The complete chloroplast genomes and comparative study of the two tung trees of Vernicia (Euphorbiaceae)

Jianjun Yang
Xinjiang University
Zhenzhou Chu
Xinjiang University
Haoyu Miao
Xinjiang University
Bingbing Liu
Shanxi University
XINMIN TIAN (tianxm06@lzu.edu.cn)
Xinjiang University

Research Article

Keywords: chloroplast, comparative genomics, Vernicia, conservation genetic

Posted Date: April 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1503656/v1

Abstract

Vernicia montana and *Vernicia fordii* are considered to have great economic importance in terms of having industrial oil and ornamental greening properties. However, wild resources of *Vernicia* trees have been reduced because of their ecology and habitat destruction. To better understand the genetic differences between *V. montana* and *V. fordii*, collecting more molecular data was necessary for further analysis. Here, we sequenced, assembled, and annotated the complete chloroplast (CP) genome of two tung trees based on the genome skimming approach. The whole CP genomes of *V. montana* and *V. fordii* were 160,906 bp and 161,494 bp in length, respectively, both of which comprised 135 genes, including 81 protein-coding genes, 29 tRNA genes, 4 rRNA genes, and 21 duplicated genes. The overall GC contents in *V. montana* and *V. fordii* CP genome were 36.13% and 36.03%, respectively. By comparing the CP genome sequences between *V. montana* and *V. fordii*, a total of 2,280 SNPs and 257 indels were identified. Among them, 1,807 SNPs and 153 indels were within the large single copy region, while 93 SNPs and 74 indels were within the small single copy region. The phylogenic analysis showed that *V. montana* was closely related to *V. fordii* and is considered a sister group. Hence, this study will be useful in investigating the conservation genetics and potential breeding applications of this oil shrub.

Introduction

In plant cells, chloroplasts (CPs) are the most important organelle as they conduct photosynthesis to provide energy for plants (Yang et al 2010; Terakami et al 2012). CPs have been derived from the endosymbiosis between independent living cyanobacteria and a non-photosynthetic host (Dyall et al 2004). It has been commonly considered that the CP genomes are prominently conserved in terms of size, structure, and gene content among the majority of flowering plant species (Palmer 1985; Jansen et al 2008). The CP genome is small and consists of a circular double-stranded DNA molecule. In angiosperms, CP genomes are 120 to 217 kb in length with 110 to 130 distinct genes, of which approximately 80 genes code for proteins, while other genes code for rRNAs and tRNAs (Raubeson and Jansen 2005). Usually, CP genomes consist of four distinct regions, including a pair of inverted repeats (IRs), a small single copy region (SSC), and a large single copy region (LSC) (Wang et al 2016; Yao et al 2016). To a certain degree, the CP DNA sequences are versatile tools for studying species identification, molecular evolution, and phylogenies (Gao et al 2019). In recent years, genomic research has rapidly developed with more than 2,000 entire organellar and nuclear genomes being efficiently sequenced.

Vernicia contains three species, two of which, *V. montana* and *V. fordii* occured in the tropical areas of South China (Li 1964; Wu 2007; Xu et al 2011). The two *Vernicia* species, which are also called tung trees, are of important economic value for extracting industrial oil and are considered one of the four most important woody oil trees in China (Chen et al 2010). It has been cultivated in South China for a thousand years, and especially in Southwest China. The tung oil contains 80% eleostearic acid and can be used for painting, varnishing, and in other chemicals. In recent years, the tung oil displayed a large potential for biodiesel production because of its high oil yields (Chen et al 2010). Furthermore, *Vernicia* species possess high commercial value in the landscaping industry, which have been planted widely for the

afforestation of city (Guo et al 2019). The tung tree also rapidly grows and yields fruits in three years due to its high photosynthetic efficiency (Li et al 2017). In fact, *V. fordii* is more widely planted than *V. montana*, which is distinguished by its wrinkled fruits and its big glands in the petiole (Wu 2007). Additionally, *V. montana* is evergreen and unisexual, while *V. fordii* is deciduous and bisexual (Wu 2007).

To understand the fast growing and high photosynthestic efficiency mechanisms, Li et al sequenced and annotated the CP complete genome of *V. fordii* (Li et al 2017). However, as a sister species of *V. fordii*, *V. montana* was not typically focused on and only a few studies mentioned it. To answer the questions of what is the genetic difference between the two-sister groups and how many genes are contained in the CP genome of *V. montana*, it was necessary to collect more molecular data on both species for further analysis. Simultaneously, the wild resources of *V. montana* trees have been decreasing dramatically in recent years because of their ecology and habitat destruction. Hence, it is necessary to take effective methods to conserve this important oil tree. The deep comparisons of the closely related plant species can dramatically improve the sensitivity of the evolutionary inferences when including the genomic structural knowledge to elucidate the mechanisms or rates of CP genome evolution. Thus, the genus *Vernicia* is an ideal group for understanding the diversity and evolution of CP genes as well as the genomes of flowering plants.

In this study, we sequenced, assembled, and annotated the complete CP genome of the two tung trees based on the Illumina pair-end sequencing data and the genome skimming approach. To deeply understand the sequences and the data, we compared the genome variation of two tung trees. Additionally, we also compared the tung trees genome with other known Euphorbiaceae species CP genomes, while aiming to determine the phylogenetic relationships among the angiosperms. Furthermore, the complete CP genome studies would be useful for developing the conservation strategy for these oil trees.

Materials And Methods

Plant materials and DNA sequencing

The fresh leaves of *V. montana* and *V. fordii* (identified by Xinmin Tian) were collected from Malipo (Yunnan, China) and used for DNA extraction. Vouchers were deposited at the Plant Herbarium of Xinjiang University, Urumqi, Xinjiang (Accession Number Shili011 and Shili012, respectively). Here we must declare that we have obtained local government permission to collect the plant materials we needed. The total genomic DNA was extracted using the CTAB method (Doyle 1987). The complete-genome sequencing was conducted using 150 bp pair-end reads on the Illumina Hiseq Platform (Illumina, Inc.; USA). In total, approximately 50 million high quality clean reads for each species with trimmed adaptors and filtered for high quality were collected by the genome skimming technique.

Genome Assembly And Annotation

Since the raw sequence data was mixed with non-CP DNA from the nucleus and the mitochondria, we isolated the CP sequence based on the known CP genome sequences and the filtered CP sequence was used to assemble the CP genomes by the program NOVOPlasty v3.7.2 (Nicolas 2017). Annotation analysis was performed with a recently developed command-line application called Plastome Annotator (Plann), which is suitable for the annotation of plastomes (Huan and Cronk 2015). Next, we analyzed and corrected the annotations with Geneious v10.0.5 (Kearse et al 2012). Then, we generate a physical map of the genome using OGDRAW v1.3.1 (http://ogdraw.mpimp-golm.mpg.de/) (Marc et al 2013). Finally, the complete CP genome sequences, together with gene annotations, were submitted to GenBank under the accession numbers of MT857744 for *V. montana* and MT857743 for *V. fordii.*

Repeat Elements Analysis

The simple sequence repeats (SSR) in *V. montana* and *V. fordii* CP genomes were identified using the MISA software (http://pgrc.ipk-gatersleben.de/misa/) (Thiel et al 2003). The SSRs were counted, whose repeating sequence units were arranged from 2 to 6 bp and repeated more than three times. Repeat sequences with lengths ≥ 22 bp were considered as long repeat sequences. Two SSRs with interruption lengths of ≤ 100 bp were considered as compound microsatellite repeats. The REPuter was utilized to identify the forward (direct) repeats, reverse sequences, complementary, and palindromic sequences with at least 21 bp in length and 90% sequence identity (Kurtz et al 2001).

Genome-wide analyses of genomic variants

To detect and characterize the genomic variation (SNPs and indels), we applied the mVISTA program to identify similarities among *Vernicia* CP genomes, followed by manual examinations confirm the data (Mayor et al 2000). Insertions and deletions (indels) as well as nucleotide substitutions and inversions were scored as single independent characters. In addition, the contraction/expansion regions of the inverted repeat (IR) were compared among *V. montana*, *V. fordii*, *J. curcas* (NC012224), *M. esculenta* (EU117376), *H. brasiliensis* (HQ285842), and *D. Tonkinensis* (MH933861) (Daniellr et al 2008; Tangphatsornruang et al 2011; Wang et al 2019).

Phylogenomic Analysis

To analyze the *Vernicia* phylogenetic position, we aligned 19 complete CP genome sequences of Euphorbiaceae and other related family species. All the species representing seven orders and *Vitis vinifera* were selected as the outgroup. All sequences were aligned with the software MEGA v6.06 (Tamura et al 2013). Alignments were manually adjusted and refined with high quality sequences that were kept for the downstream analyses. The PAUP* v4.010b software was used for conducting maximum parsimony (MP) phylogenetic trees. To construct the MP trees, we selected the MULTREES and tree bisection reconnection branch swapping as the heuristic searches. A total of 100 repetitions of random

sequence additions were used to calculate the starting trees and the saved trees, which were obtained by stepwise addition. To evaluate the node support, bootstrap values (BP) were applied and calculated.

Results

General features of CP genome

In each species, a total of 10.0 GB of raw data was obtained and approximately 10 million high quality clean reads were collected after filtering out the low quality reads by the genome skimming sequencing technique on the Illumina HiSeq 2500 Platform (Illumina, Inc., USA). Then, the filtering sequence data were used to construct the CP genome by comparing the Euphorbiaceae species in the National Center for Biotechnology Information database, which displayed high sequence consensus. The program NOVOPlasty was used to assemble the genome. After the genome assembly, the complete CP genomes of *V. montana* and *V. fordii* were 160,906 bp and 161,494 bp in length, respectively. In *V. montana*, the complete CP genome contained a large single-copy (LSC) region of 88,606 bp (55.1%) and a small single-copy (SSC) region of 18,733 bp (11.6%), which were separated by two copies of an inverted repeat (IR) of 26,784 bp (16.6%) (Fig. 1, Table 1). In *V. fordii*, the LSC, SSC, and IR were 89,098 bp (55.2%), 18,761 bp (11.6%), and 26,818 bp (16.6%) (Table 1). The annotation analysis showed that both species had a total of 135 genes that included 4 unique rRNAs, 29 tRNAs, 81 protein-coding genes, and 21 genes that were duplicated in the IRs. The overall GC content in the *V. montana* and *V. fordii* CP genomes were 36.13% and 36.03%, respectively (Table 1). In addition, the overall A, T, G, and C contents in *V. montana* were 32.38%, 31.48%, 18.37%, and 17.74%, respectively. However, the GC content varied across different genomic regions in *V. montana* and *V. fordii*. For example, the GC content in the protein-coding, intergenic spacers (IGS), tRNA, and rRNA regions were 37.37%, 30.07%, 53.13%, and 55.51%, respectively, but generally, the GC content in different regions were very similar in both *V. montana* and *V. fordii* (Table 1).

Table 1
Comparison of the plastome genomes of the two studied *Vernicia* species, including details of the intergenic spacers (IGS), large single copy (LSC), small single copy (SSC), and inverted repeats (IR) regions.

Genome features	*Vernicia fordii*	*Vernicia montana*
Size (bp)	161494	160906
LSC length (bp)	89098	88606
SSC length (bp)	18760	18732
IR length (bp)	53638	53568
Coding regions (bp)	92073	92053
Protein-coding regions (bp)	80283	80262
Total genes	135	135
Protein-coding genes	81	81
rRNA genes	4	4
tRNA genes	29	29
Genes with introns	16	16
Genes duplicated by IR	21	21
Overall GC content (%)	36.03	36.12
Overall A content (%)	32.41	32.38
Overall T content (%)	31.55	31.48
Overall G content (%)	18.33	18.37
Overall C content (%)	17.69	17.74
GC content in protein-coding regions (%)	37.38	37.37
GC content in IGS (%)	29.83	30.07
GC content in introns (%)	37.02	36.99
GC content in tRNA (%)	53.17	53.13
GC content in rRNA (%)	55.41	55.51

Microsatellites Analysis Of Two Cp Genome

Using the MISA analysis, 94 microsatellites were found in the *V. montana* CP genome, which included 51 SSRs, 18 long repeat sequences, and 25 compound microsatellite repeats (Supplementary table S1-3). At

the same time, 126 microsatellites were found in the *V. fordii* CP genome, which included 70 SSRs, 25 long repeat sequences, and 31 compound microsatellite repeats (Supplementary table S4-6). We then analyzed the SSRs of the CP genomes, which are important molecular markers in plant population genetics as well as in evolutionary and ecological studies. The most abundant SSRs consisted of the mono- and dinucleotides repeats, which accounted for approximately 75–81% of the total SSRs, which was followed by the tri- and tetranucleotides. The penta- and hexanucleotides were very rare across the CP genomes. Compound microsatellites are a special variation of microsatellites in which two or more individual microsatellites can be found directly adjacent to one another. Of these, the most abundant were < 100 bp long, which accounted for approximately 72–81% of the total compound microsatellites.

Structural Variation Analysis Between V. Montana And V. Fordii

By comparing the CP genome sequences between *V. montana* and *V. fordii*, a total of 2,280 SNPs and 257 indels were identified. Among them, 1,807 SNPs and 153 indels were within the LSC region, while 93 SNPs and 74 indels were within the SSC region (Table 2). Sequence identity comparisons between *V. montana* and *V. fordii* CP genomes were generated by mVISTA using the annotated *V. montana* sequence as a reference (Fig. 3). Both *Vernicia* CP genomes were found to be similar between 160,906 to 163,856 bp (Table 1).

Table 2
Structural variation of the plastome genomes among the two *Vernicia* species.

Sequence region	SNPs	Indels
in LSC region	1807	153
in SSC region	93	74
in IR region	380	30
Total	2280	257

Irs Expansion

The expansion and contraction of the IR regions and the SSC boundary regions can result in overall length variations in the plastid genomes. The IR/SSC and IR/LSC borders as well as the adjacent genes were compared across the six *Euphorbiaceae* CP genomes (Fig. 4). The *V. montana* CP genome had the shortest LSC region (88,606 bp), whereas *J. curcas* had the longest LSC region (91,756 bp). Both *Vernicia* CP genomes showed similar LSC/IR borders, where the LSC/IRb boundary was located between the genes *rps*19 and *rpl*22, while the LSC/IRa boundary was located between the genes *rps*19 and *trn*H.

However, in *Deutzianthus tonkinensis* and *Manihot esculenta*, the *rps*19 gene overlapped with the LSC/IR borders, while the SSC/IR boundary overlapped with the *ycf*1 gene across all species (Fig. 4).

Phylogenetic Analysis

We aligned 19 complete CP genome sequences to perform the phylogenetics analysis, which represented seven angiosperm orders, where *Vitis vinifera* was selected as the outgroup. The phylogenetic trees were constructed using machine learning and the MP method, resulting in both trees generating similar topologies (Fig. 5). *V. montana* was placed as a sister species to *V. fordii* with high bootstrap values (BP = 100), while the *Vernicia* branches were grouped with *Deutzianthus tonkinensis*, which is a member of the *Trib Aleuritideae*. In addition, all Euphorbiaceae species formed a monophylogenetic group that had high bootstrap values (BP = 100) (Fig. 5).

Discussion

In our study, the complete CP genomes of *V. montana* and *V. fordii* were sequenced, assembled, and annotated, which were 160,906 bp and 161,494 bp in length, respectively. Compared to *V. fordii*, the genome of *V. montana* was smaller, but it contained the same genes. As shown in figureures 2 and 3, the CP genomes of *V. montana* and *V. fordii* are typically circular with four regions that consisted of a pair of IRs of 53,568 kb and 53,638 in *V. montana* and *V. fordii*, respectively, which is separated by an SSC region of 18,732 kb and 18,760 in *V. montana* and *V. fordii*, respectively, and a LSC region of 88,606 kb and 89,098 in *V. montana* and *V. fordii*, respetively (Table 1). In both species, 135 genes were annotated in those regions, including 81 protein genes, 4 tRNA genes, 29 rRNA genes, and 21 genes that were duplicated in the IRs (Table 1). Furthermore, a total of 2,280 SNPs and 257 indels were identified within the *V. montana* and *V. fordii* genomes. The CP genome sequence similarity between the two species was 98.4%, which suggested a close relationship between *V. montana* and *V. fordii* that was also indicated by the phylogenetic analysis. Through these comparative analyses, the similarity between the two *Vernicia* species in CP genome was easily identified.

It was reported that the CP genomes in most land plants contain two identical IR regions, which would have lower nucleotide substitution rates and fewer indels than in the LSC and SSC regions (Li et al 2017; Kim and Lee 2004). Similarly, in our study, 30 indels and 380 SNPs were identified in the IR regions of the *Vernicia* CP genome (Table 2), where there are more duplicated genes (21 in *Vernicia*) than in other *Euphorbiaceae* species. In contrast, the IGSs and intron regions had more indels than the protein-coding genes, and thus, seemed to have evolved more quickly than the protein-coding genes. Traditionally, the nucleotide substitutions and indels in the CP genomes were used as DNA and barcoding markers in the phylogenetic analysis of many land plants (Clegg et al 1994; Morton and Clegg 1995; Katayama and Uematsu 2005). Hence, certified indel site information can be an important resource in future studies.

Repeat sequences are useful for studying genome rearrangements, while acting as important molecular markers in plant population genetics as well as evolutionary and ecological studies (Cavalier-Smith

2002). In the CP genome of *V. montana*, 94 microsatellites were identified, which included 51 SSRs, 18 long repeat sequences, and 25 compound microsatellite repeats (Supplemental tables S1-3). Meanwhile, 126 microsatellites were detected in the *V. fordii* CP genome, which included 70 SSRs, 25 long repeat sequences, and 31 compound microsatellite repeats (Supplemental tables S4-6). The lengths of the repeat units ranged between 10 to 233 bp, where a large number of repeats were distributed within the IGS regions, while the IRs accounted for the majority of the repeats. In addition, we found many repeats in the *ycf*2 gene, including two forward repeats and four palindromic repeats in both *Vernicia* CP genomes. Additionally, most of the repeats were found in the non-coding regions of the tung tree CP genome. Hence, the non-coding regions in CP genomes can act as important molecular markers for future phylogenetic studies (Small et al 1998).

Organelle genome sequencing is becoming an important approach for phylogenetic and taxonomic studies at low taxonomic levels (Yang et al 2013). Based on the current developments in the technology, thousands of organelle genomes have been sequenced, which can greatly mitigate the current reliance of phylogenetic research on relatively short sequences (Parks et al 2009). Whole CP genome sequences could provide more adequate information for phylogenetic and population-based studies, improving the discrimination efficiency when identifying species (Yang et al 2013). In fact, phylogenomic studies have currently gained popularity, where the possibility of using organelle-scale “barcodes” has recently been widely considered and applied (Parks et al 2009; Kuang et al 2011; Yang et al 2013). These CP genomes contained moderate variations that could provide sufficient phylogenetic information for resolving evolutionary relationships (Yang et al 2013). In the Euphorbiaceae family, several studies have analyzed the phylogenetic relationships based on CP DNA sequences (Daniell et al 2008; Tangphatsornruang et al 2011; Li et al 2017). Furthermore, to the best of our knowledge, no study has focused on the phylogenetic relationships between *V. motana* and *V. fordii*, and here, we have sequenced two *Vernicia* species using the Illumina sequencing technology, while performing a phylogenomic analysis. As expected, both *V. montana* and *V. fordii* included a sister group that had high bootstrap values (PP = 100). In addition, *Vernicia* and the four other species of the family Euphorbiaceae were clustered as a monophyly with high bootstrap values, which is consistent with previous studies (Li et al 2017). Additionally, *Vernicia* was suggested to be more closely related to *Deutzianthus* than other taxonomical groups.

Declarations

Acknowledgements XM Tian conceived and designed the experiments. ZZ Chu and HY Miao analyzed the data. XM Tian contributed reagents/materials/analysis tools. JJ Yang, XM Tian and BB Liu wrote and revised the paper. All authors reviewed and agreed to the published version of the manuscript.

Funding This study was supported by the National Natural Science Foundation of China (grant No 31601782 to Xinmin Tian) and Open project of Xinjiang Key Laboratory of Biological Resources and Genetic Engineering (grant 2020D04033 to Xinmin Tian)

Conflict of interest The authors have no confict of interest to declare.

References

1. Cavalier ST (2002) Chloroplast evolution: secondary symbiogenesis and multiple losses. Curr Biol 12:R62–R64
2. Chen YH, Chen JH, Chang CY, Chang CC (2010) Biodiesel production from tung (*Vernicia montana*) oil and its blending properties in different fatty acid compositions. Bioresour Technol 101:9521–9526
3. Clegg MT, Gaut BS, Learn GH, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci 91:6795–6801
4. Daniell H, Wurdack KJ, Kanagaraj A, Lee SB, Saski C et al (2008) The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atp*F in Malpighiales: RNA editing and multiple losses of a group II intron. Theor Appl Genet 116:723–737
5. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
6. Dyall SD, Brown MT, Johnson PJ (2004) Ancient invasions: from endosymbionts to organelles. Science 304:253–257
7. Guo L, Jinxiang HE, Wang X, Lin C, Chengxin HE (2009) Research progress on exploitation of main oil plants of Euphorbiaceae. China Oils and Fats 34:57–61
8. Gao LZ, Liu YL, Zhang D, Li W, Eichler EE (2019) Evolution of *Oryza* chloroplast genomes promoted adaptation to diverse ecological habitats. Commun Biol 2:1–13
9. Huan DI, Cronk QC (2015) Plann: A Command-Line Application for Annotating Plastome Sequences. Appl Plant Sci 3:1500026
10. Jansen RK, Cai Z, Raubeson LA, Daniell H, Boore JL (2008) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104:19369–19374
11. Katayama H, Uematsu C (2005) Structural analysis of chloroplast DNA in *Prunus* (Rosaceae): evolution, genetic diversity and unequal mutations. Theor Appl Genet 111:1430–1439
12. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M et al (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649
13. Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res 11:247–261
14. Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ et al (2011) Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54:663–673
15. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J et al (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642

16. Li BT (1994) Flora of China, Euphorbiaceae. Beijing Science Press, Beijing
17. Li Z, Long HX, Zhang L, Liu ZM, Cao HP et al (2017) The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): organization and phylogenetic relationships with other angiosperms. Sci Rep 7:1–11
18. Marc L, Oliver D, Sabine K, Ralph B (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
19. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM et al (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046–1047
20. Morton BR, Clegg MT (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J Mol Evol 41:597–603
21. Nicolas D, Patrick M, Guillaume S (2017) Novoplasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45:e18
22. Palmer JD (1985) Comparative Organization of Chloroplast Genomes. Annu Rev Genet 19:325–354
23. Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol 7:1–17
24. Raubeson L, Jansen R (2005) Chloroplast genomes of plants, Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Cabi Publishing, London
25. Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85:1301–1315
26. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729
27. Tangphatsornruang S, Uthaipaisanwong P, Sangsrakru D, Chanprasert J, Yoocha T et al (2011) Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene 475:104–112
28. Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y et al (2012) Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. Tree Genet Genomes 8:841–854
29. Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422
30. Wang L, Wuyun TN, Du H, Wang D, Cao D (2016) Complete chloroplast genome sequences of *Eucommia ulmoides*: genome structure and evolution. Tree Genet Genomes 12:1–15
31. Wang QY, Qu ZZ, Tian XM (2019) Complete chloroplast genome of an endangered oil tree, *Deutzianthus tonkinensis* (Euphorbiaceae). Mitochondrial DNA Part B-Resour 4:299–300
32. Wu ZY (2007) Flora of China, Euphorbiaceae. Beijing Science Press, Beijing

33. Xu W, Yang Q, Huai HY, Liu AZ (2011) Microsatellite marker development in tung trees (*Vernicia montanaand* v. fordii, Euphorbiaceae). Am J Bot 98:e226–e228
34. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ (2013) Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol Biol 13:1–12
35. Yang M, Zhang XW, Liu GM, Yin YX, Chen KF et al (2010) The complete chloroplast genome sequence of Date Palm (*Phoenix dactylifera* L.). PLoS ONE 5:e12762
36. Yao X, Tan YH, Liu YY, Song Y, Yang JB et al (2016) Chloroplast genome structure in *Ilex* (Aquifoliaceae). Sci Rep 6:28559

Figures

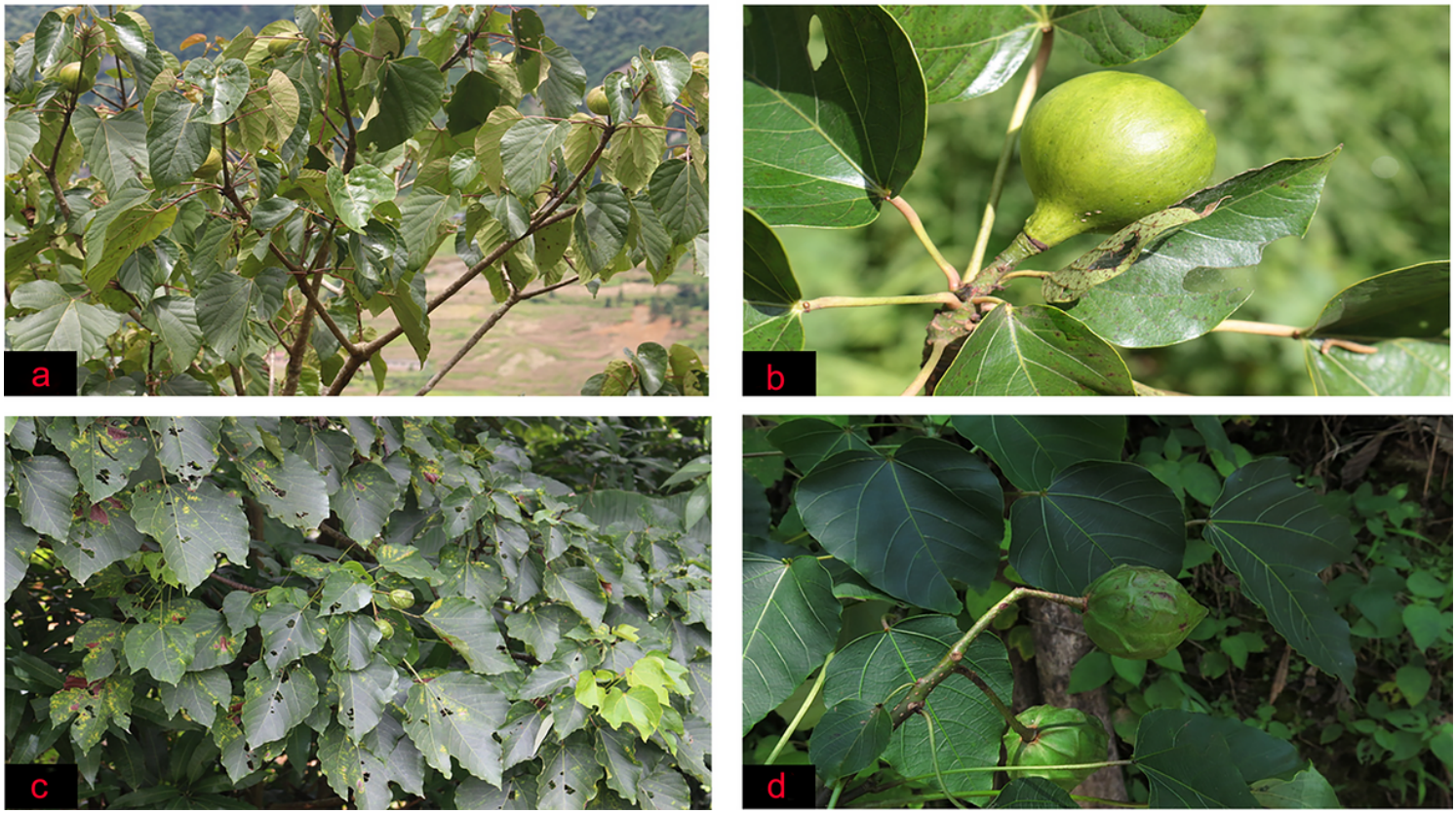

Figure 1

Images of the two studied *Vernicia* species, including (a-b) *Vernicia fordii* and (c-d) *Vernicia montana.*

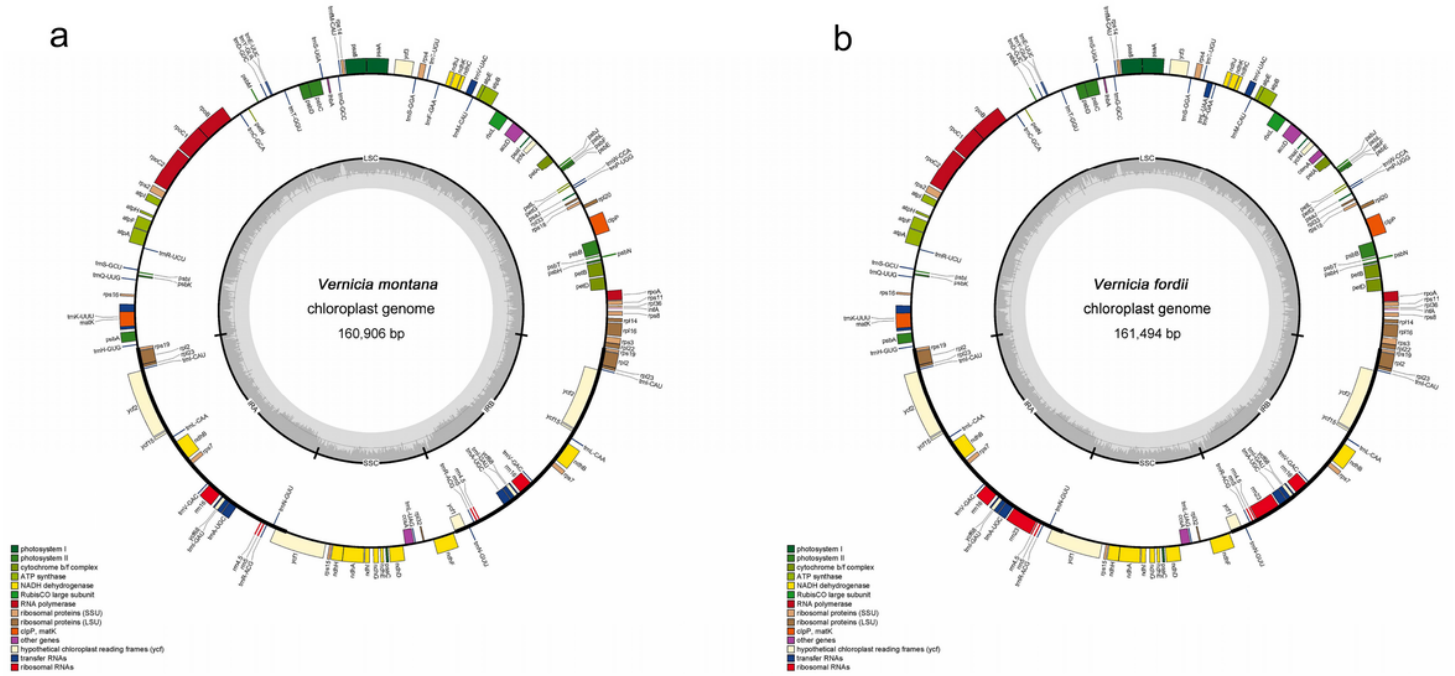

Figure 2

The genetic map of the (a) *Vernicia montana* and (b) *Vernicia fordii* chloroplast genomes, where the different colors represent different components of the genome.

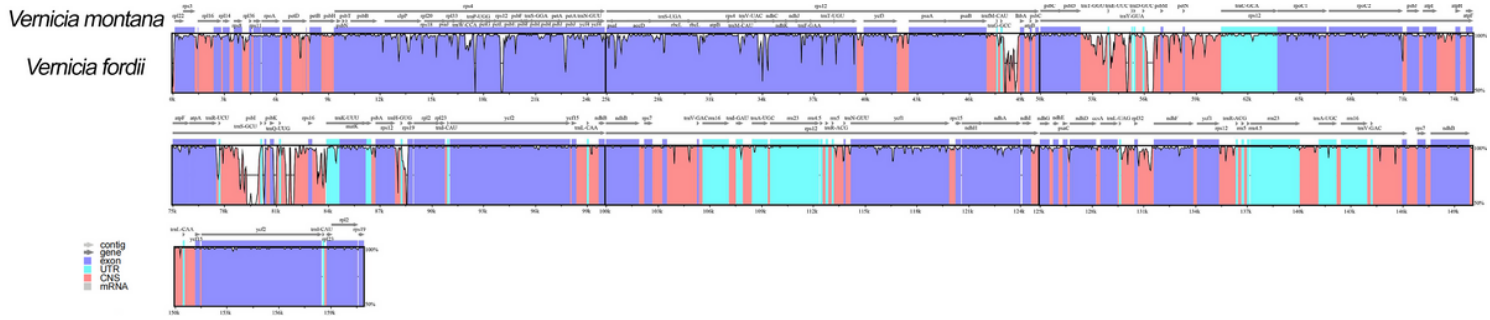

Figure 3

Comparison of the chloroplast genomes of the two studied *Vernicia* species using the mVISTA program. The gray arrows and thick black lines located above the alignment indicate the gene orientation. Purple and blue bars represent the exons and the introns, respectively (shown: 50-100%).

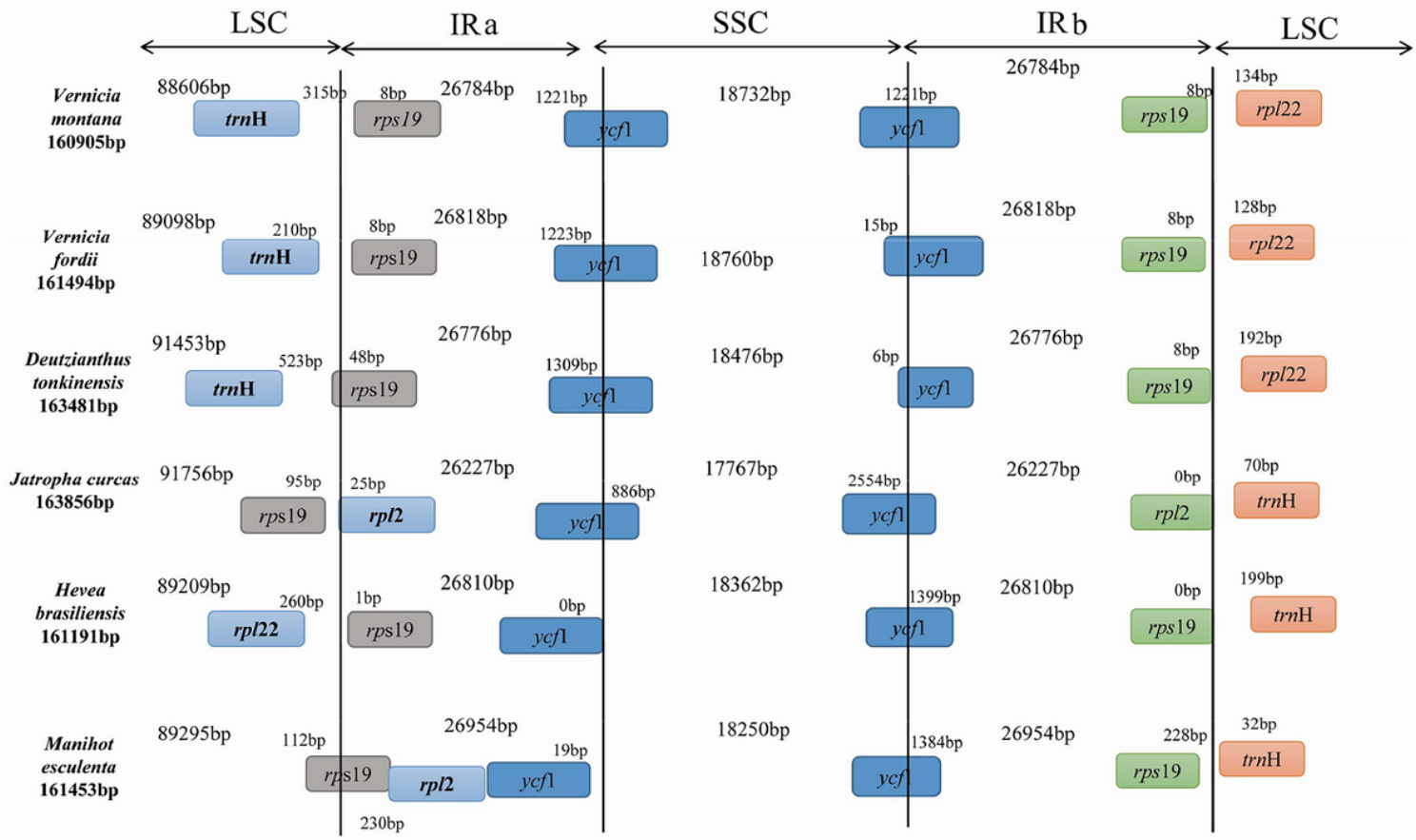

Figure 4

Comparison of the border positions of the large single copy (LSC), small single copy (SSC), and inverted repeats (IR) regions among the six chloroplast genomes.

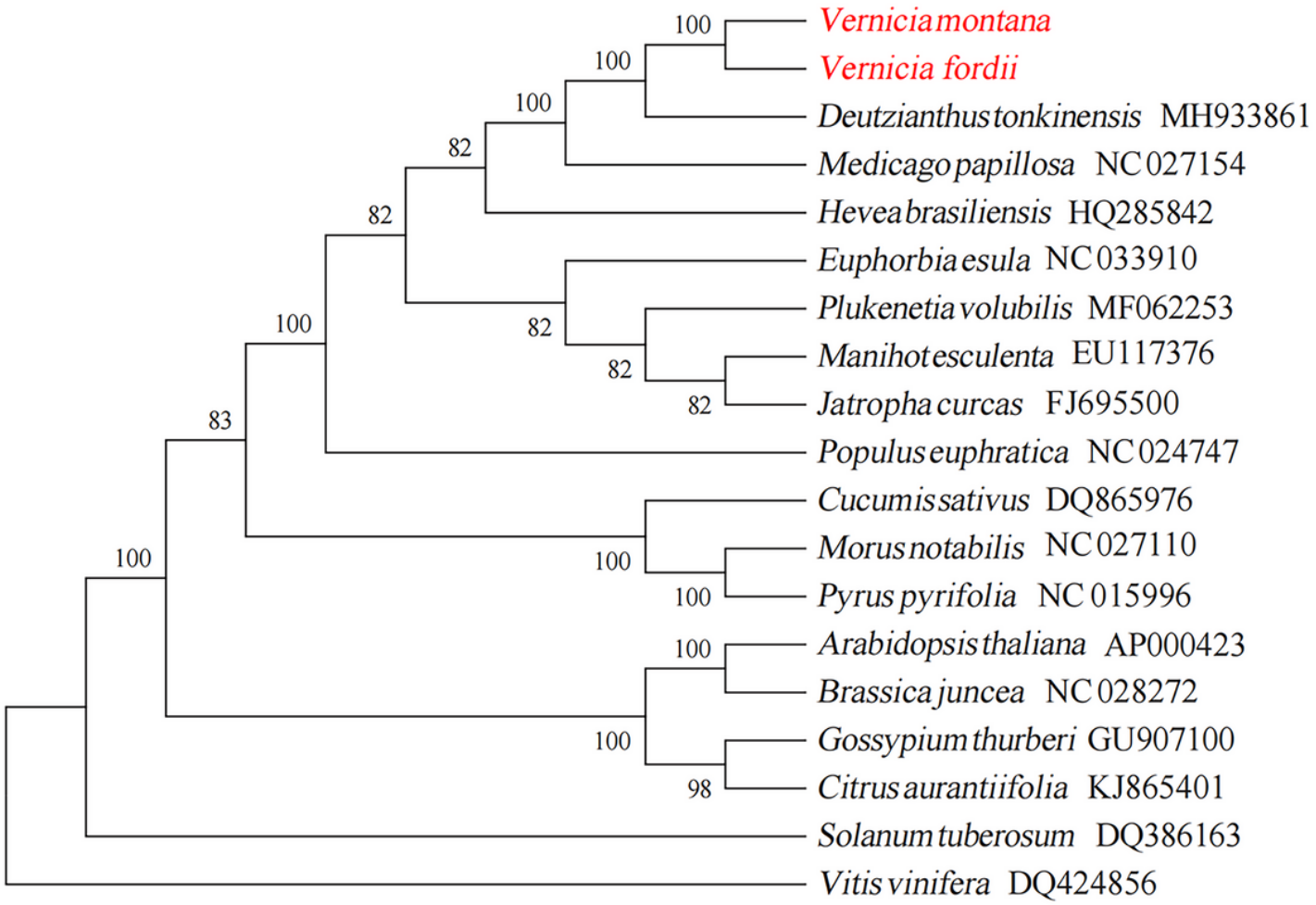

Figure 5

The phylogenetic tree showing the two studied *Vernicia* species that was based on the 19 complete sequences of different chloroplast genomes. Accession numbers: *Manihot esculenta* (EU117376), *Jatropha curcas* (FJ695500), *Euphorbia esula* (NC033910), *Plukenetia volubilis* (MF062253), *Hevea brasiliensis* (HQ285842), *Deutzianthus tonkinensis* (MH933861), *Vernicia fordii* (MT857743), *Vernicia montana* (MT857744), *Cucumis sativus* (DQ865976), *Arabidopsis thaliana* (AP000423), *Brassica juncea* (NC028272), *Gossypium thurberi* (GU907100), *Citrus aurantiifolia* (KJ865401), *Medicago papillosa* (NC027154), *Morus notabilis* (NC027110), *Pyrus pyrifolia* (NC015996), *Populus euphratica* (NC024747), *Solanum tuberosum* (DQ386163), *Vitis vinifera* (DQ424856)

Supplementary Files

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

- Supplementarytables.docx