblica*. These genes are involved in biosynthesis of flavonoids such as isoflavones, flavones, anthocyanins and flavonols that have antioxidant properties and provide tolerance against various abiotic and biotic stresses (Verhoeyen et al., 2002; Agati et al., 2012). The evolutionary selection of these flavonoid associated genes might be responsible for the high antioxidant property and stress tolerance of *P. emblica*.

Being a stone fruit, a lignified endocarp is a trait found in the fruits of *P. emblica*, thus the evolution of lignin biosynthesis pathway was one of the major findings. Lignin is important for the stony seed coat formation in drupes that provides rigidity for its protection. The lignin biosynthesis genes were observed to be highly expanded and among the MSA genes in *P. emblica* genome, which hints towards the evolutionary significance of lignified endocarp in this stone fruit. The lignin biosynthesis gene families were also reported to be expanded in the other stone fruit genomes such as pear and *Populus*, which are economically important due to their fruit and wood, respectively, where lignin is the main content of pear's stone cells and poplar's wood (Wu et al., 2013; Li et al., 2022).

P. emblica also produces a large variety of secondary metabolites that provide tolerance against plant stresses. The expansion of gene families and MSA in genes involved in biosynthesis of various secondary metabolites and pathogen resistance against abiotic and biotic stresses were found that indicates the evolution of stress tolerance genes in this genome. Among the genes related to plant stress tolerance, the genes involved in ROS regulation and detoxification were also evolved in *P. emblica*.

Taken together, it is apparent that the adaptive evolution in genes involved in ascorbate biosynthesis, glutathione metabolism, flavonoid biosynthesis, and ROS detoxification are associated with the high antioxidant potential of *P. emblica*, which makes it a valuable herbal plant for use in traditional and modern medicine, horticulture, food and cosmetic products. Further, the high concentration of vitamin C in the amla fruit and the large production (up to 100 kg) of fruits per tree compared to other vitamin C rich fruits like *Malpighia glabra* (15-30 Kg/tree) and *Myrciaria dubia* (25-30 Kg/tree), makes it the perfect choice in switching from synthetic to natural supplementation of vitamin C (Rodrigues et al., 2001; Orwa et al., 2009; Carr and Vissers, 2013).

Further, this plant also shows high genetic diversity and easy adaptation to various climatic zones and environmental conditions (Liu et al., 2020). The availability of the first draft genome of this economically important plant is likely to help in developing improved nutraceuticals, food, cosmetics and pharmaceutical products, and for further horticultural and genomic studies.

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 accession- PRJNA947813 and BioSample accession SAMN33867225.

Author contributions

VKS conceived and coordinated the project. SM performed sample collection, DNA-RNA extraction, prepared samples for sequencing, performed long read sequencing, species identification assays, functional annotation of gene sets, and constructed all the figures. MSB and AC designed computational framework of the study, and performed all the computational analyses presented in the study. SM, MSB, AC and VKS analysed the data and interpreted the results. SM and VKS wrote the first draft of manuscript. SM, AC, MSB and VKS wrote and prepared the final manuscript. All the authors have read and approved the final version of the manuscript.

Acknowledgments

SM and AC thank Council of Scientific and Industrial Research (CSIR) for fellowship. MSB thanks Ministry of Education, Govt. of India for Prime Minister Research Fellowship (PMRF). The authors also thank the sequencing facilities at Central Instrumentation Facility, IISER Bhopal and the intramural research funds provided by IISER Bhopal.

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/fpls.2023.1210078/full#supplementary-material

References

Abeysuriya, H. I., Bulugahapitiya, V. P., and Loku Pulukkuttige, J. (2020). Total vitamin C, ascorbic acid, dehydroascorbic acid, antioxidant properties, and iron content of underutilized and commonly consumed fruits in Sri Lanka. *Int. J. Food Sci.* 2020. doi: 10.1155/2020/4783029

Abo Ghanima, M. M., Aljahdali, N., Abuljadayel, D. A., Shafi, M. E., Qadhi, A., Abd El-Hack, M. E., et al. (2023). Effects of dietary supplementation of Amla, Chicory and Leek extracts on growth performance, immunity and blood biochemical parameters of broilers. *Ital. J. Anim. Sci.* 22 (1), 24–34. doi: 10.1080/1828051X.2022.2156932

Agati, G., Azzarello, E., Pollastri, S., and Tattini, M. (2012). Flavonoids as antioxidants in plants: location and functional significance. *Plant Sci.* 196, 67–76. doi: 10.1016/j.plantsci.2012.07.014

Agius, F., González-Lamothe, R., Caballero, J. L., Muñoz-Blanco, J., Botella, M. A., and Valpuesta, V. (2003). Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nat. Biotechnol.* 21 (2), 177–181. doi: 10.1038/nbt777

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., et al. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *BioRxiv* 2021, 2021.11.18.469135. doi: 10.1101/2021.11.18.469135

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ammal, E. J., and Raghavan, R. S. (1958). "Polyploidy and vitamin C in Emblica officinalis Gaertn," in *Proceedings/Indian Academy of Sciences*. (New Delhi: Springer India) 47, 312–314.

Badejo, A. A., Wada, K., Gao, Y., Maruta, T., Sawa, Y., Shigeoka, S., et al. (2012). Translocation and the alternative D-galacturonate pathway contribute to increasing the ascorbate level in ripening tomato fruits together with the D-mannose/L-galactose pathway. *J. Exp. Bot.* 63 (1), 229–239. doi: 10.1093/jxb/err275

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28 (1), 45–48. doi: 10.1093/nar/28.1.45

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32 (suppl_1), D138–D141. doi: 10.1093/nar/gkh121

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolser, D., Staines, D. M., Pritchard, E., and Kersey, P. (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Plant Bioinf.* 1374, 115–140. doi: 10.1007/978-1-4939-3167-5_6

Bouman, R. W., Keßler, P. J., Telford, I. R., Bruhl, J. J., Strijk, J. S., Saunders, R. M., et al. (2021). Molecular phylogenetics of Phyllanthus sensu lato (Phyllanthaceae): Towards coherent monophyletic taxa. *Taxon* 70 (1), 72–98. doi: 10.1002/tax.12424

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinf.* 48 (1), 4.11. 11–14.11. doi: 10.1002/0471250953.bi0411s48

Carr, A. C., and Maggini, S. (2017). Vitamin C and immune function. *Nutrients* 9 (11), 1211. doi: 10.3390/nu9111211

Carr, A. C., and Vissers, M. C. (2013). Synthetic or food-derived vitamin C—are they equally bioavailable? *Nutrients* 5 (11), 4284–4304. doi: 10.3390/nu5114284

Caruso, P., Russo, M. P., Caruso, M., Guardo, M. D., Russo, G., Fabroni, S., et al. (2021). A transcriptional analysis of the genes involved in the ascorbic acid pathways

based on a comparison of the juice and leaves of navel and anthocyanin-rich sweet orange varieties. *Plants* 10 (7), 1291. doi: 10.3390/plants10071291

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Gene Prediction* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chavhan, A. (2017). Comparative studies on ascorbic acid content in various fruits, vegetables and leafy vegetables. *Int. J. @ Life Sci.* 5 (4), 667–671.

Chen, Z., Young, T. E., Ling, J., Chang, S.-C., and Gallie, D. R. (2003). Increasing vitamin C content of plants through enhanced ascorbate recycling. *Proc. Natl. Acad. Sci.* 100 (6), 3525–3530. doi: 10.1073/pnas.0635176100

Cruz-Rus, E., Amaya, I., Amaya, I., Sanchez-Sevilla, J. F., Botella, M. A., and Valpuesta, V. (2011). Regulation of L-ascorbic acid content in strawberry fruits. *J. Exp. Bot.* 62 (12), 4191–4201. doi: 10.1093/jxb/err122

Cruz-Rus, E., Botella, M. A., Valpuesta, V., and Gomez-Jimenez, M. C. (2010). Analysis of genes involved in L-ascorbic acid biosynthesis during growth and ripening of grape berries. *J. Plant Physiol.* 167 (9), 739–748. doi: 10.1016/j.jplph.2009.12.017

Dardick, C., and Callahan, A. M. (2014). Evolution of the fruit endocarp: molecular mechanisms underlying adaptations in seed protection and dispersal strategies. *Front. Plant Sci.* 5, 284. doi: 10.3389/fpls.2014.00284

Dasaroju, S., and Gottumukkala, K. M. (2014). Current trends in the research of Emblica officinalis (Amla): A pharmacological perspective. *Int. J. Pharm. Sci. Rev. Res.* 24 (2), 150–159.

Department of Agriculture & Farmers Welfare, M. o. A. F. W and Government of India, India (2021) *Area and Production of Horticulture crops for 2021-22 (3rd Advance Estimates)*. Available at: https://agricoop.nic.in/en/StatHortEst#gsc.tab=0.

Di Matteo, A., Sacco, A., Anacleria, M., Pezzotti, M., Delledonne, M., Ferrarini, A., et al. (2010). The ascorbic acid content of tomato fruits is associated with the expression of genes involved in pectin degradation. *BMC Plant Biol.* 10 (1), 1–11. doi: 10.1186/1471-2229-10-163

Dorion, S., Ouellet, J. C., and Rivoal, J. (2021). Glutathione metabolism in plants under stress: Beyond reactive oxygen species detoxification. *Metabolites* 11 (9), 641. doi: 10.3390/metabo11090641

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20 (1), 1–14. doi: 10.1186/s13059-019-1832-y

Fenech, M., Amaya, I., Valpuesta, V., and Botella, M. A. (2019). Vitamin C content in fruits: Biosynthesis and regulation. *Front. Plant Sci.* 9, 2006. doi: 10.3389/fpls.2018.02006

Feng, C., Feng, C., Lin, X., Liu, S., Li, Y., and Kang, M. (2021). A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (Psidium guajava). *Plant Biotechnol. J.* 19 (4), 717–730. doi: 10.1111/pbi.13498

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39 (suppl_2), W29–W37. doi: 10.1093/nar/gkr367

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.* 117 (17), 9451–9457. doi: 10.1073/pnas.1921046117

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi: 10.1093/bioinformatics/bts565

Gallie, D. R. (2013). L-ascorbic acid: a multifunctional molecule supporting plant growth and development. *Scientifica* 2013, 1–24. doi: 10.1155/2013/795964

Gantait, S., Mahanta, M., Bera, S., and Verma, S. K. (2021). Advances in biotechnology of Emblica officinalis Gaertn. syn. Phyllanthus emblica L.: a nutraceuticals-rich fruit tree with multifaceted ethnomedicinal uses. *3 Biotech.* 11, 1–25. doi: 10.1007/s13205-020-02615-5

Geethangili, M., and Ding, S.-T. (2018). A review of the phytochemistry and pharmacology of phyllanthus urinaria L. *Front. Pharmacol.* 9, 1109. doi: 10.3389/fphar.2018.01109

Gómez-García, M., and Ochoa-Alejo, N. (2016). Predominant role of the l-galactose pathway in l-ascorbic acid biosynthesis in fruits and leaves of the Capsicum annuum L. chili pepper. *Braz. J. Bot.* 39 (1), 157–168. doi: 10.1007/s40415-015-0232-0

Griffiths-Jones, S., Saini, H. K., Saini, H. K., Van Dongen, S., and Enright, A. J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36 (suppl_1), D154–D158. doi: 10.1093/nar/gkm952

Group, A. P. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* 161 (2), 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

Gul, M., Liu, Z.-W., Rabail, R., Faheem, F., Walayat, N., Nawaz, A., et al. (2022). Functional and nutraceutical significance of amla (Phyllanthus emblica L.): A review. *Antioxidants* 11 (5), 816. doi: 10.3390/antiox11050816

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

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8 (8), 1494–1512. doi: 10.1038/nprot.2013.084

Han, X., Zhang, Y., Zhang, Q., Ma, N., Liu, X., Tao, W., et al. (2022). Two haplotype-resolved, gap-free genome assemblies for Actinidia latifolia and Actinidia chinensis shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol. Plant* 16 (2), 452–470. doi: 10.1016/j.molp.2022.12.022

Hasanuzzaman, M., Bhuyan, M. B., Anee, T. I., Parvin, K., Nahar, K., Mahmud, J. A., et al. (2019). Regulation of ascorbate-glutathione pathway in mitigating oxidative damage in plants under abiotic stress. *Antioxidants* 8 (9), 384. doi: 10.3390/antiox8090384

Hoang, M. T. T., Almeida, D., Chay, S., Alcon, C., Corratge-Faillie, C., Curie, C., et al. (2021). AtDTX25, a member of the multidrug and toxic compound extrusion family, is a vacuolar ascorbate transporter that controls intracellular iron cycling in Arabidopsis. *New Phytol.* 231 (5), 1956–1967. doi: 10.1111/nph.17526

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* 34 (8), 2115–2122. doi: 10.1093/molbev/msx148

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., et al. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinf.* 19 (1), 1–10. doi: 10.1186/s12859-018-2425-6

Jaiswal, S. K., Mahajan, S., Chakraborty, A., Kumar, S., and Sharma, V. K. (2021). The genome sequence of Aloe vera reveals adaptive evolution of drought tolerance mechanisms. *Iscience* 24 (2), 102079. doi: 10.1016/j.isci.2021.102079

Jombart, T., Balloux, F., Balloux, F., and Dray, S. (2010). Adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* 26 (15), 1907–1909. doi: 10.1093/bioinformatics/btq292

Kathriarachchi, H., Samuel, R., Hoffmann, P., Mlinarec, J., Wurdack, K. J., Ralimanana, H., et al. (2006). Phylogenetics of tribe Phyllantheae (Phyllanthaceae; Euphorbiaceae sensu lato) based on nrITS and plastid matK DNA sequence data. *Am. J. Bot.* 93 (4), 637–655. doi: 10.3732/ajb.93.4.637

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

Kawakita, A., and Kato, M. (2017). Diversity of Phyllanthaceae plants. In: M Kato and A Kawakita (Eds) *Obligate Pollination Mutualism, Ecological Research Monographs*, (Tokyo: Springer), 81–116. doi: 10.1007/978-4-431-56532-1_4

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi: 10.1038/nmeth.3317

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37 (5), 540–546. doi: 10.1038/s41587-019-0072-8

Kubola, J., Siriamornpun, S., and Meeso, N. (2011). Phytochemicals, vitamin C and sugar content of Thai wild fruits. *Food Chem.* 126 (3), 972–981. doi: 10.1016/j.foodchem.2010.11.104

Kumar, A., Kumar, S., Bains, S., Vaidya, V., Singh, B., Kaur, R., et al. (2016). *De novo* transcriptome analysis revealed genes involved in flavonoid and vitamin C biosynthesis in Phyllanthus emblica (L.). *Front. Plant Sci.* 7, 1610. doi: 10.3389/fpls.2016.01610

Kumar, A., and Singh, K. (2012). Isolation of high quality RNA from Phyllanthus emblica and its evaluation by downstream applications. *Mol. Biotechnol.* 52 (3), 269–275. doi: 10.1007/s12033-011-9492-5

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., et al. (2022). TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* 39 (8), msac174. doi: 10.1093/molbev/msac174

Laetsch, D. R., and Blaxter, M. L. (2017). KinFin: software for Taxon-Aware analysis of clustered protein sequences. *G3: Genes Genomes Genet.* 7 (10), 3349–3357. doi: 10.1534/g3.117.300233

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. doi: 10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H., Huang, W., Wang, G.-L., Wang, W.-L., Cui, X., and Zhuang, J. (2017). Transcriptomic analysis of the biosynthesis, recycling, and distribution of ascorbic acid during leaf development in tea plant (Camellia sinensis (L.) O. Kuntze). *Sci. Rep.* 7 (1), 1–11. doi: 10.1038/srep46212

Li, X., Lin, Y., Jiang, Y., Wu, B., and Yu, Y. (2022). Aqueous Extract of Phyllanthus emblica L. Alleviates Functional Dyspepsia through Regulating Gastrointestinal Hormones and Gut Microbiome *In Vivo*. *Foods* 11 (10), 1491. doi: 10.3390/foods11101491

Li, C., Xing, H., Li, C., Ren, Y., Li, H., Wan, X.-Q., et al. (2022). Chromosome-scale genome assembly provides insights into the molecular mechanisms of tissue development of Populus wilsonii. *Commun. Biol.* 5 (1), 1125. doi: 10.1038/s42003-022-04106-0

Liao, G., Chen, L., He, Y., Li, X., Lv, Z., Yi, S., et al. (2021). Three metabolic pathways are responsible for the accumulation and maintenance of high AsA content in kiwifruit (Actinidia eriantha). *BMC Genomics* 22 (1), 1–11. doi: 10.1186/s12864-020-07311-5

Liu, X., Ma, H., Li, T., Li, Z., Wan, Y., Liu, X., et al. (2018). Development of novel EST-SSR markers for Phyllanthus emblica (Phyllanthaceae) and cross-amplification in two related species. *Appl. Plant Sci.* 6 (7), e01169. doi: 10.1002/aps3.1169

Liu, X., Ma, Y., Wan, Y., Li, Z., and Ma, H. (2020). Genetic diversity of Phyllanthus emblica from two different climate type areas. *Front. Plant Sci.* 11, 580812. doi: 10.3389/fpls.2020.580812

Luo, X., Zhang, B., Pan, Y., Gu, J., Tan, R., and Gong, P. (2022). Phyllanthus emblica aqueous extract retards hepatic steatosis and fibrosis in NAFLD mice in association with the reshaping of intestinal microecology. *Front. Pharmacol.* 13. doi: 10.3389/fphar.2022.893561

Mao, X., Wu, L.-F., Guo, H.-L., Chen, W.-J., Cui, Y.-P., Qi, Q., et al. (2016). The genus Phyllanthus: an ethnopharmacological, phytochemical, and pharmacological review. *Evidence-Based Complement. Altern. Med* 2016, 7584952. doi: 10.1155/2016/7584952

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27 (6), 764–770. doi: 10.1093/bioinformatics/btr011

Martini, E. (2003). How did Vasco da Gama sail for 16 weeks without developing scurvy? *Lancet* 361(9367), 1480. doi: 10.1016/S0140-6736(03)13131-5

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families.". *Bioinformatics* 36 (22-23), 5516–5518. doi: 10.1093/bioinformatics/btaa1022

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35 (suppl_2), W182–W185. doi: 10.1093/nar/gkm321

Murugesan, S., Kottekad, S., Crasta, I., Sreevathsan, S., Usharani, D., Perumal, M. K., et al. (2021). Targeting COVID-19 (SARS-CoV-2) main protease through active phytocompounds of ayurvedic medicinal plants–Emblica officinalis (Amla), Phyllanthus niruri Linn.(Bhumi Amla) and Tinospora cordifolia (Giloy)–A molecular docking and simulation study. *Comput. Biol. Med.* 136, 104683. doi: 10.1016/j.compbiomed.2021.104683

Muzaffar, K., Sofi, S. A., Makroo, H. A., Majid, D., and Dar, B. (2022). Insight about the biochemical composition, postharvest processing, therapeutic potential of Indian gooseberry (amla), and its utilization in development of functional foods—A comprehensive review. *J. Food Biochem.* 46 (11), e14132. doi: 10.1111/jfbc.14132

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31 (13), 3812–3814. doi: 10.1093/nar/gkg509

Nguse, M., Yang, Y., Fu, Z., Xu, J., Ma, L., and Bu, D. (2022). Phyllanthus emblica (Amla) fruit powder as a supplement to improve preweaning dairy calves' Health: effect on antioxidant capacity, immune response, and gut bacterial diversity. *Biology* 11 (12), 1753. doi: 10.3390/biology11121753

Nguyen, Y. T. H., Hoang, H. T. T., Mai, A. T. H., Nguyen, L. T. N., Nguyen, Q. H., Pham, N. T. T., et al. (2021). The Aconitum carmichaelii f3′ 5′ h gene overexpression increases flavonoid accumulation in transgenic tobacco plants. *Horticulturae* 7 (10), 384. doi: 10.3390/horticulturae7100384

Orwa, C., Mutua, A., Mutua, A., Kindt, R., Jamnadass, R., and Simons, A. (2009). Agroforestree Database: a tree reference and selection guide. Version 4.

Paciolla, C., Fortunato, S., Dipierro, N., Paradiso, A., De Leonardis, S., Mastropasqua, L., et al. (2019). Vitamin C in plants: from functions to biofortification. *Antioxidants* 8 (11), 519. doi: 10.3390/antiox8110519

Pandey, K., Lokhande, K. B., Lokhande, K. B., Swamy, K. V., Nagar, S., Dake, M., et al. (2021). In silico exploration of phytoconstituents from Phyllanthus emblica and Aegle marmelos as potential therapeutics against SARS-CoV-2 RdRp. *Bioinf. Biol. Insights* 15, 11779322211027403. doi: 10.1177/11779322211027403

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., and Birol, I. (2015). Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinf.* 16 (1), 1–8. doi: 10.1186/s12859-015-0663-4

Perry, B. A. (1943). Chromosome number and phylogenetic relationships in the Euphorbiaceae. *Am. J. Bot.* 30, 527–543. doi: 10.1002/j.1537-2197.1943.tb14796.x

Rahman, M. S., Sultana, S. S., Sultana, S. S., and Hassan, M. A. (2021). Variable chromosome number and ploidy level of five phyllanthus species in Bangladesh. *Cytologia* 86 (2), 143–148. doi: 10.1508/cytologia.86.143

Ranallo-Benavidez, T. R., Jaron, K. S., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11 (1), 1–10. doi: 10.1038/s41467-020-14998-3

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9 (2), 173–175. doi: 10.1038/nmeth.1818

Rigano, M. M., Lionetti, V., Raiola, A., Bellincampi, D., and Barone, A. (2018). Pectic enzymes as potential enhancers of ascorbic acid production through the D-galacturonate pathway in Solanaceae. *Plant Sci.* 266, 55–63. doi: 10.1016/j.plantsci.2017.10.013

Rodrigues, R. B., De Menezes, H. C., Cabral, L. M., Dornier, M., and Reynes, M. (2001). An Amazonian fruit with a high potential as a natural source of vitamin C: the camu-camu (Myrciaria dubia). *Fruits* 56 (5), 345–354. doi: 10.1051/fruits:2001135

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17 (2), 155–158. doi: 10.1038/s41592-019-0669-3

Ruggieri, V., Sacco, A., Calafiore, R., Frusciante, L., and Barone, A. (2015). Dissecting a QTL into candidate genes highlighted the key role of pectinesterases in regulating the ascorbic acid content in tomato fruit. *Plant Genome* 8 (2), plantgenome2014.2008.0038. doi: 10.3835/plantgenome2014.08.0038

Saini, R., Sharma, N., Oladeji, O. S., Sourirajan, A., Dev, K., and Zengin, G. (2022). Traditional uses, bioactive composition, pharmacology, and toxicology of Phyllanthus emblica fruits: A comprehensive review. *J. ethnopharmacol.* 282, 114570. doi: 10.1016/j.jep.2021.114570

Sarin, B., Verma, N., Martín, J. P., and Mohanty, A. (2014). An overview of important ethnomedicinal herbs of Phyllanthus species: present status and future prospects. *Sci. World J.* 2014. doi: 10.1155/2014/839172

Simão, 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 (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Simpson, J. T., and Durbin, R. (2012). Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* 22 (3), 549–556. doi: 10.1101/gr.126953.111

Sodeyama, T., Nishikawa, H., Harai, K., Takeshima, D., Sawa, Y., Maruta, T., et al. (2021). The d-mannose/l-galactose pathway is the dominant ascorbate biosynthetic route in the moss Physcomitrium patens. *Plant J.* 107 (6), 1724–1738. doi: 10.1111/tpj.15413

Soontornchainaksaeng, P., and Chaiyasut, K. (1999). Cytogenetic investigation of some Euphorbiaceae in Thailand. *Cytologia* 64 (3), 229–234. doi: 10.1508/cytologia.64.229

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30 (9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34 (suppl_2), W435–W439. doi: 10.1093/nar/gkl200

Teh, B. T., Lim, K., Yong, C. H., Ng, C. C.Y., Rao, S. R., Rajasegaran, V., et al. (2017). The draft genome of tropical fruit durian (Durio zibethinus). *Nat. Genet.* 49 (11), 1633–1641. doi: 10.1038/ng.3972

Van Doan, H., Lumsangkul, C., Sringarm, K., Hoseinifar, S. H., Dawood, M. A., El-Haroun, E., et al. (2022). Impacts of Amla (Phyllanthus emblica) fruit extract on growth, skin mucosal and serum immunities, and disease resistance of Nile tilapia (Oreochromis niloticus) raised under biofloc system. *Aquacult. Rep.* 22, 100953. doi: 10.1016/j.aqrep.2021.100953

Varnasseri, M., Siahpoosh, A., Hoseinynejad, K., Amini, F., Karamian, M., Yad, M. J. Y., et al. (2022). The effects of add-on therapy of Phyllanthus emblica (Amla) on laboratory confirmed COVID-19 Cases: a randomized, double-blind, controlled trial. *Complement. Therapies Med.* 65, 102808. doi: 10.1016/j.ctim.2022.102808

Verhoeyen, M., Bovy, A., Collins, G., Muir, S., Robinson, S., De Vos, C., et al. (2002). Increasing antioxidant levels in tomatoes through modification of the flavonoid biosynthetic pathway. *J. Exp. Bot.* 53 (377), 2099–2106. doi: 10.1093/jxb/erf044

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9 (11), e112963. doi: 10.1371/journal.pone.0112963

Wang, Y.-S., Xu, Y.-J., Xu, Y.-J., Gao, L.-P., Yu, O., Wang, X.-Z., He, X.-J., et al. (2014). Functional analysis of flavonoid 3′, 5′-hydroxylase from tea plant (Camellia sinensis): critical role in the accumulation of catechins. *BMC Plant Biol.* 14, 1–14. doi: 10.1186/s12870-014-0347-7

Wang, J.-P., Yu, J.-G., Yu, J.-G., Li, J., Sun, P.-C., Wang, L., Yuan, J.-Q., et al. (2018). Two likely auto-tetraploidization events shaped kiwifruit genome and contributed to establishment of the Actinidiaceae family. *Iscience* 7, 230–240. doi: 10.1016/j.isci.2018.08.003

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J., et al. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* 4 (1), s13742–13015-10076-13743. doi: 10.1186/s13742-015-0076-3

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27 (5), 757–767. doi: 10.1101/gr.214874.116

Wheeler, G., Ishikawa, T., Pornsaksit, V., and Smirnoff, N. (2015). Evolution of alternative biosynthetic pathways for vitamin C following plastid acquisition in photosynthetic eukaryotes. *Elife* 4, e06369. doi: 10.7554/eLife.06369.021

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., et al. (2013). The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res.* 23 (2), 396–408. doi: 10.1101/gr.144311.112

Xia, Z., Huang, D., Zhang, S., Wang, W., Ma, F., Wu, B., et al. (2021). Chromosome-scale genome assembly provides insights into the evolution and flavor synthesis of passion fruit (Passiflora edulis Sims). *Horticult. Res.* 8. doi: 10.1038/s41438-020-00455-1

Xiong-fang, L., Tai-qiang, L., Zheng-hong, L., You-ming, W., Xiu-xian, L., Xu, Z., et al. (2018). Transcriptome analysis for Phyllanthus emblica distributed in dry-hot valleys in Yunnan, China. 林业科学研究 31 (5), 1–8. doi: 10.13275/j.cnki.lykxyj.2018.05.001

Xu, Q., Chen, L.-L., Ruan, X., Chen, D., Zhu, A., Chen, C., et al. (2013). The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.* 45 (1), 59–66. doi: 10.1038/ng.2472

Xu, G.-C., Xu, T.-J., Zhu, R., Zhang, Y., Li, S.-Q., Wang, H.-W., et al. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience* 8 (1), giy157. doi: 10.1093/gigascience/giy157

Yan, X., Li, Q., Jing, L., Wu, S., Duan, W., Chen, Y., et al. (2022). Current advances on the phytochemical composition, pharmacologic effects, toxicology, and product development of Phyllanthi Fructus. *Front. Pharmacol.* 13. doi: 10.3389/fphar.2022.1017268

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24 (8), 1586–1591. doi: 10.1093/molbev/msm088

Yeo, S., Coombe, L., Warren, R. L., Chu, J., and Birol, I. (2018). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 34 (5), 725–731. doi: 10.1093/bioinformatics/btx675

Zhang, S. V., Zhuo, L., and Hahn, M. W. (2016). AGOUTI: improving genome assembly and annotation using transcriptome data. *GigaScience* 5 (1), s13742–13016-10136-13743. doi: 10.1186/s13742-016-0136-3

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 January 2024
DOI 10.3389/fpls.2023.1295186

OPEN ACCESS

EDITED BY
Gao Jihai,
Chengdu University of Traditional Chinese Medicine, China

REVIEWED BY
Kewang Xu,
Nanjing Forestry University, China
Sabulal Baby,
Jawaharlal Nehru Tropical Botanic Garden and Research Institute, India

*CORRESPONDENCE
Ming Lei
leiming@gxyyzwy.com

†These authors have contributed equally to this work

RECEIVED 15 September 2023
ACCEPTED 22 December 2023
PUBLISHED 12 January 2024

CITATION
Wan L, Huang Q, Li C, Yu H, Tan G, Wei S, El-Sappah AH, Sooranna S, Zhang K, Pan L, Zhang Z and Lei M (2024) Integrated metabolome and transcriptome analysis identifies candidate genes involved in triterpenoid saponin biosynthesis in leaves of *Centella asiatica* (L.) Urban.
Front. Plant Sci. 14:1295186.
doi: 10.3389/fpls.2023.1295186

Integrated metabolome and transcriptome analysis identifies candidate genes involved in triterpenoid saponin biosynthesis in leaves of *Centella asiatica* (L.) Urban

Lingyun Wan[1,2,3†], Qiulan Huang[4†], Cui Li[2,3], Haixia Yu[1,2,3], Guiyu Tan[1,2,3], Shugen Wei[1,2,3], Ahmed H. El-Sappah[4,5], Suren Sooranna[3,6], Kun Zhang[1,2,3], Limei Pan[1,2,3], Zhanjiang Zhang[1,2,3] and Ming Lei[2,3*]

[1]Guangxi Key Laboratory for High-Quality Formation and Utilization of Dao-Di Herbs, Guangxi Botanical Garden of Medicinal Plants, Nanning, China, [2]National Center for Traditional Chinese Medicine (TCM) Inheritance and Innovation, Guangxi Botanical Garden of Medicinal Plants, Nanning, China, [3]National Engineering Research Center for the Development of Southwestern Endangered Medicinal Materials, Guangxi Botanical Garden of Medicinal Plants, Nanning, China, [4]Faculty of Agriculture, Forestry and Food Engineering, Yibin University, Yibin, China, [5]Genetics Department, Faculty of Agriculture, Zagazig University, Zagazig, Egypt, [6]Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom

Centella asiatica (L.) Urban is a well-known medicinal plant which has multiple pharmacological properties. Notably, the leaves of *C. asiatica* contain large amounts of triterpenoid saponins. However, there have only been a few studies systematically elucidating the metabolic dynamics and transcriptional differences regarding triterpenoid saponin biosynthesis during the leaf development stages of *C. asiatica*. Here, we performed a comprehensive analysis of the metabolome and transcriptome to reveal the dynamic patterns of triterpenoid saponin accumulation and identified the key candidate genes associated with their biosynthesis in *C. asiatica* leaves. In this study, we found that the key precursors in the synthesis of terpenoids, including DMAPP, IPP and β-amyrin, as well as 22 triterpenes and eight triterpenoid saponins were considered as differentially accumulated metabolites. The concentrations of DMAPP, IPP and β-amyrin showed significant increases during the entire stage of leaf development. The levels of 12 triterpenes decreased only during the later stages of leaf development, but five triterpenoid saponins rapidly accumulated at the early stages, and later decreased to a constant level. Furthermore, 48 genes involved in the MVA, MEP and 2, 3-oxidosqualene biosynthetic pathways were selected following gene annotation. Then, 17 CYP450s and 26 UGTs, which are respectively responsible for backbone modifications, were used for phylogenetic-tree construction and time-specific expression analysis. From these data, by integrating metabolomics and transcriptomics analyses, we identified *CaHDR1* and *CaIDI2* as the candidate genes associated with DMAPP and IPP synthesis, respectively, and *CaβAS1* as the one regulating β-amyrin synthesis. Two genes from the CYP716 family were confirmed as *CaCYP716A83* and *CaCYP716C11*. We also selected two UGT73 families as

candidate genes, associated with glycosylation of the terpenoid backbone at C-3 in *C. asiatica*. These findings will pave the way for further research on the molecular mechanisms associated with triterpenoid saponin biosynthesis in *C. asiatica*.

KEYWORDS

***Centella asiatica*, transcriptome, metabolome, triterpenoid saponin, candidate gene**

Introduction

Centella asiatica (L.) Urban, a stoloniferous perennial herb of the Umbelliferae family found in tropical and sub-tropical areas (Nguyen et al., 2019), is a well-known medicinal plant used in China, India and Southeast Asia (Bhavna and Jyoti, 2011; Weng et al., 2012). It has been registered in different Pharmacopoeia of the world such as China, India and Europe (Brinkhaus et al., 2000; Kalita et al., 2015). Experimental and clinical investigations show that *C. asiatica* is useful in the treatment of varicose veins, eczema, wound healing disturbances, intellectual disability, ulcer, lupus and striae gravidarum (Liu et al., 2008; Prakash et al., 2017). This is attributed to an enormous amount of triterpenoid saponins, mainly asiaticoside, madecassoside, centelloside and sceffoleoside, which are collectively known as centelloids in *C. asiatica* (Matsuda et al., 2001; James and Dubery, 2009). Therefore, the triterpenoid saponins of *C. asiatica* are vital resources for their important and unique medicinal values. Identification of the relevant biosynthetic pathways will thus be helpful for our further understanding and utilization of these triterpenoid saponins. However, most of the previous studies on these compounds have only focused on their properties, bioactivities and pharmacology, and only a few genes involved in the triterpenoid saponin biosynthesis in *C. asiatica* have been identified, such as HMG-CoA reductase (Kalita et al., 2018), farnesyl diphosphate synthase (Kim et al., 2005a), squalene synthase (Kim et al., 2005b), oxidosqualene cyclase (Kim et al., 2005c), CYP450s (Fukushima et al., 2011; Kim et al., 2018) and UGTs (Kim et al., 2017; Han et al., 2022) and so on. Therefore, the biosynthetic pathways involved have yet to be fully characterized and need to be further investigated at the molecular level.

It is reported that there are more than 300 types of triterpenoid carbon skeletons (Xu et al., 2004; Thimmappa et al., 2014), whose chemical structures are formed from triterpene aglycones (sapogenins) and one or more sugar groups connected by glycosidic bonds (Phillips et al., 2006). Further evidences have supported the notion that there are similar routes of the triterpenoid saponin biosynthesis in different plants (Augustin et al., 2011; Cárdenas et al., 2019), and the core of triterpenoid saponin biosynthetic pathway is well-understood (Xu et al., 2022).

The whole triterpenoid saponin pathway involves three major stages: an initial stage, followed by terpenoid backbone construction and the modification stage (Sawai and Saito, 2011; Zhao et al., 2022). In the first stage, isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP), are formed from two independent pathways: IPP biosynthesis by initially condensing three units of acetyl-CoA in the cytosolic mevalonate (MVA) pathway (Huang et al., 2014), and DMAPP biosynthesis starting from the condensation of pyruvate and phosphoglyceraldehyde in the plastidial methylerythritol phosphate (MEP) pathway (Luo et al., 2016). IPP can be partially isomerized into DMAPP by IPP isomerase (IDI) (Wen et al., 2017). When the terpenoid backbone is constructed, farnesyl diphosphate synthase (FPS) condenses and catalyzes IPP and DMAPP into farnesyl pyrophosphate (FPP). Subsequently, two FPP molecules form squalene by the action of squalene synthase (SS). Under the action of squalene epoxidase (SE), epoxidation of squalene at the second and third carbon positions yields 2,3-oxidosqualene, which is a direct precursor to the construction of the triterpenoid backbones (Haralampidis et al., 2002). Next, oxidosqualene cyclases (OSCs) catalyze the cyclitization of 2,3-oxidosqualene into various triterpenoid backbones, which is thus a critical step of the triterpenoid biosynthesis pathway (Aharoni et al., 2006). During the modification stage, the triterpenoid backbones are chemically modified by cytochrome P450 monooxygenases (CYP450s) through a series of hydroxylation/oxidation reactions and then these are changed by UDP-glycosyltransferases (UGTs) through glycosylation reactions, all of which finally result in the formation of various triterpenoid saponins (Seki et al., 2015; Yang et al., 2018).

Transcriptomics is an effective technical means of mining the key regulatory genes associated with the active ingredients of biosynthetic pathways (Kuwahara et al., 2019). In an earlier study, an analysis of transcriptome sequences of leaves of *C. asiatica* by Illumina (Solexa sequencing technology) was conducted, and this contributed to the mining of the targeted genes related to triterpenoid saponin biosynthesis (Sangwan et al., 2013). The development of long-read sequencing technologies and assembly pipelines has allowed the completion of the whole genome assembly of *C. asiatica* at the chromosome-scale (Pootakham et al., 2021). These complete, accurate and contiguous representative genome sequences can provide valuable references to further understand the triterpenoid saponin biosynthesis in *C. asiatica*. Additionally, the metabolomics database can then be used to provide information

regarding the accumulation of various secondary metabolites during the biosynthetic pathways of bioactive ingredient formation (Saito and Matsuda, 2010), and it can also provide reference information for gene mining.

The comprehensive analysis of transcriptomics and metabolomics has been widely applied to delineate the relationships between the dynamic changes of secondary metabolites and the differential expression of corresponding genes during the bioactive ingredients biosynthetic pathways in numerous medicinal plants, such as *Panax notoginseng* (Wei et al., 2018), *Sapindus mukorossi* (Xu et al., 2022), *Platycodon grandiflorus* (Chang et al., 2022), *Dendrobium huoshanense* (Yuan et al., 2022) and *Carthamus tinctorius* (Wang et al., 2021). The medicinal properties of *C. asiatica* are largely attributed to the centelloids that present predominantly in the leaves of *C. asiatica* (Sangwan et al., 2013). However, there have been few studies systematically elucidating the metabolic dynamics and transcriptional differences regarding the triterpenoid saponin biosynthesis during the leaf developmental stages of *C. asiatica* based on triterpenoid saponin metabolism and transcriptomics data.

In this study, we identified the key candidate genes and metabolites by analyzing the expression patterns of differentially expressed genes (DEGs) by transcriptomics and the differentially accumulated metabolites (DAMs) via metabolomics in *C. asiatica* leaves at different growth stages. This study will pave the way for future research on the molecular mechanisms associated with triterpenoid saponin biosynthesis in *C. asiatica*.

Materials and methods

Plant materials

The plants of *C. asiatica* were artificially cultivated in the greenhouse at the Guangxi Botanical Garden of Medicinal Plants. The plants were cultivated using an illumination intensity of 300 μmol m^{-2} s^{-1}, a relative humidity of 60 ± 2% at 26 ± 2°C with a light/dark cycle of 16/8 h. The leaves of *C. asiatica* were collected from four growth stages and representative examples are shown in Supplementary Figure S1. The samples were at the 10, 20, 30 and 40 day leaf-ages with three replicates at each stage and were designated as JXC1-1, JXC1-2, JXC1-3, JXC2-1, JXC2-2, JXC2-3, JXC3-1, JXC3-2, JXC1-3, JXC4-1, JXC4-2 and JXC4-3, respectively. The leaf age was determined by attaching a label to each leaf indicating the date when the leaf first emerged. The number of leaves collected at each stage was subjected to their size, and all the samples were immediately frozen by using liquid nitrogen and then stored at −80°C for subsequent extraction of RNA and metabolites.

Metabolomics determination and analysis

A wide targeted metabolomics analysis was conducted on the samples from each leaf growth stage to explore the dynamic patterns of metabolites during the developmental stages of *C. asiatica* leaves. This analysis was carried out by MetWare Biotechnology Co., Ltd., Wuhan, China.

Initially, the leaf samples were freeze-dried by using a vacuum freeze-dryer (Scientz-100F, Ningbo, China), and then zirconia beads were added and the mixture were grinded in a mixer mill (MM 400, Retsch, Haan, Deutschland) for 1.5 min, followed by 70% methanol extraction. The suspension was centrifuged (at 2,000 rpm for 3 min) (Anpel, Shanghai, China), and then filtered through a 0.22 μm econofilter. The filtrates were analyzed using a UPLC-ESI-MS/MS system (UPLC, SHIMADZU Nexera X2, MS, Applied Biosystems 4500 Q TRAP). An Agilent SB-C18 column (1.8 μm, 2.1 mm×100 mm) (California, USA) was used for separation at 40° C. Mobile phases of solvents A and B were used and these were composed of pure water with 0.1% formic acid and acetonitrile with 0.1% formic acid, respectively. The gradient conditions were as follows: Phase B started with 5%, followed by 95% at 0 ~ 10 min, and maintained at 95% for another 1.0 min, and finally 5% at 11.1 ~ 14 min,. For each sample, the injection volume was 4 μL at a flow rate of 0.35 mL/min. The effluent was collected and assessed after connection to electrospray ionization (ESI)-triple quadrupole-linear ion trap (QTRAP)-MS.

The operating parameters for ESI were as follows: the source temperature was 550°C and the ion spray voltage (IS) was set to 5,500 V (positive ion mode)/-4,500 V (negative ion mode). The ion source gases I and II as well as the curtain gas were set at 50, 60 and 25 psi, respectively, and the collision-activated dissociation was set to high. Ten and 100 μmol/L polypropylene glycol solutions were used to tune the instrument and to perform mass calibration in the triple quadrupole (QQQ) as well as the linear ion trap modes, respectively. Scans of the QQQ were acquired in the form of multiple reaction monitoring (MRM) experiments with nitrogen as the collision gas and this was set to medium. The declustering potential (DP) and collision energy (CE) for each MRM transition were performed after further optimizations. A specific set of MRM transitions was used to monitor each period based on the metabolites found to be eluted within that period.

To assess the metabolite diversity between and within groups, all the metabolic data were subjected to multivariate statistical analysis, including unsupervised principal component (PCA) and orthogonal partial least squares discriminant analyses (OPLS-DA). After evaluation of these results, the metabolites with $|\mathrm{Log_2}\ (FC)| \geq 1$, p-value < 0.05, and variable importance in the project (VIP) ≥ 1 were identified as DAMs.

RNA extraction, Illumina sequencing and determination of DEGs

RNA-seq was carried out by MetWare Biotechnology Co., Ltd. (Wuhan, China). Total RNA was extracted from the four stages of *C. asiatica* leaves and its quality was assessed by running on 1% agarose gels and using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using the Qubit® RNA Assay Kit in Qubit®2.0 Flurometer (Life Technologies, CA, USA). RNA integrity was determined using the

RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). One μg of RNA from each sample was used for sample preparation. cDNA libraries were generated using the NEBNext®UltraTM RNA Library Prep Kit purchased from Illumina® (NEB, USA) by following the manufacturer's instructions. Raw sequence reads with nucleotides with uncertain base information as well as those with low-quality sequence reads were filtered using FASTP v 0.19.3 to produce clean reads. These clean reads were then mapped to the whole genome of *C. asiatica* (unpublished data) to acquire the genes using HISAT v2.1.0.

The expression levels of the genes were calculated after normalization to the number of fragments per kilobase of transcript per million fragments (FPKM). Genes were further analyzed using functional annotations and pathway analyses to seven public databases: KEGG, KOG, GO, Pfam, Tremble, NR and Swiss-Prot. The levels of DEGs were analyzed using the original count data from each sample which was obtained by their expression quantification with DESeq2 software. Those genes with Benjamini-Hochberg-adjusted p value < 0.05 and $|\log_2 (FC)| > 1$ were identified as DEGs.

Gene identification and phylogenetic analysis of the CYP450 and UGT families

Two methods were used for gene identification of the CYP450 and UGT families in the *C. asiatica* genome. Firstly, the protein sequences from the chromosome-scale genome of *C. asiatica* (unpublished data) were used to construct a local protein database. The known protein sequences of *Arabidopsis thaliana* for 242 CYP450s and 117 UGTs were collected from the Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org/). These were used to search the local protein database of *C. asiatica* by using BLASTP, which is a basic local environment alignment search tool for proteins. The cut-off E-value used was 10^{-10}. Subsequently, the Hidden-Markov Model (HMM) files for P450.hmm (PF00067) and UGT.hmm (PF00201) which were downloaded from the Pfam database (https://pfam.xfam.org/), were used to search the local protein database of *C. asiatica* to find all predicted CYP450 and UGT family members via the HMMER 3.0 (https://www.hmmer.org/) program with a cut-off E-value of 10^{-10}. After integrating the results of the above two methods, the remaining sequences were assumed as the candidate CYP450 and UGT family members of *C. asiatica*, after removing the redundant sequences. The candidate family members were further verified via PfamScan (e-value = 0.001, https://www.ebi.ac.uk/Tools/pfa/pfamscan/).

Phylogenetic analysis was performed to align the full-length protein sequences of CYP450 and UGT from *Arabidopsis* and CYP450 and UGT from *C. asiatica* by using the MEGA software (version 7.0) with default parameters, respectively. A phylogenetic tree was constructed by using the MEGA software with the 1,000-replicate neighbor-joining (NJ) method and the results were illustrated by using the interactive tree of life online software (https://itol.embl.de/help.cgi).

Integration of transcriptome and metabolome data by WGCNA and Pearson correlation analysis

To better outline relationships of the data between the genes and related metabolites involved in triterpenoid saponin biosynthesis, weighted gene coexpression network analysis (WGCNA) and Pearson correlation analysis were conducted with the default parameters in the free online data analysis, Metware Cloud Platform (https://cloud.metware.cn). The soft threshold, the network heatmap plot of genes, the module heatmap and the trait-related modules were constructed by using WGCNA. The correlation results obtained from Pearson correlation analysis were mapped in a correlation cluster heatmap.

Real-time quantitative PCR analysis

Eighteen DEGs associated with the triterpenoid saponin biosynthesis were selected for real-time quantitative PCR (RT-qPCR) analysis relative to the reference gene, *CaGAPDH*. cDNA was synthesized using the SuperScript® III Reverse Transcriptase kit by following the manufacturer's protocols. RT-qPCR was performed on a Bio-Rad CFX96 RT-qPCR platform (Bio-Rad Laboratories, Hercules, CA, USA) using SYBR® Premix Ex Taq™ kit (Takara, Dalian, China) according to the manufacturer's protocols. The relative expression levels of DEGs were calculated using the $2^{-\Delta\Delta CT}$ method and then normalized against *CaGAPDH*. The primers used are listed in Supplementary Table S1.

Results

Metabolic profiles of *C. asiatica* leaves at four different growth stages

In order to evaluate the metabolite variations in the leaves at different growth stages, a wide targeted metabolomics analysis was performed using samples collected at four time-points during leaf development. A clustering heatmap analysis was performed from the whole metabolite database, which showed that the three replicates used in this study had a high similarity (Supplementary Figure S2A). PCA analysis also showed a distinct separation in leaf samples from the four different stages, indicating that these metabolites presented a dynamic change pattern with the leaf growth (Supplementary Figure S2B). These results indicated that the metabolite detection methods used in this study were reliable.

A total of 802 metabolites were obtained, mainly containing terpenoids (68), flavonoids (84), alkaloids (44), amino acids and their derivatives (96), phenolic acids (135), nucleotides and their derivatives (49), lignans (18), coumarins (26), tannins (1), organic acids (68) and lipids (108) (Supplementary Table S2). To further understand the complexity of the different metabolites at each leaf growth stage, these metabolites were divided into four clusters by using a k-mean clustering method. Terpenoids were mainly found

in clusters 3 and 2, with the highest accumulation of metabolites belonging to these clusters being from JXC1 and JXC2, respectively (Figure 1).

The 263 metabolites were mapped to the reference pathways according to KEGG annotations, and these were annotated to 96 pathways into three categories. A total of 91 pathways were associated with metabolism, and three and two pathways were involved with environmental information and genetic information processing, respectively. For the "metabolism" term, the top priority was "global and overview maps" (228 metabolites), followed by "amino acid metabolism" (89 metabolites), "carbohydrate metabolism" (69 metabolites) and "biosynthesis of other secondary metabolites" (67 metabolites) (Supplementary Figure S3A).

Analysis of the metabolites involved in the biosynthesis of triterpenoid saponins in the *C. asiatica* leaves at four different growth stages

The key precursors in the biosynthesis of terpenoids, including DMAPP (Wmjp001948), IPP (Wmjn000963) and β-amyrin (Wmjn007463), were detected and they showed differential accumulation in the *C. asiatica* leaves at the four different growth stages (Supplementary Table S3). Gradual increased trends were observed in their contents from JXC1 to JXC4 with a peak at JXC4 (Supplementary Table S3).

In the downstream metabolites of triterpenoid saponin biosynthesis, a total of 32 triterpenes and 14 triterpenoid saponins were tentatively detected (Supplementary Table S3). Hierarchical cluster analysis (HCA) based on the relative levels of triterpenes (Supplementary Figure S4A) and triterpenoid saponins (Supplementary Figure S4B) in the *C. asiatica* leaves at the four growth stages showed that the levels of most of these compounds were markedly changed during leaf development, and this further confirmed the PCA results. Additionally, a total of 22 triterpenes were found to be differentially accumulated (Supplementary Table S4), which were divided into 7 groups based on their accumulation levels. The 12 triterpenes in cluster A had their levels significantly decreased from stage JXC3, and three of these compounds in cluster B showed high accumulations at stage JXC3 (Supplementary Figure S5A). A total of eight triterpenoid saponins showed differential accumulation (Supplementary Table S4), which were divided into four groups. In cluster A, the five triterpenoid saponins showed high accumulation at stage JXC1 and then tapered off from stages JXC2 to JXC4. For cluster B, the accumulation of one triterpenoid saponin showed a downtrend trend, with the lowest level seen at stage JXC3, but then it increased by stage JXC4. Cluster C was similar to cluster B, where lower levels were seen at stages JXC2 and JXC3. The content of triterpenoid saponins in cluster D appeared to increase continuously during leaf growth (Supplementary Figure S5B).

Transcriptome profiles of *C. asiatica* leaves at four different growth stages

In this study, in order to explore gene expression patterns during the four different stages of leaf development, and to identify genes participating in the biosynthetic pathways of triterpenoid saponins at the transcriptional level, 12 cDNA libraries, in batches of three, were prepared and analyzed. The 12 leaf samples generated 56.59 million high-quality clean reads which constituted 84.89 GB of cDNA sequence data. The Q20 and Q30 values ranged from 98.09 to 98.46% and 94.03 to 95.00%, respectively, and the GC content ranged from 43.50 to 44.92%. The reads in each library were aligned to the *C. asiatica* reference genome with an average match ratio of 82.71% (Table 1).

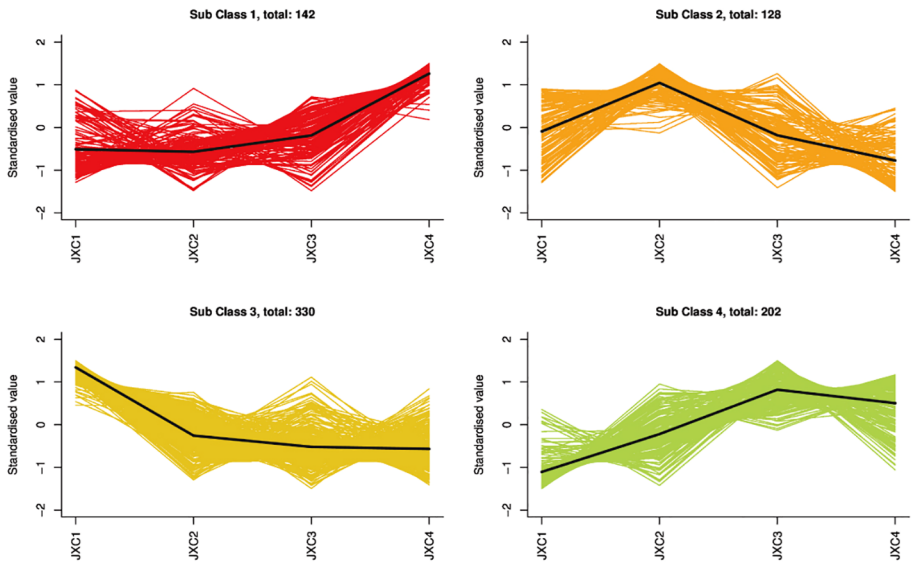


FIGURE 1
The k-means analysis of metabolites.

TABLE 1 Summary statistics of the RNA-seq results.

Sample	Raw Reads	Clean Reads	Clean Base(G)	Q20 (%)	Q30 (%)	GC Content(%)	Mapped Reads	Unique mapped
JXC1-1	44495124	43049490	6.46	98.13	94.18	44.07	85.65%	83.70%
JXC1-2	50987344	49197288	7.38	98.09	94.03	43.99	84.86%	82.99%
JXC1-3	45120822	44072352	6.61	98.17	94.24	44.02	84.39%	82.42%
JXC2-1	54430828	51992362	7.8	98.34	94.75	44.18	84.83%	82.25%
JXC2-2	48504278	46719306	7.01	98.13	94.14	43.73	78.94%	76.89%
JXC2-3	49317248	47190664	7.08	98.29	94.55	44.02	80.63%	78.35%
JXC3-1	49342226	47165158	7.07	98.46	95	43.84	82.78%	80.64%
JXC3-2	53689036	51481462	7.72	98.35	94.67	43.75	83.03%	80.71%
JXC3-3	50530798	49049278	7.36	98.32	94.61	43.92	83.96%	81.60%
JXC4-1	48555886	46378168	6.96	98.34	94.65	43.63	81.49%	79.45%
JXC4-2	46447748	44272732	6.64	98.34	94.77	43.69	83.52%	81.45%
JXC4-3	47132200	45318910	6.8	98.31	94.52	43.5	78.42%	76.61%

Identification of DEGs

A total of 22,594 genes expressed were obtained in at least one sample with FPKM > 0 (Supplementary Table S5). The PCA results showed most of the samples at the same stages were highly correlated, indicating that the gene expression patterns were also highly reproducible (Supplementary Figure S6). Six comparative groups were constructed to demonstrate the differentiation at the transcriptional level at different growth stages of *C. asiatica* leaves: JXC1_vs_JXC2, JXC1_vs_JXC3, JXC1_vs_JXC4, JXC2_vs_JXC3, JXC2_vs_JXC4 and JXC3_vs_JXC4, and a total of 3889 DEGs were identified (Supplementary Table S6). Among these DEGs, the numbers of up- and down-regulated DEGs for each comparison group are shown in Table 2. There were more down-regulated DEGs than up-regulated ones in all the compared groups except for JXC3_vs_JXC4. Compared with JXC4, JXC3 contained a smaller number of DEGs, with 253 and 238 up- and down-regulated DEGs, respectively. The highest number of DEGs were identified in the JXC1_vs_JXC4 comparison, which showed 945 and 1,960 up- and down-regulated DEGs, respectively. In summary, there were significant differences in the number of DEGs presented in the six comparative groups.

GO enrichment and KEGG pathway analysis

The functions of DEGs were annotated by GO enrichment analysis, which were enriched within three main ontologies: biological process (BP), cellular component (CC) and molecular function (MF). Finally, the 3,389 DEGs of all samples were annotated to 3,623 GO terms (Supplementary Figure S7). Among them, the BP category contained 2,151 GO terms (59.37%), of which the major terms were the cellular (GO: 0009987), metabolic (GO: 0008152), single-organism (GO: 0044699) and biological processes (GO: 0050789)). In addition, some of these DEGs also responded to stimuli (GO: 0050896), biological regulation (GO: 0065007) and localization (GO: 0051179). The CC category contained 387 GO terms, including the membrane (GO: 0016020 and 0044425), cell (GO: 0005623 and 0044464) and organelle (GO: 0043226). The MF category contained 1,085 GO terms (29.95%), including mainly catalytic (GO: 0003824), binding (GO: 0005488), transporter (GO: 0005215) and nucleic acid binding transcription factor activities (GO: 0001071).

Moreover, the DEGs were also mapped to the KEGG pathways by using the KEGG database. The 3,389 DEGs in all samples were annotated to 126 KEGG pathways into five categories. There were three, four, 18, two and 99 pathways for cellular processes, environmental information processing, genetic information processing, organismal systems and metabolism, respectively (Supplementary Figure S3B). The main enriched pathways associated with secondary metabolism in each comparison group in our study were: biosynthesis of secondary metabolites (ko01110), metabolic pathways (ko01100), sesquiterpenoid and triterpenoid biosynthesis (ko00940), terpenoid backbone biosynthesis (ko00900), phenylpropanoid biosynthesis (ko00940) and flavonoid biosynthesis (ko00941). These enriched pathways

TABLE 2 The counts of differentially expressed genes.

Group	Total	Down	Up
JXC1_vs_JXC2	1800	1236	564
JXC1_vs_JXC3	2281	1555	726
JXC1_vs_JXC4	2905	1960	945
JXC2_vs_JXC3	498	338	160
JXC2_vs_JXC4	1372	829	543
JXC3_vs_JXC4	491	238	253

provided valuable information for mining candidate genes associated with triterpenoid saponin biosynthesis.

DEGs expression patterns associated with triterpenoid backbone biosynthesis

To further explore the molecular mechanism of triterpene and triterpenoid saponin biosynthesis in the *C. asiatica* leaves, the expression levels of genes associated with their biosynthesis were analyzed. From the transcriptomics analysis, we found 48 genes were associated with the terpenoid backbone biosynthetic pathways, all of which could be classified into the MVA, MEP and 2, 3-oxidosqualene pathways (Figure 2).

The KEGG pathway analysis showed that a total of 12 genes encoding six enzymes were associated with the MVA pathway: two acetyl-CoA C-acetyltransferase (AACTs), two 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthetases (HMGSs), four

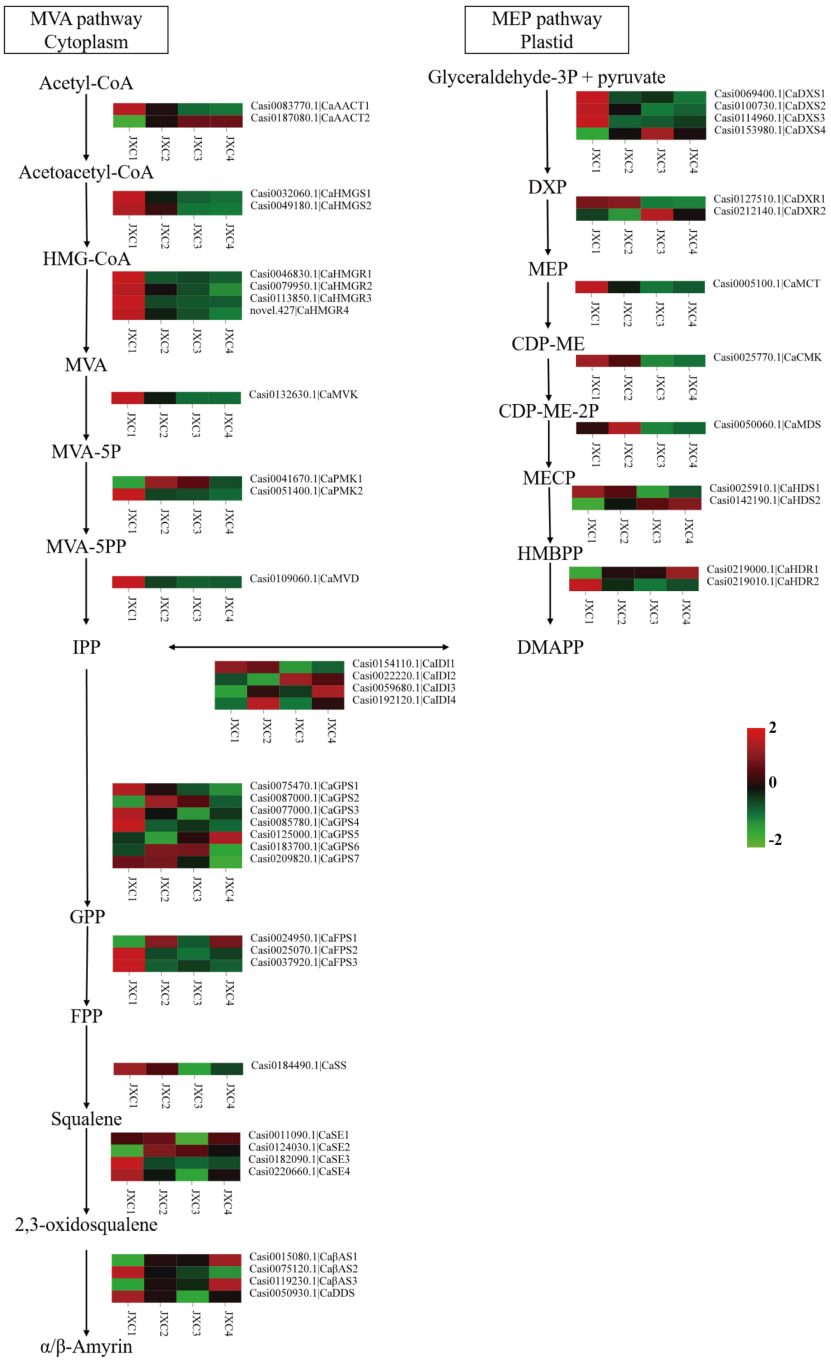


FIGURE 2
The proposed pathways for triterpenoid saponin biosynthesis in *C. asiatica*. The expression levels of the gene encoding enzymes that catalyze each step of the triterpenoid saponin biosynthesis pathway are shown.

HMG-CoA reductases (HMGRs), one mevalonate kinase (MVK), two phosphomevalonate kinases (PMKs) and one metholvalic-5-diphosphate decarboxylase (MVD). The MEP pathway, starting with pyruvate and glyceraldehyde3-phosphate, consisted of 17 genes encoding eight enzymes: four 1-deoxy-D-xylulose-5-phosphate (DXP) synthetases (DXSs), two DXP reductoisomerases (DXRs), one 4-cytidine diphosphate-2-C-methyl-D-erythritol (CDP-ME) synthetase (MCT), 1 CDP-ME kinase (CMK), one 2-C-methyl-D-erythritol-2,4-cyclic phosphate synthetase (MDS), two (E)-4-hydroxy-3-methyl-2-butenyl-pyrophosphate (HMBPP) synthetases (HDSs), two HMBPP reductases (HDRs) and four isopentenyl pyrophosphate (IPP) isomerases (IDIs). The expression levels were highest in JXC1 for most of these genes, except for *CaAACT2*, *CaMVK* and *CaPMK1* which were involved in the MVP pathway and for *CaDXS4*, *CaDXR2*, *CaMDS*, *CaHDS2*, *CaHDR1* and 3 *CaIDIs* involved in the MEP pathway.

Additionally, 19 genes encoding six enzymes were involved in carbocyclic biosynthesis: seven geranyl pyrophosphate (GPP) synthetases (GPSs), three farnesyl pyrophosphate(FPP) synthetase (FPSs), one squalene synthase (SS), four squalene epoxidases (SEs), three β-amyrin synthases (βASs) and one dammarenediol-II synthase (DDS). However, these genes had variable expression levels at the different developmental stages.

Identification and phylogenetic analysis of CYP450s and UGTs associated with backbone modifications

Through our genome-wide analysis, we identified 231 CYP450s in the *C. asiatica*. By using the CYP450 reference database, they were classified into 9 clades, 41 families, with 48.05 and 51.95% being the A- and non-A-types of CYP450s, respectively (Supplementary Table S7). For further analysis of their functional roles and evolutionary relationships in *C. asiatica*, 231 CYP450s along with 242 CYP450s from *A. thaliana* were employed in generating phylogenetic trees for the A- (Supplementary Figure 8A) and non-A-types (Supplementary Figure 8B) CYP450s, respectively. All the A-types of CYP450s were represented by the CYP71 clans. These consisted of 111 genes belonging to 18 families, which were CYP71, 73, 75-79, 81-P84, 89, 93, 98, 701, 703, 706 and 712. The members of non-A-type CYP450s were divided into eight clans according to the phylogenetic tree, including five single-family clans (CYP51, 74, 710, 97 and 711) and three multifamily clans (CYP72, 85 and 86). CYP72, 85 and 86 were the largest three clans, containing 48, 37 and 27 CYP450s, respectively.

By using genome-wide analysis, we identified 140 UGTs in the *C. asiatica*. We aligned the *C. asiatica* UGT amino acid sequences with the 117 UGTs obtained from *Arabidopsis* and was able to construct a phylogenetic tree (Supplementary Table S8). We ultimately classified these UGTs into 15 groups and 20 families, among which three with amino acid sequences less than 40% identities to any representative sequences could not be classified. The UGT85 family of Group G was the largest family with 19 genes, and the next largest group was the UGT73 family of Group D with 18 genes (Supplementary Figure 9).

It has been reported that the members of the CYP85, CYP72 and CYP71 families are primary CYP450s which are associated with triterpenoid saponin biosynthesis in dicotyledonous plants. We identified five *CYP716s* of CYP85 family genes (one *CYP716C* gene, two *CYP716E* genes and two *CYP716A* genes), six CYP72 family genes (two *CYP72A* genes and four *CYP714E* genes) and six *CYP71D* genes by using the transcriptomic data obtained from *C. asiatica*. According to the results of the phylogenetic analysis (Figure 3A), Casi0095280.1, Casi0032820.1 and Casi0108660.1

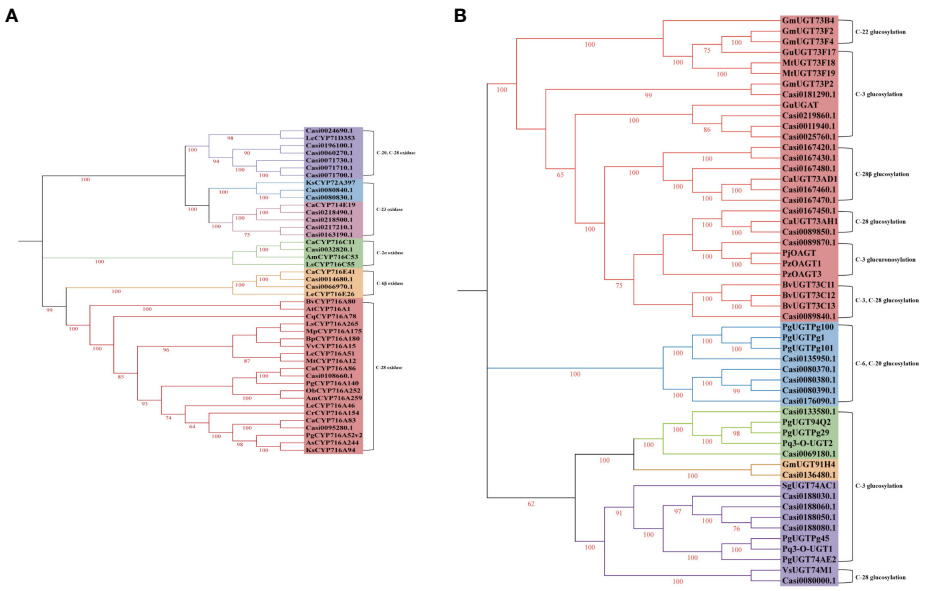


FIGURE 3
The phylogenetic tree of previously characterized triterpenoid biosynthesis CYP450s **(A)**, UGTs **(B)** and those CYP450s **(A)** and UGTs **(B)** isolated from *C. asiatica* in this study.

10.3389/fpls.2023.1295186

had 100.00, 100.00 and 87.62% identities to CaCYP716A83, CaCYP716C11 and CaCYP716A86, respectively. Casi0014680.1 and Casi0066970.1 exhibited 100.00 and 69.58% sequence identities with CaCYP716E41, respectively. Casi0218490.1, Casi0218500.1, Casi0163190.1 and Casi0217210.1 exhibited 94.79, 79.23, 42.27 and 40.32% sequence identities with CaCYP714E19, respectively. Casi0080830.1 and Casi0080840.1 exhibited 64.63 and 63.85% sequence identities with KsCYP72A397, respectively.

In recent studies, several UGT family members (UGT71, UGT73, UGT74, UGT91 and UGT94) were shown to be associated with triterpenoid saponin biosynthesis. Here, we screened five UGT71, 13 UGT73, five UGT74, one UGT91 and two UGT94 family gene members via the transcriptomic data obtained from *C. asiatica*. According to the results of the phylogenetic analysis (Figure 3B), Casi0167460.1, Casi0167470.1, Casi0167480.1, Casi0167430.1 and Casi0167420.1 exhibited 99.60, 85.11, 72.95, 52.92 and 47.29% sequence identities with CaUGT73AD1, respectively. Casi0089850.1 and Casi0167450.1 exhibited 100.00 and 64.88% sequence identities with CaUGT73AH1, respectively.

Construction of correlated gene co-expression networks by using WGCNA

To identify genes related to triterpenoid saponin biosynthesis, low-expressed genes (FPKM value ≤ 0) were filtered out, and a total of 15,808 genes obtained were used for the construction of gene co-expression networks by WGCNA. Our results showed that a soft threshold power of 16 was the lowest power with a scale-free topology model fit index of 0.90 and relatively high mean connectivity (Figure 4A). Gene clustering dendrogram was constructed and detected 13 gene co-expression modules labeled with different colors. Genes in the same module indicated that their expression patterns were similar (Figures 4B, C).

To determine the correlation relationship between gene co-expression modules and 33 metabolites involved in the triterpenoid saponin biosynthesis, the heatmap of module-traits relationships was constructed (Figure 4D). We observed that the 'turquoise' module had the highest correlation with three key precursors in the biosynthesis of terpenoids, DMAPP (Wmjp001948) ($r = 0.82$, $p = 0.0011$), IPP (Wmjn000963) ($r = 0.91$, $p < 0.001$) and β-amyrin (Wmjn007463) ($r = 0.95$, $p < 0.001$). The 'brown' module was associated with 18 triterpenes and had the highest correlation with 13 triterpenes ($r > 0.83$, $p < 0.001$). The 'blue' module was associated with 7 triterpenoid saponins, of which 5 had the highest correlation ($r > 0.84$, $p < 0.001$).

Pearson correlation analysis between the concentrations of triterpenoid saponins and expressions of their biosynthesis-related genes

The expression levels of the 68 candidate genes and the metabolite concentrations involved in triterpenoid saponin biosynthesis were analyzed by Pearson correlation (Figure 5). The

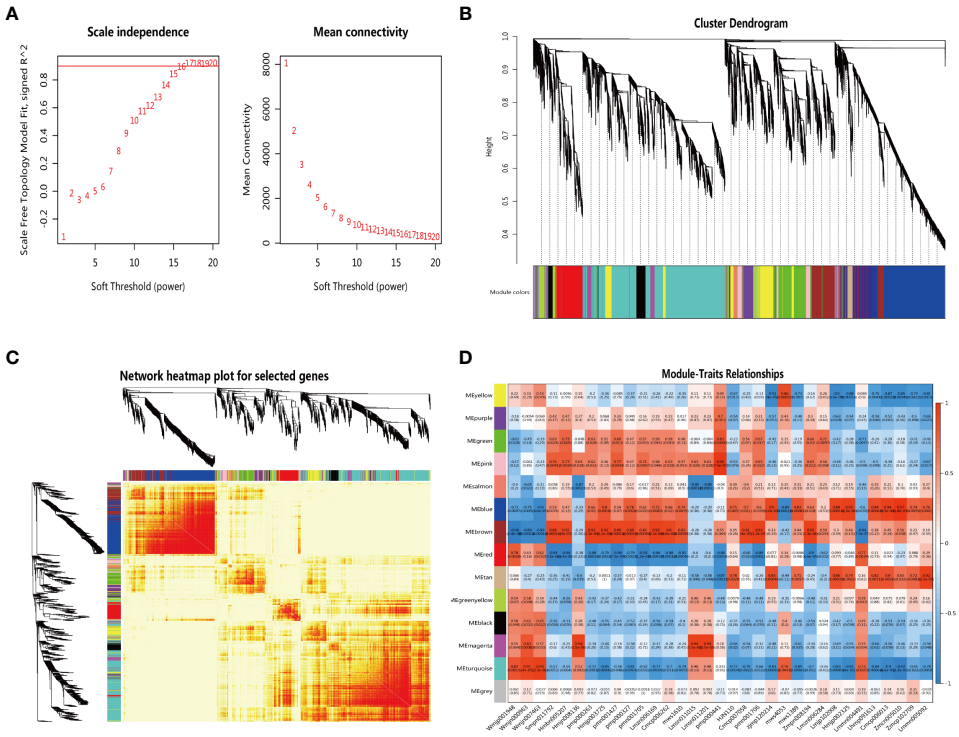


FIGURE 4
The network topology analysis for different soft threshold powers **(A)** The gene clustering dendrogram with adjacency-based dissimilarity, together with assigned module colors **(B)** The construction of module heatmap **(C)** The construction of the relationships between modules and traits **(D)**.

results indicated that DMAPP (Wmjp001948), IPP (Wmjn000963) and β-amyrin (Wmjn007463), the key precursors in the biosynthesis of terpenoids, were significantly and positively correlated with *HDR1* (Casi0219000.1) (r = 0.78, p = 0.0026), *IDI2* (Casi0022220.1) (r = 0.79, p = 0.0023) and *CaβAS1* (Casi0015080.1) (r = 0.91, p < 0.001). Our results showed that Casi0095280.1 and Casi0108660.1 had 100.00 and 87.62% identities to CaCYP716A83 and CaCYP716A86, respectively, and these CYPs catalyzed α-amyrin/β-amyrin to ursolic acid/oleanolic acid. Ursolic acid (mws4053) was positively correlated with Casi0095280.1. In addition, corosolic (mws1610) and maslinic acids (Zmpn008194) showed positive correlations with Casi0032820.1, which had 100.00% identity to CaCYP716C11. Finally, we found that saikosaponin L (Zmcp102709) and medicagenic acid-3-O-glucosyl-(1,6)-glucosyl-(1,3)-glucoside (Lmmn004491) were significantly and positively correlated with Casi0219860.1 (r = 0.85, p = 0.0004) and Casi0011940.1 (r = 0.93, p < 0.001), respectively, both of which were members of UGT73 family and had high degrees of identity with GuUGA.

Verification of DEGs related to triterpenoid saponins by qRT-PCR measurements

To further verify the transcriptomics data obtained, the relative expression levels of 18 genes associated with triterpene and triterpenoid saponin biosynthesis in *C. asiatica* were selected for analysis by RT-qPCR. These included six genes (*CaAACT1*, *CaHMGS1*, *CaHMGR2*, *CaMVK*, *CaPMK2* and *CaMVD*) involved in the MVP pathway, four genes (*CaDXS1*, *CaDXR1*, *CaHDS1* and

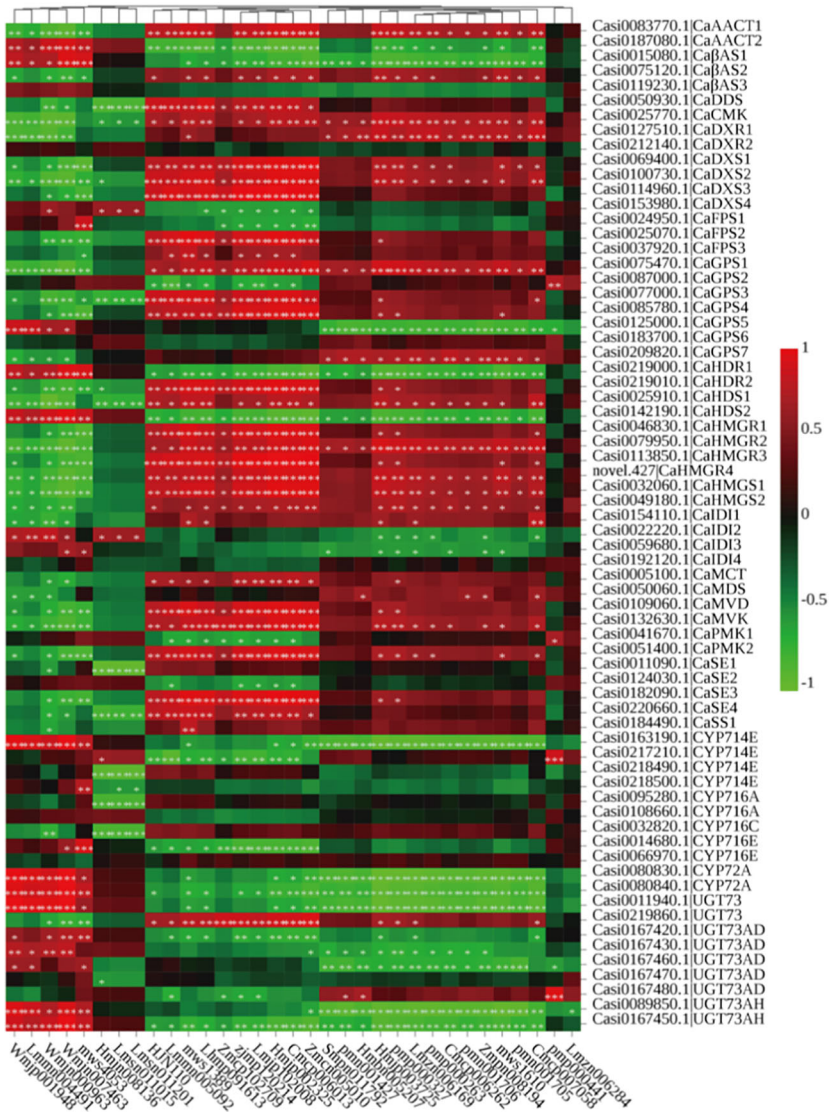


FIGURE 5
The Pearson correlation analysis between genes and metabolites in the triterpenoid saponin pathway. Significant ones p < 0.05 marked with *, p < 0.01 marked with **, p < 0.001 marked with ***.

CaHDR2) involved in the MEP pathway, four genes (*CaGPS1*, *CaFPS2*, *CaSE3* and *CaβAS1*) involved in the terpenoid backbone construction stage and four genes (*CaCYP714E*, *CaCYP716A*, *CaCYP716C* and *CaUGT73*) involved in terpenoid backbone modification. The results showed that the majority of these genes exhibited consistent expression patterns with the transcriptomics data (Figure 6), which supported the analysis we performed.

Discussion

The triterpenes and triterpenoid saponins of *C. asiatica* have been regarded as its essential bioactive compounds due to their important medicinal values in the treatment of various chronic disorders (Prakash et al., 2017; Puttarak et al., 2017). However, previous studies regarding these compounds from *C. asiatica* have

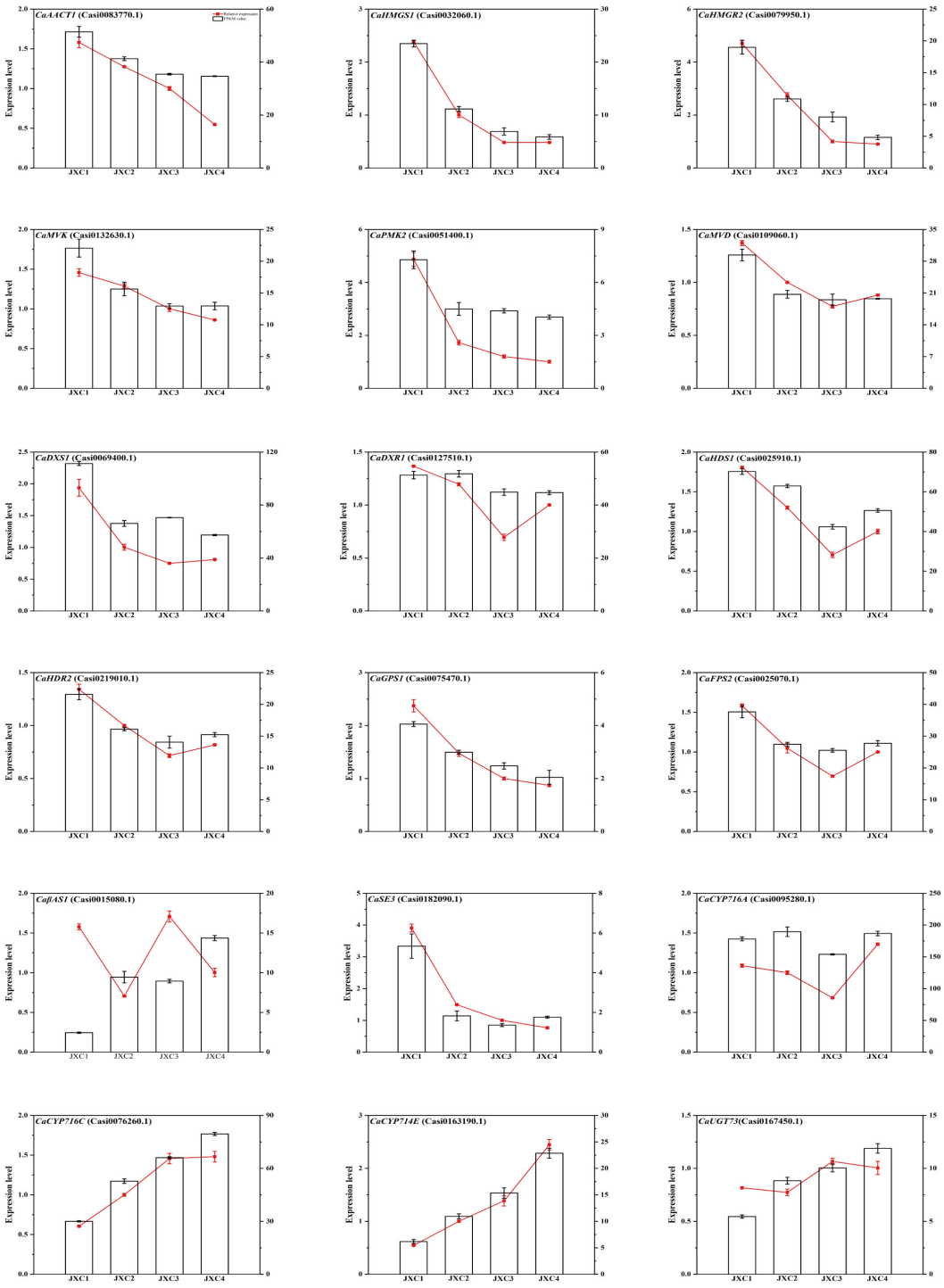


FIGURE 6
The qRT-PCR analysis of the potential genes involved in triterpenoid saponin biosynthesis.

only focused on their chemical structures, bioactivity assessments and cultivation (Prasad et al., 2019). Relatively, few studies regarding the accumulation of metabolites as well as the molecular mechanisms of their biosynthesis in *C. asiatica* have been conducted. Therefore, it is important to assess the biosynthetic pathways of triterpenes and triterpenoid saponins in *C. asiatica* in order to achieve the optimal production of these bioactive compounds from the plants. The majority of the triterpenes and triterpenoid saponins are known to be predominantly accumulated in the leaves of *C. asiatica* (Sangwan et al., 2013). Here, we identified the metabolites and genes of *C. asiatica* leaves at different growth stages based on transcriptomics and metabolomics data, in order to gain novel insights into the accumulation and biosynthesis of triterpenes and triterpenoid saponins.

Structurally, the triterpenes and triterpenoid saponins in the *C. asiatica* belong to the terpenoid group, and they have a similar precursor synthesis pathway to those of the other terpenoids. These are thought to be derived through a cross-talk between the MVA and MEP pathways (Patel et al., 2016; Li et al., 2022). We identified 12 genes in the MVA pathway and these involved six key enzymes and 17 genes in the MEP pathway as well as the involvement of eight other key enzymes from our transcriptomics analysis. These genes were further analyzed in combination with the metabolomics analysis, and the expression levels of *CaHDR1* (Casi0219000.1) and *CaIDI2* (Casi0022220.1) were found to be consistent with the concentrations of DMAPP (Wmjn000963) and IPP (Wmjp001948), respectively.

HDR, the limiting enzyme in the MEP pathway, catalyzes HMBPP into a mixture of 5:1 IPP and DMAPP (Adam et al., 2002). Overexpression of the *HDR* genes in *Arabidopsis* has been shown to increase the production of IPP and DMAPP (Botella-Pavía et al., 2010). Both IPP and DMAPP are the essential precursors at regulatory branching points in the biosynthesis of terpenoids (Sun et al., 2010), but the relatively unreactive IPP can be interconverted to its highly electrophilic isomer, DMAPP, by the catalysis of IDI. This compound provides both the electrophilic primer substrate and the condensation substrate for terpenoid biosynthesis (Chen et al., 2012). Therefore, the function of IDI is considered to be the critical first step that controls the overall biosynthesis of all terpenoids (Ramos-Valdivia et al., 1997; Zhu et al., 2014). Previous studies reported that overexpression of the *IDI* gene generated an accumulation of several related terpenoids by acting downstream in the MEP pathway (Kajiwara et al., 1997; Sun et al., 1998). All these reports suggested that IDI may be a target enzyme for regulating terpenoid biosynthesis (Chen et al., 2012). Our results also showed that the up-regulation of *CaHDR1* and *CaIDI2* promoted the accumulation of IPP and DMAPP, suggesting that these genes play pivotal roles in the regulation of the biosynthesis of these compounds in *C. asiatica*.

The OSCs are considered to play key roles in the formation of the different triterpenoid skeletons (Basyuni et al., 2007), because they are the key initial steps for catalyzing 2,3-oxidosqualene cyclization into diverse forms of the triterpenoid skeletons (Mao et al., 2023). This is supported by experiments whereby the enzyme activities were altered and differential effects on triterpene synthesis were observed (Morita et al., 2000). The down-regulation of the expression of the *β-AS* gene in *Panax ginseng* hairy roots has also been shown to reduce β-amyrin and oleanane-type ginsenoside (Zhao et al., 2015). In addition, α-AS from *Pisum sativum*, a multifunctional triterpene synthase, generated α- and β-amyrins by heterologous expression in yeast (Iturbe-Ormaetxe et al., 2003). Similarly, in our study, the expression level of *CaβAS1* (Casi0015080.1) was consistent with the concentration of β-amyrin (Wmjn007463). In addition, the sequence of *CaβAS1* had a high identity (79.79%) with the known *C. asiatica βAS*, which indicated that *CaβAS1* may catalyze 2,3-oxidosqualene to β-amyrin in *C. asiatica*.

The triterpenoid skeletons result from carboxyl modification by the CYP450s and then the saccharide side chains are normally introduced by the UGTs leading to the biosynthesis of different triterpene saponins (Xu et al., 2022). The members of CYP450s associated with triterpenoid saponin biosynthesis are mainly the CYP85, CYP72 and CYP71 families. For instance, CYP716s of the CYP85 family catalyze the successive three-step oxidation of α-amyrin, lupel and β-amyrin at C-28 (Shang and Huang, 2019). CYP72s (CYP72A and CYP714E) catalyze the hydroxylation of oleanolic acid/ursolic acid at C-23 (Han et al., 2018) and the members of CYP71D subfamily are involved in the successive three-step oxidation of 20-hydroxylupeol in *Lotus japonicus* to produce 20-hydroxybetulinic acid (Krokida et al., 2013). It has been shown that CYP716As catalyze the successive three-step oxidation of the α-amyrin/β-amyrin/lupeol backbone to form hydroxyl, aldehyde and carboxyl moieties at C-28 (Xu et al., 2022). In this study, two CYP716As, Casi0095280.1 and Casi0108660.1, showed 100.00 and 87.62% identities to CaCYP716A83 and CaCYP716A86, respectively, and these can catalyze α-amyrin/β-amyrin to ursolic acid/oleanolic acid (Miettinen et al., 2017). Our results also showed that ursolic acid (mws4053) was more abundant at the 20 and 40 day stages of leaf development, which was similar to the expression pattern of Casi0095280.1. In addition, Casi0032820.1 had 100.00% identity to CaCYP716C11, which is known to catalyze the hydroxylation of ursolic acid/oleanolic acid/6β-hydroxy-oleanolic acid at C-2α to corosolic acid/maslinic acid/6β-hydroxy-maslinic acid (Miettinen et al., 2017). In our study, corosolic (mws1610) and maslinic acids (Zmpn008194) from our metabolomics data also showed they accumulated at the 10 and 20-day stages of leaf development, which was consistent with the up-regulation of Casi0032820.1 at JXC1 and JXC2.

It was reported that CaCYP716E41 can catalyze the hydroxylation of ursolic acid/oleanolic acid/maslinic acid at C-6β to 6β-hydroxy-ursolic acid/6β-hydroxy-oleanolic acid/6β-hydroxy-maslinic acid in *C. asiatica* (Miettinen et al., 2017). Although the products of CaCYP716E41 were not found in our metabolomics data, Casi0014680.1 and Casi0066970.1 exhibited 100.00% and 69.58% sequence identities with CaCYP716E41, respectively, according to our results of the phylogenetic analysis. Therefore, Casi0066970.1 may also be a candidate gene involved in triterpene biosynthesis in *C. asiatica*.

We also found that four candidate CYP72As (Casi0218490.1, Casi0218500.1, Casi0163190.1 and Casi0217210.1) exhibited 94.79, 79.23, 42.27 and 40.32% sequence identities with CaCYP714E19, respectively. In addtion, Casi0080830.1 and Casi0080840.1

exhibited 64.63 and 63.85% sequence identities with KsCYP72A397, respectively, and it catalyzes the hydroxylation of ursolic acid/oleanolic acid at C-23. This suggested that these CYP72As may catalyze ursolic acid/oleanolic acid to hederagenin/23-hydroxy ursolic acid in *C. asiatica* (Han et al., 2018; Liu et al., 2019). We also found an accumulation of hederagenin (Hmbn005207). However, the change in expression pattern of this compound was significantly and negatively correlated with Casi0163190.1, Casi0080830.1 and Casi0080830.1. A previous study had shown similar results, where it was found that CqCYP716A15 levels were negatively correlated with the triterpene content of *Chenopodium quinoa* (Zhao et al., 2022). These data inferred that CYP72As had multiple regulatory roles during the biosynthesis of triterpenes in *C. asiatica.*

The UGT73 clan is the best candidate family for triterpenoid saponin biosynthesis. We identified five genes (*Casi0167460.1*, *Casi0167470.1*, *Casi0167480.1*, *Casi0167430.1* and *Casi0167420.1*) belonging to the UGT73 clan which showed high degrees of identity with CaUGT73AD1. This enzyme transfers glucuronic acid at C-28β of asiatic acid/madecassic acid to form 28β-Glc-asiatic acid/28β-Glc-madecassic acid (De Costa et al., 2017), suggesting that they catalyze the glucuronosylation of the C28β-hydroxyl group for the biosynthesis of triterpenoid saponins in *C. asiatica*. In addition, the other two members of the UGT73 clan (Casi0089850.1 and Casi0167450.1) exhibited 100.00 and 64.88% sequence identities with CaUGT73AH1, respectively. This enzyme is known to transfer glucuronic acid at C-28 of asiatic acid/UDP-Glc to form 28-Glc-asiatic acid (Kim et al., 2017). This indicates that these genes may catalyze the glucuronosylation of the C28-hydroxyl group for the biosynthesis of triterpenoid saponins in *C. asiatica*. However, the expression levels of the above seven UGT73 candidates were not significantly different during the different stages of leaf development. We found 10 triterpenoid saponins glycosylated at the C-28 position from metabolomics data, most of which also showed no significant difference during the different developmental stages. Therefore, these candidate UGT73 genes and their key intermediates involved in the triterpenoid saponin biosynthetic pathway of *C. asiatica* need to be studied further.

Our study also found two candidate UGT73 DEGs (*Casi0219860.1* and *Casi0011940.1*) which catalyzed glucuronic acid at the C-3 position and showed high degrees of identity to GuUGA (Xu et al., 2016). The expression pattern of *Casi0219860.1*, which showed a high expression level and accumulation at the 10 day-stage of leaf development, was consistent with the level found in saikosaponin L (Zmcp102709). In addition, the expression pattern of *Casi0011940.1*, which showed high expression levels and accumulation at the 30 and 40 day-stages of leaf development, was consistent with the level of medicagenic acid-3-O-glucosyl-(1,6)-glucosyl-(1,3)-glucoside (Lmmn004491). This result infers that Casi0219860.1 and Casi0011940.1 are most likely to be involved in the biosynthesis of triterpenoid saponins that are glycosylated at the C-3 position in *C. asiatica*.

Conclusions

Despite the multiple pharmacological properties associated with triterpenes and triterpenoid saponins found in *C. asiatica* leaves, relatively few systematic studies have been undertaken to elucidate the biosynthesis of these compounds during their leaf developmental stages. In our investigation, we used the transcriptome and metabolome data to perform a comprehensive analysis to reveal the dynamic patterns of triterpenoid saponin accumulation and the identity of the key candidate genes associated with their biosynthesis in *C. asiatica* leaves. We found that the decrease in levels of the majority of triterpenes occurred only at the late stages of leaf development. However, the majority of triterpenoid saponins rapidly accumulated at early stages of leaf development, and then decreased but remained constant during the later periods. The levels of triterpenes and triterpenoid saponins showed different dynamic patterns at different leaf growth stages, indicating that the whole triterpenoid saponin biosynthetic pathway is relatively complex. Furthermore, 48 genes involved in the synthesis of the terpenoid backbone as well as 17 CYP450s and 26 UGTs genes were found to be associated with the backbone modifications of these compounds. By integrating metabolomics and transcriptomics analyses, we identified five candidate genes as having crucial roles in the biosynthesis of triterpenoid saponins in *C. asiatica*. The relative expression levels of these key genes were further verified by qRT-PCR. Overall, these results will provide a valuable reference for further research on the molecular mechanisms associated with triterpenoid saponin biosynthesis in *C. asiatica* as well as other medicinal plants.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repositories and accession number can be found at the Genome Sequence Archive (GSA), accession number: CRA012378 by using following link: https://ngdc.cncb.ac.cn/search/?dbId=gsa&q=CRA012378.

Author contributions

LW: Conceptualization, Data curation, Funding acquisition, Investigation, Writing – original draft. QH: Writing – original draft. CL: Investigation, Writing – review & editing, Funding acquisition. HY: Investigation, Writing – review & editing. GT: Data curation, Funding acquisition, Investigation, Writing – review & editing. SW: Data curation, Writing – review & editing. AE-S: Writing – review & editing. SS: Writing – review & editing. KZ: Investigation, Writing – review & editing. LP: Investigation, Writing – review & editing. ZZ: Data curation, Funding acquisition, Writing – review & editing. ML: Conceptualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Natural Science Foundation of Guangxi Province (2023GXNSFAA026330 and 2020GXNSFAA159151), Guangxi Science and Technology Base and Special Talents (Guike AD22035026), Scientific Research Funding Project of Guangxi Botanical Garden of Medicinal Plants (GYJ202008), Innovative Team for Traditional Chinese Medicinal Materials Quality of Guangxi (GZKJ2305) and the Key Laboratory Construction Program of Guangxi Health commission (ZJC2020003).

Acknowledgments

We are very grateful to the kind administration of Guangxi Botanical Garden of Medicinal Plants, Nanning, China for providing us such a prestigious and well-equipped platform for research and development.

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/fpls.2023.1295186/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The *C. asiatica* leaves at four different growth stages.

SUPPLEMENTARY FIGURE 2
A correlation heatmap of all the samples **(A)**; The metabolomics sample replicates resulting from PCA analysis **(B)**.

SUPPLEMENTARY FIGURE 3
The KEGG pathway annotation of the metabolites detected **(A)**; The KEGG pathway annotations of the differentially expressed genes **(B)**.

SUPPLEMENTARY FIGURE 4
The hierarchical clustering for the dynamic change pattern of triterpenes **(A)** and triterpenoid saponins **(B)** in the *C. asiatica* at four leaf growth stages.

SUPPLEMENTARY FIGURE 5
Upper: temporal expression patterns of the differentially accumulated triterpenes, Lower: a heatmap of triterpenes accumulation levels in different clusters **(A)**; Upper: temporal expression patterns of the differentially accumulated triterpenoid saponins, Lower: a heatmap of the triterpenoid saponins accumulation levels in different clusters **(B)**.

SUPPLEMENTARY FIGURE 6
The transcriptomic sample replicates resulting from PCA analysis.

SUPPLEMENTARY FIGURE 7
The GO classification of the differentially expressed genes.

SUPPLEMENTARY FIGURE 8
The phylogenetic tree of A-type **(A)** and non-A-type **(B)** CYP450s from *C. asiatica* and Arabidopsis. The representative CYP450 family members from *C. asiatica* and Arabidopsis are marked with red circles and blue triangles, respectively.

SUPPLEMENTARY FIGURE 9
The phylogenetic tree of UGTs from *C. asiatica* and Arabidopsis. The representative UGT family members from *C. asiatica* and Arabidopsis are marked with red circles and blue triangles, respectively.

References

Adam, P., Hecht, S., Eisenreich, W., Kaiser, J., Gräwert, T., Arigoni, D., et al. (2002). Biosynthesis of terpenes: studies on 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase. *Proc. Natl. Acad. Sci.* 99, 12108–12113. doi: 10.1073/pnas.182412599

Aharoni, A., Jongsma, M. A., Kim, T. Y., Ri, M. B., Giri, A. P., Verstappen, F. W. A., et al. (2006). Metabolic engineering of terpenoid biosynthesis in plants. *Phytochem. Rev.* 5, 49–58. doi: 10.1007/s11101-005-3747-3

Augustin, J. M., Kuzina, V., Andersen, S. B., and Bak, S. (2011). Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochem.* 72, 435–457. doi: 10.1016/j.phytochem.2011.01.015

Basyuni, M., Oku, H., Tsujimoto, E., Kinjo, K., Baba, S., and Takara, K. (2007). Triterpene synthases from the Okinawan mangrove tribe, Rhizophoraceae. *FEBS J.* 274, 5028–5042. doi: 10.1111/j.1742-4658.2007.06025.x

Bhavna, D., and Jyoti, K. (2011). *Centella asiatica*: the elixir of life. *Int. J. Res. Ayurveda. Pharm.* 2, 431–438.

Botella-Pavía, P., Besumbes, Ó., Phillips, M. A., Carretero-Paulet, L., Boronat, A., and Rodríguez-Concepción, M. (2010). Regulation of carotenoid biosynthesis in plants: evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. *Plant J.* 40, 188–199. doi: 10.1111/j.1365-313X.2004.02198.x

Brinkhaus, B., Lindner, M., Schuppan, D., and Hahn, E. G. (2000). Chemical, pharmacological and clinical profile of the East Asian medical plant *Centella aslatica*. *Phytomedicine* 7, 427–448. doi: 10.1016/s0944-7113(00)80065-3

Cárdenas, P. D., Almeida, A., and Bak, S. (2019). Evolution of structural diversity of triterpenoids. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01523

Chang, A., Pei, W., Li, S., Wang, T., Song, H., Kang, T., et al. (2022). Integrated metabolomic and transcriptomic analysis reveals variation in the metabolites and genes of *Platycodon grandiflorus* roots from different regions. *Phytochem. Anal.* 33, 982–994. doi: 10.1002/pca.3153

Chen, R., Harada, Y., Bamba, T., Nakazawa, Y., and Gyokusen, K. (2012). Overexpression of an isopentenyl diphosphate isomerase gene to enhance trans-polyisoprene production in *Eucommia ulmoides* Oliver. *BMC Biotechnol.* 12, 78. doi: 10.1186/1472-6750-12-78

De Costa, F., Barber, C. J., Kim, Y., Reed, D. W., Zhang, H., Fett-Neto, A. G., et al. (2017). Molecular cloning of an ester-forming triterpenoid: UDP-glucose 28-O-glucosyltransferase involved in saponin biosynthesis from the medicinal plant *Centella asiatica*. *Plant Sci.* 262, 9–17. doi: 10.1016/j.plantsci.2017.05.009

Fukushima, E. O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., et al. (2011). CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant Cell Physiol.* 52 (12), 2050–2061. doi: 10.1093/pcp/pcr146

Han, J. Y., Chun, J.-H., Oh, S. A., Park, S.-B., Hwang, H.-S., Lee, H., et al. (2018). Transcriptomic analysis of *Kalopanax septemlobus* and characterization of KsBAS, CYP716A94 and CYP72A397 genes involved in hederagenin saponin biosynthesis. *Plant Cell Physiol.* 59, 319–330. doi: 10.1093/pcp/pcx188

Han, X. Y., Zhao, J. Y., Chang, X. C., Li, Q. Y., Deng, Z. X., and Yu, Y. (2022). Revisiting the transcriptome data of *Centella asiatica* identified an ester-forming triterpenoid: UDP-glucose 28-O-glucosyltransferase. *Tetrahedron* 129, 133136. doi: 10.1016/j.tet.2022.133136

Haralampidis, K., Trojanowska, M., and Osbourn, A. E. (2002). Biosynthesis of triterpenoid saponins in plants. *Adv. Biochem. Eng. Biotechnol.* 75, 31–49. doi: 10.1007/3-540-44604-4_2

Huang, Z. J., Liu, D. Q., Ge, F., and Chen, C. Y. (2014). Advances in studies on key post-modification enzymes in triterpenoid saponins biosynthesis. *Acta Botanica Boreali-Occidentalia Sin.* 34, 2137–2144. doi: 10.7606/j.issn.1000.2014.10.2137

Iturbe-Ormaetxe, I., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and *Lotus japonicus*. *Plant Mol. Biol.* 51, 731–743. doi: 10.1023/A:1022519709298

James, J. T., and Dubery, I. A. (2009). Pentacyclic triterpenoids from the medicinal herb, *Centella asiatica* (L.) Urban. *Molecules* 14, 3922–3941. doi: 10.3390/molecules14103922

Kajiwara, S., Fraser, P. D., Kondo, K., and Misawa, N. (1997). Expression of an exogenous isopentenyl diphosphate isomerase gene enhances isoprenoid biosynthesis in *Escherichia coli*. *Biochem. J.* 324 (Pt 2), 421–426. doi: 10.1042/bj3240421

Kalita, R., Modi, M. K., and Sen, P. (2018). RNAi mediated silencing of 3-hydroxy-3-methylglutaryl-CoA reductases (HMGR) in *Centella asiatica*. *Gene Rep.* 11, 52–57. doi: 10.1016/j.genrep.2018.02.004

Kalita, R., Patar, L., Shasany, A. K., Modi, M. K., and Sen, P. (2015). Molecular cloning, characterization and expression analysis of 3-hydroxy-3-methylglutaryl coenzyme A reductase gene from *Centella asiatica* L. *Mol. Biol. Rep.* 42, 1431–1439. doi: 10.1007/s11033-015-3922-6

Kim, O. T., Ahn, J. C., Hwang, S. J., and Hwang, B. (2005a). Cloning and expression of a farnesyl diphosphate synthase in *Centella asiatica* (L.) Urban. *Mol. Cell* 19, 294–299.

Kim, O. T., Jin, M. L., Lee, D. Y., and Jetter, R. (2017). Characterization of the asiatic acid glucosyltransferase, UGT73AH1, involved in asiaticoside biosynthesis in *Centella asiatica* (L.) Urban. *Int. J. Mol. Sci.* 18, 2630. doi: 10.3390/ijms18122630

Kim, O. T., Kim, M. Y., Huh, S. M., Bai, D. G., Ahn, J. C., and Hwang, B. (2005c). Cloning of a cDNA probably encoding oxidosqualene cyclase associated with asiaticoside biosynthesis from *Centella asiatica* (L.) Urban. *Plant Cell Rep.* 24, 304–311. doi: 10.1007/s00299-005-0927-y

Kim, O. T., Seong, N. S., Kim, M. Y., and Hwang, B. (2005b). Isolation and characterization of squalene synthase cDNA from *Centella asiatica* (L.) Urban. *J. Plant Biol.* 48, 263–269. doi: 10.1007/BF03030521

Kim, O. T., Um, Y., Jin, M. L., Kim, J. U., Hegebarth, D., Busta, L., et al. (2018). A novel multifunctional C-23 oxidase, CYP714E19, is involved in asiaticoside biosynthesis. *Plant Cell Physiol.* 59 (6), 1200–1213. doi: 10.1093/pcp/pcy055

Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Peña-Rodríguez, L. M., et al. (2013). A metabolic gene cluster inLotus japonicusdiscloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.* 200, 675–690. doi: 10.1111/nph.12414

Kuwahara, Y., Nakajima, D., Shinpo, S., Nakamura, M., Kawano, N., Kawahara, N., et al. (2019). Identification of potential genes involved in triterpenoid saponins biosynthesis in *Gleditsia sinensis* by transcriptome and metabolome analyses. *J. Nat. Med.* 73, 369–380. doi: 10.1007/s11418-018-1270-2

Li, Y., Zhao, J., Chen, H., Mao, Y., Yang, Y., Feng, L., et al. (2022). Transcriptome level reveals the triterpenoid saponin biosynthesis pathway of *Bupleurum falcatum* L. *Genes* 13, 2237. doi: 10.3390/genes13122237

Liu, M., Dai, Y., Li, Y., Luo, Y., Huang, F., Gong, Z., et al. (2008). Madecassoside isolated from *Centella asiatica* herbs facilitates burn wound healing in mice. *Planta Med.* 74, 809–815. doi: 10.1055/s-2008-1074533

Liu, Q., Khakimov, B., Cárdenas, P. D., Cozzi, F., Olsen, C. E., Jensen, K. R., et al. (2019). The cytochrome P450 CYP72A552 is key to production of hederagenin-based saponins that mediate plant defense against herbivores. *New Phytol.* 222, 1599–1609. doi: 10.1111/nph.15689

Luo, Z. L., Zhang, K. L., Ma, X. J., and Guo, Y. H. (2016). Research progress in synthetic biology of triterpen saponins. *Chin. Tradit. Herbal Drugs* 47, 1806–1814. doi: 10.7501/j.issn.0253-2670.2016.10.029

Mao, Y., Chen, H., Zhao, J., Li, Y., Feng, L., Yang, Y., et al. (2023). Molecular cloning, functional characterization and expression of the β-amyrin synthase gene involved in saikosaponin biosynthesis in *Bupleurum chinense* DC. *J. Plant Biochem. Biotechnol.* 32, 284–295. doi: 10.1007/s13562-022-00804-2

Matsuda, H., Morikawa, T., Ueda, H., and Yoshikawa, M. (2001). Medicinal foodstuffs. XXVII. Saponin constituents of gotu kola (2): structures of new ursane- and oleanane-type triterpene oligoglycosides, centellasaponins B, C, and D, from *Centella asiatica* cultivated in Sri Lanka. *Chem. Pharm. Bull.* 49, 1368–1371. doi: 10.1248/cpb.49.1368

Miettinen, K., Pollier, J., Buyst, D., Arendt, P., Csuk, R., Sommerwerk, S., et al. (2017). The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nat. Commun.* 8, 14153. doi: 10.1038/ncomms14153

Morita, M., Shibuya, M., Kushiro, T., Masuda, K., and Ebizuka, Y. (2000). Molecular cloning and functional expression of triterpene synthases from pea (*Pisum sativum*) New α-amyrin-producing enzyme is a multifunctional triterpene synthase. *Eur. J. Biochem.* 267, 3453–3460. doi: 10.1046/j.1432-1327.2000.01357.x

Nguyen, K. V., Pongkitwitoon, B., Pathomwichaiwat, T., Viboonjun, U., and Prathanturarug, S. (2019). Effects of methyl jasmonate on the growth and triterpenoid production of diploid and tetraploid *Centella asiatica* (L.) Urb. hairy root cultures. *Sci. Rep.* 9, 18665. doi: 10.1038/s41598-019-54460-z

Patel, K., Mishra, R., and Patel, D. K. (2016). A review on phytopharmaceutical importance of asiaticoside. *J. Coast. Life Med.* 4, 1000–1007. doi: 10.12980/jclm.4.2016J6-161

Phillips, D. R., Rasbery, J. M., Bartel, B., and Matsuda, S. P. (2006). Biosynthetic diversity in plant triterpene cyclization. *Curr. Opin. Plant Biol.* 9, 305–314. doi: 10.1016/j.pbi.2006.03.004

Pootakham, W., Naktang, C., Kongkachana, W., Sonthirod, C., Yoocha, T., Sangsrakru, D., et al. (2021). *De novo* chromosome-level assembly of the *Centella asiatica* genome. *Genomics* 113 (4), 2221–2228. doi: 10.1016/j.ygeno.2021.05.019

Prakash, V., Jaiswal, N., and Srivastava, M. (2017). A review on medicinal properties of *Centella asiatica*. *Asian J. Pharm. Clin. Res.* 10, 69–74. doi: 10.22159/ajpcr.2017.v10i10.20760

Prasad, A., Mathur, A. K., and Mathur, A. (2019). Advances and emerging research trends for modulation of centelloside biosynthesis in *Centella asiatica* (L.) Urban-A review. *Ind. Crops Prod.* 141, 111768. doi: 10.1016/j.indcrop.2019.111768

Puttarak, P., Dilokthornsakul, P., Saokaew, S., Dhippayom, T., and Chaiyakunapruk, N. (2017). Effects of *Centella asiatica* (L.) Urb. on cognitive function and mood related outcomes: A Systematic Review and Meta-analysis. *Sci. Rep.* 7, 10646. doi: 10.1038/s41598-017-09823-9

Ramos-Valdivia, A. C., van der Heijden, R., and Verpoorte, R. (1997). Isopentenyl diphosphate isomerase: a core enzyme in isoprenoid biosynthesis. A review of its biochemistry and function. *Nat. Prod. Rep.* 14, 591–603. doi: 10.1039/NP9971400591

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489. doi: 10.1146/annurev.arplant.043008.092035

Sangwan, R. S., Tripathi, S., Singh, J., Narnoliya, L. K., and Sangwan, N. S. (2013). *De novo* sequencing and assembly of *Centella asiatica* leaf transcriptome for mapping of structural, functional and regulatory genes with special reference to secondary metabolism. *Gene* 525, 58–76. doi: 10.1016/j.gene.2013.04.057

Sawai, S., and Saito, K. (2011). Triterpenoid biosynthesis and engineering in plants. *Front. Plant Sci.* 2. doi: 10.3389/fpls.2011.00025

Seki, H., Tamura, K., and Muranaka, T. (2015). P450s and UGTs: key players in the structural diversity of triterpenoid saponins. *Plant Cell Physiol.* 56, 1463–1471. doi: 10.1093/pcp/pcv062

Shang, Y., and Huang, S. (2019). Multi-omics data-driven investigations of metabolic diversity of plant triterpenoids. *Plant J.* 97, 101–111. doi: 10.1111/tpj.14132

Sun, Z., Cunningham, F. X.Jr., and Gantt, C. E. (1998). Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte. *Proc. Natl. Acad. Sci. U.S.A* 95, 14585–14585. doi: 10.1073/pnas.95.19.11482

Sun, J., Zhang, Y. Y., Liu, H., Zou, Z., Zhang, C.-J., Zhang, X. H., et al. (2010). A novel cytoplasmic isopentenyl diphosphate isomerase gene from tomato (*Solanum lycopersicum*): Cloning, expression, and color complementation. *Plant Mol. Biol. Rep.* 28, 473–480. doi: 10.1007/s11105-009-0174-4

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* 65, 225–257. doi: 10.1146/annurev-arplant-050312-120229

Wang, R., Ren, C., Dong, S., Chen, C., Xian, B., Wu, Q., et al. (2021). Integrated metabolomics and transcriptome analysis of flavonoid biosynthesis in safflower (*Carthamus tinctorius* L.) with different colors. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.712038

Wei, G., Dong, L., Yang, J., Zhang, L., Xu, J., Yang, F., et al. (2018). Integrated metabolomic and transcriptomic analyses revealed the distribution of saponins in *Panax notoginseng*. *Acta Pharm. Sin. B* 8, 458–465. doi: 10.1016/j.apsb.2017.12.010

Wen, L., Yun, X., Zheng, X., Xu, H., Zhan, R., Chen, W., et al. (2017). Transcriptomic comparison reveals candidate genes for triterpenoid biosynthesis in two closely related *Ilex* species. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00634

Weng, X. X., Zhang, J., Gao, W., Cheng, L., Shao, Y., and Kong, D. Y. (2012). Two new pentacyclic triterpenoids from *Centella asiatica*. *Helv. Chim. Acta* 95, 255–260. doi: 10.1002/hlca.201100287

Xu, G., Cai, W., Gao, W., and Liu, C. (2016). A novel glucuronosyltransferase has an unprecedented ability to catalyse continuous two-step glucuronosylation of glycyrrhetinic acid to yield glycyrrhizin. *New Phytol.* 212, 123–135. doi: 10.1111/nph.14039

Xu, R., Fazio, G. C., and Matsuda, S. P. (2004). On the origins of triterpenoid skeletal diversity. *Phytochem.* 65, 261–291. doi: 10.1016/j.phytochem.2003.11.014

Xu, Y., Zhao, G., Ji, X., Liu, J., Zhao, T., Gao, Y., et al. (2022). Metabolome and transcriptome analysis reveals the transcriptional regulatory mechanism of triterpenoid saponin biosynthesis in soapberry (*Sapindus mukorossi* Gaertn.). *J. Agric. Food Chem.* 70, 7095–7109. doi: 10.1021/acs.jafc.2c01672

Yang, J. L., Hu, Z. F., Zhang, T. T., Gu, A. D., Gong, T., and Zhu, P. (2018). Progress on the studies of the key enzymes of ginsenoside biosynthesis. *Molecules* 23, 589. doi: 10.3390/molecules23030589

Yuan, Y., Zuo, J., Zhang, H., Li, R., Yu, M., and Liu, S. (2022). Integration of transcriptome and metabolome provides new insights to flavonoids biosynthesis in *Dendrobium huoshanense*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.850090

Zhao, Y., Ma, Y., Li, J., Liu, B., Liu, X., Zhang, J., et al. (2022). Transcriptomics–metabolomics joint analysis: New highlight into the triterpenoid saponin biosynthesis in quinoa (*Chenopodium quinoa* Willd.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.964558

Zhao, C., Xu, T., Liang, Y., Zhao, S., Ren, L., Wang, Q., et al. (2015). Functional analysis of β-amyrin synthase gene in ginsenoside biosynthesis by RNA interference. *Plant Cell Rep.* 34, 1307–1315. doi: 10.1007/s00299-015-1788-7

Zhu, X., Zeng, X., Sun, C., and Chen, S. (2014). Biosynthetic pathway of terpenoid indole alkaloids in *Catharanthus roseus*. *Front. Med.* 8, 285–293. doi: 10.1007/s11684-014-0350-2

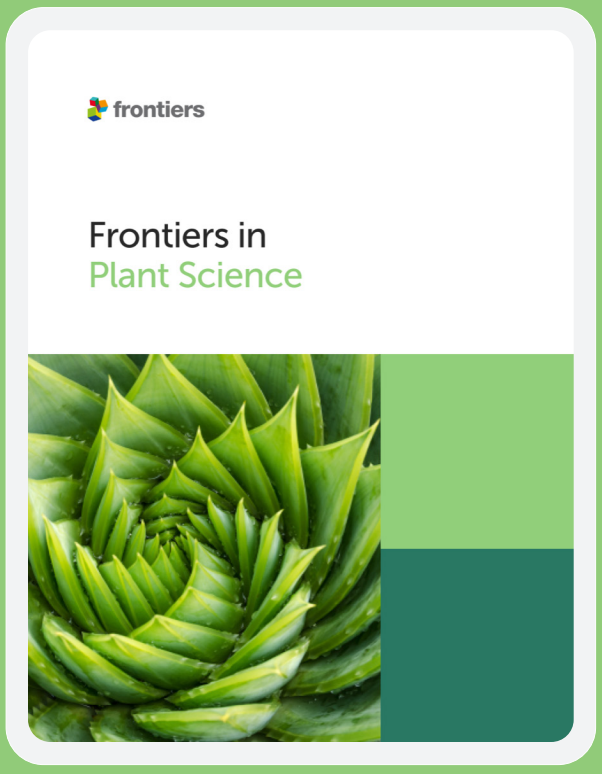
Frontiers in Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

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