PLOS ONE

RESEARCH ARTICLE

Comparative study on chloroplast genomes of three *Hansenia forbesii* varieties (Apiaceae)

Chenghao Zhu[1], **Yuan Jiang**[1], **Yu Bai**[1], **Shengjian Dong**[2], **Sun Zhirong**[1]*

1 School of Chinese Materia Medica, Beijing University of Chinese Medicine. Beijing, Beijing, China,
2 College of Applied Technology, Gansu Agricultural University. Lanzhou, Gansu, China

* szrbucm67@163.com

Citation: Zhu C, Jiang Y, Bai Y, Dong S, Zhirong S (2023) Comparative study on chloroplast genomes of three *Hansenia forbesii* varieties (Apiaceae). PLoS ONE 18(6): e0286587. https://doi.org/10.1371/journal.pone.0286587

Editor: Shyamal Krishna Talukder, Texas A&M University and Texas A&M Agrilife Research, UNITED STATES

Received: February 19, 2023

Accepted: May 18, 2023

Published: June 1, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0286587

Data Availability Statement: The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at

Abstract

To find the gene hypervariable regions of three varieties of *Hansenia forbesii* H. Boissieu and determine their phylogenetic relationship, the chloroplast (cp) genome of these three varieties were firstly sequencing by the Illumina hiseq platform. In this study, we assembled the complete cp genome sequences of *Hansenia forbesii* LQ (156,954 bp), *H. forbesii* QX (157,181 bp), *H. forbesii* WQ (156,975 bp). They all contained 84 protein-coding genes, 37 tRNAs, and 8 rRNAs. The hypervariable regions between three cp genomes were *atp*F-*atp*H, *pet*D, and *rps*15-*ycf*1. Phylogenetic analysis showed that *H. forbesii* LQ and *H. forbesii* WQ were closely related, followed by *H. forbesii* QX. This study showed that the three varieties of *H. forbesii* could be identified by the complete cp genome and specific DNA barcode (*trn*C-GCA-*pet*N) and provided a new idea for germplasm identification of similar cultivated varieties.

Introduction

Hansenia forbesii H.Boissieu is a rare and endangered perennial herb belonging to the genus *Hansenia* (Apiaceae), and the Chinese name is "Kuan-ye-qiang-huo". The main chemical components isolated from *H. forbesii* are volatile oils, terpenoids, coumarins, sugars and glycosides, alkynes, phenolic acids, steroids, flavonoids, and alkaloids. It has anti-inflammatory, antibacterial, antioxidant, antiviral, anticancer, antipyretic, and analgesic activities, and has a significant impact on the cardiovascular and cerebrovascular, digestive, respiratory, and central nervous systems [1–4]. In the traditional production areas of *Hansenia*, the loss of germplasm resources is an extremely serious problem. As the problem of artificial breeding has not been solved, it is difficult to collect enough seeds, roots, and other reproductive materials in the genuine production areas and even enough genetic standard materials, showing that the severity of the risk of endangerment of *Hansenia* cannot be ignored. The wild-growing to domestic cultivation for *Hansenia* and the establishment of a standardized production base are the most effective ways to ensure the quality of medicinal materials, meet the medicinal requirements, protect wild resources and achieve sustainable development. Due to the wide distribution range and low altitude of *H. forbesii*, the requirements for environmental and soil conditions are relatively loose, the root system is well developed and the medicinal value is the same as that of *H. weberbaueriana*. Therefore, vigorously promoting the artificial cultivation of *H. forbesii* can reduce the pressure on this resource.

[https://www.ncbi.nlm.nih.gov] under accession no. ON 208134, ON208135, ON208136.

Funding: the Ministry of Finance and Ministry of Agriculture and Rural Areas: special subsidy of national modern agricultural industrial technology system (cars-21); Gansu Dingxi Science and Technology Project (dx2018n04). The sponsor is responsible for publishing decisions and guiding manuscript revisions in this study.

Competing interests: We declare that there is no conflict of interest regarding the publication of this article.

The main distribution types of *H. forbesii* are purple stem type, green stem type, large leaf type, small leaf type, root single branch type, root dispersed type, and other mixed species. The excellent varieties of *H. forbesii* have not been popularised and applied in production and cultivation, which cannot meet the needs of its industrial development. Moreover, there is a serious problem of mixed varieties in field production, which has a significant impact on its high-yield cultivation and is the bottleneck for the development of the *H. forbesii* industry. Before this study, we tried to address these issues by the following actions. Taking the wild *H. forbesii* (Wei-Qiang) resources in Gansu Province as the original materials, using the single plant optimization method and taking the plant agronomic characters, susceptibility, and drug grade as the indicators, we have selected two new varieties of *H. forbesii* (Long-Qiang and Qiang-Xuan) with high quality and high yield through line identification, line comparison, multi-point test, and production test. But it is not clear what kind of variation exists at the gene level and what is their phylogenetic relationship.

Chloroplast is an essential organelle in plant cells and it plays an important role in photosynthesis, carbon sequestration, and amino acid synthesis. Its complete genome contains a large amount of genetic information and its structure is highly conserved, including gene content, gene organization, and intron content within genes [5]. Chloroplast (cp) genome has also undergone intron loss, gene loss, gene duplication, gene rearrangement, pseudogenization, as well as uneven contraction and expansion of IRs regions. These genomic events can in turn become synapomorphies for entire clades [6]. In addition, many mutational events take place within the chloroplast genome including substitutions, insertions-deletions, inversions, microstructural changes, and oligonucleotide repeats [7–9]. The cp genome is slow evolving [10] and inherited from a single parent: maternally in most angiosperms [11] and paternally in some gymnosperms [12]. It is widely used in the study of plant molecular evolution and phylogeny, as well as in genetic transformation, genetic engineering, and molecular breeding [13–15]. Recently, with the gradual popularization and low cost of high-throughput sequencing technology, some scientists have proposed that cp genome data can be used for species identification and interspecific differentiation research [16–20]. There is no significant difference in some appearance indicators among the three *H. forbesii*. Through comparative analysis of their cp genome, the differences between them can be identified at the genetic level, and polymorphism markers can be developed to achieve variety identification. This is necessary for solving the problem of mixed varieties of *H. forbesii*, and has good significance for promoting excellent variety.

To find the gene hypervariable regions of three varieties of *H. forbesii* and determine their phylogenetic relationship, we used Illumina hiseq technology to sequence the cp genome. And we compared the characteristics of the cp genome, screened the different nucleotide sequences, and deciphered the genetic relationship to provide the basis for the differentiation of target varieties with excellent and optimal shapes.

Materials and methods

Materials and pretreatment

The fresh, tender, and healthy leaves of three *H. forbesii* were collected from the germplasm resource garden in Taoyang Town, Lintao County, Gansu Province in China on 30 September 2021. Professor Zhirong Sun, a scientist in the national traditional Chinese medicine system, identified that all the samples were *Hansenia forbesii* H.Boissieu (Apiaceae). The plant materials produced and used in this study comply with China guidelines and legislation. All the experiments were carried out to national and international guidelines. The aboveground forms of three *H. forbesii*, are shown in Fig 1 and described in S1 Table, which are named QX,

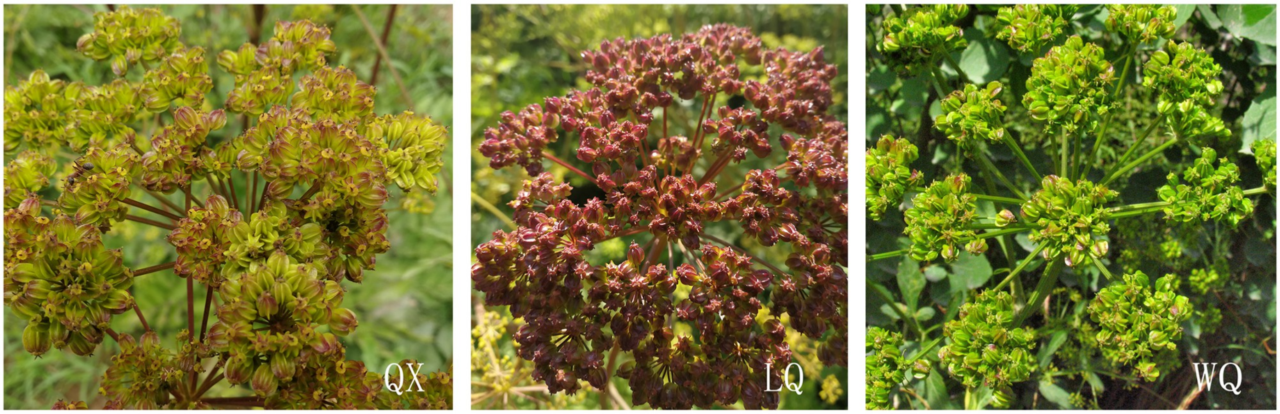


Fig 1. The aboveground forms of three *H. forbesii*.

https://doi.org/10.1371/journal.pone.0286587.g001

LQ, and WQ according to their morphological characteristics and breeding process. After cutting off the veins of leaves with clean scissors, we put the remaining leaves into molecular bags, added silica gel, and saved them. The voucher specimens (QH210930) of all materials were kept in the herbal medicine library of the School of Chinese Materia Medica, Beijing University of Chinese Medicine.

Total DNA extraction and library preparation

The total DNA was extracted using a small plant DNA Extraction Kit (Omega Bio-Tek, USA). Then we randomly broke qualified DNA samples using an ultrasonic crusher (Covaris, USA), and then completed the whole library preparation through the steps of end repair, adding A-tailing, adding sequencing connector, purification, and PCR amplification [21]. The qualified library was sequenced using the Illumina high-throughput sequencing platform (Illumina, NovaSeq 6000, USA).

Chloroplast genome assembly and annotation

The high-quality reads were assembled using Get Organelle v.4.0 and then annotated by CpGAVAS2 [22]. The annotations of tRNA genes were confirmed by using tRNAscan-SE v.2.03 [23].

The cp genome of LQ, QX, and WQ were submitted to GenBank at the National Center of Biotechnology Information (NCBI), and the accession numbers were ON 208134, ON208135, and ON208136, respectively. Fully annotated plastome circular diagrams were drawn by OrganellarGenomeDRAW (OGDRAW) [24, 25].

Codon usage

The protein-coding genes were extracted by Phylosuite v.1.2.2 [26]. Relative synonymous codon usage (RSCU) and codon usage values were analyzed by CodonW v.1.4.2. Moreover, the RSCU values were shown in a heatmap by TBtools [27].

Repeat analysis and comparative analyses

Repetitive sequence analyses were performed using CPGAVAS2 analysis. Tandem repeats were identified using default settings by Tandem Repeats Finder [28]. The Misa. pl was used to screen the simple sequence repeats (SSRs) [29]. The scattered repetitive sequences were found by using VMATCH. The REPuter was used to determine the size and location of the oligonucleotide repeats (ORs) [30]. The complete cp genome of three cultivated varieties was compared by mVISTA [31], and the genome of *H. forbesii* (NC035054) was used as the reference sequence for annotation. Sliding window analysis was conducted to assess the nucleotide diversity (Pi) values of the cp genome by DnaSP v6 (window length = 300 bp, step size = 25 bp). IRscope [32] was used to analyze inverted repeated traction and expansion at cp genome junctions.

Identification and validation of barcode for varieties discrimination

According to the results of DNAsp, we chose the high variation region to distinguish the three varieties. Primers to discriminate between the three varieties under study were designed on the variable intergenic regions using Snapgene 6.2.1 (Snapgene from Insightful Science, available at http://www.snapgene.com, last used in 2023). PCR amplifications were performed in a final volume of 20 μL with 10 μL 2×Taq PCR Master Mix, 0.5 μM of each primer, 5 μL template DNA, and 4 μL ddH$_2$O following the manufacturer's instructions (Mei5 Biotechnology, Co., Ltd). All amplifications were carried out in a Pro-Flex PCR system (Applied Biosystems, Waltham, MA, USA) under the following conditions: denaturation at 95 ˚C for 3 min, followed by 36 cycles of 94 ˚C for 25 s and 55 ˚C for 10 s, and 72 ˚C for 2 min as the final extension following the manufacturer's instructions (Mei5 Biotechnology, Co., Ltd). PCR amplicons were visualized on 1% agarose gels, purified, and then subjected to bidirectional Sanger sequencing on an ABI 3730 XL instrument (Applied Biosystems, USA) using the same set of primers used for PCR amplification with BigDye v3.1 chemistry (Applied Biosystems) following manufacturer's instructions. All amplifications were repeated three for each variety.

Phylogenetic analysis

Phylogenetic analysis was performed based on 37 complete cp genomes, including the three assembled sequences in our study, 34 cp genomes (30 Genus, 2 Family) downloaded from the NCBI and *Cornus officinalis* (NC042746) as outgroup. The shared protein-coding genes were extracted using Phylosuite v.1.2.2 software, and the MAFFT v.7.307 was used for alignment [33]. Subsequently, the alignment was conducted using the GTR+I+Gevolution model based on Bayesian inference (BI) in MrBayes [34]. The parameter was set to run for five million generations and sampled every 1,000 generations, with all other settings left at their defaults, and the first 25% of each run was discarded as burn-in. The alignment was also evaluated using bootstrap analysis on 1000 in a maximum likelihood (ML) by RAxML [35], with parameters: raxmlHPC-PTHREADS-SSE3 -fa -N 1000-m GTRGAMMA -x 551,314,260 -p 551,314,260 -o *Cornus_officinalis* _NC_042746-T 20, 1000 replications and best-fit model selection.

Results

Chloroplast genome information

The basic information of the cp genome of three *H. forbesii* is shown in Table 1 and compared with other species of the *Hansenia* genus and *Arabidopsis thaliana* (S2 Table). The length of the cp genome of *H. forbesii* QX was the longest in three varieties (157,181 bp), the total GC content of three cp genomes was about 37.6%, and they all had a typical four-region structure, in which the LSC (large single-copy region) was 86,167 bp (LQ), 86,361 bp (QX), and 86,188 bp (WQ), the SSC (small single-copy region) was 17,755 bp (LQ), 17,786 bp (QX) and 17,755

Table 1. Basic information of cp genome of three varieties of *H. forbesii*.

Name	Overall length	LSC	SSC	IR	GC content (%)	protein-coding genes	tRNA genes	rRNA genes
H.f. LQ	156,954	86,167	17,755	26,516	37.68	84	37	8
H.f. QX	157,181	86,361	17,786	26,517	37.66	84	37	8
H.f. WQ	156,975	86,188	17,755	26,516	37.68	84	37	8

GC: Percentage of G/C base number in total base number.

https://doi.org/10.1371/journal.pone.0286587.t001

bp (WQ) and the IR (inverted repeats, including IRa and IRb) was 26,516 bp (LQ), 26,517 bp (QX), and 26,516 bp (WQ) (Fig 2). A total of 129 genes were found in these varieties, including 84 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Table 2). The number and types of introns were similar among the three *H. forbesii*. The four types of cp genes contained 45 photosynthesis-related genes, 29 self-replication-related genes, six other genes, and four unknown genes. Amongst them, *ndh*B, *rpl*2, *rpl*23, *rps*12, *rps*7, and *ycf*2 appeared twice. Eleven genes each contained one intron, including *rpl*2 (×2) and *ndh*B (×2), which were located in the IR, and the genes (*rps*16, *atp*F, *rpo*C1, *pet*B, *pet*D, and *rpl*16) were located in the LSC, and the *ndh*A was the only present in the SSC region. In addition, the *ycf*3 and *clp*P comprise two introns.

Comparison of chloroplast introns and exons

By analyzing the cp genome annotation files of the three varieties of *H. forbesii*, we found that the introns and exons of 21 genes in the cp genome of the three varieties were different. QX had one more bp than LQ and WQ in terms of the introns in the *trn*K-UUU, *rps*16, *atp*F, *ycf*3, *clp*P, and *rpl*16, but one less bp in *rpo*C1 (S3 Table). Studies have shown that introns can reduce gene splicing and translation, slow down cell metabolism and reduce energy consumption [36] and the *Clp*P gene is also related to plant disease resistance, which may be the reason why QX has better disease resistance than LQ and WQ.

Codon usage

The cp genome of three *H. forbesii* contained 64 codons encoding 20 amino acids. The result of the RSCU revealed that 32 codons were used frequently in these cultivars, with the highest frequency of UUA followed by AGA (Fig 3). Moreover, the codon exhibited a strong bias toward an A or T at the third position. The codons that contain A/T at the 3' end mostly have RSCU ≥1, whereas the codons are having C or G at the 3' end mostly have RSCU ≤1. Amino acid frequency analyses revealed the highest frequency of Leucine, whereas Tryptophane was a rare amino acid. In general, we found high similarities in codon usage and amino acid frequency among the three cultivated varieties, and both contain high AT content.

Repeat analysis

Our analyses identified SSRs per genome composed of mono- to di- nucleotide repeating units (Fig 4a). The number and type of SSRs in LQ and WQ were similar, with 34 single nucleotide repeats and 2 dinucleotide repeats. QX contains 39 single nucleotide repeats more than LQ and WQ. Moreover, the main type of mononucleotide repeats was T. Oligonucleotide repeats in three varieties. Fig 4b showed that the number of repeats varied in three varieties. We discovered 49 repeats in three varieties (LQ: 26 forward and 23 palindromic, QX: 30 forward and 19 palindromic, and WQ: 26 forward and 23 palindromic). Most of the repeats ranged in size from

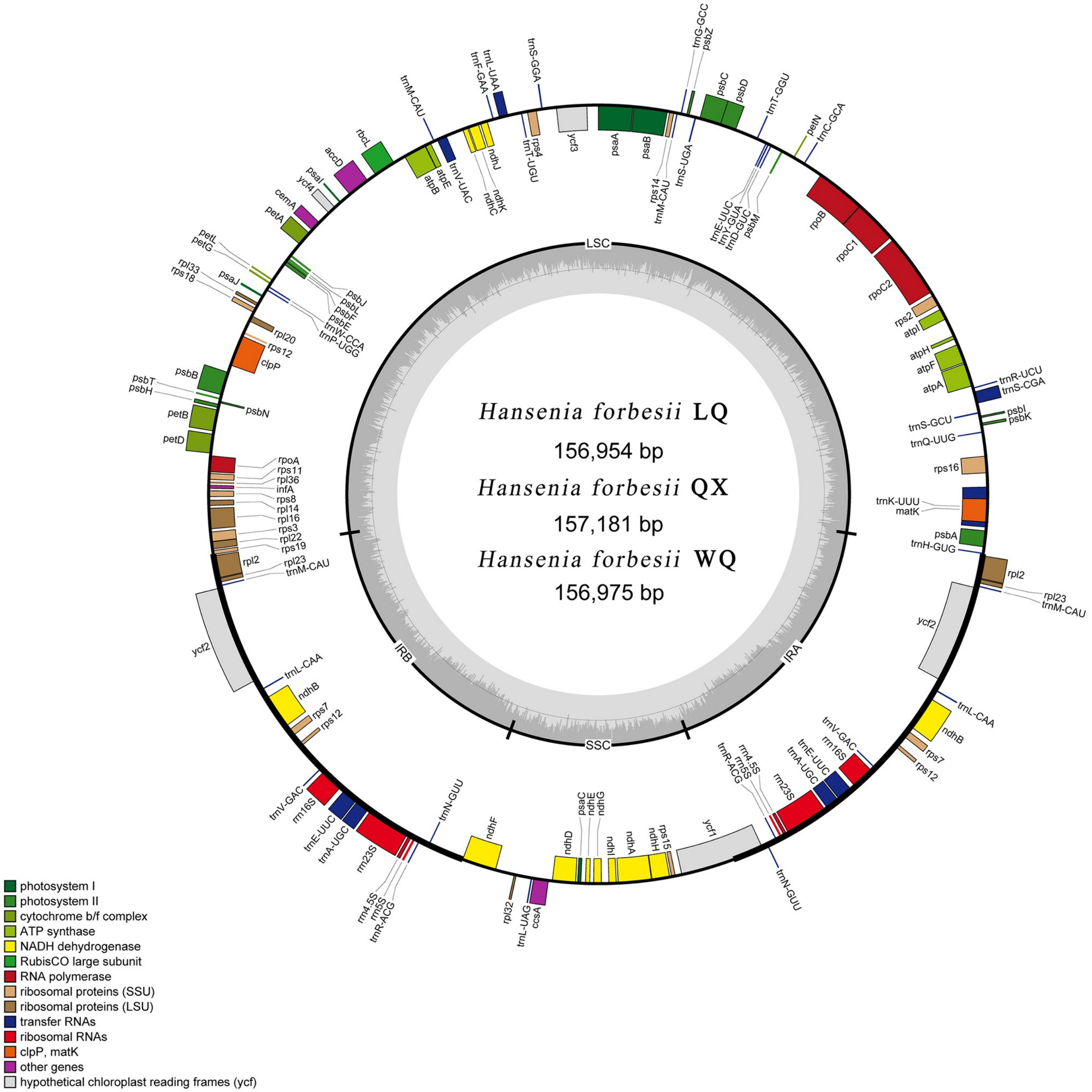


Fig 2. Cp genome map of three varieties of *H. forbesii*. Genes lying outside the circle are transcribed in the counterclockwise direction, while those insides are transcribed in the clockwise direction. The colored bars indicate different functional groups. The darker grey area in the inner circle denotes GC content, while the lighter grey corresponds to the AT content of the genome.

https://doi.org/10.1371/journal.pone.0286587.g002

30 to 40 bp. This result showed that QX was more different than LQ and WQ. We also evaluated the number of repeats about the species' phylogenetic position using the topology in Fig 9.

Inverted repeats contraction and expansion

The inverted repeats contraction and expansion revealed variation at LSC/IRs/SSC junctions (Fig 5). In three varieties, a truncated copy of the *rps*19 gene was found at the IRb/LSC

Table 2. List of genes in the cp genome of three varieties of *H. forbesii*.

Category of genes	Group of genes	Name of genes	Number
rRNA	rRNA	rrn16S (×2), rrn23S (×2), rrn5S (×2), rrn4.5S (×2)	8
tRNA	tRNA	trnY-GUA, trnW-CCA, *trnV-UAC, trnV-GAC(×2), trnT-UGU, trnT-GGU(×2), trnS-UGA, trnS-GGA, trnS-GCU, *trnS-CGA, trnR-UCU, trnR-ACG(×2), trnQ-UUG, trnP-UGG, trnN-GUU (×2), trnM-CAU(×4), trnL-UAG,* trnL-UAA, trnL-CAA(×2),* trnK-UUU, trnH-GUG, trnG-GCC, trnF-GAA, *trnE-UUC(×2), trnD-GUC, trnC-GCA,*trnA-UGC(×2)	37
Genes for photosynthesis	Subunits of ATP synthase	atpA, atpB, atpE, *atpF, atpH, atpI	6
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, **ycf3	15
	Subunits of NADH-dehydrogenase	*ndhA, *ndhB (×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK	12
	Subunits of cytochrome b/f complex	petA, *petB, *petD, petG, petL, petN	6
	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ	5
	Subunit of rubisco	rbcL	1
Self replication	Large subunit of ribosome	rpl14, *rpl16, *rpl2(×2), rpl20, rpl22, rpl23(×2), rpl32, rpl33, rpl36	11
	DNA dependent RNA polymerase	rpoA, rpoB, *rpoC1, rpoC2	4
	Small subunit of ribosome	rps11, rps12(×2), rps14, rps15, *rps16, rps18, rps19, rps2, rps3, rps4, rps7(×2), rps8	14
Other genes	Subunit of Acetyl-CoA-carboxylase	accD	1
	c-type cytochrom synthesis gene	ccsA	1
	Envelop membrane protein	cemA	1
	Protease	**clpP	1
	Translational initiation factor	infA	1
	Maturase	matK	1
Unkown	Conserved open reading frames	ycf1, ycf2 (×2), ycf4	4

*Gene with one intron,

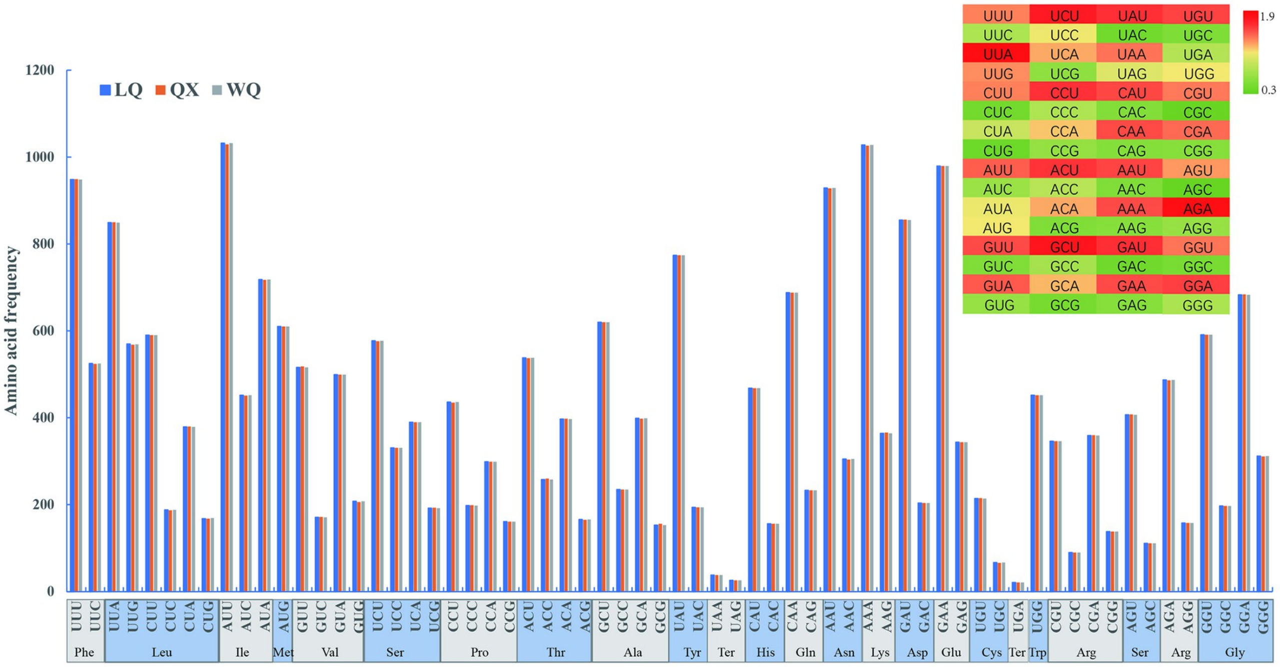
**Gene with two introns.

https://doi.org/10.1371/journal.pone.0286587.t002

junction; the *rpl*22 gene was found entirely in the LSC region; and the *rpl*2 gene was found entirely in the IRb region. Another truncated copy of the *ndh*F gene was found at the junction of IRb/SSC in all species, which starts in IRb regions and integrates into the SSC region. Moreover, a truncated copy of *ycf*1 was found in the SSC/IRa junction, which was longer in IRa of *H. forbesii*. In three varieties, *trn*H completely exists in LSC and only 3 bp from the junction of IRa/LSC. These results showed that the chloroplast genome of these three varieties did not expand or contract.

Comparative cp genomic analysis

The cp genome of the three varieties was compared by mVISTA, and the *H. forbesii* (NC035054) was used as the reference sequence for annotation. The three varieties exhibited similar variation sites and degrees of variation (Fig 6). The coding regions (CDS) were more conserved than the intergenic spacers (IGS). The high divergence area in IGS was found in *pet*N-*psb*M, *trn*Y(GUA)-*trn*T(GGU), *ndh*C-*trn*V(UAC), *pet*A-*psb*J, *rps*8-*rpl*14, *rrn*4.5-*rrn*5. Furthermore, some mutations of CDS were found in *rpo*C1 and *rpl*2. Moreover, the result showed that IR regions had lower sequence divergence than LSC and SSC regions.



UUU	UCU	UAU	UGU
UUC	UCC	UAC	UGC
UUA	UCA	UAA	UGA
UUG	UCG	UAG	UGG
CUU	CCU	CAU	CGU
CUC	CCC	CAC	CGC
CUA	CCA	CAA	CGA
CUG	CCG	CAG	CGG
AUU	ACU	AAU	AGU
AUC	ACC	AAC	AGC
AUA	ACA	AAA	AGA
AUG	ACG	AAG	AGG
GUU	GCU	GAU	GGU
GUC	GCC	GAC	GGC
GUA	GCA	GAA	GGA
GUG	GCG	GAG	GGG

Fig 3. The RSCU values of all protein-coding genes for three varieties. Color key: the red values indicate higher RSCU values and the green values indicate lower RSCU values.

https://doi.org/10.1371/journal.pone.0286587.g003

We used the DnaSP software to compare the nucleotide variation values (Pi) between the whole cp genome of the three varieties. The hypervariable regions were detected, and the sequence differences were analyzed. Sliding window analysis revealed that the nucleotide diversity values varied from 0 to 0.008. Three mutational hotspots in the LSC and SSC regions were identified, including *atp*F-*atp*H, *pet*D, and *rps*15-*ycf*1 (Fig 7). These can be used as potential sites for studying population genetics and the identification of these three varieties.

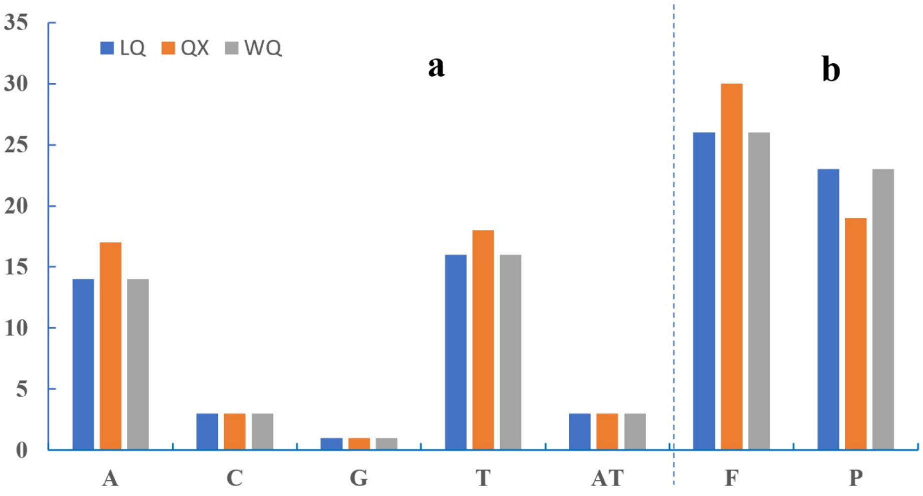


Fig 4. Comparison of repeats in three varieties. (a). SSR distributed situation in the cp genome of five species. (b). Long repeats classification of five species. P-palindromic repeats; F- forward repeats.

https://doi.org/10.1371/journal.pone.0286587.g004

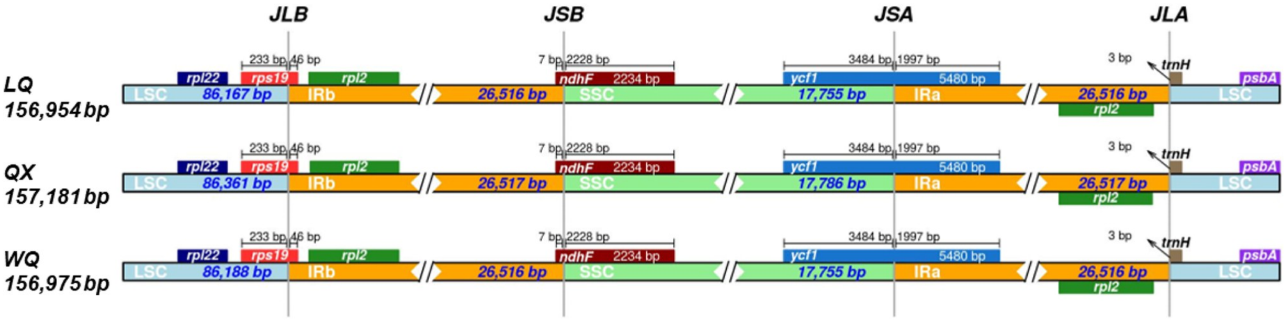


Fig 5. Comparison of quadripartite junction sites in three varieties cp genome. Gene transcribed clockwise are presented below the track, whereas transcribed counterclockwise are presented on top of the track. The start and end of each gene from the junctions have been shown with arrows. The T scale bar above or below the track shows genes integrated from one region of the cp to another. JLB (IRb /LSC), JSA (SSC/IRa), JSB (IRb/SSC), and JLA (IRa/LSC) denote the junction sites between the quadripartite regions of the genome.

https://doi.org/10.1371/journal.pone.0286587.g005

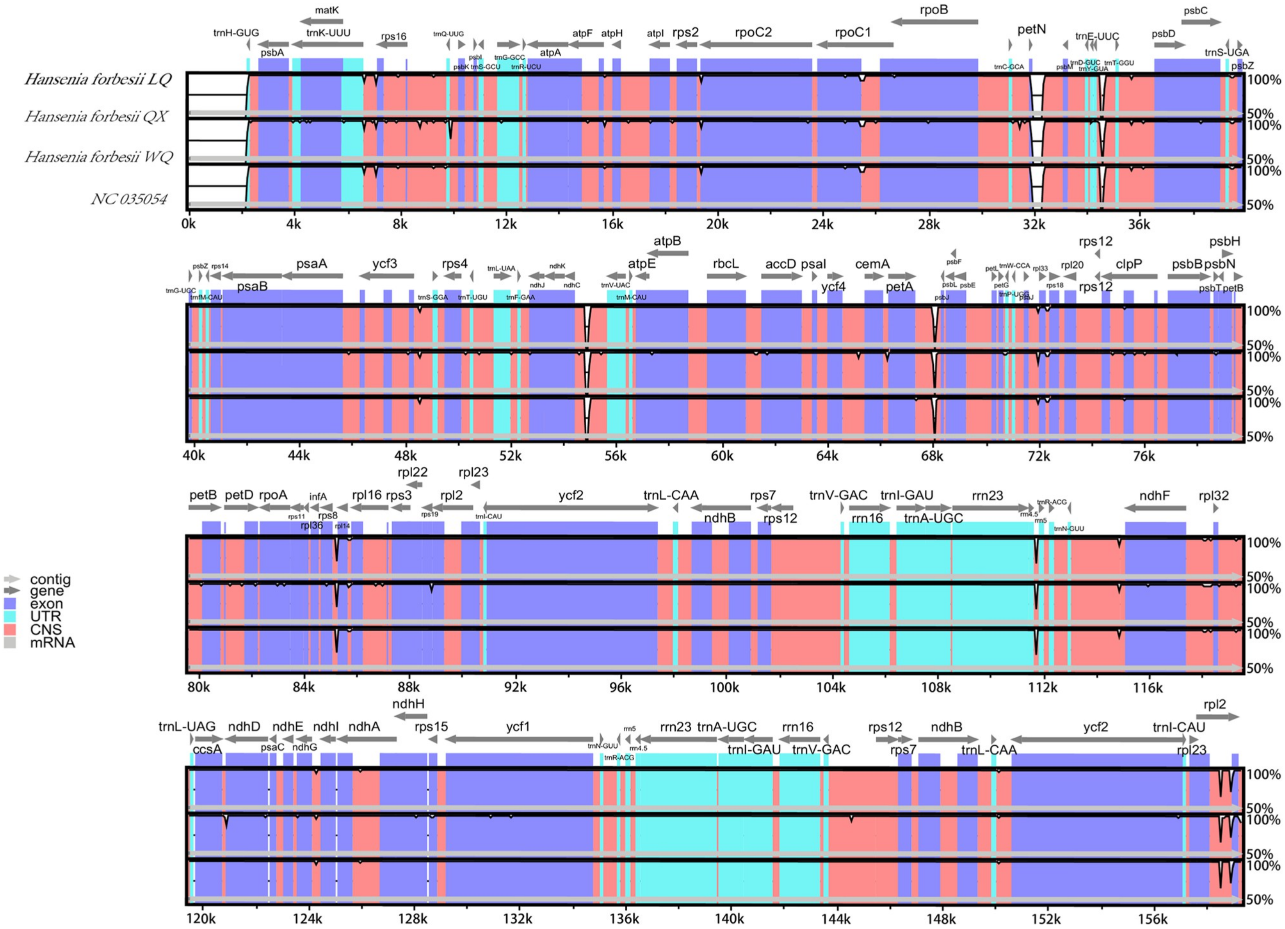


Fig 6. Comparison of three varieties of cp genome using *H. forbesii* (NC035054) annotation as a reference. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The horizontal axis shows the coordinates within the cp genome. Genome regions are color-coded as exons, introns, and intergenic spacer (IGS), and the Gray arrows indicate the direction of transcription of each gene. Annotated genes are displayed along the top.

https://doi.org/10.1371/journal.pone.0286587.g006

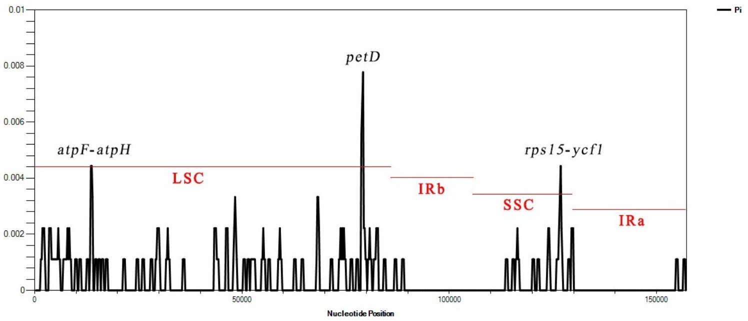


Fig 7. Sliding window test of nucleotide diversity (Pi) in the multiple alignments of the three *H*. *forbesii* cp genomes. Peak regions with a Pi value of>0.004 were labeled with loci tags of the genic or intergenic region names.

https://doi.org/10.1371/journal.pone.0286587.g007

Specific DNA barcode maker design for three varieties

To discriminate the three varieties, we selected two hypervariable regions, *trn*C-GCA—*pet*N (29,207–29,893 bp) and *trn*Q-UUG—*psb*K (7,894–8,266 bp), to develop specific DNA barcode. The related primer sequence is shown in S4 Table. PCR amplification of total DNA from three varieties resulted in products having the expected size (S1 Fig). In terms of the amplification of primer 2, QX exhibited more polymorphic loci compared to LQ and WQ. Under the amplification of primer 1, all three varieties showed a single bright band. The DNA fragments were extracted from each band and then subjected to Sanger sequencing. The sequencing results were identical to the expected sequences (Fig 8). The barcode has two specific Indel loci. These two variable loci can be used to differentiate three varieties.

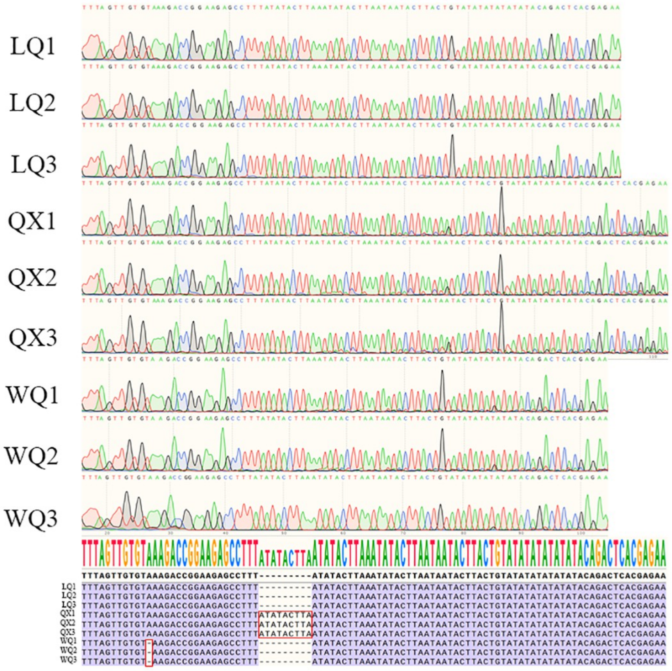


Fig 8. Sequencing chromatograms of the barcode regions from three varieties with consensus sequence and alignment.

https://doi.org/10.1371/journal.pone.0286587.g008

Phylogenetic analysis

The phylogenetic trees (ML and BI) were constructed by the shared protein-coding genes to investigate the phylogeny of the three *H. forbesii*, and *C. officinalis* (NC042746) was the outgroup. Most nodes were supported with maximum support (100% bootstrap support) on the consensus trees, and the topological structure of the two phylogenetic trees had perfect consistency (Fig 9). Three varieties of *H. forbesii* clustered with *Hansenia forbesii* (NC035054) and then merged with *H. forrestii* (NC035056) and *H. oviformis* (NC035055), indicating that LQ and WQ were closely related, followed by QX, and they were closely related to *Hansenia forbesii* (NC035054). Finally, they became a clade with *Changium smyrnioides* (NC053938), *Tongoloa silaifolia* (NC049062), *H. weberbaueriana* (NC035053), *Haplosphaera himalayensis* (NC056096). This clade had a close genetic relationship, in which QX was relatively distant from LQ and WQ.

Discussion and conclusion

This study found that the three varieties of *Hansenia forbesii* had similar genetic relationships. Only the length of introns of some *H. forbesii* QX genes was longer than *H. forbesii* LQ and *H.*

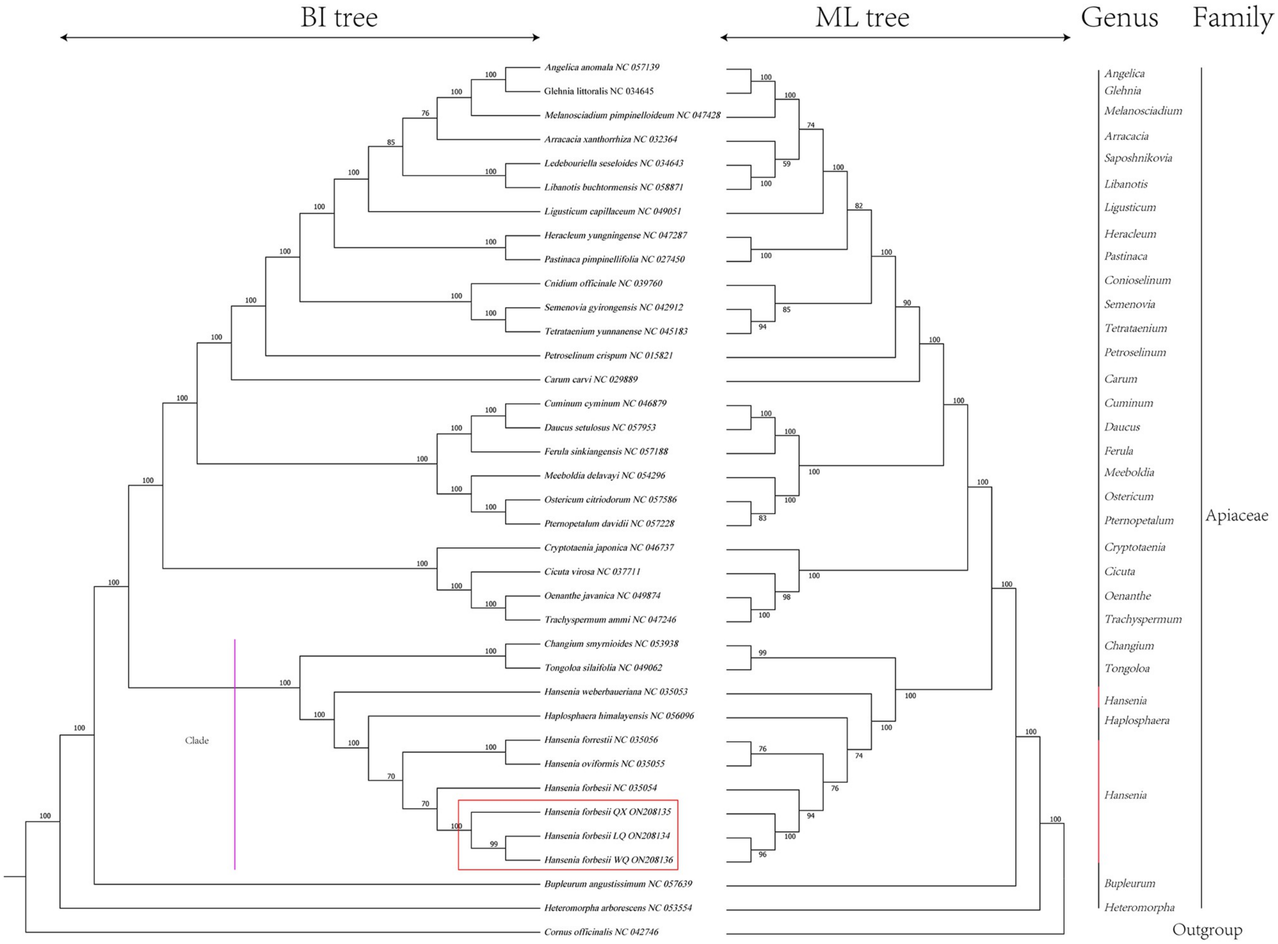

BI tree
ML tree
Genus
Family
Clade
Angelica anomala NC 057139
Glehnia littoralis NC 034645
Melanosciadium pimpinelloideum NC 047428
Arracacia xanthorrhiza NC 032364
Ledebouriella seseloides NC 034643
Libanotis buchtormensis NC 058871
Ligusticum capillaceum NC 049051
Heracleum yungningense NC 047287
Pastinaca pimpinellifolia NC 027450
Cnidium officinale NC 039760
Semenovia gyirongensis NC 042912
Tetrataenium yunnanense NC 045183
Petroselinum crispum NC 015821
Carum carvi NC 029889
Cuminum cyminum NC 046879
Daucus setulosus NC 057953
Ferula sinkiangensis NC 057188
Meeboldia delavayi NC 054296
Ostericum citriodorum NC 057586
Pternopetalum davidii NC 057228
Cryptotaenia japonica NC 046737
Cicuta virosa NC 037711
Oenanthe javanica NC 049874
Trachyspermum ammi NC 047246
Changium smyrnioides NC 053938
Tongoloa silaifolia NC 049062
Hansenia weberbaueriana NC 035053
Haplosphaera himalayensis NC 056096
Hansenia forrestii NC 035056
Hansenia oviformis NC 035055
Hansenia forbesii NC 035054
Hansenia forbesii QX ON208135
Hansenia forbesii LQ ON208134
Hansenia forbesii WQ ON208136
Bupleurum angustissimum NC 057639
Heteromorpha arborescens NC 053554
Cornus officinalis NC 042746
Angelica
Glehnia
Melanosciadium
Arracacia
Saposhnikovia
Libanotis
Ligusticum
Heracleum
Pastinaca
Conioselinum
Semenovia
Tetrataenium
Petroselinum
Carum
Cuminum
Daucus
Ferula
Meeboldia
Ostericum
Pternopetalum
Cryptotaenia
Cicuta
Oenanthe
Trachyspermum
Changium
Tongoloa
Hansenia
Haplosphaera
Hansenia
Bupleurum
Heteromorpha
Apiaceae
Outgroup

Fig 9. BI and ML phylogenetic tree reconstruction containing the cp genome of 37 plants. The number on the branch indicates that the support rate of the ML tree is greater than 50%; The *C. officinalis* were set as the outgroup.

https://doi.org/10.1371/journal.pone.0286587.g009

forbesii WQ. We found three hypervariable regions between the cp genome of the three varieties of *H*. *forbesii*, including *atp*F-*atp*H, *pet*D, and *rps*15-*ycf*1. So the character differences between the three varieties of *H*. *forbesii* could be judged by the complete cp genome. Studies have shown that introns can reduce gene splicing and translation, slow down cell metabolism and reduce energy consumption [36]. QX had one more bp than LQ and WQ in terms of the introns in the *trn*K-UUU, *rps*16, *atp*F, *ycf*3, *clp*P, and *rpl*16, but one less bp in *rpo*C1. Taking Caseinolytic protease (*Clp*P) as an example, an ATP-dependent and highly conserved serine protease [37], it is essential for the regulation of the cellular stress response, cell homeostasis, and bacterial virulence [38, 39]. When encoding *Clp*P, the increase in introns length may affect gene splicing and translation, thereby affecting the expression of *Clp*P in QX, resulting in better disease resistance. Amino acid frequency analyses revealed the highest frequency of Leucine, Leucine in plants can enhance resistance to salt stress, enhance pollen vitality and germination, and enhance fragrance. The highest frequency of Leucine may also be one of the reasons why the three varieties have aromatic aromas.

The IR regions exhibited lower sequence divergence than LSC and SSC regions and were consistent with the previous comparisons of the cp genome [40]. Intergenic spacer regions are the most frequently used cp markers for phylogenetic studies at lower taxonomic levels in plants [41], as they are regarded as more variable and could provide more phylogenetically informative characters. In a previous study, *rps*15-*ycf*1 was identified as a mutational hotspot for identifying *Isodon rubescens* [18]. The *Pet*D located in the LSC region, as a protein-coding gene, contains an intron, which is consistent with previous genetic studies [40] and can also be used as a highly variable region for species identification. Phylogenetic analysis further showed that *H*. *forbesii* LQ and *H*. *forbesii* WQ were closely related, followed by *H*. *forbesii* QX, and they were closely related to *Hansenia forbesii* (NC035054), this also indirectly indicated the reliability of the data of three varieties of *H*. *forbesii*.

The Cp genome has a moderate rate of nucleotide evolution, which results in its suitability for species identification and phylogenetic studies at different taxonomic levels [42, 43]. Chen et al. identified the genus *Clerodendrum* with no significant difference in external morphology using the cp genome, they found two intergenic regions: *trn*H-GUG-*psb*A, *nhd*D-*psa*C, exhibiting a high degree of variations and showed that the complete cp genome can be used as a super-barcode to identify these *Clerodendrum* species [44]. Wu et al. compared five *Crataegus* cp genomes and found five highly variable regions, namely *pet*G-*trn*W-*CCA*, *trn*H-GUG-*psb*A, *trn*R-UCU-*atp*A, *pet*A-*psb*J, and *trn*T-GGU-*psb*D, indicating that the complete cp genome could also be used as a super-barcode to accurately authenticate the five Crataegus species [45]. Fu et al. obtained 48 cp genome from 16 *Taxus* species and *Pseudotaxus chienii* and constructed the ML evolutionary tree, which showed that the cp genome can successfully distinguish all *Taxus* species and can be used as a super barcode to identify species [46]. Lan et al. analyzed five cp genomes from *Artemisia* species, and showed that the cp genome can provide distinguishing features to help identify closely related *Artemisia* species and had the potential to serve as a universal super barcode for plant identification [20]. Most studies on the application of the cp genome have pointed out that a complete cp genome can be used as super barcodes for species identification. We found that the internal reason for the difference in the external phenotypes of the three varieties of *H*. *forbesii* can be reflected in the highly variable region of the cp genome, so the complete cp genome of three varieties of *H*. *forbesii* can be used as the DNA super barcode for identifying these three varieties. And *trn*C-GCA—*pet*N can be used as a specific DNA barcode to effectively identify the three varieties. This assists in solving the problem of mixed varieties in the actual production of *H*. *forbesii*, and can also be taken into account by other plants, which has great significance. In addition, the genetic relationship of the *H*. *forbesii* QX variety is far away from the other two varieties, so it can be selected as the first-generation variety for single plant optimization and pure breeding to ensure its excellent phenotypes, and it can be popularized and applied on a large area. This study can provide a reference for the sustainable utilization of *Hansenia forbesii* resources and the identification of different varieties with different phenotypes of the same species.

Supporting information

S1 Table. The aboveground forms of three *H. forbesii*.
(XLSX)

S2 Table. Comparison of chloroplast genome basic information between three *H. forbesii* and the genus *Hansenia* and *Arabidopsis thaliana*.
(XLSX)

S3 Table. The genes with introns of the cp genome from three varieties and the length of exons and introns.
(XLSX)

S4 Table. The primers capturing hypervariable regions.
(XLSX)

S1 Fig. The gel electrophoresis results of the PCR products. Lane M was the marker of DL 2000. The lanes from left to right corresponded to products of LQ1, LQ2, LQ3, QX1, QX2, QX3, WQ1, WQ2, and WQ3, respectively.
(TIF)

Author Contributions

Data curation: Chenghao Zhu.

Formal analysis: Chenghao Zhu, Yuan Jiang, Yu Bai.

Funding acquisition: Shengjian Dong, Sun Zhirong.

Methodology: Yuan Jiang.

Resources: Chenghao Zhu, Yu Bai.

Software: Yuan Jiang.

Writing – original draft: Chenghao Zhu.

Writing – review & editing: Yuan Jiang, Shengjian Dong, Sun Zhirong.

References

1. Li YH, Jiang SY, Guan YL, Liu X, Zhou Y, Li LM, et al. Quantitative determination of the chemical profile of the plant material "Qiang-huo" by LC-ESI-MS-MS. Chromatographia. 2006; 64(7–8):405–11. https://doi.org/10.1365/s10337-006-0052-2

2. Shi J, Jiang SY, Ma XJ, Sun H, Zhou Y. [Studies on seeds germination and seedlings growth of Notopterygium incisum]. Zhongguo Zhong Yao Za Zhi. 2007; 32(18):1841–4. PMID: 18051884.

3. Xu KJ, Jiang SY, Zhou Y, Zhang YX, Xia B, Xu XM, et al. Discrimination of the seeds of Notopterygium incisum and Notopterygium franchetii by validated HPLC-DAD-ESI-MS method and principal component analysis. J Pharmaceut Biomed. 2011; 56(5):1089–93. https://doi.org/10.1016/j.jpba.2011.07.034 PMID: 21856104.

4. Azietaku JT, Ma HF, Yu XA, Li J, Oppong MB, Cao J, et al. A review of the ethnopharmacology, phytochemistry and pharmacology of Notopterygium incisum. J Ethnopharmacol. 2017; 202:241–55. https://doi.org/10.1016/j.jep.2017.03.022 PMID: 28336469.

5. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17(1):134. https://doi.org/10.1186/s13059-016-1004-2 PMID: 27339192.

6. Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M, Lobo FP, et al. Chloroplast genomes of Byrsonima species (Malpighiaceae): comparative analysis and screening of high divergence sequences. Sci Rep. 2018; 8(1):2210. https://doi.org/10.1038/s41598-018-20189-4 PMID: 29396532.

7. Mehmood F, Abdullah, Shahzadi I, Ahmed I, Waheed MT, Mirza B. Characterization of Withania somnifera chloroplast genome and its comparison with other selected species of Solanaceae. Genomics. 2020; 112(2):1522–30. https://doi.org/10.1016/j.ygeno.2019.08.024 PMID: 31470082.

8. Yu XL, Tan W, Zhang HY, Gao H, Wang WX, Tian XX. Complete Chloroplast Genomes of Ampelopsis humulifolia and Ampelopsis japonica: Molecular Structure, Comparative Analysis, and Phylogenetic Analysis. Plants-Basel. 2019; 8(10). https://doi.org/10.3390/plants8100410 PMID: 31614980.

9. Iram S, Hayat MQ, Tahir M, Gul A, Abdullah, Ahmed I. Chloroplast Genome Sequence of Artemisia scoparia: Comparative Analyses and Screening of Mutational Hotspots. Plants-Basel. 2019; 8(11). https://doi.org/10.3390/plants8110476 PMID: 31698805.

10. Palmer JD. Comparative organization of chloroplast genomes. Annu Rev Genet. 1985; 19:325–54. https://doi.org/10.1146/annurev.ge.19.120185.001545 PMID: 3936406.

11. Daniell H. Transgene containment by maternal inheritance: effective or elusive? Proc Natl Acad Sci U S A. 2007; 104(17):6879–80. https://doi.org/10.1073/pnas.0702219104 PMID: 17440039.

12. Neale DB, Sederoff RR. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. Theor Appl Genet. 1989; 77(2):212–6. https://doi.org/10.1007/BF00266189 PMID: 24232531.

13. Jin Y, Zhang T, Liu B, Zheng C, Huo H, Zhang J. Comparative and phylogenetic analysis of the complete chloroplast genome sequences of Allium mongolicum. Sci Rep. 2022; 12(1):21676. https://doi.org/10.1038/s41598-022-26354-0 PMID: 36522492.

14. Fahrenkrog AM, Matsumoto GO, Toth K, Jokipii-Lukkari S, Salo HM, Haggman H, et al. Chloroplast genome assemblies and comparative analyses of commercially important Vaccinium berry crops. Sci Rep. 2022; 12(1):21600. https://doi.org/10.1038/s41598-022-25434-5 PMID: 36517490.

15. Zhang Q, Belyakov EA, Lapirov AG, Zhao Y, Freeland J, Xu X. A reappraisal of the phylogeny and historical biogeography of Sparganium (Typhaceae) using complete chloroplast genomes. BMC Plant Biol. 2022; 22(1):588. https://doi.org/10.1186/s12870-022-03981-3 PMID: 36522739.

16. Wang J, Qian J, Jiang Y, Chen XC, Zheng BJ, Chen SL, et al. Comparative Analysis of Chloroplast Genome and New Insights Into Phylogenetic Relationships of Polygonatum and Tribe Polygonateae. Front Plant Sci. 2022; 13. https://doi.org/10.3389/fpls.2022.882189 PMID: 35812916.

17. Kuang Q, Sheng WT. The entire chloroplast genome sequence of Asparagus setaceus (Kunth) Jessop: Genome structure, gene composition, and phylogenetic analysis in Asparagaceae. Open Life Sci. 2022; 17(1):1541–54. https://doi.org/10.1515/biol-2022-0497 PMID: 36474705.

18. Zhou ZY, Wang J, Pu TT, Dong JJ, Guan Q, Qian J, et al. Comparative analysis of medicinal plant Isodon rubescens and its common adulterants based on chloroplast genome sequencing. Front Plant Sci. 2022; 13. https://doi.org/10.3389/fpls.2022.1036277 PMID: 36479509.

19. Zeng QW, Chen M, Wang SC, Xu XX, Li T, Xiang ZH, et al. Comparative and phylogenetic analyses of the chloroplast genome reveal the taxonomy of the Morus genus. Front Plant Sci. 2022; 13. https://doi.org/10.3389/fpls.2022.1047592 PMID: 36507423.

20. Lan ZH, Shi YH, Yin QG, Gao RR, Liu CL, Wang WT, et al. Comparative and phylogenetic analysis of complete chloroplast genomes from five Artemisia species. Front Plant Sci. 2022; 13. https://doi.org/10.3389/fpls.2022.1049209 PMID: 36479523.

21. Fang H, Dai G, Liao B, Zhou P, Liu Y. Application of chloroplast genome in the identification of Phyllanthus urinaria and its common adulterants. Front Plant Sci. 2022; 13:1099856. https://doi.org/10.3389/fpls.2022.1099856 PMID: 36684764.

22. Shi LC, Chen HM, Jiang M, Wang LQ, Wu X, Huang LF, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019; 47(W1):W65–W73. https://doi.org/10.1093/nar/gkz345 PMID: 31066451.

23. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021; 49(16):9077–96. https://doi.org/10.1093/nar/gkab688 PMID: 34417604.

24. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019; 47(W1):W59–W64. https://doi.org/10.1093/nar/gkz238 PMID: 30949694.

25. Henriquez CL, Abdullah, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, et al. Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). Genomics. 2020; 112(3):2349–60. https://doi.org/10.1016/j.ygeno.2020.01.006 PMID: 31945463.

26. Zhang D, Gao FL, Jakovlic I, Zou H, Zhang J, Li WX, et al. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020; 20(1):348–55. https://doi.org/10.1111/1755-0998.13096 PMID: 31599058.

27. Chen CJ, Chen H, Zhang Y, Thomas HR, Frank MH, He YH, et al. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plant. 2020; 13(8):1194–202. https://doi.org/10.1016/j.molp.2020.06.009 PMID: 32585190.

28. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27 (2):573–80. https://doi.org/10.1093/nar/27.2.573 PMID: 9862982.

29. Beier S, Thiel T, Munch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017; 33(16):2583–5. https://doi.org/10.1093/bioinformatics/btx198 PMID: 28398459.

30. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29(22):4633–42. https://doi.org/10.1093/nar/29.22.4633 PMID: 11713313.

31. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004; 32:W273–W9. https://doi.org/10.1093/nar/gkh458 PMID: 15215394.

32. Amiryousefi A, Hyvonen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018; 34(17):3030–1. https://doi.org/10.1093/bioinformatics/bty220 PMID: 29659705.

33. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690.

34. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. Bayesian phylogenetic analysis of combined data. Syst Biol. 2004; 53(1):47–67. https://doi.org/10.1080/10635150490264699 PMID: 14965900.

35. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623.

36. Parenteau J, Maignon L, Berthoumieux M, Catala M, Gagnon V, Abou Elela S. Introns are mediators of cell response to starvation. Nature. 2019; 565(7741):612-+. https://doi.org/10.1038/s41586-018-0859-7 PMID: 30651641.

37. Ni T, Ye F, Liu X, Zhang J, Liu H, Li J, et al. Characterization of Gain-of-Function Mutant Provides New Insights into ClpP Structure. ACS Chem Biol. 2016; 11(7):1964–72. https://doi.org/10.1021/acschembio.6b00390 PMID: 27171654.

38. Gaillot O, Pellegrini E, Bregenholt S, Nair S, Berche P. The ClpP serine protease is essential for the intracellular parasitism and virulence of Listeria monocytogenes. Mol Microbiol. 2000; 35(6):1286–94. https://doi.org/10.1046/j.1365-2958.2000.01773.x PMID: 10760131.

39. Frees D, Qazi SN, Hill PJ, Ingmer H. Alternative roles of ClpX and ClpP in Staphylococcus aureus stress tolerance and virulence. Mol Microbiol. 2003; 48(6):1565–78. https://doi.org/10.1046/j.1365-2958.2003.03524.x PMID: 12791139.

40. Jiang Y, Miao YJ, Qian J, Zheng Y, Xia CL, Yang QS, et al. Comparative analysis of complete chloroplast genome sequences of five endangered species and new insights into phylogenetic relationships of Paris. Gene. 2022; 833. https://doi.org/10.1016/j.gene.2022.146572 PMID: 35609799.

41. Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS, Miller J, et al. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot. 2005; 92 (1):142–66. https://doi.org/10.3732/ajb.92.1.142 PMID: 21652394.

42. Aecyo P, Marques A, Huettel B, Silva A, Esposito T, Ribeiro E, et al. Plastome evolution in the Caesalpinia group (Leguminosae) and its application in phylogenomics and populations genetics. Planta. 2021; 254(2):27. https://doi.org/10.1007/s00425-021-03655-8 PMID: 34236509.

43. Li L, Hu Y, He M, Zhang B, Wu W, Cai P, et al. Comparative chloroplast genomes: insights into the evolution of the chloroplast genome of Camellia sinensis and the phylogeny of Camellia. Bmc Genomics. 2021; 22(1):138. https://doi.org/10.1186/s12864-021-07427-2 PMID: 33637038.

44. Chen H, Chen H, Wang B, Liu C. Conserved chloroplast genome sequences of the genus Clerodendrum Linn. (Lamiaceae) as a super-barcode. Plos One. 2023; 18(2):e0277809. https://doi.org/10.1371/journal.pone.0277809 PMID: 36757949.

45. Wu L, Cui Y, Wang Q, Xu Z, Wang Y, Lin Y, et al. Identification and phylogenetic analysis of five Crataegus species (Rosaceae) based on complete chloroplast genomes. Planta. 2021; 254(1):14. https://doi.org/10.1007/s00425-021-03667-4 PMID: 34180013.

46. Fu CN, Wu CS, Ye LJ, Mo ZQ, Liu J, Chang YW, et al. Prevalence of isomeric plastomes and effectiveness of plastome super-barcodes in yews (Taxus) worldwide. Sci Rep. 2019; 9(1):2773. https://doi.org/10.1038/s41598-019-39161-x PMID: 30808961.