Comparative analysis of whole chloroplast genomes of *Ligusticum sinense* and *L. jeholense*, Umbelliferae

Shijie Wang
Liaoning University of Traditional Chinese Medicine
Tingting Zhang
Liaoning University of Traditional Chinese Medicine
Liang Xu (861364054@qq.com)
Liaoning University of Traditional Chinese Medicine https://orcid.org/0000-0002-8623-7708
Zhilai Zhan
Chinese Academy of Traditional Chinese Medicine
Guihua Bao
Inner Mongolia University for Nationalities
Tong Zhang
Liaoning University of Traditional Chinese Medicine
Zixuan Ding
Liaoning University of Traditional Chinese Medicine
Nan Sun
Liaoning University of Traditional Chinese Medicine
Shaobin Sun
Liaoning University of Traditional Chinese Medicine
Ming Xie
Liaoning University of Traditional Chinese Medicine
Tingguo Kang
Liaoning University of Traditional Chinese Medicine

Research

Keywords: Ligusticum sinense, Ligusticum jeholense, Chloroplast genome, Phylogeny

Posted Date: August 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-58085/v1

Abstract

Background

Gaoben has a long history of application as medicine. There are few records of Liaogaoben in ancient books. The varieties of *Ligusticum* in practical application are confused. Therefore, it is very important to identify *Ligusticum* accurately. The phylogenetic position of *Ligusticum* in Umbelliferae needs to be determined. It is also of great significance to analyze the phylogeny of *Ligusticum* in Umbelliferae the difference of somatic genome.

Methods

Chloroplast (cp) genomic DNA was extracted from two species of *Ligusticum* and sequenced on Hiseq4000 light source. The sequence was assembled into contrags by soapenovo 2.04, aligned with reference genome by blast, and then corrected manually. Genereanation is performed by online dogma tools. The general characteristics of cp genomes of two species were analyzed, and compared with the relative species. The DNA of chloroplasts of higher plants is double stranded covalently closed loop molecule, and its length varies with species. According to the assembly genome sequence of the sequenced samples, combined with the prediction results of the coding genes, the genome of the samples was displayed in circles. After alignment, the evolutionary tree was constructed based on the single nucleotide polymorphism (SNP) of cp genome in 31 species by ML method.

Results

The whole chloroplast genome of Gaoben and Liaogaoben is 148,515 bp and 148,493 bp, both of which contain IR (IRa and IRb) LSC and SSC. 127 genes have been annotated, including 83 protein coding genes, 8 tRNA and 36 rRNA of all species, and 28 coding genes in IR region. There are six genes in the ATPase subunit of photosynthesis gene group, and there are obvious differences in the types of introns in NADH dehydrogenase subunit between them. In the comparative analysis of Pi value, two significant gene variation points are *petG* gene and *psaL-ycf4* gene. The phylogenetic tree of the whole cp genome of SNP Umbelliferae was constructed, including 28 Umbelliferae and 3 *Ligusticum*.

Conclusions

In this study, the cp genomic characteristics of *Ligusticum sinense* and *L. jeholense* were identified, which provided a theoretical basis and documentation for the identification and phylogenetic analysis of *Ligusticum*.

Background

Gaoben it has a long cultivation history in China. It has been recorded in “Shennong Classic of Materia Medica”, which has the effect of dispelling wind and dampness, dispersing cold and relieving pain [1]. Gaoben comes from *Ligusticum sinense* and *L. jeholense*. Gaoben is mainly distributed in the south of China, while Liaogaoben mostly occurs in the north. There are obvious differences in root morphology and microscopic characteristics between the two. The traditional Chinese medicine Gaoben is commonly used in the treatment of wind cold headache, pain on the top of the mountain, wind cold dampness, etc. Gaoben is included in the traditional Chinese medicine, while Liaogaoben as the source of *Ligusticum* is recorded in it, not included in it [2]. Because of the close relationship between them, it is difficult to distinguish them, so we use the methods of chloroplast genome to identify them.

In recent years, with the rapid development of plant molecular systematics, new methods have been widely used in the localization of chloroplast (cp) restriction sites in cp genome, but limited to the applicable classification level. The change of gene sequence and content in cp genome is rare, which is usually caused by a part of genome or gene deletion. For example,

differences in gene sequence and content [3]. With a few exceptions, the chloroplast genome contains two reverse repeat sequences (IR). These repeat regions separate the rest of the molecule into large single copy (LSC) and small single copy (SSC) regions [4]. Recent studies on more than 200 flowering plants in 30 families have shown that the chloroplast genome is highly conserved in its size, tissue and primary sequence [5]. Umbelliferae is a large and complex family of flowering plants Phylogenetic relationships, especially at the lower taxonomic level, are usually fuzzy based on the molecular markers currently widely used. Therefore, it is very advantageous to attach the information of phylogenetic utility of molecular markers at these levels [6].

The cp genomes of two *Ligusticum* species were studied. To explore the differences between the molecular level, and to analyze and identify *L. sinense* and *L. jeholense* by cp genome and other content. The relationship between *L. sinense* and *L. jeholense* was analyzed.

Methods

DNA extraction and sequencing

DNA extraction from plant tissues is a very critical process. Many time-consuming steps in plant molecular biology are involved [7]. We collected the fresh leaves and roots of *L. sinense* and *L. jeholense* from the key varieties protection park of Liaoning Province in Dalian campus of Liaoning University of traditional Chinese medicine (N 39°06´, E 121°87´, Dalian, Liaoning Province, China). Five mg fresh leaves and 5 mg roots were collected by improved extraction method for cp DNA separation [8]. After DNA separation, 1 μg of purified DNA was segmented and used to construct a short insertion Library (430 bp insertion size) according to the manufacturer's instructions (Illumina), and then sequenced on Illumina Hiseq 4000 [9].

Genome assembly and annotation

Using Illumina truseq Gamma The library was constructed by nano DNA sample prep kit method. Using assembly software to assemble clean data, adjusting parameters and filling holes for many times to obtain the genome assembly sequence cp genome is reconstructed by using the combination of denovo and reference guided assembly. Before assembly, first filter the original reading. This filtering step is performed to remove reads with adapters, reads with a display quality score of less than 20 (Q < 20), reads that contain a percentage of an uncalled base ("n" character) equal to or greater than 10%, and repeat sequences. The following three steps are used to assemble the cp genome [10]. First, use soapdenovo 2.04 [11] to assemble the filtered reads into contrags. Secondly, we use blast to compare the contigs with the reference genomes of the two species, and rank the contigs according to the comparison of the reference genomes (both similarity and query coverage are greater than 80%). Third, map clean reads to the assembled cp genome sketch to correct the wrong base and fill in most of the gaps through local assembly. Cp gene uses online dogma tool annotation [12]. using default parameters to predict protein coding genes, transfer RNA (tRNA) genes and ribosomal RNA (rRNA) genes. The general function database compares the gene function annotation information of the samples, and classifies the specific functions. The sequencing data and gene annotations were then submitted to GenBank and assigned accession numbers (*L. sinens*e: MT512897, *L. jeholense*: MT512888).

The cp genomes of two *L. sinense* species were derived in GenBank format, and the cp gene map was drawn by organelle genome draw (ogdraw). Max-Planck Institute of Molecular PlantPhysiology, AmMühlenberg, Potsdam, Germanyhttp://OGDRAW.mpimp-golm.mpg.de/index.Shtml [13].

Comparative analysis of genomes

The differences between samples and reference genomes were compared from two levels of genome and gene. SSR sequences were identified by SSR software microsatellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/), and the tandem repeats of 1-6 nucleotides were regarded as microsatellites. For single, two, three, four, five, and six nucleotides, the minimum number of repeats is set to 10, 6, 5, 5, 5, and 5, respectively. The maximum base number of two SSR interruptions in composite microsatellites is 100. We compared these data with *Ligusticum tenuissimum* (NC_029394), and focused on the

perfect repeat sequence [14]. The long repeat sequences were analyzed by using the network reputer (http://bibiserv.techfak.uni-bielefeld.de/reputer/), including forward, reverse, complement and palindrome repeat sequences. The minimum sequence length was 30 bp, and the editing distance was 3 bp [15]. The variation characteristics of two kinds of *Ligusticum* were screened out. The average number of nucleotide differences and the total number of mutations were measured by dnasp v5.10 to analyze nucleotide diversity (polymorphic information, Pi) [16].

Phylogenetic analysis

In order to analyze the phylogenetic relationship between *Ligusticum* and othergenera of Umbelliferae, phylogenetic tree was constructed from cp genome sequences of 25 species of Umbelliferae, 23 of which were downloaded from GenBank. Among the 28 plants, there are three exotaxa: *Ginkgo biloba* (NC_016986), *Panax ginseng* (NC_006290) and *Arabidopsis thaliana* (NC_000932). The phylogeny includes 25 species of Umbelliferae, 1 species of Cruciferae, 1 species of Ginkgo family and 1 species of Acanthopanax family. Single nucleotide polymorphisms (SNPs) in the whole cpgenome of 28 species of plants were analyzed. Phymlv3.0 software was constructed by maximum likelihood method (ML), GTR + I + GML model was established, bootstrap value was calculated by 100 bootstrap replications [17].

Results And Discussion

Chloroplast genomic characteristics of L. sinense and L. jeholense

The number and sequence of coding genes in chloroplast genome of higher plants are highly conserved [18]. The variation of genome composition may be related to the change of base selection at the third codon, which affects the composition of amino acids, but it seems to have little effect on the overall physical and chemical properties of amino acids [19]. The average GC content of the whole gene, as well as the average GC content at three codon positions, the nuclear gene is higher than the cp gene, indicating that the pressure of genome organization and mutation of nuclear gene and cp gene is different [20]. The total cp gene length of *L. sinense* was 148,515 bp, and that of *L. jeholense* was 148,493 bp. The difference between them was small. The content of GC in *L. sinense* and *L. jeholense* were 36.67% and 37.25%. These are the distinct features of the two genomes. The GC content of cp DNA in sweet potato is 38.45%, which is similar to other cp genomes reported in Convolvulaceae [21]. The difference of cp genome length is mainly caused by the change of LSC length. The length of SSC is 17,607 bp and 17,629 bp. The length of IR is 17,607 bp and 17,629 bp (Table 1). The regional constraints strongly affect the sequence evolution of the cp genomes, while the functional constraints weakly affect the sequence evolution of cp genomes [22].

Table 1
Comparison of general characteristics of chloroplast genomes of two species in Umbelliferae

Type	Size(bp)	GC Content(%)	LSC length (bp)	SSC length (bp)	IR length (bp)	Gene number	Gene number in IR regions	Protein-coding gene number	rRNA gene number	tRNA gene number
L. sinense	148,515	36.67	93,978	17,607	51,781	127	28	83	8	36
L. jeholense	148,493	37.25	93,932	17,629	36,932	127	28	83	8	36

Among them, *L. sinense* and *L. jeholense* are obviously different in IR region, and 8 rRNA and 36 tRNA are in IR region. By sequencing, we found that the size of LSC region of cp genome of *L. sinense* and *L. jeholense* is very similar, and the number of total genes and coding protein genes are basically the same, which proves that there is a close relationship between them. Organization of the spacer in Umbelliferae is consistent with a general pattern evident for angiosperms [23]. Cp DNA has been used extensivelyto infer plant phylogenies at different taxonomic levels [24].

The genetic types of *L. sinense* and *L. jeholense* are the same, the difference of bp length between them is very small, which proves that the genetic relationship between them is very close. In general, cp are divided into a large single copy (LSC) area, short single copy (SSC) area and two reverse repeat (IR) areas [25]. Specifically, the length of the IRb region of *L. sinense* is longer than that of *L. jeholense*, which shows that all the *ycf2* genes of *L. jeholense* are in the LSC region, while in *L. sinense*, part of the *ycf2* genes are in the LSC region and part of the IRb region. According to the symmetry of IRb and IRA, *L. sinense*is has a blank gene in IRa area. The remaining LSC and SSC areas are identical. This also explains the reason why the total length of *L. sinense*is 148,515 bp longer than that of *L. jeholense* (Fig. 1–2). Prangos fedtschenkoi (Regel & schmalh.) Korovin and P. lipskyi Korovin (Apiaceae) also combined new DNA in LSC region near LSC/IRa connection [26]. The *ycf94* predicted protein has a distinct transmembrane domain but with no sequence homology to other proteins with known function [27].

In the cp gene of *L. sinense* and *L. jeholense*, 127 genes were detected, among which 6 genes (*atpA*, *atpB*, *atpE*, *atpF*, *atph*, *atpI*) were found in the gene group of photosynthesis, The Chlamydomonas reinhardtii chloroplast *atpB* mRNA contains sequences at its 3' end that can form a complex stem/loop structure [28]. Twelve genes (*ndhA*, *ndhB*, *ndhc*, *ndhd*, *ndhe*, *ndhF*, *ndhg*, *ndhk*) were found in NADH dehydrogenase subunits, including 2 *ndhB* genes and 6 cytochrome subunits *petA*, *petB**, *petD**, *petG*, *petL*, *petN* gene, among which 20 coding genes were detected in light system, including 6 genes (*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psaL*) in light system I and 14 genes (*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbK, psbL*, *psbM*, *psbN*, *psbT*, *psbZ)* in light system II. The *psbA* gene has been located in a large single copy fragment in some monocotyledons, which is very close to the end of a reverse repeat sequence in the cp genome of most terrestrial plants [29]. Cp intergenic *psbA-trnH* spacer has recently become a popular tool in plant molecular phylogenetic studies at low taxonomic level and as suitable for DNA barcoding studies [23]. It includes *accD* gene of acetylCoA carboxylase subunit, *accD* gene of type C cytochrome, *ccsA* gene of synthetic gene, *cemA* gene of envelope protein, *clpP* gene of protease and *matK* gene of maturity. There are also 75 self replicating genes and 7 unknown genes (*ycf1*, *ycf2*, *ycf3*, *ycf4*, *ycf5*, *ycf15*) (Table 2). The sequence and structure of chloroplast genes are highly conserved [30, 31]. Among 127 genes in *L. sinense*, one intron gene is *rps16*, *atpf*, *ropc1*, *petb*, *petd*, *rpl16*, *ndhb-d2*, *ndhH* genes and two introns of *ndhB* gene are *ycf3*, *clpP* and *rps12* gene. Among the 127 genes with one intron compared with *L. sinense*, *L. jeholense* does not contain the intron gene of *ndhH*, in addition, there is one more intron gene of *ndhB* compared with *L. sinense*. The two introns have the same kind and quantity. Exons showed more random behaviors than introns [32] (Additional file 1: Table S1).

Table 2
List of the genes in the chloroplast genomes of two species of *Ligusticum*

Gene category	Gene group	Gene name
Photosynthesis	Subunits of ATP synthase (6)	*atpA*, *atpB*, *atpE*, *atpF* *, *atpH*, *atpI*
	Subunits of NADH dehydrogenase (12)	*ndhA*, *ndhB**(x2), *ndhC*, *ndhD*, *ndhE*, *ndhJ*, *ndhF*, *ndhH*, **ndhG*, *ndhJ*, *ndhK*
	Subunits of cytochrome (6)	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	Subunits of photosystem I (6)	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psaL*
	Subunits of photosystem II (14)	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Subunit of rubisco (1)	*rbcL*
Other genes	Subunit of Acetyl-CoA-carboxylase (1)	*accD*
	c-type cytochrome synthesis gene (1)	*ccsA*
	Envelop membrane protein (1)	*cemA*
	Protease (1)	*clpP***
	Maturase (1)	*matK*
Self-replication	Large subunit of ribosome (9)	*rpl2*, *rpl14*, *rpl16**, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*,
	DNA dependent RNA polymerase (4)	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
	Small subunit of ribosome (13)	*rps2*, *rps3*, *rps4*, *rps7* (x2), *rps8*, *rps11*, *rps12***(x2), *rps15*, *rps16**, *rps18*, *rps19*
	rRNA Genes (8)	*rrn4.5* (x2), *rrn5* (x2), *rrn16* (x2), *rrn23* (x2)
	tRNA Genes (36)	*trnA-UGC* (x2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC, trnG-GCC*, *trnH-GUG*, *trnI-CAU*, *trnI-CAA*, *trnK-UUU*, *trnL-UAG*, *trnL-UAA*, *trnL-GAU* (x2), *trnL-CAA*, *trnfM-CAU*, *trnM-CAU*, *trnN-GUU* (x2), *trnP-UGG*, *trnQ-UUG*, *trnR-UCU*, *trnR-ACG* (x2), *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC* (x2), *trnV-UAC*, *trnW-CCA*, *trnY-GUA*
Unknown function	Conserved open reading frames (7)	*ycf 1*, *ycf 2*, *ycf 3***, *ycf4* (x2), *ycf5*, *ycf15*

* contains one intron

** contains two introns

Numbers in brackets behind name of gene group give number ofrepetitive genes

Repeat sequences analysis

SSR sequence analysis

Site-specific recombinase (SSR) technology allows the manipulation of gene structure to explore gene function and has become an integral tool ofmolecular biology [33]. When SSR technology is used to analyze closely related genotypes, the high polymorphism of many microsatellites has special value, just like the casein breeding project working in the narrow sense adaptive gene pool [34]. Polymorphic SRAPs and SSRs were abundant in genetic diversity analysis among closely related cultivars [35]. Because of the characteristics of neutral markers, the highly variable numbers of repeats and the relative conservatism of flanking sequences of SSRs, it is widely distributed in the genome of organisms. The technique is easy to operate and has high repeatability and codominant inheritance among alleles. SSRs marker technique is the best choice for evaluating genetic diversity of crops [36, 37] (Additional file 2: Table S2).

The availability of complete sequences of chloroplast genomes enhances their use for genetic engineering. In chloroplast transformation, finding appropriate intergenic spacer regions is very important for efficient integration of transgenes [38]. There are 169 SSRs in the cp genome of *L. sinense* and 166 SSRs in the cp genome of *L. jeholense*. The main difference between them is the number of single nucleotide. The cp genome of *L. tenuissimum* contains 174 SSRs. There are six SSR types in *L. sinense* and *L. jeholense*, including single nucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide (Table 3, Fig. 3). In addition, *L. tenuissimum* does not contain six nucleotides for identification. The number of dinucleotide of *L. tenuissimum* is different from that of the other two. Dinucleotide repeat sequence shows high polymorphism in eukaryotic DNA. These sequences are convenient as genetic markers and can be analyzed by PCR [39] (Additional file 3: Table S3, Additional file 4: Table S4).

Table 3
SSRs in the chloroplast genomes in two species of *Ligusticum*

Unit size	Mononucleotide	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
L. sinense	136	19	3	7	2	2
L. jeholense	134	19	3	7	1	2
L. tenuissimum	137	23	3	7	4	0

Large Repeat Analysis

Many repeats occur in whole cp genes. In this study, *L. sinense* and *L. jeholense* have 39 and 37 pairs of large repeat sequences, respectively. It was found in the cp genome with sequence identity exceeding 90%. The large repeat range of *L. sinense* is 30 to 102 bp, and *L. jeholense* is 30 to 66 bp. A total of 16 and 13 large repeats were located in the genic regions of the two *Ligusticum*, respectively (Additional file 5: Table S5, Additional file 6: Table S6).

Analysis Of The LSC, SSC, And IR Border Regions

Inversion repeat (IR) is a feature of most plant cp genomes [40]. The cp is roughly divided into three regions, LSC, IR and SSC, among which IR is divided into IRa and IRb, which are symmetrical. The results showed that there was little difference in the length of infrared region among the three species of *L. sinense*, among which the length of *ycf2* gene was different between *L. sinense* and *L. tenuifolia* at the boundary of LSC and IRb. In addition, all *ndhF* gene fragments of them were in SSC region. The length of *ycf1* gene at the boundary between SSC and IRa was 17,607 bp in the SSC region of *L. sinense* and 17,692 bp in the SSC region of *L. jeholense*. The length of *ycf1* gene in SSC region of *L. tenuissimum* is 17,661 bp (Fig. 4). *L. tenuissimum* and the other two are quite different at the boundary of LSC and IRb. After connecting *rps3* gene and *rpl2* gene 2 in LSC region, the SSC region connects *rps19* gene and *rpl2* gene. As in Escherichia coli and Euglena, the *C. reinhardtii rps12* gene is continuous, in contrast to its trans-spliced structure in higher plants [41]. In addition, *L. tenuissimum* separated *ycf1* gene from SSC region in IRb region, and connected 7 bp *trnH* gene with LSC region after *rps19* gene ended at the border of IRa region. In the evolutionary process of cp genome, the length of IR region is not constant [42]. The intron boundary sequence does not follow the G-U / A-G rule, but is similar to the tobacco cp division gene trnagly (UCC) and ribosomal proteins L2 and S12 [43] .

Nucleotide Diversity Analysis

The coding region and noncoding region of gene can better reflect homologous kinship, and these regions are highly variable, according to the coding region and noncoding region of genome to explore and solve the relationship between the same genus of plants. Additionally, fewer SSRs are distributed in the protein-coding sequences compared to the non-coding regions, indicating an uneven distribution of SSRs within the cp genomes [44]. In this experiment, the coding region and noncoding region of two kinds of *L. sinense* were compared and analyzed. 151 coding genes (Fig. 5a) and 152 noncoding genes (Fig. 5b) were generated in cp genome of two kinds of *Ligusticum.* Through the comparative analysis of Pi value (Additional file 7: Table S7), it was found that the Pi value of coding genes was almost zero between 0-0.0087719 (*petG* gene), and 66 genes were found. In the non coding region, the change of Pi value in LSC region is more than that in LSC region, indicating that the coding region is more stable and conservative. The genes of IRa, SSC and IRb in the non-coding region range from 0 to 0.0044843 (*ccsA* < *ndhD* gene), and most of the Pi values are zero, including 51 genes. The first two significant gene mutations were *petG* gene and *psaI*, *ycf4* gene in LSC region.

Conclusion

By comparing the cp genomes of *L. sinense* umbelliforme and *L. jeholense*. We found that they have the same cp genome sequence, mainly different in the length of a single gene and the length of IR region. There are also differences between the two annotated genes, including different types and numbers of an intron, which can be used to identify the two. The difference of cp length of the total gene is that *L. sinense* in IR region isolates the *ycf2* gene at the boundary of LSC and IRb region, while *L. jeholense* completely exists in LSC region, and 83 coding genes and 44 non-coding genes are produced in cp genome of both *L. sinense* and *L. jeholense.* There are obvious differences in the *ycf2* gene which can be used for marker development. There are 88 repeats in the repeats, which can be used to develop markers. Based on the cp genome of Umbelliferae and SNPs of shared coding proteins, the phylogeny of two *L. sinense* species in Umbelliferae was analyzed, which provided a reference for further study on the evolutionary history of Umbelliferae.

Abbreviations

cp: chloroplast; LSC: large single copy; SSC: small single copy; IR: inverted repeat; tRNA: transfer RNA; rRNA: ribosomal RNA.

Declarations

Availability of data and materials

All data generated or analyzed during the course of this study are included in this document or obtained from the appropriate author(s) at reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by National Natural Science Foundation of China (General Program, Grant Numbers 81773852), 2019 Liaoning Provincial Department of Education Scientific Research Project, China (Grant Number L201942), National Key Research and Development in the 13th Five-Year Plan (Grant Number 2018YFC1708200), Major Special Fund for Science and Technology of Inner Mongolia Autonomous Region (Grant Number 2019ZD004), Liaoning University of Traditional Chinese Medicine University Student Innovation and Entrepreneurship Training Program (Grant Number 201910162012), Natural Science Fund Project of Liaoning Province (Grant Number 2020-MS-224) and Key Project at Central Government Level: The Ability Establishment of Sustainable Use for Valuable Chinese Medicine Resources (Grant Number 2060302).

Authors' contributions

Conceptualization, SW; Methodology, SW; Software, TZ; Validation, TZ and ZZ; Formal analysis, ZD; Investigation, TZ; Resources, GB; Data curation, TZ; Writing—original draft, TZ; Writing—review and editing, SS; Visualization, MX; Supervision, SN; Project administration, LX; Funding acquisition, LX and TK. All authors read and approved the final manuscript.

Acknowledgements

We thank the Shanghai BIOZERON Biotechnology Co., Ltd. for processing the raw sequencing data.

Author details

[1] School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian, China. [2] Liaoning Quality Monitoring and Technology Service Center for Chinese Materia Medica Raw Materials, Dalian, China. [3] Traditional Chinese Medicine Resource Center, Chinese Academy of Traditional Chinese Medicine, Beijing, China. [4] School of Mongol Medicine, Inner Mongolia University for Nationalities, Tongliao, China.

Publisher's Note

References

1. Chen DX, Zhang YP, Xu LL, Sun XY. Shennong bencaojing. Fujian: Science and Technology Press; 2012. p. 54.

2. Chinese Pharmacopoeia Commission. Chinese Pharmacopoeia. 342nd ed. Beijing: China Medical Science and Technology Press; 2015. p. 986.
3. Olmstead RG, Palmer JD. Chloroplast DNA systematics: a review of methods and data analysis. Am J Bot. 1994;81:1205–24.
4. Stephen RD, Jeffrey DP. Use of Chloroplast DNA Rearrangements in Reconstructing Plant Phylogeny. Molecular Systematics of Plants. 1992;2:14–35.
5. Jansen RK, Palmer JD. Chloroplast DNA from lettuce and Barnadesia (Asteraceae): structure, gene localization, and characterization of a large inversion. Curr Genet. 1987;11:6–7.
6. Logacheva MD, Valiejo-Roman CM, Degtjareva GV, Stratton JM, Downie SR, Samigullin TH. A comparison of nrDNA ITS and ETS loci for phylogenetic inference in the Umbelliferae: An example from tribe. Tordylieae Molecular Phylogenetics Evolution. 2010;57:471–6.
7. Fulton TM. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol. Biol. Rept. 1995; p. 13.
8. Mcpherson H, Van der MM, Delaney, Edwards SK, Henry MA, Malntosh RJ. E, et al. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. 2013;13:8.
9. BorgstrÖm E, Lundin S, Lundeberg J. Large scale library generation for high throughput sequencing. PLoS ONE. 2011;6:e19119.
10. Cronn R, Liston A, Parks M, Gernandt DS, Shen RK, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 2008;36:e122.
11. Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan ZY, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience. 2012;1:18.
12. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20:3252–5.
13. Lohse M, Drechsel O, Bock R. Organellar genome DRAW (OGDRAW): atool for the easy generation of high-quality custom graphical maps ofplastid and mitochondrial genomes. Curr Genet. 2007;52:267–74.
14. Mayer C, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in Daphnia pulex-a comparative approach. BMC Genome. 2010;11:277.
15. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. RePuter: the manifold applications of repeat analysis on a genomicscale. Nucleic Acids Res. 2001;29:4633–42.
16. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 2000;16:1046–7.
17. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. Newalgorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
18. Minoru U, Masaru F, Shin-ichi A, Jin M, Nobuhiro T, Koh-ichi K. Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene. 2007;402:51–6.
19. Chen W, Xie T, Shao Y, Chen FS. Genomic characteristics comparisons of 12 food-related filamentous fungi in tRNA gene set, codon usage and amino acid composition. Gene. 2012;497:116–24.
20. Liu Q, Xue Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. J Genet. 2005;84:55–62.
21. Yan L, Lai X, Li X, Li XD, Wei CH, Tan XM, et al. Analyses of the Complete Genome and Gene Expression of Chloroplast of Sweet Potato [Ipomoea batata]. PLOS ONE. 2015;10:e0124083.
22. Yi DK, Kim KJ. Complete Chloroplast Genome Sequences of Important Oilseed Crop *Sesamum indicum* L. PLoS ONE. 2012;7:e35872.

23. Degtjareva GV, Logacheva MD, Samigullin TH. Organization of chloroplastpsbA-trnH intergenic spacer in dicotyledonous angiosperms of the family umbelliferae. Biochemistry. 2012;77:1056–64.
24. Gielly L, Taberlet P. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. Molecular biology evolution. 2018;5:769–77.
25. Song M, Guo LY, Pryer KM, Huiet L, Rothfels CJ, Li FW, et al. A novel chloroplast gene reported for flagellate plants. Am J Bot. 2018;105:117–21.
26. Mustafina FU, Dong KY, Choi K, Shin CH, Tojibaev KS, Downie SR. A comparative analysis of complete plastid genomes from Prangos fedtschenkoi and Prangos lipskyi (Apiaceae). Ecology Evolution. 2018;9:7.
27. Grennan CP. Molecular phylogenetic analyses of the chloroplast sequences of an herbaceous bamboo (Cryptochloa strictiflora) and a panic grass (Microcalamus convallarioides). Azarbe Revista Internacional De Trabajo Social Y Bienestar. 2008;64:21–9.
28. Stern DB, Radwanski ER, Kindle KL. A 3' stem/loop structure of the Chlamydomonas chloroplast atpB gene regulates mRNA accumulation in vivo. Plant Cell. 1991;3:285–97.
29. Livore A, Scheuring C, Magill C. The ricepsb-A chloroplast gene has a standard location. Current Genetics. 1989; 16 : 447 – 51.
30. Temnykh S, Declerck G, Lukashova A, Dogu T. Computational and experimental analysis of microsatellites in rice(*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001;11:1441–52.
31. Leseberg CH, Duvall MR. The complete chloroplast genome of Coix lacryma-jobi and a comparative molecular evolutionary analysis of plastomes in cereals. J Mol Evol. 2009;69:311–8.
32. Roy M, Biswas S, Barman S. Identification and analysis of coding and non-coding regions of a DNA sequence by positional frequency distribution of nucleotides (PFDN) algorithm International Conference on Computers & Devices for Communication. 2001; 11: 1441-52.
33. Bucholtz. Frank. Principles of Site-Specific Recombinase (SSR) Technology. Journal of Visualized Experiments. 2008;29:718.
34. Livore A, Scheuring C, Magill C. The ricepsb-A chloroplast gene has a standard location. Curr Genet. 1989;16:447–51.
35. Li HZ, Yin YP, Zhang CQ, Zhang M. Comparison of characteristics of SRAP and SSR markers in genetic diversity analysis of cultivars in Allium fistulosum L. Seed Science Technology. 2008;36:423–34.
36. Mccouch SR, Leonid T, Yunbi X, Katarzyna BL, Karen C, Mark W, et al. Development and Mapping of 2240 New SSR Markers for Rice (*Oryza sativa* L.). DNA Res. 2002;9:199–207.
37. BindlerG, van der Hoeven R, Gunduz I, Plieske J, Ganal M, Rossi L, et al. A microsatellite marker based linkage map of tobacco. TheorAppl Genet. 2007;114:341–9.
38. Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H. The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. Bmc Plant Biology. 2006;6:21–30.
39. Aitman TJ, Hearne CM, Mcaleer MA, Todd JA. Mononucleotide repeats are an abundant source of length variants in mouse genomic DNA. Mammalian Genome Official Journal of the International Mammalian Genome Society. 1991;1:206.
40. Aii J, Kishima Y, Mikami T, Wei CH, Tan XM, Zhang YZ. Expansion of the IR in the chloroplast genomes of buckwheat species is due to incorporation of an SSC sequence that could be mediated by an inversion. Curr Genet. 1997;31:276–9.
41. Liu XQ, Gillham NW, Boynton JE. Chloroplast ribosomal protein gene rps12 of Chlamydomonas reinhardtii. Wild-type sequence, mutation to streptomycin resistance and dependence, and function in Escherichia coli. J Biol Chem. 1989;264:16100–8.
42. Kato A, Kato H, Shida T, Saito T, Komeda Y. Evolutionary Process of the Genomic Sequence Around the 100 Map Unit of Chromosome 1 in Arabidopsis thaliana. Journal of Plant Biology. 2009;52:616–24.

43. Sugita M, Shinozaki K, Sugiura M. Tobacco chloroplast tRNALys (UUU) gene contains a 2.5-kilobase-pair intron: An open reading frame and a conserved boundary sequence in the intron. Proceedings of the National Academy of ences. 1985; 82: 3557-61.
44. Jun Q, Jingyuan S, Huanhuan G, Zhu YZ, Xu J, Pang XH. The Complete Chloroplast Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. Plos One. 2013;8:e57607.
45. Yao H, Song J, Ma X, Ma XY, Liu C, Li Y, et al. Identification of DendrobiumSpecies by a Candidate DNA Barcode Sequence: The ChloroplastpsbA-trnHIntergenic Region. Planta Med. 2009;75:667–9.
46. Ravi V, Khurana JP, Tyagi AK, Khurana P. An update on chloroplast genomes. Plant Syst Evol. 2008;271:101–22.
47. Huelsenbeck JP, Bollback JP, Levine AM. Inferring the Root of a Phylogenetic Tree. Syst Biol. 2002;51:32–43.
48. Plunkett GM, Downie SR. Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. Syst Bot. 2000;25:648–67.
49. Lee TH, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genom. 2014;15:162.

Additional Files

Additional file 1: Table S1. Location and length of intron-containing cp genes within the two Umbelliferae species.

Additional file 2: Table S2. Type and abundance of different SSRs in two species of *Ligusticum*.

Additional file 3: Table S3. SSRs distribution of the *Ligusticum sinense* cp genome.

Additional file 4: Table S4. SSRs distribution of the *Ligusticum jeholense* cp genome.

Additional file 5: Table S5. Large repeats identified in the *Ligusticum sinense* cp genome.

Additional file 6: Table S6. Large repeats identified in the *Ligusticum jeholense* cp genome.

Additional file 7: Table S7. Pi values of the coding and no-coding regions in the two *Ligusticum* cp genomes.

Figures

Figure 1

Circular gene map of L. sinense genes on the outside circle are transcribed counterclockwise, while genes on the inside circle presented clockwise. LSC, large single copy; SSC, small single copy; IRa, inverted repeat A; IRb, inverted repeat B.

Figure 2

Circular gene map of L. jeholense genes on the outside circle are transcribed counterclockwise, while genes on the inside circle presented clockwise. LSC, large single copy; SSC, small single copy; IRa, inverted repeat A; IRb, inverted repeat B

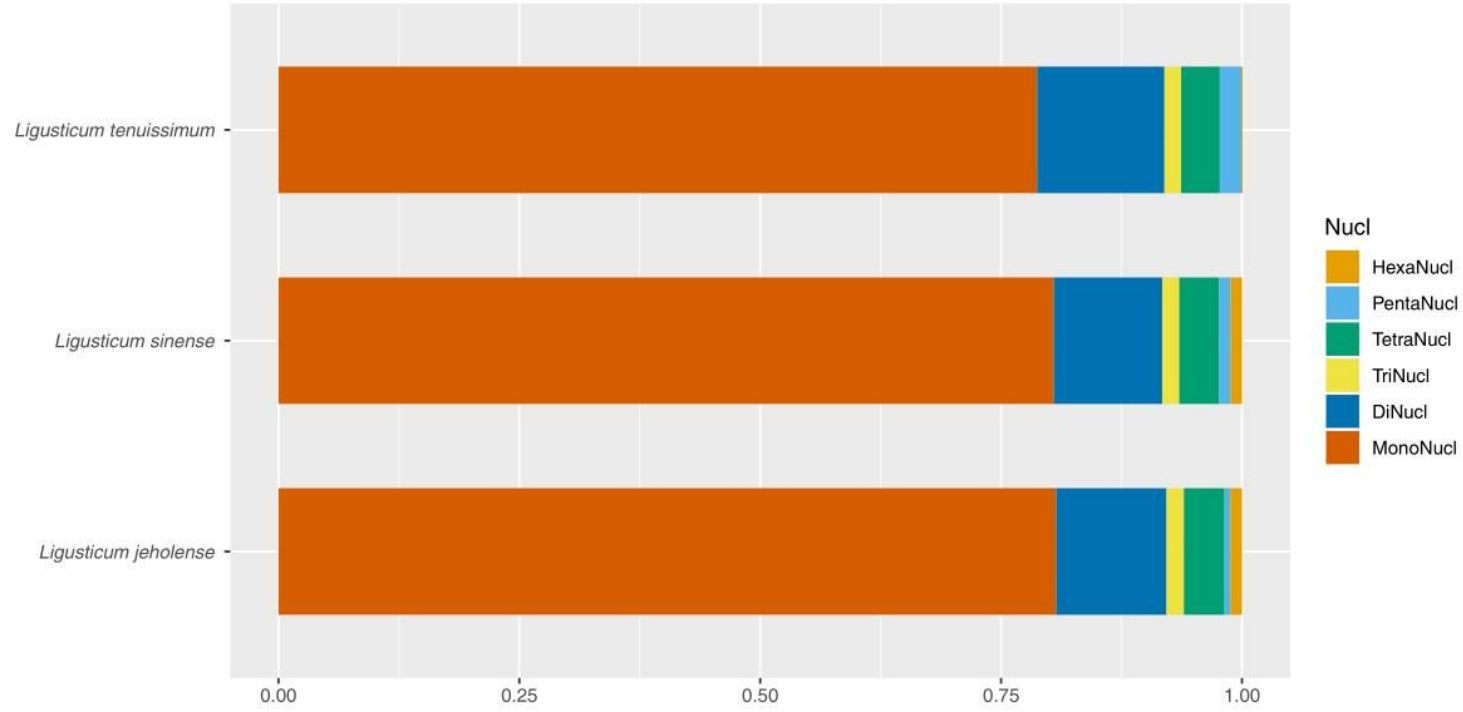

Figure 3

SSRs in the chloroplast genomes of 3 species in Umbelliferae. MonoNucl represents mononucleotide repeats, DiNucl represents dinucleotide repeats, TriNucl represents trinucleotide repeats, TetraNucl represents tetranucleotide repeats, PentaNucl represents pentanucleotide repeats and HexaNucl represents hexalnucleotide repeats

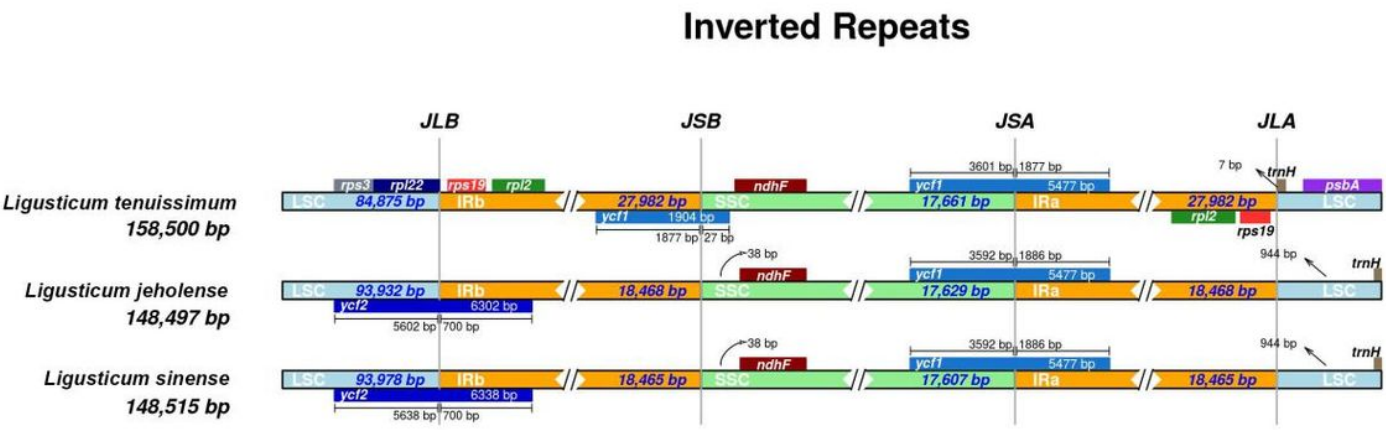

Figure 4

Comparisons of LSC, SSC, and IR border regionsamong the chloroplast genomes of 3 species in Umbelliferae. Number above the gene features means the distance between the ends of genes and the borders sites

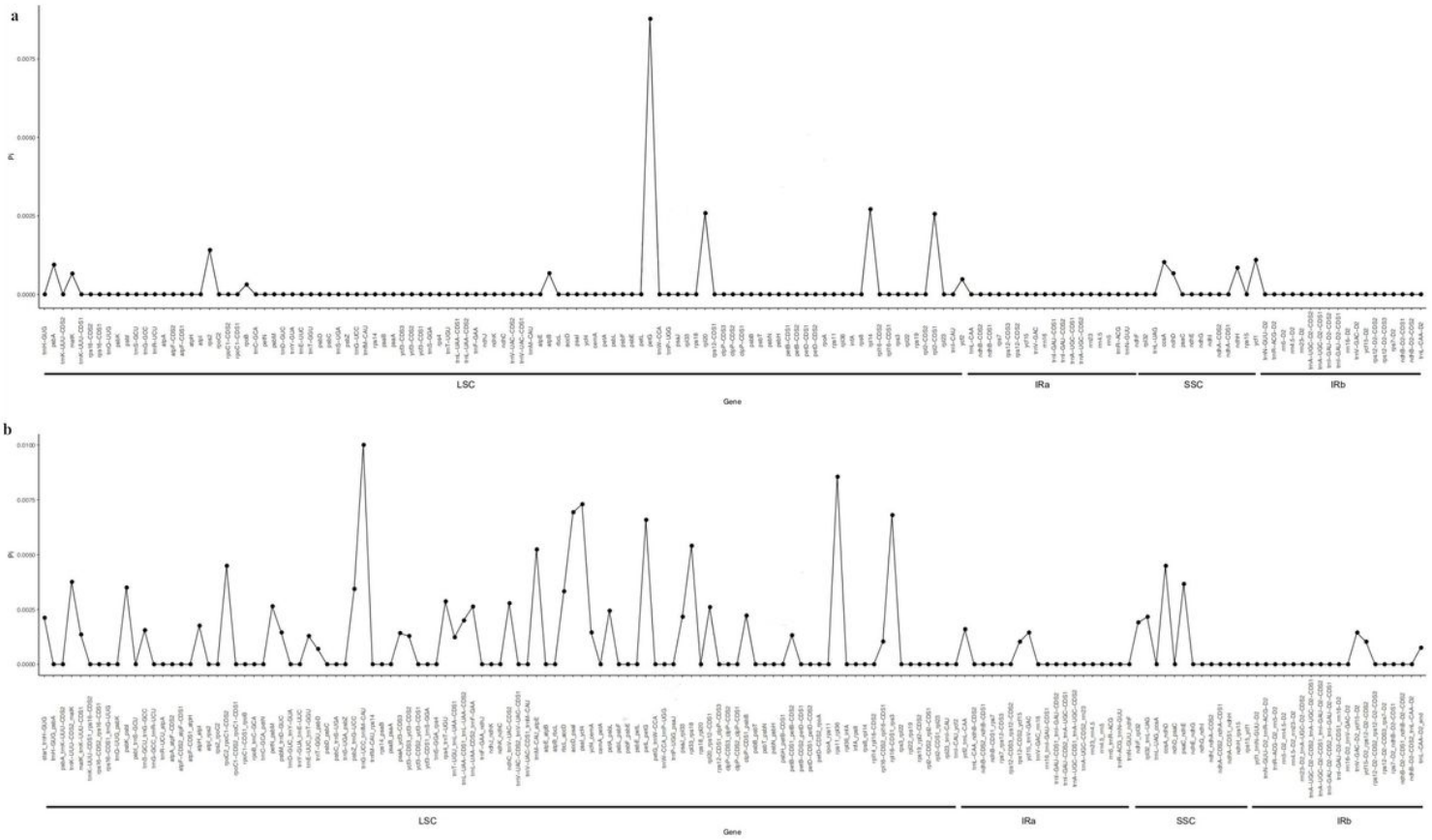

Figure 5

Comparative analysis of the nucleotide variability by Pi values of two speciesof Ligusticum (a coding region, b non-coding region)

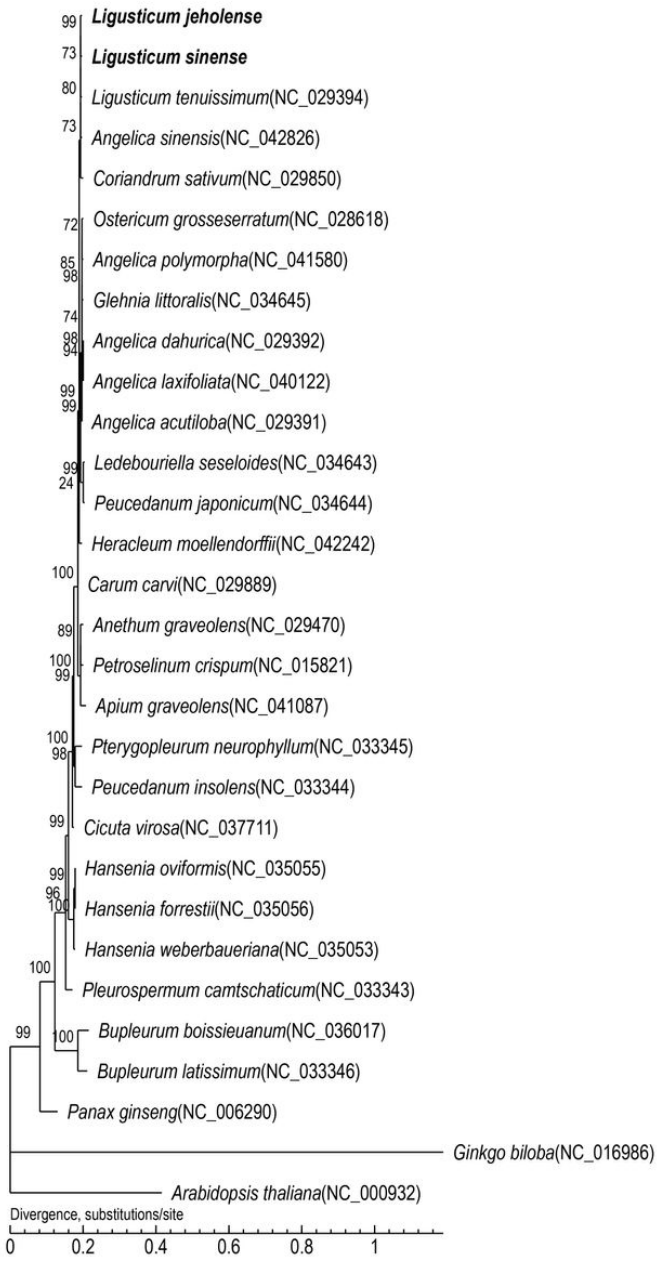

Figure 6

Molecular phylogenetic tree of 30 species of whole genome SNPs. The tree was constructed using a maximum likelihood analysis using PhyML v3.0 with a bootstrap of 1000 replications

Supplementary Files

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

- Additionalfile7.docx
- Additionalfile6.docx
- Additionalfile5.docx
- Additionalfile4.docx
- Additionalfile3.docx

- Additionalfile2.docx
- Additionalfile1.docx