RESEARCH

Plant super-barcode: a case study on genome-based identification for closely related species of *Fritillaria*

Lan Wu[1†], Mingli Wu[1†], Ning Cui[1], Li Xiang[1], Ying Li[2], Xiwen Li[1*] and Shilin Chen[1*]

Abstract

Background: Although molecular analysis offers a wide range of options for species identification, a universal methodology for classifying and distinguishing closely related species remains elusive. This study validated the effectiveness of utilizing the entire chloroplast (cp) genome as a super-barcode to help identify and classify closely related species.

Methods: We here compared 26 complete cp genomes of ten *Fritillaria* species including 18 new sequences sequenced in this study. Each species had repeats and the cp genomes were used as a whole DNA barcode to test whether they can distinguish *Fritillaria* species.

Results: The cp genomes of *Fritillaria* medicinal plants were conserved in genome structure, gene type, and gene content. Comparison analysis of the *Fritillaria* cp genomes revealed that the intergenic spacer regions were highly divergent compared with other regions. By constructing the phylogenetic tree by the maximum likelihood and maximum parsimony methods, we found that the entire cp genome showed a high discrimination power for *Fritillaria* species with individuals of each species in a monophyletic clade. These results indicate that cp genome can be used to effectively differentiate medicinal plants from the genus *Fritillaria* at the species level.

Conclusions: This study implies that cp genome can provide distinguishing differences to help identify closely related *Fritillaria* species, and has the potential to be served as a universal super-barcode for plant identification.

Keywords: Species identification, Closely related species, Chloroplast genome, Super-barcode, Genome comparison, *Fritillaria*

Background

Although many biological studies depend on accurate species identification and delimitation, such as the implement of biodiversity conservation, therapy of disease and the identification of invasive species, taxonomic expertise is collapsing [1, 2]. Morphology-based identification of plant species has remained elusive due to the scarcity and ambiguity of diagnostic characters. Fortunately the advent of molecular markers made an impact on species identification, and undoubtedly has made a substantial contribution to systematics. However, currently none of the available DNA loci work for all species, especially for the closely related species. Moreover, multiple closely related species that occupy the same area have always posed insurmountable barriers to the goal of current highly accurate identification [3]. Therefore, a new method is required in the search for a universal marker for taxon recognition.

*Correspondence: xwli@icmm.ac.cn; slchen@icmm.ac.cn
†Lan Wu and Mingli Wu contributed equally to this work
[1] Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China
Full list of author information is available at the end of the article

Wu *et al. Chin Med*   (2021) 16:52

Page 2 of 11

The chloroplast (cp) genome is a versatile tool for phylogenetics. During the past decade, there have been many analyses addressing phylogenetic questions at deep nodes based upon the complete sequences of cp genomes [4–6]. As plant biologists enter the era in which comparative genomics promises to address in-depth questions, the inestimable effectiveness of cp genome in systematic studies quickly become clear. The entire cp genome contains approximately as much information as does the COI gene used in animals and it has the potential to provide distinguishing differences that can help molecularly identify closely related species [7]. With advances in high-throughput sequencing, achieving cp genome is easily acquirable at a large-scale with lower costs. This has promoted studies of systematics using cp genome in *Epimedium* [8], *Paris* [9] and *Sanguisorba* [10]. Because of the low discrimination power of general molecular markers in plants and their closely related species, researchers have proposed the entire cp genome as a super-barcode to discriminate closely related species [11].

The bulbs of *Fritillaria* species (called BeiMu, BM) have been used medicinally for more than 2000 years, specifically in the treatment of dry cough and blood-stained sputum. Due to the over exploitation of natural resources, the availability of Chuan BeiMu continues to decline [12]. Currently, most *Fritillaria* species used as Chuan BeiMu are in the list of wild protected species (level 3) in China, and the price of high-quality Chuan BeiMu can be as high as ~500$/kg. With the decreased availability and high price, Chuan BeiMu is often adulterated by other cheaper bulbs from other *Fritillaria* species, with a market survey reporting the adulteration rate of Chuan BeiMu to be as high as 20% [13].

Presently, *Fritillaria* bulbs are identified by morphological features [14] and chemical properties [15]. Unfortunately, different species can be morphologically similar and they always have the similar chemical constituents, which make the identification of *Fritillaria* difficult at the species level using traditional methods. Although DNA barcoding provided accurate identification for plants, it is insufficient in the authentication of *Fritillaria* species. Luo and Xiang et al. [16, 17] reported that ITS2 sequence could not provide monophyletic clades for the genus *Fritillaria* at the species level. Meanwhile, Sharifi [18] and TÜRKTAŞ et al. [19] constructed the phylogenetic trees based on the *trnH-psbA* and *trnL-trnF* regions using 22 Iranian *Fritillaria* species and ten Turkey *Fritillaria* species respectively, and the phylogenetic trees showed that it is impossible to distinguish these *Fritillaria* species. Rønsted et al. [20] presented the same result based on *matK*, *rpl16*, *trnK*, and ITS sequences for *Fritillaria*. Therefore, these findings demonstrate that the single-locus markers have low resolution for *Fritillaria* due to high sequence similarities.

Compared with the most frequently used and predicted genus-specific DNA barcodes, cp genome contains more variations with a significantly higher resolution of phylogenies, which is valuable to reveal phylogenetic relationships between closely related species [12]. Cp genome has been widely applied in phylogenetic analyses [21–24], plant population studies [25], and plant identification [7]. The phylogenetic tree constructed based on complete cp genomes has a higher supporting rate and discrimination power [26]. Li et al. [11] therefore proposed to use the entire cp genome as a super-barcode to accurately identify closely related species.

Here, we compared 26 complete cp genomes, including 18 newly sequenced genome sequences for this study, from ten *Fritillaria* species that are included in the Chinese Pharmacopoeia 2020. We performed a comprehensive analysis of the complete cp genomes of the *Fritillaria* species, which are difficult to be identified by morphology and taxonomy alone. The aims of our study were as follows: (1) to verify the hypothesis whether super-barcode can be used as a universal barcode to identify closely related species, (2) to present 18 new complete cp genomes from ten *Fritillaria* species and explore polymorphic regions within *Fritillaria* cp genomes, and (3) to evaluate the discrimination power of cp genomes in the genus *Fritillaria* at the species level. The results demonstrated that the cp genome could be used to identify *Fritillaria* at species level. The entire cp genome was found to be a most promising universal DNA marker in identification of closely related species.

Methods

DNA extraction

Twenty-six cp genomes from ten *Fritillaria* species were used in this study (see Additional file 1: Table S1). Fresh leaves of 18 individuals from nine *Fritillaria* species were collected. The cp genomes of eight additional individuals were downloaded from GenBank. Total genomic DNA of each sample was isolated from ~200 mg of fresh leaf using the DNeasy Plant Mini Kit (QIAGEN, Germany), according to manufacturer's instructions. To meet the quality requirements for sequencing, we assessed the quality and quantity of each DNA sample using a Qubit2.0 Fluorometer (Thermo Scientific, USA) and a NanoDrop 2000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA), respectively.

Genome sequencing, assembly and annotation

The shotgun libraries (450 bp) were constructed using ~2 μg of total DNA according to the manufacturer's instructions (Illumina Inc., San Diego, CA). A total

Wu *et al. Chin Med* (2021) 16:52

Page 3 of 11

of 11 cp genomes from seven *Fritillaria* species were sequenced using Illumina HiSeq X platform (Illumina, San Diego, CA, USA), and we obtained >2 Gb data for each sample. Raw reads were filtered using the Fastqc trim tool (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Thereafter, contigs were extracted by BLASTs [27], and the cp genomes of six published *Fritillaria* species (Accession No.: KF769143, KF712486, KY646166, KC713823, KF769142, and KY646165) were set as reference sequences. The contigs were assembled using SOAPdenovo [28]. Sequence extension was performed using SSPACE [29], and the gap fillers were excluded by GapCloser [30]. Other seven cp genomes from six *Fritillaria* species were sequenced using Roche 454 titanium sequencing platform and assembled using Newbler sequence assembler. The four junctions between IRs (inverted repeats) and SC (large single-copy region, LSC; small single-copy region, SSC) were validated by PCR amplification and Sanger sequencing with specific primers, as listed in Additional file 2: Table S2. The initial gene annotation was performed using CpGAVAS [31]. Identified tRNA genes were confirmed by tRNAscan-SE [32, 33]. Circular cp genome maps were drawn using OGDRAW software (http://ogdraw.mpimp-golm.mpg.de/) [34]. GC content was analyzed using MEGA5.0. [35]. The validated entire 18 cp genome sequences were deposited in NCBI (Accession No. were listed in Additional file 1: Table S1).

Genome comparison and divergent analyses

Comparison of the sequence divergence in the cp genomes of ten *Fritillaria* species was performed using the mVISTA [36, 37] program in the Shuffle-LAGAN mode, and the annotation of *F. unibracteata* (MN148410) was used as the reference. In addition, simple sequence repeats (SSRs) were detected using MISA (http://pgrc.ipk-gatersleben.de/misa/) [38] with thresholds of repeat numbers of eight, four, four, three, three and three for mono-, di-, tri-, tetra-, penta- and hexa-nucleotides, respectively. Then, insertions/deletions (indels) were counted using LASTZ software and single nucleotide polymorphisms (SNPs) were analysed by MUMmer. All SNPs in the coding sequence were detected whether it affects the protein sequence and were distinguished from synonymous and non-synonymous SNPs. Variations were visualized by Circos software [39] including A-G layers. The discrimination ability of highly variable loci selected in this study was tested using 26 samples from 10 species.

Species identification

To evaluate the effectiveness of super-barcode in identification for closely related species, 26 complete cp genome sequences were aligned using the MAFFT program [40], and then adjusted manually in Bioedit. Phylogenetic trees were constructed by the maximum likelihood (ML) and maximum parsimony (MP) methods. *Lilium brownie* (accession no.: KY748296) and *Cardiocrinum giganteum* (Accession No.: KX528334) were set as outgroups. ML analyses were conducted using RAxML-HPC2 on XSEDE at the CIPRES Science Gateway website (https://www.phylo.org/) [41] with the GTR+I+G model as the best-fitting model, which was selected by jModelTest 2.1.4 [42]. MP analyses were performed using PAUP*4.0b10 [43]. The branch support of the tree was estimated in 1000 bootstrap replicates.
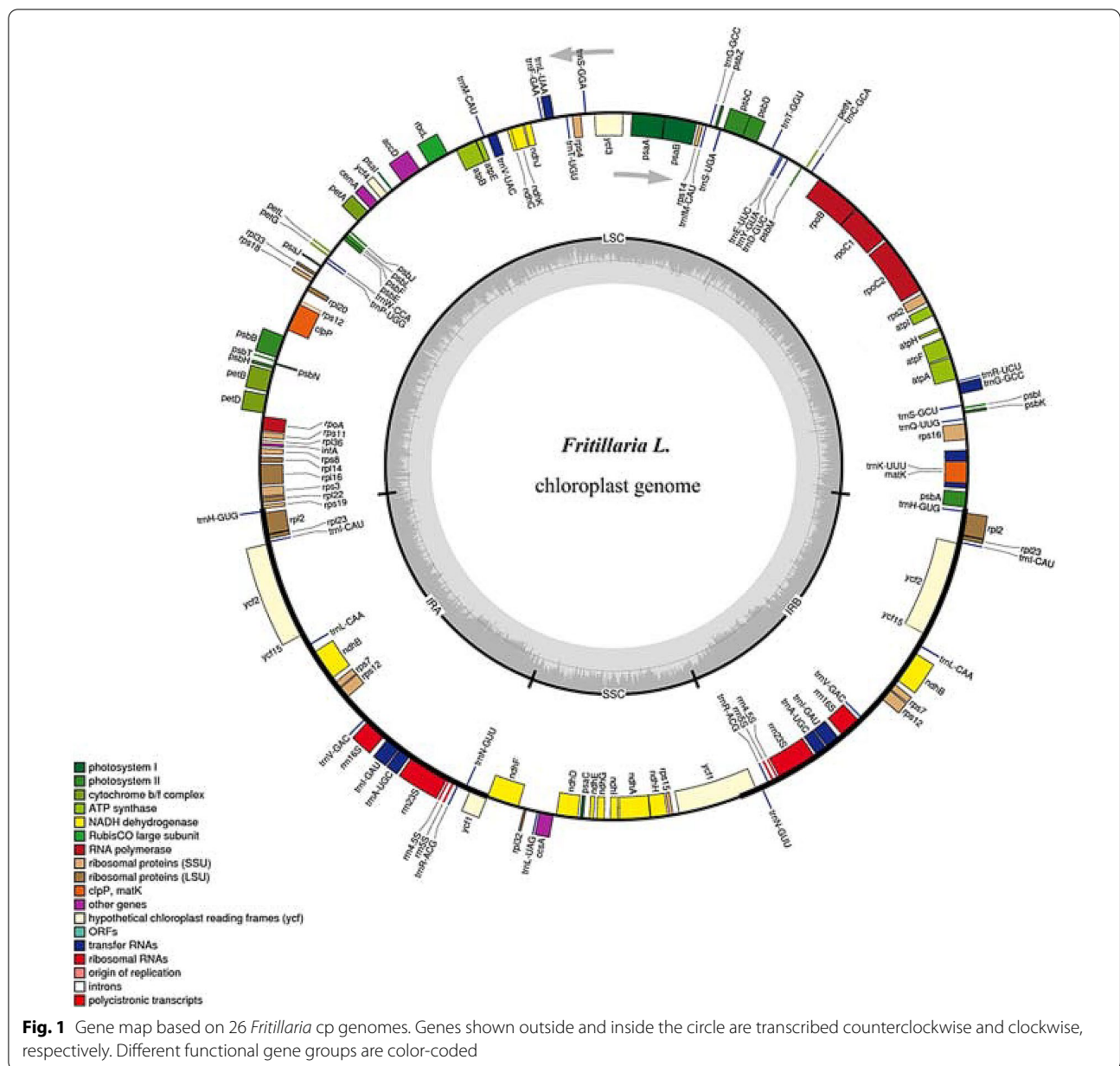
Results

Genome features

All the 26 cp genomes were similar in length, among which the shortest was *F. unibracteata* (150,764 bp) and the longest was *F. hupehensis* (152,186 bp), with the typical quadripartite structure of angiosperms. They contained a LSC (81,182–81,926 bp) and a SSC (17,114–17,586 bp), separated by a pair of IRs (26,024–26,390 bp) (see Fig. 1 and Additional file 1: Table S1). The GC content was unevenly distributed throughout *Fritillaria* cp genomes. In the *F. cirrhosa* (MN148400) cp genome, for example, GC content of the IR region (42.5%) was significantly higher than that of the LSC region (34.7%) or the SSC region (30.5%). This may be a reason that the conservation is divergent between the IR and LSC/SSC regions [10, 44]. The coding regions accounted for 52.5% of the genome, and therefore, the non-coding regions, such as the pseudogenes, introns, and intergenic spacers, accounted for 47.5%. The 26 *Fritillaria* cp genomes possessed 114 unique genes (Fig. 1) that included 80 protein-coding genes, 30 tRNA genes and four rRNA genes. In addition, we identified two pseudogenes (*infA* and *ycf15*). The *rps12* is a trans-spliced gene in which two 3′ end residues are located within the IR region and the 5′ end is located within the LSC region.

Introns are critical for the regulation of alternative splicing in the genome [45]. Similar to other angiosperm [16, 46], we identified 18 intron-containing genes in each of the 26 *Fritillaria* cp genomes, which included 12 protein-coding genes and six tRNA genes. Fifteen out of the 18 genes contained a single intron, whereas the remaining three genes (*ycf3*, *clpP*, and *rps12*) contained two introns.

SSR analysis

SSRs are short (1–6 nucleotide repeat units) tandemly repeated sequences that are widely distributed across the entire cp genome, and they are important for the population studies in plants. We here analyzed the distribution

Wu *et al. Chin Med*   (2021) 16:52

Page 4 of 11

Fig. 1 Gene map based on 26 *Fritillaria* cp genomes. Genes shown outside and inside the circle are transcribed counterclockwise and clockwise, respectively. Different functional gene groups are color-coded

of SSRs in 26 *Fritillaria* cp genomes. The number of SSRs ranged from 179 in *F. cirrhosa* to 195 in *F. pallidiflora*. Most of the SSRs were mononucleotide repeats ranging from 113 in *F. unibracteata* to 122 in *F. hupehensis*. The number of di-, tri- and tetranucleotide repeats was 57–63, 1–5 and 6–10, respectively. The number of penta- and hexanucleotide repeats were few, and none were detected in most *Fritillaria* cp genomes.

Comparative analysis of cp genome

Using the *F. unibracteata* (MN148410) as reference, cp genomes of ten *Fritillaria* species were compared and

analyzed to show the sequence divergence, which is relevant to further phylogeny and species authentication analyses. The genome comparison showed that there was a high similarity among these cp genomes. The sequences in LSC regions were more divergent than in the SSC and IR regions. The high divergences occurred in *trnS-GCU-trnG-GCC, trnG-GCC-trnR-UCU, trnE-UUC-trnT-GGU, trnT-GGU-psbD, atpH-atpI, trnT-UGU-trnL-UAA* and *psbE-petL* (see Additional file 3: Figure S1). The identification efficien of these seven loci were tested in this study and data showed that they

Wu *et al. Chin Med*   (2021) 16:52

Page 5 of 11

could not distinguish all 26 samples from ten *Fritillaria* species (see Additional file 4: Table S3).

In addition, sequence variability was estimated using SNPs and indels (Fig. 2). Based on cp genome-wide investigations, a total of 2449 SNPs and 565 Indels were detected among ten *Fritillaria* species (Table 1). Most of the variants (838 SNPs and 358 Indels) were located in intergenic spacers. Analysis of the distribution of genetic variability revealed that the most variable protein-coding region was the *rps19* gene. It is located in the Large Single Copy with a length of

279 bp containing 10 SNP sites. Among the non-coding regions, the highest frequency of polymorphism was found in the *rpl22-rps19* spacer. Within the 127 bp-long *rpl22-rps19* region, 14 SNPs and 10 indels were identified. Relatively high variability was also characteristic for the *rpl16-rps3* spacer.

Nowadays, cp DNA regions have been widely used in studies on species identification and phylogenetic analysis. However, none of cp markers work for *Fritillaria* species in previous studies [47, 48] as well as in this study.

Fig. 2 SNP and indel variation among cp genomes of *Fritillaria*. Track A represent nonsynonymous SNP occurrence in genes. Track B shows identified SNPs (cyan histogram) while track C represents identified indels (magenta histogram), with 100 bp shift per 600 bp window size. Track D and E represent percent of SNPs and indels per CDS length, respectively. Track F and G show percent of SNPs and indels per noncoding region length, respectively

Wu *et al. Chin Med*   (2021) 16:52

Page 6 of 11

Table 1 The SNP and Indels in 26 cp genomes of the ten *Fritillaria* species

Region	Coding region		Intergenic spacer		Intron		Summary	
	SNP	InDel	SNP	InDel	SNP	InDel	SNP	InDel
LSC	600	14	838	358	254	87	1692	459
SSC	383	10	107	29	31	16	521	55
IR$_a$	47	3	60	20	12	6	119	29
IR$_b$	45	2	60	15	12	5	117	22
Total	1075	29	1065	422	309	114	2449	565

Species identification

In this study, MP and ML trees were constructed based on 26 complete cp genomes from ten *Fritillaria* species with two species (*Lilium brownie* and *Cardiocrinum giganteum*) from the same family as outgroup. Our results showed that phylogenetic trees constructed by the entire cp genome presented a high discrimination power for the *Fritillaria* species, and different individuals in each species were resolved as a monophyletic clade. At first, *Fritillaria* medicinal plants were divided into two large branches, and *F. ussuriensis* was separated out with an approval rate of 100%. Thereafter, *F. walujewii* and *F. pallidiflora* were integrated into one big branch and separated from the other *Fritillaria* medicinal plants, with a supporting rate of 100%. These two medicinal species were used as Fritillariae pallidiflorae bulbus (Yi BeiMu), which was recorded in the Chinese Pharmacopoeia 2020. In addition, individuals of these two species were separated into two monophyletic clades. Lastly, the original species of Fritillariae cirrhosae bulbus (Chuan BeiMu) were gathered into one branch and separated from the other two types of *Fritillaria* bulbs, which included Fritillariae thunbergii bulbus (Zhe BeiMu) and Fritillariae hupehensis bulbus (Hubei BeiMu), with an approval rate of 100%. In addition, individuals of all species from Fritillariae cirrhosae bulbus (Chuan BeiMu) were separated into a monophyletic clade for each species, respectively. For the five types of *Fritillaria* bulbs, Fritillariae cirrhosae bulbus (Chuan BeiMu) had the closest relationship with Fritillariae thunbergii bulbus (Zhe BeiMu) and Fritillariae hupehensis bulbus (Hubei BeiMu). In total, all individuals of the original species from each *Fritillaria* bulbs recorded in the Chinese Pharmacopoeia 2020 were clustered and separated from other *Fritillaria* bulbs with a high branch supporting rate.

Ten *Fritillaria* species covered five different types of traditional medicinal materials: "PingBeiMu (PBM)" which original plant species is from *F. ussuriensis*, "ChuanBeiMu (CBM)" which original plant species are from *F. unibracteata var. wabuensis, F. unibracteata, F. taipaiensis, F. cirrhosa* and *F. delavayi*, "ZheBeiMu (ZBM)" from *F. thunbergii*, "HuBeiBeiMu (HBBM)" from *F. hupehensis*, "YiBeiMu (YBM)" from *F. pallidiflora* and *F. walujewii*. Phylogenetic trees in this study formed into five groups which branches were drawn in different colors. We found that the five monophyletic clades were consistent with five types of BM material medica. Another interesting finding is that the topologies of phylogenetic trees also formed into five major groups: PBM, ZBM, HBBM, CBM, and YBM. The five groups belong to five different ecological and geographical regions (see Additional file 6: Figure S3). PBM is mainly distributed in the plain of Mid-temperate zone in Northeast China. ZBM is mainly distributed in the subtropical plain and close to the ocean. HBBM is mainly loacted in plain mountain with a subtropical climate. CBM is mainly located in plateau mountain of western China, and YBM is mainly distributed in Mid-temperate zone of western plateau.

Discussion

This study investigated the feasibility of developing a cp-genome based identification method for closely related plants at lower taxonomic levels. Although DNA barcoding provides accurate identification for plants, it remains a significant challenge for authentication of *Fritillaria* species. Firstly we analyzed the cp genome of *Fritillaria*. They were highly conserved in gene structure, gene order and gene content. The average GC content was ~37.0%, which was similar to the published cp genomes of Liliales species [48, 49]. We also investigated introns in 26 *Fritillaria* cp genomes. In this study, most of the protein-coding genes had the standard ATG as the initiator codon, but *rpl2, ndhD* and *rps19* genes started with AUG, ATC and GUG, respectively. This variation, which may have been caused by RNA editing [50], has been reported in other cp genomes as well [51, 52]. As valuable molecular markers, SSRs are widely used in studies of population genetics, molecular breeding and species identification because of high polymorphisms [53, 54]. In this study, the distribution of SSRs in *Fritillaria* cp genomes was different but the interspecific variations were higher than intraspecific variations. In addition, most SSRs were located in the LSC. The content of polyA/T was

Wu *et al. Chin Med*   (2021) 16:52

Page 7 of 11

greater than that of polyG/C. We speculate that the richness of A/T SSRs may be related to the AT abundance in these *Fritillaria* genomes [22, 55]. Comparative analysis showed a high similarity among 26 *Fritillaria* cp genomes.

In order to find ideal molecular markers in *Fritillaria*, many studies tried to select highly variable regions based on cp genomes as genus-specific DNA barcodes for species identification [56–64]. Li et al. [56] found eight genes which had abundant variations among species by comparing four *Fritillaria* cp genomes. And Li et al. [57] performed multiple sequence alignment analysis on gene and intergenic regions respectively using clustalw2 and chose 20 highly variable genes and 20 highly variable intergenic regions. They found that both genes and intergenic regions in *Fritillaria* were relatively conservative compared with other species. Only seven hypervariable intergenic regions were selected as potential specific DNA barcode based on comparison of four *Fritillaria* cp genomes. Similar results existed in other studies [58–64], except that the number of highly variable regions was different. Unfortunately, none of the above regions have been verified by further experiments. We tested all 57 highly variable loci selected by published works except regions over 2 k bp in length using 26 samples from 10 *Fritillaria* species (see Additional file 4: Table S3 and Additional file 7). A total of seven hypervariable sites were screened and verified in this study. We found none of the loci could identify all these 10 species (see Additional file 4: Table S3). Therefore, these findings demonstrated that traditional molecular methods including DNA barcoding could not solve the problem of species identification in *Fritillaria* due to high sequence similarities.
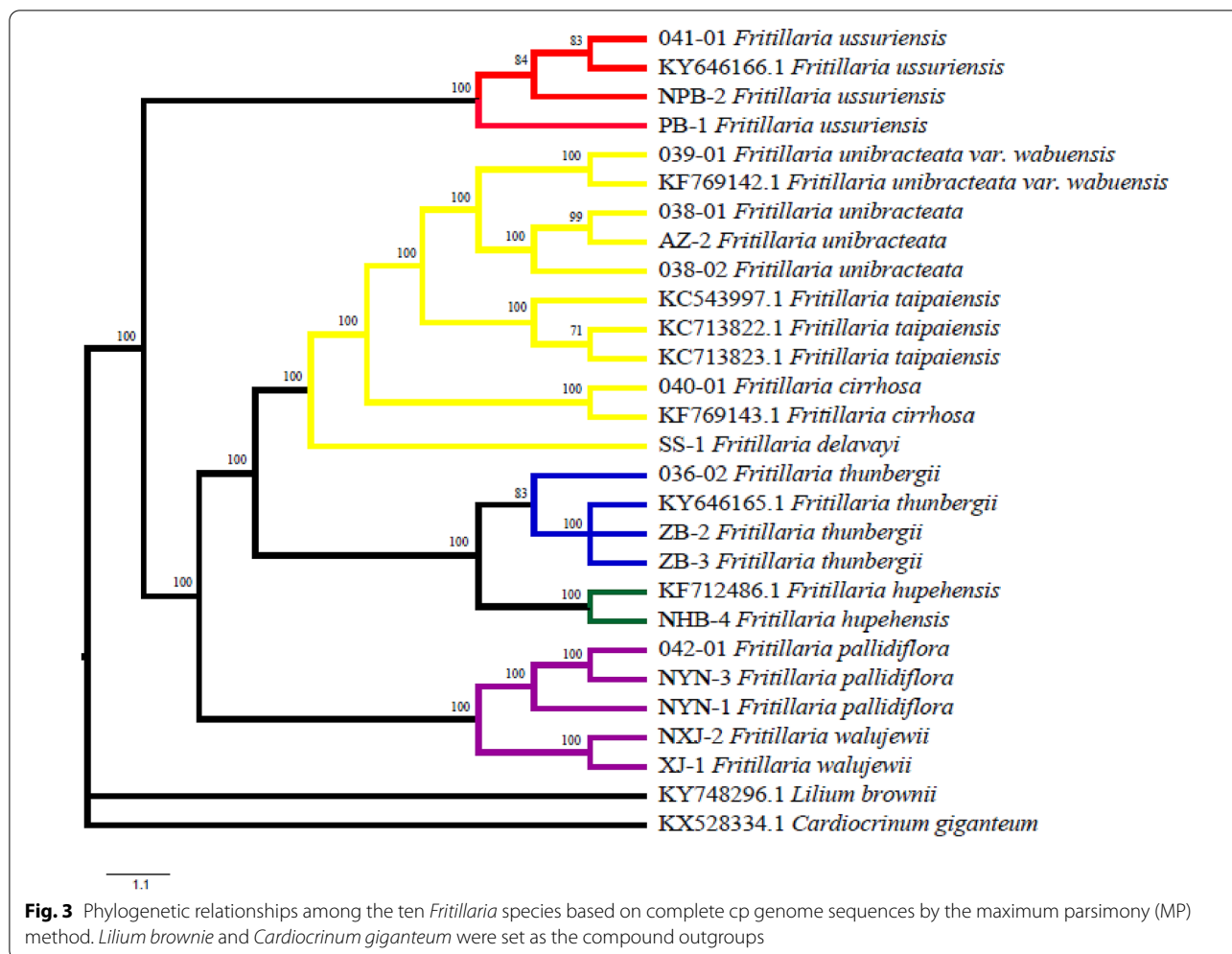
Because of the low discriminatory ability of general molecular markers in plants and their closely related species, researchers have placed high hopes on the use of plastid genome sequences in plant identification [7, 65, 66]. Some authors have performed tentative studies to test the potential of cp genomes in certain plant groups of closely related species. Bayly et al. [67] presented a phylogenetic analysis in three genera (*Eucalyptus*, *Corymbia* and *Angophora*) and demonstrated that cp genome was useful in lower-level genetic studies. Yang et al. [68] found that the cp genome lighted the species identification as organelle scale-scale "barcodes". Li et al. [11] then put forward that cp genome can be regarded as a super-barcode for closely related species. Xia et al. [69] and Chen et al. [70] tested the ability of super-barcode in *Chrysanthemum* and *Ligularia* respectively. However, the above two studies lacked sufficient species number and intraspecific samples. We here extended earlier investigations on a large scale to evaluate the feasibility of using the cp-genome to discriminate closely related species of *Fritillaria*.

Our results showed that the two topologies of MP and ML were identical with high support values (see Fig. 3 and Additional file 5: Figure S2). Both phylogenetic trees constructed by the entire cp genome presented a high discrimination power for the *Fritillaria* species, and different individuals from same species were formed into a monophyletic clade whatever in species level and in subspecies level. The cp-genome possesses the basic qualifications to be a universal marker compared with traditional molecular identification markers. First, Chloroplasts are haploid and non-recombining and cp-genome sequences are highly conserved, so they can act as a single locus [66]. Second, in contrast to a single gene, they have more variation and have the potential to identify closely related species at lower taxonomic levels [7]. Third, in our study, chloroplast sequence data has really shown 100% identification efficiency in *Fritillaria*. Because the results of screening genus-specific barcodes with different species groups were distinct, plant identification of closely related species based on super-barcode using DNA barcoding may no longer need to choose between more loci or more taxa.

Some authors disagreed with this approach for several reasons, including high expenditures, compared to Sanger sequencing, and the lack of close reference sequences for assembly [71]. With the development of next generation sequencing, the cost for cp-genome sequencing and assembly is almost the same as that of PCR-based sequencing on average. And for most plants, close reference for assembly has become less important than before. Taking the entire cp genome as a super-barcode becomes feasible for accurate species identification, since it has been demonstrated that cp genome could provide a higher resolution in species authentication in species and even population level [7, 25, 26].

Our results showed that the super-barcode based on the full length of the cp genome sequence could successfully distinguish the *Fritillaria* species recorded in the Chinese Pharmacopoeia 2020. According to the pharmacophylogenetic theory of Xiao et al. [72], the species that had the closest phylogenetic relationships were also similar in chemical components and curative effects. This confirmed the accuracy and reliability of the original species division for the five types of *Fritillaria* bulbs in the Chinese Pharmacopoeia 2020. Furthermore, this study demonstrated that the super-barcode based on the full length of the *Fritillaria* cp genome sequence could not only successfully distinguish the *Fritillaria* species in species level but also reflect the characteristics of biogeography. The analysis of phylogenetic relationship was coincident

Wu *et al. Chin Med* (2021) 16:52

Page 8 of 11

Fig. 3 Phylogenetic relationships among the ten *Fritillaria* species based on complete cp genome sequences by the maximum parsimony (MP) method. *Lilium brownie* and *Cardiocrinum giganteum* were set as the compound outgroups

with geographical distribution of BM (see Additional file 6: Figure S3), which provided a way for alternative resource discovery in natural drug development.

Although super-barcode has many advantages, it is not suitable for plant species identification when DNA extraction is difficult. For dried, cooked or decocting material medica, DNA is degraded seriously. We may not extract enough DNA or longer DNA fragments. It is not easy to obtain whole cp genome sequence by assembly. Compared with a single-locus barcode, the cost of super-barcode is higher and data analysis is complex using windows software. In fact, we don't recommend superbarcode if commonly used DNA barcode can make accurate identification. Therefore super-barcode is a useful supplement to the current molecular identification. It can show its advantages when traditional DNA barcoding is limited to plant identification of some closely related species.

Conclusions

The cp genome is now a reasonable option for increasing the resolution of plant identification in closely related species. In this study, we firstly verified the hypothesis that the cp genome could be used as a super-barcode to actually identify closely related species. Secondly, we analyzed and compared 26 complete cp genomes of the ten *Fritillaria* species, including 18 newly sequenced genomes. Finally, the phylogenetic analysis constructed by the 26 complete cp genomes strongly showed that the medicinal plants from the genus *Fritillaria* can effectively be distinguished at the species level. Recent advances in sequencing strategies make an unprecedented depth and scale of plastid genome sampling possible. Plastome sequencing is now a reasonable option for increasing the resolution of plant identification studies at low taxonomic levels. We are encouraged by the fact that species identification based upon the cp-genome is generally straightforward. Although

Wu *et al. Chin Med*   (2021) 16:52

Page 9 of 11

there are some issues to be solved, i.e., intraspecies sampling remains sporadic, and discrimination is not rapid, we believe that super-barcode is a good choice for identification of closely related species especially when DNA barcoding encounters difficulties.

Abbreviations
IRs: Inverted repeats regions; ITS/ITS2: Internal transcribed spacer/2; SSC: Small single-copy regions; SSRs: Simple sequence repeats; SNPs: Single nucleotide polymorphisms; LSC: Large single-copy regions; ML: Maximum likelihood; MP: Maximum parsimony; NJ: Neighbour-joining.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s13020-021-00460-z.

 Additional file 1: Table S1. Sequence information of 26 individuals from ten *Fritillaria* species cp genomes.

Additional file 2: Table S2. Specific primers used for validation in assembly.

Additional file 3: Figure S1. Comparison of the ten *Fritillaria* species cp genomes using mVISTA.

Additional file 4: Table S3. Species resolution of selected highly variable regions in related literatures and this study.

Additional file 5: Figure S2. Phylogenetic relationships among the ten *Fritillaria* species based on complete cp genome sequences by the maximum likelihood (ML) method. *Lilium brownie* and *Cardiocrinum giganteum* were set as the compound outgroups.

Additional file 6: Figure S3. Ecological and geographical regions of five BM material medica.

Additional file 7: Verification of discrimination ability of 57 highly variable loci selected by published works.

Availability of data and materials
All data generated in this study has been submitted to the NCBI under the following Accession Numbers: MN148400-MN148416, and MN126570.

Declarations

Prof. Shilin Chen serves as Associate Editor of *Chinese Medicine***.**

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China. [2]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China.

References
1. Struck TH, Feder JL, Bendiksby M, Birkeland S, Cerca J, Gusarov VI, et al. Finding evolutionary processes hidden in cryptic species. Trends Ecol Evol. 2018;33:153–63.
2. Hebert PD, Cywinska A, Ball SL, Dewaard JR. Biological identifications through DNA barcodes. P Roy Soc B Biol Sci. 2003;270:313–21.
3. Chase MW, Fay MF. Barcoding of plants and fungi. Science. 2009;325:682–3.
4. Dong W, Xu C, Wu P, Cheng T, Yu J, Zhou S, et al. Resolving the systematic positions of enigmatic taxa: manipulating the chloroplast genome data of Saxifragales. Mol Phylogenet Evol. 2018;126:321–30.
5. Zhai W, Duan X, Zhang R, Guo C, Li L, Xu G, et al. Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. Mol Phylogenet Evol. 2019;135:12–21.
6. Raman G, Park S. The complete chloroplast genome sequence of *Ampelopsis*: gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. Front Plant Sci. 2016;7:341.
7. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:84.
8. Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, et al. The complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analyses. Front Plant Sci. 2016;7:306.
9. Huang Y, Li X, Yang Z, Yang C, Yang J, Ji Y. Analysis of complete chloroplast genome sequences improves phylogenetic resolution in Paris (Melanthiaceae). Front Plant Sci. 2016;7:1797.
10. Meng X-X, Xian Y-F, Xiang L, Zhang D, Shi Y-H, Wu M-L, et al. Complete chloroplast genomes from *Sanguisorba*: identity and variation among four species. Molecules. 2018;23:2137.
11. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90:157–66.
12. Li X-W. Study on conservation biology of *Fritillaria cirrhosa*. Beijing: Chinese Academy of Medical Sciences and Peking Union Medical College. 2009.
13. Wang B, Zhou Y, Qin G, Hu Z, Hu J. An authenticity survey of traditional Chinese medicines from Hubei Province using DNA barcoding. World Sci Technol. 2018;20(2):276–80.
14. Tekşen M, Aytac Z, Pınar N. Pollen morphology of the genus *Fritillaria* L. (Liliaceae) in Turkey. Turk J Bot. 2010;34(5):397–416.
15. Li Y, Zhang L, Wu H, Wu X, Ju L, Zhang Y. Metabolomic study to discriminate the different Bulbus fritillariae species using rapid resolution liquid chromatography-quadrupole time-of-flight mass spectrometry coupled with multivariate statistical analysis. Anal Methods. 2014;6(7):2247–59.
16. Luo K, Pei M, Hui Y, Jingyuan S, Keli C, Yimei L. Molecular identification of Fritillariae Cirrhosae Bulbus and its adulterants. World Sci Technol. 2012;1:1153–8.
17. Xiang L, Su Y, Li X, Xue G, Wang Q, Shi J, Wang L, Chen S. Identification of Fritillariae bulbus from adulterants using ITS2 regions. Plant Gene. 2016;7:42–9.
18. Sharifi-Tehrani M, Advay M. Assessment of relationships between Iranian *Fritillaria* (Liliaceae) species using chloroplast trnH-psbA sequences and morphological characters. J Genet Resour. 2015;1(2):89–100.

Wu *et al. Chin Med* (2021) 16:52

Page 10 of 11

19. Türktaş M, Aslay M, Kaya E, Ertuğrul F. Molecular characterization of phylogenetic relationships in *Fritillaria* species inferred from chloroplast trnL-trnF sequences. Turk J Biol. 2012;36(5):552–60.

20. Rønsted N, Law S, Thornton H, Fay MF, Chase MW. Molecular phylogenetic evidence for the monophyly of *Fritillaria* and *Lilium* (Liliaceae; Liliales) and the infrageneric classification of *Fritillaria*. Mol Phylogenet Evol. 2005;35(3):509–27.

21. Huang H, Shi C, Liu Y, Mao S-Y, Gao L-Z. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol. 2014;14(1):151.

22. Guo H, Liu J, Luo L, Wei X, Zhang J, Qi Y, Zhang B, Liu H, Xiao P. Complete chloroplast genome sequences of *Schisandra chinensis*: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. Sci China Life Sci. 2017;60(11):1286–90.

23. Yu X, Zuo L, Lu D, Lu B, Yang M, Wang J. Comparative analysis of chloroplast genomes of five *Robinia* species: genome comparative and evolution analysis. Gene. 2019;689:141–51.

24. Yang Z, Zhao T, Ma Q, Liang L, Wang G. Comparative genomics and phylogenetic analysis revealed the chloroplast genome variation and interspecific relationships of *Corylus* (Betulaceae) Species. Front Plant Sci. 2018;9:927.

25. Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res. 2011;18:93–105.

26. Zeng C-X, Zhang Y-X, Triplett JK, Yang J-B, Li D-Z. Large multi-locus plastid phylogeny of the tribe Arundinarieae (Poaceae: Bambusoideae) reveals ten major lineages and low rate of molecular divergence. Mol Phylogenet Evol. 2010;56:821–39.

27. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:W5-9.

28. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.

29. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2010;27:578–9.

30. Acemel RD, Tena JJ, Irastorza-Azcarate I, Marlétaz F, Gómez-Marín C, de la Calle-Mustienes E, et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. Nat Genetics. 2016;48:336.

31. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, et al. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics. 2012;13:715.

32. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:W686–9.

33. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.

34. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 2007;52:267–74.

35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.

36. 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.

37. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–9.

38. Yang X-M, Sun J-T, Xue X-F, Zhu W-C, Hong X-Y. Development and characterization of 18 novel EST-SSRs from the western flower thrips, *Frankliniella occidentalis* (Pergande). Int J Mol Sci. 2012;13:2863–76.

39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

40. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

41. Miller MA, Pfeiffer W, Schwartz T. The CIPRES science gateway: a community resource for phylogenetic analyses. In: Proceedings of the 2011 TeraGrid Conference: extreme digital discovery. ACM. 2011: 41.

42. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25:1253–6.

43. Swofford DL. Paup*: phylogenetic analysis using parsimony (and other methods) 4.0. B5. Sunderland: Sinauer; 2001.

44. Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L, Yitao W. Complete chloroplast genome sequence of poisonous and medicinal plant datura stramonium: organizations and implications for genetic engineering. PLoS ONE. 2014;9: e110656.

45. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 2001;17:100–7.

46. Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV, Boore JL, Jansen RK. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). Mol Phylogenet Evol. 2007;45:547–63.

47. Xiang L, Su Y, Li X, Xue G, Wang Q, Shi J, et al. Identification of Fritillariae bulbus from adulterants using ITS2 regions. Plant Gene. 2016;7:42–9.

48. Do HDK, Kim JS, Kim J-H. Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae). Gene. 2013;530:229–35.

49. Kim JS, Kim J-H. Comparative genome analysis and phylogenetic relationship of order Liliales insight from the complete plastid genome sequences of two Lilies (*Lilium longiflorum* and *Alstroemeria aurea*). PLoS ONE. 2013;8: e68180.

50. Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A. RNA editing in plants and its evolution. Annu Rev Genet. 2013;47:335–52.

51. Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, et al. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS ONE. 2010;5: e12762.

52. Zhao Y, Yin J, Guo H, Zhang Y, Xiao W, Sun C, et al. The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. Front Plant Sci. 2015;5:696.

53. Chmielewski M, Meyza K, Chybicki IJ, Dzialuk A, Litkowiec M, Burczyk J. Chloroplast microsatellites as a tool for phylogeographic studies: the case of white oaks in Poland. Iforest. 2015;8:765.

54. Jiao Y, Jia H-m, Li X-w, Chai M-l, Jia H-j, Chen Z, et al. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). BMC Genomics. 2012;13:201.

55. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, et al. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant Artemisia annua. Molecules. 2017;22:1330.

56. Li Q, Li Y, Song J, et al. High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. New Phytol. 2015. https://doi.org/10.1111/nph.12966.

57. Ying L, Hui Y, Jingyuan S, et al. Screening *Fritillaria* genus-specific DNA barcodes based on complete chloroplast genome sequences. World Sci Technol. 2016;18(01):24–8.

58. Türktaş M, Aslay M, Kaya E, et al. Molecular characterization of phylogeneticrelationships in *Fritillaria* species inferred from chloroplast trnL-trnF sequences. Turk J Biol. 2012;36:552–60.

59. Park I, Kim WJ, Yeo S-M, et al. The complete chloroplast genome sequences of *Fritillaria ussuriensis* Maxim. and *Fritillaria cirrhosa* D. Don, and comparative analysis with other *Fritillaria* species. Molecules. 2017;22:982.

60. Bi Y, Zhang M-F, Xue J, et al. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on *Fritillaria*. Sci Rep. 2018;8:1184.

61. Li Y, Zhang Z, Yang J, et al. Complete chloroplast genome of seven Fritillaria species, variable DNA markers identification and phylogenetic relationships within the genus. PloS One. 2018, 13(3): e0194613.

62. Moon BC, Park I, Kim WJ, et al. The complete chloroplast genome sequence of *Fritillaria thunbergii* Miq., an important medicinal plant, and

Wu *et al. Chin Med*   (2021) 16:52

Page 11 of 11

identification of DNA markers to authenticate Fritillariae Bulbus. Hortic Environ Biotechnol. 2018;59:71–80.

63. Lu RS, Yang T, Chen Y, et al. Comparative plastome genomics and phylogenetic analyses of Liliaceae. Bot J Linn Soc. 2021. https://doi.org/10.1093/botlinnean/boaa109.

64. Chen Q, Wu X, Zhang D. Phylogenetic analysis of *Fritillaria cirrhosa* D. Don and its closely related species based on complete chloroplast genomes. Peer J. 2019;7: e7480.

65. Nikolaus JS, Maria CC. Genome-Based approaches to the authentication of medicinal plants. Planta Med. 2008;74:603–23.

66. Nock CJ, Waters DLE, Edwards MA, et al. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J. 2010;9:328–33.

67. Michael BJ, Rigault P, Spokevicius A, et al. Chloroplast genome analysis of Australian eucalypts—*Eucalyptus*, *Corymbia*, *Angophora*, *Allosyncarpia* and *Stockwellia* (Myrtaceae). Mol Phylogenet Evol. 2013;69:704–16.

68. Yang JB, Tang M, Li H, et al. Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol Biol. 2013;13:84.

69. Xia Y, Hu Z, Li X, et al. The complete chloroplast genome sequence of *Chrysanthemum indicum*. Mitochondrial DNA. 2016;27:4668–9.

70. Chen X, Zhou J, Cui Y, et al. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. Front Pharmacol. 2018;9:695.

71. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS ONE. 2011;5: e19254. https://doi.org/10.1371/journal.pone.0019254.

72. Xiao P, Yan J, Ping L, Yi-Bo L, Yong L. The botanical origin and pharmacophylogenetic treatment of Chinese materia medica Beimu. Acta Phytotaxonomica Sinica. 2007;45(4):473–87.

Publisher's Note