Chloroplast genomes elucidate diversity, phylogeny, and taxonomy of *Pulsatilla* (Ranunculaceae)

Qiu-jie Li¹², Na Su¹², Ling Zhang³, Ru-chang Tong¹², Xiao-hui Zhang⁴, Jun-ru Wang¹², Zhao-yang Chang¹², Liang Zhao¹²* & Daniel Potter⁵

*Pulsatilla* (Ranunculaceae) consists of about 40 species, and many of them have horticultural and/or medicinal value. However, it is difficult to recognize and identify wild *Pulsatilla* species. Universal molecular markers have been used to identify these species, but insufficient phylogenetic signal was available. Here, we compared the complete chloroplast genomes of seven *Pulsatilla* species. The chloroplast genomes of *Pulsatilla* were very similar and their length ranges from 161,501 to 162,669 bp. Eighty highly variable regions and potential sources of molecular markers such as simple sequence repeats, large repeat sequences, and single nucleotide polymorphisms were identified, which are valuable for studies of infra- and inter-specific genetic diversity. The SNP number differentiating any two *Pulsatilla* chloroplast genomes ranged from 112 to 1214, and provided sufficient data for species delimitation. Phylogenetic trees based on different data sets were consistent with one another, with the IR, SSC regions and the barcode combination *rbcL + matK + trnH-psbA* produced slightly different results. Phylogenetic relationships within *Pulsatilla* were certainly resolved using the complete cp genome sequences. Overall, this study provides plentiful chloroplast genomic resources, which will be helpful to identify members of this taxonomically challenging group in further investigation.

DNA barcoding is an effective tool to identify many plant species rapidly and accurately¹⁴. However, there is no single universal barcode that can be successfully used to identify all plants to the species level⁵. Consequently, two alternative strategies have been proposed to distinguish among plant species: the first one is the use of complete chloroplast genomes⁶, named ‘super-barcoding’, and the second one is an approach that involves searching for mutational hotspots⁷, or using comparative plastid analyses to find loci with suitable species-level divergence⁸.⁹. Analyses of entire chloroplast genome sequences provide an effective way to develop both of these strategies.

In most angiosperms, the chloroplast genomes are inherited maternally and have a consistent structure, including two inverted repeats (IR), one large (LSC) and one small (SSC) single copy region. The chloroplast genome always contains 110–130 genes that exhibit a range of levels of polymorphism⁴.⁵. Thus, chloroplast genome sequence data are extremely valuable for studies of plant population genetics, phylogeny reconstruction, species identification, and genome evolution⁸.⁹.

The Ranunculaceae is a large family, which includes approximately 59 genera and 2500 species. Many plants of *Ranunculaceae* are pharmaceutically important¹⁰. The genus *Pulsatilla* Adans. consists of about 40 species which are distributed in temperate subarctic and mountainous areas of the Northern Hemisphere¹⁰. There are always long, soft hairs covering plants of *Pulsatilla* species. Most of the flowers of *Pulsatilla* are large and showy, and therefore the genus has horticultural importance¹¹.¹². The flowers are solitary and bisexual. In one flower, there are always six tepals, numerous stamens and carpels, with the outermost stamens resembling degenerated petals, excluding *P. kostyczewii*¹¹,¹³–¹⁵.

In China, there are eleven species of *Pulsatilla*. Some species of *Pulsatilla* have been used in traditional Chinese medicine for many years, such as for “detoxification” or “blood-cooling”, because *Pulsatilla* species contain

¹College of Life Sciences, Northwest A&F University, Yangling 712100, China. ²Herbarium of Northwest, A&F University, Yangling 712100, China. ³College of Life Sciences, Tarim University, Alaer 843300, China. ⁴College of Life Science, Shaanxi Normal University, Xi’an 710062, China. ⁵Department of Plant Sciences, MS2, University of California, Davis, CA 95616, USA. ⁶These authors contributed equally: Qiu-jie Li and Na Su. ⁷Email: biology_zhaoliang@126.com
Pulsatilla 161,501 bp (P. grandis) cp genomes ranged from 151.5 × to 503.4 × coverage. These seven novel Pulsatilla cp genome sequences were submitted to GenBank (Tables 1, 2). Their quadripartite structure is similar to the majority of cp genomes of land plants, which are composed of a pair of IRs (31,184–31,416 bp), separated by the LSC (81,615–82,149 bp) and SSC (of 17,431–17,908 bp) regions26,27 (Fig. 1; Table 2). Previous studies of other angiosperm groups have found that chloroplast genomes are conserved28 or highly polymorphic29,30. These genomes which we reported are highly conserved in gene order, gene content and intron number, which is in agreement with previous studies25.

**Results and discussion**

**Chloroplast genome features.** We have obtained 1.95 Gb of average NGS clean data for each species, with minimum and maximum values of 1.14 Gb (P. dahurica), and 3.56 Gb (P. alpina alpina), respectively. The read number for each species ranged from 6,468,944 (P. dahurica) to 15,816,765 (P. alpina). The average length of the reads was 150 bp on the Illumina Sequencing System. The seven new Pulsatilla cp genomes ranged from 161,501 bp (P. grandis) to 162,669 bp (P. alpina) in length and 151.5 × to 503.4 × coverage. These seven novel Pulsatilla cp genome sequences were submitted to GenBank (Tables 1, 2). Their quadripartite structure is similar to the majority of cp genomes of land plants, which are composed of a pair of IRs (31,184–31,416 bp), separated by the LSC (81,615–82,149 bp) and SSC (of 17,431–17,908 bp) regions26,27 (Fig. 1; Table 2). Previous studies of other angiosperm groups have found that chloroplast genomes are conserved25 or highly polymorphic29,30. These genomes which we reported are highly conserved in gene order, gene content and intron number, which is in agreement with previous studies25.

**Table 1.** Voucher information and GenBank accession numbers for *Pulsatilla* and outgroups. Species with asterisks were collected by this study, whereas others were obtained from Genbank. NA not applicable.

| Taxon              | Location                  | Date    | Herbarium | Accession   | SRA accession |
|--------------------|---------------------------|---------|-----------|-------------|---------------|
| P. grandis*        | Europe Moaiju                 | 1904 | US   | MN025344    | SRR12822481  |
| P. multifida*      | U.S.R.                     | 1957 | US   | MN025347    | SRR12822474  |
| P. alpina*         | America Graubunden            | 1936 | US   | MN025343    | SRR12822486  |
| P. occidentalis*   | America Siskiyon California  | 1943 | US   | MN025348    | SRR12822484  |
| P. ludoviciana*    | America Albany county      | 1898 | US   | MN025346    | SRR12822477  |
| P. hirsutissima*   | America                        | US   | MN025345    | SRR12822480  |
| P. dahurica*       | China                        | 2014 | WUK | MN025349    | SRR12822482  |
| P. patens          | NA                        | NA    | NA   | KR_297058   | NA            |
| P. pratensis       | NA                        | NA    | NA   | KR_297060   | NA            |
| P. vernalis        | NA                        | NA    | NA   | KR_297062   | NA            |
| P. chinensis       | NA                        | NA    | NA   | MG_001341   | NA            |
| Anemoclema glaucifolium | NA                   | NA    | NA   | MH_205609   | NA            |

**Table 2.** Summary of complete chloroplast genomes of *Pulsatilla*.

| Species          | Number of reads | Average depth of coverage (x) | Size (bp) | Length (bp) | Coding (bp) | Non-coding (bp) | GC% |
|------------------|-----------------|-------------------------------|-----------|-------------|-------------|-----------------|-----|
| P. grandis       | 13,012,108      | 194.8                         | 161,501   | 81,672      | 17,431      | 31,199          | 78.377 | 37.6 |
| P. multifida     | 11,405,519      | 151.5                         | 161,743   | 81,653      | 17,648      | 31,221          | 81,800 | 37.6 |
| P. alpina        | 15,816,765      | 367.1                         | 162,669   | 82,149      | 17,688      | 31,416          | 81,246 | 37.6 |
| P. occidentalis  | 12,554,200      | 276.9                         | 161,764   | 81,615      | 17,755      | 31,197          | 79,089 | 37.6 |
| P. ludoviciana   | 9,251,830       | 257.5                         | 162,051   | 81,860      | 17,771      | 31,210          | 79,280 | 37.6 |
| P. hirsutissima  | 13,337,339      | 503.4                         | 161,936   | 81,866      | 17,702      | 31,184          | 80,206 | 37.6 |
| P. dahurica      | 6,468,944       | 410.1                         | 162,064   | 81,688      | 17,908      | 31,234          | 79,114 | 37.5 |

In Europe, some species of *Pulsatilla* are rare, endangered and endemic. Those taxa are protected due to their small populations and disappearing localities, and those species have been placed on the Red Lists of Endangered Species25. Taxonomically, *Pulsatilla* is an especially complex and challenging group. In all treatments published before, three subgenera have been recognized: subgenus *Kostyczewianae* (only one species), subgenus *Preonanthus*, and the largest subgenus *Pulsatilla*. However, the intragenic morphological variability of *Pulsatilla* was especially complicated25. The recognition and identification of wild *Pulsatilla* species is particularly difficult based on traditional approaches.

Molecular markers are significant to explore the phylogenetic relationships of the genus *Pulsatilla*. Phylogenetic relationships between *Pulsatilla* and closely related genera have been dedicated during the past years21–25. Previous studies have attempted to identify these species among *Pulsatilla* with universal molecular markers, but the species resolution was relatively low25.

In this study, we present seven complete cp genomes from two subgenera of *Pulsatilla* obtained through next-generation sequencing (NGS) and genomic comparative analyses with four previously published cp genome sequences of *Pulsatilla* from NCBI, with *Anemoclema glaucifolium* as the outgroup. We identify microsatellites (SSRs), larger repeat sequences, and highly variable regions, with the aim of developing DNA barcodes and testing the feasibility of phylogenetic analyses of *Pulsatilla* using the chloroplast genome.
In accordance with the results from many other taxa\textsuperscript{26,27}. However, in some taxa, e.g. \textit{Amorphophallus} of Araceae, some genes were lost\textsuperscript{31}, and in others, e.g. \textit{Pelargonium}\textsuperscript{32}, the structure and gene order diverges from what is reported here and in most other angiosperms. The cp genomes of \textit{P. alpina}, \textit{P. grandis}, \textit{P. hirsutissima}, \textit{P. ludoviciana}, \textit{P. multifida} and \textit{P. occidentalis} had the same GC content of 37.6%, while \textit{P. dahurica} had a subtle difference (37.5%) compared with the others.

**Chloroplast genome comparison.** In most angiosperms, the IR regions of cp genomes of angiosperms are highly conserved, but the expansion and contraction of IR region boundaries are ever present\textsuperscript{33,34}. At the same time, several lineages of land plant chloroplast genomes show great structural rearrangement, even loss of IR regions or some gene families\textsuperscript{35}. The expansion and contraction in IRs are significant evolutionary events, because they can change gene content and chloroplast genome size\textsuperscript{30,36}. Expansion of the IRs has been reported in Araceae\textsuperscript{36,37}. Sometimes, the size of LSC increases and that of SSC decreases, becoming only 7000 bp in \textit{Pothos}\textsuperscript{38}. At the same time, a linear chloroplast genome was also reported in some groups, e.g. maize\textsuperscript{35,39}. Expansion and contraction of the IR regions can also lead to duplication of certain genes or conversion of duplicate genes to single copy, respectively\textsuperscript{36,38}. Changes in the size of the IRs can also cause rearrangement of the genes in the SSC as recently observed in \textit{Zantedeschia}\textsuperscript{36}.

The \textit{Pulsatilla} chloroplast genomes were compared to previously published data and showed typical Anemoineae (Ranunculaceae) genome structure\textsuperscript{37,39}. As reported for \textit{Anemonelea}, \textit{Anemone}, \textit{Clematis} and \textit{Hepatica}, the
IR regions of genus *Pulsatilla* are roughly 4.4 kb longer than those of other genera of the family Ranunculaceae, such as *Aconitum*, *Coptis*, *Thalictrum*, *Megaleranthis*, *Ranunculus*, and *Trollius*[^37][^39]. The gene orders located within the IR-SSC and IR-LSC boundaries are similar among tribe Anemoneae but different from those of other genera of Ranunculaceae (Fig. 2, Fig. S1).

We compared the IR/SC boundary regions of *Pulsatilla*, and the junction positions are very similar and conserved within genus *Pulsatilla*. In the four boundary regions (LSC, IRa, SSC, IRb) of seven *Pulsatilla* cp genomes, the LSC/IRa and IRb/LSC border was in the intergenic region, and the adjacent genes is *rps36*, *rps8* and *rps4*, respectively. The genes *ycf1* and *ψycf1* have crossed the SSC/IRb and IRa/SSC boundary, respectively, which was also found in Monsteroideae (Araceae)^[^28]. The pseudogene *ycf1* has been found in other groups^[^30][^36]. The IR regions were highly conserved, with nucleotide diversity values in those regions less than 2%.

**Sequence divergence.** Multiple alignments of plastid genomes were performed to investigate levels of genome divergence. Based on MAFFT analysis, there are three inversions in LSC of *Pulsatilla*, same as the tribe Anemoneae (Fig. 1, Fig. S1)^[^37]. The mVISTA analysis has revealed high sequence similarity across the coding region and there exists more variability in non-coding regions. Sequence identity among the seven species was 96.68–98.66%. The number of nucleotide substitutions and sequence distance (Pi) were the highest (1214, 0.0063) between *P. alpine* and *P. dahurica*, with the lowest (112, 0.0005) between *P. vernalis* and *P. patens* (Fig. 3; Table 3).

**Identification of highly variable regions.** Chloroplast genome markers, especially several universal chloroplast regions, have been widely used in plant systematics and identification at multiple taxonomic levels. Highly suitable polymorphic chloroplast loci have been identified and designed as unique markers in different groups[^28][^40]. However, relationships within the genus *Pulsatilla* have not been well resolved because of the low

---

[^37]: Figure 2. Comparisons of LSC, SSC, and IR region borders among the seven *Pulsatilla* chloroplast genomes.
polymorphism of these universal markers\textsuperscript{15}. In order to facilitate identification of closely related species of \textit{Pulsatilla}, we sought to identify highly variable regions of the chloroplast genome, as previously described\textsuperscript{9,27,41–44}. As a result, we identified nine divergent hotspot regions, including six intergenic spacer regions (\textit{rps4-rps16}, \textit{rps16-matK}, \textit{ndhC-trnV}, \textit{psbE-petL}, \textit{ndhD-ccsA}, \textit{ccsA-ndhF}) and four protein-coding regions (\textit{ycf1}, \textit{ndhF}, \textit{ndhI}) (Fig. 4; Table 4). Most commonly employed loci, e.g. \textit{trnL-trnF}, \textit{trnH-psbA} were not selected in our finding. The nine highly variable regions included 684 variable sites, including 181 indels. However, these indels are not suitable for the phylogenetic inference because Maximum likelihood model used only substitutions not indels\textsuperscript{28}. Their nucleotide diversity values ranged from 0.00802 to 0.02212. The region of \textit{ccsA-ndhF} showed the highest variability, the next most variable regions were \textit{rps4-rps16}, \textit{ndhC-trnV}, and \textit{psbE-petL}. The diversity level of two protein-coding regions (\textit{ycf1}, \textit{ndhF}) was the lowest. Among the nine divergent hotspot regions, the \textit{ndhI} is difficult to align. There are large numbers of indels in \textit{ndhI} and the intergenic spacer between \textit{ndhI} and \textit{ndhG}, these regions were not considered suitable for the phylogenetic inference of \textit{Pulsatilla}. Thus, we selected eight regions, four (\textit{rps4-rps16}, \textit{rps16-matK}, \textit{ndhC-trnV}, \textit{psbE-petL}) in the LSC and four (\textit{ndhD-ccsA}, \textit{ccsA-ndhF}, \textit{ycf1}, \textit{ndhF}) in the SSC, with relatively high variability as potential molecular markers for the study of species identification and phylogeny in \textit{Pulsatilla}. Five hotspots were found in chloroplast genome of Veroniceae (Plantaginaceae), and two universal marker, \textit{trnH-psbA} and \textit{matK} were identified, respectively\textsuperscript{45,46}. Ten highly variable regions were selected as potential molecular markers for \textit{Fritillaria}, including \textit{ycf1}\textsuperscript{44,46}, which was also selected in this study. Sequences of these variable regions founded in this study could be regarded as potential molecular markers for species identification and evolutionary studies and have been shown to be valuable for studies in other groups (e.g., \textit{Fritillaria})\textsuperscript{44}. **SSRs and large repeat sequences.** Oligonucleotide repeats play an important role for generating indels, inversion and substitutions\textsuperscript{29}. Repeat sequences in the chloroplast genome could provide valuable information for understanding not only the sequence divergence but the evolutionary history of the plant\textsuperscript{48–50}. We have detected five types of large repeats (forward, reverse, palindromic, complement and tandem repeats) in the seven \textit{Pulsatilla} cp genomes. Among them, the most common repeat types are forward and palindromic repeats, followed by reverse repeats, and only little complement repeats were found in \textit{Pulsatilla} cp genomes (Fig. 5A). Most of the repeats were short, ranging from 30–49 bp (Fig. 5B). We also identified multiple microsatellite repeats, also known as simple sequence repeats (SSR) or short tandem repeats (STR)\textsuperscript{49}. Due to their codominant inheritance and high variability, SSRs are robust and effective.
markers for species identification and population genetic analyses. Most of the mononucleotide repeats were composed of A/T. The other microsatellites types were also dominated by AT/TA, with very little G/C (Fig. 5C). In this study, plentiful microsatellite loci were found through the comparative analysis of *Pulsatilla* cp genome sequences. In total, we detected six types of microsatellite (mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeats) based on the comparison of seven *Pulsatilla* cp genomes (Fig. 5D). Each *Pulsatilla* cp genome had 69–87 microsatellites. The lengths of repeat motifs of these microsatellites ranged from 10 to 21 bp (Fig. 5E). Among the four structural regions in the cp genomes, most of the repeats and microsatellites were distributed in LSC, followed by SSC, and fewest in IRa/IRb (Fig. 5F), which were also reported in other studies in angiosperms. These SSRs and repeat sequences are uncorrelated with genome size and phylogenetic position of the species, but will provide important information for further studies of phylogenetic reconstruction and infra- and inter-specific genetic diversity.

**Phylogenetic analyses.** Chloroplast genomes have been widely used and have made significant contributions to phylogeny reconstruction at different taxonomic levels in plants. To better clarify the evolutionary
relationships within *Pulsatilla*, we used each data set to construct phylogenetic trees using the ML analytical methods. We also construct phylogenetic trees with those eight highly variable regions using the ML, MP analytical methods. All tree topology structures were identical. Therefore, here we presented the phylogenetic studies using the ML tree with the support values from the MP analyses recorded at the corresponding nodes (Fig. 6).

The phylogenetic tree based on all data sets (except the IR and SSC regions) from the complete plastid genome sequences yielded the same topology. The phylogenetic tree based on chloroplast genome differed from that of Table 4. Sequence characteristics of eight high variable regions among eleven complete cp genomes of *Pulsatilla*.

| Region       | Aligned length | Variable sites | No | %   | Indels | No | Length range | Nucleotide diversity (Pi) |
|--------------|----------------|----------------|----|-----|--------|----|--------------|--------------------------|
| rps14-rps16  | 986            | 22             | 2.23 | 11 | 1–44   | 0.01497 |
| rps16-matK   | 1985           | 62             | 3.12 | 23 | 1–79   | 0.00998 |
| ndhC-trnV    | 1484           | 61             | 4.11 | 18 | 1–34   | 0.01502 |
| psbE-psbL    | 1469           | 70             | 4.77 | 24 | 1–139  | 0.01140 |
| ndhD-ccfA    | 732            | 39             | 5.33 | 10 | 1–87   | 0.04368 |
| ccsA-ndhF    | 2757           | 196            | 7.11 | 69 | 1–43   | 0.02212 |
| ycf1         | 5807           | 148            | 2.55 | 18 | 1–24   | 0.00802 |
| ndhF         | 2345           | 86             | 3.67 | 8  | 1–18   | 0.00813 |

Figure 5. Analyses of repeated sequences in seven newly sequenced chloroplast genomes. (A) Number of five repeat types; (B) frequency of four repeats by length; (C) frequency of microsatellites by base composition; (D) frequency of microsatellites by types; (E) frequency of microsatellites by length; (F) number of all repeats by location.
by sequencing the complete plastid genome using the Illumina Miseq platform (Illumina, San Diego, CA, USA). We extracted total genomic DNA with the DNeasy Plant mini kits (QIAGEN, Guangzhou, China). Fresh leaves of P. dahurica were collected from Jilin province of China and dried with silica gel. Dry leaves of other six Pulsatilla species were taken from herbarium specimens. We extracted total genomic DNA with the DNeasy Plant mini kits (QIAGEN, Guangzhou, China). The genomic DNA was sequenced using the Illumina Miseq platform (Illumina, San Diego, CA, USA).

Materials and methods

Plant material, DNA extraction and sequencing. Fresh leaves of P. dahurica were collected from Jilin province of China and dried with silica gel. Dry leaves of other six Pulsatilla species were taken from herbarium specimens. We extracted total genomic DNA with the DNeasy Plant mini kits (QIAGEN, Guangzhou, China). The genomic DNA was sequenced using the Illumina Miseq platform (Illumina, San Diego, CA, USA).

Chloroplast genome assembly and annotation. Whole chloroplast genome sequencing was done for the seven species of Pulsatilla. For each species, high-quality Illumina sequencing reads were assembled into scaffolds with de novo sequence assembly software Spades, SOAPdenovo and CLC Genomics Workbench v.6.5 (CLC Bio), respectively. We checked the contigs against the reference genome of P. chinensis (MG001341), using BLAST (https://blast.ncbi.nlm.nih.gov/) and oriented aligned contigs according to the reference genome. We mapped all the raw reads back to assembled sequences to check the assembly and then constructed the complete cp genomes using Geneious v.9.0.7. We submitted all the newly sequence data in raw format (fastq) and obtained SRA accessions (Table 1).

Annotations of cp genome sequences were performed using Plastid Genome Annotator (https://github.com/quxiaojian/PGA) and adjusted in Geneious v.9.0. To verify the accuracy of the annotations, we also used GeSeq to annotate each chloroplast genome in this study. We checked every boundary of tRNAs using trnascan-SE v.2.0.6. The circular genome maps were generated in OGDRAW (https://ogdraw.mpimp-golm.mpg.de/) and hotspots by DnaSP v5.

Genome comparisons. We aligned the cp genomes of Pulsatilla using multiple alignment of MAFFT v7 and manually edited in Geneious v.9.0. The contraction and expansion of inverted repeat regions were also examined among the seven species (excluded P. chinensis, MG001341) of the genus Pulsatilla using Isrscope. Then, we performed multiple alignments of the eight genomes of Pulsatilla in the mVISTA program under Shuffle-LAGAN mode, with default parameters for other options, using the annotation genome of P. chinensis as a reference, with the aim of comparing and visualized the similarities and differences among different plastid genomes.

To analyse chloroplast genome organisation and gene arrangement, we perform the analyses of collinear blocks with Mauve v 2.3.1 plugin in Geneious v.9.0, including only one copy of the IR, assuming collinear genomes for the full alignment. Detailed gene inversions were identified by comparing the gene order of Pulsatilla samples and Anemonelema to Berberis.

To observe the plastid genome divergence and determine parsimony informative sites, we conducted sliding window analysis after alignment to determine the nucleotide diversity (Pi) of the cp genome using DnaSP v5, with 200 bp of step size and 600 bp window length. We defined hotspots as those regions with a higher value of Pi. We computed the variable sites across the complete cp genomes and the sequence characteristics of hotspots by DnaSP v5.0.

Repeated sequences identification. We identified repeat sequences, including palindromic, reverse and forward repeats, using the online software REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer), with the following settings: Hamming distance of 3 and minimum repeat size of 30 bp. We used the online program
Figure 6. Phylogenetic relationships of the eleven Pulsatilla species inferred from maximum likelihood (ML). Including whole chloroplast genome, rbcL + matK + trnH-psbA, LSC region, coding region, SSC region, IR region, and the concatenation of the eight highly variable regions mentioned in Table 4 (Numbers above nodes are support values with ML bootstrap values on the left, and MP bootstrap values on the right).
Tandem Repeats Finder (https://tandem.bu.edu/trf/trf.html) to find the tandem repeat sequences, in which the similarity percentage of two repeat copies was at least 90% and the minimal repeat size was 10 bp. The alignment parameters for match, mismatch, and indels were set at 2, 7, and 7, respectively. We identified microsatellites (SSRs) by MISA with thresholds of 10, 5, 4, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide, respectively.

**Phylogenetic analyses.** For the purpose of reconstructing the phylogenetic relationships, four published complete cp genome sequences from the genus *Pulsatilla* and *Anemoclema* were also included in our analyses. The monotypic genus *Anemoclema* (MH205609) was selected as the outgroup. Because molecular evolutionary rates among the different cp genome regions are diverse, analyses of phylogenetic relationships were performed based on the following seven datasets: (a) the complete cp genome sequences; (b) coding genes (CDS); (c) one inverted repeat (IR) region (IRb); (d) the large single copy region (LSC); (e) the small single copy region (SSC); (f) the consensus sequences of eight highly variable regions; and (g) the DNA barcodes combination (*rbcL* + *matK* + *trnH-psbA*). We applied Maximum Likelihood (ML) analysis for each of the seven datasets to construct tree-sets. Maximum Parsimony (MP) analyses were also applied for the consensus sequences of eight highly variable regions and the DNA barcodes combination.

We conducted ML analyses with RAXMLHCPC2 v.8.0.9 on the Cyberinfrastructure for Phylogenetic Research (CIPRES) Science Gateway v.3.3. Then the analysis of 1000 rapid bootstrap replicates (-x) was followed by a search for the best-scoring ML tree in one program (-f a). The best-fit model for nucleotide and amino acid sequences were evaluated using jModelTest 2. We applied the GTR+G model to nucleotide data for both bootstrapping and best-tree searching phases, with other parameters as the default settings. We performed the maximum Parsimony (MP) analysis on PAUP* v.4.0b10. All the characters were treated as unordered and equally weighted. The heuristic search specified 1000 random sequence addition replicates with TBR branch swapping, saving only 10 trees per replicate. We obtained the strict consensus tree from all the most-parsimonious trees (MPTs) detected during the search. We calculated bootstrap percentages (BP) from 10,000 rapid bootstrap replicates, each comprising 10 random sequence addition replicates and saving only one tree per replicate.

Received: 17 February 2020; Accepted: 29 October 2020
Published online: 13 November 2020

**References**

1. Hebert, P. D. N., Cywinska, A., Ball, S. L. & Waard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B Biol. Sci.* **270**, 313–321 (2003).
2. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8369–8374 (2005).
3. Kress, W. J. et al. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18621–18626 (2009).
4. CBOL Plant Working Group. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12794–12797 (2009).
5. Li, X. et al. Plant DNA barcoding: from gene to genome. *Biol. Rev. Camb. Philos. Soc.* **90**, 157–166 (2015).
6. Hollingsworth, P. M., Graham, S. W. & Little, D. P. Choosing and using a plant DNA barcode. *PLoS ONE* **6**, e19254 (2011).
7. Ahmed, I. et al. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasiaesculenta* (L.) Schott (Araceae) and closely related taxa. *Mol. Ecol. Resour.* **13**, 929–937 (2013).
8. Kuang, D. Y. et al. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* **54**, 663–673 (2011).
9. Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* **7**, e35071 (2012).
10. Tamura, M. Ranunculaceae. in *Die natürlichen pflanzenfamilien*. Aufl. II. 2nd ed. 17 a IV (ed Hiepko, P.) (Duncker und Humblot, Berlin, 1995).
11. Wang, W. C. et al. *Flora of China* Vol. 6, 133–438 (Science Press, Beijing; Missouri Botanical Garden Press, St. Louis, 2001).
12. Grey-Wilson, C. *Pasque-flowers. The Genus Pulsatilla*. (The Charlotte Louise Press, Kenning Hall, 2014).
13. Ren, Y., Gu, T. Q. & Chang, H. L. Floral development of *Dichocarpum*, *Thalictrum*, and *Aquilegia* (Thalictridoeae, Ranunculaceae). *Plant Syst. Evol.* **292**, 203–213 (2011).
14. Ren, Y., Chang, H. L. & Endress, P. K. Floral development in *Anemone* (Ranunculaceae). *Bot. J. Linn. Soc.* **162**, 77–100 (2015).
15. Li, Q. J. et al. Efficient identification of *Pulsatilla* (Ranunculaceae) using DNA barcodes and micro-morphological characters. *Front. Plant Sci.* **10**, 1196 (2019).
16. China Pharmacopoeia Committee. *Pharmacopoeia of the People’s Republic of China* (China Chemical Industry Press, Beijing, 2015).
17. Xu, Q. et al. Antitumor activity of *Pulsatilla chinensis* (Bunge) Regel saponins in human liver tumor 7402 cells in vitro and in vivo. *Phytochemistry* **19**, 293–300 (2012).
18. Wang, X. W., Fan, F. G. & Cao, Q. Modified *Pulsatilla* decoction attenuates oxazolone-induced colitis in mice through suppression of inflammation and epithelial barrier disruption. *Mol. Med. Rep.* **14**, 1173–1179 (2016).
19. Sulh, S. Y. & An, W. G. Systems pharmacological approach of *Pulsatilla* radix on treating Crohn’s disease. *Evid. Based Complement. Altern. Med.* **2017**, 1–21 (2017).
20. Szczecinska, M. & Sawicki, J. Genomic resources of three *Pulsatilla* species reveal evolutionary hotspots, species-specific sites and variable plastid structure in the family Ranunculaceae. *Int. J. Mol. Sci.* **16**, 22258–22279 (2015).
21. Hoot, S. B. Phylogenetic relationships in *Anemone* (Ranunculaceae) based on morphology and chloroplast DNA. *Syst. Bot.* **19**, 169–200 (1994).
22. Hoot, S. B. Phylogenetic relationships in *Anemone* (Ranunculaceae) based on DNA restriction site variation and morphology. In *Systematics and Evolution of the Ranunculiflorae* (eds Jensen, U. & Kaderet, J. W.) 295–300 (Springer, Vienna, 1995).
23. Szczecinska, M., Gabor, S., Katarzyna, W., Jakub, S. & Dusan, G. Genetic diversity and population structure of the rare and endangered plant species *Pulsatillaplatens* (L.) Mill in east central Europe. *PLoS ONE* **11**, e0151730 (2016).
24. Jiang, N. et al. Phylogenetic reassessment of tribe *Anemoneae* (Ranunculaceae): non-monophyly of *Anemone* s.l. revealed by plastid datasets. *PLoS ONE* **12**, e0174792 (2017).
25. Sramkó, G., Laczkó, L., Volkova, P. A., Bateman, R. M. & Milnaréc, J. Evolutionary history of the Pasque-flowers (Pulsatilla, Ranunculaceae): molecular phylogenetics, systematics and rDNA evolution. *Mol. Phylogenet. Evol.* 135, 45–61 (2019).

26. Yu, X. Q., Yang, D., Guo, C. & Gao, L. M. Plant phylogenomics based on genome-partitioning strategies: progress and prospects. *Plant Divers.* 40, 158–164 (2018).

27. Tang, Y., Yukawa, T., Bateman, R. M., Jiang, H. & Peng, H. Phylogeny and classification of the East Asian *Amintostigma alliance* (Orchidaceae: Orchideae) based on six DNA markers. *BMC Evol. Biol.* 15, 96 (2015).

28. Henriquez, C. L. et al. Molecular evolution of chloroplast genomes in Monimiaceae (Malpighiaceae). *Planta* 251, 72 (2020).

29. Abdullah, et al. Correlations among oligonucleotide repeats, nucleotide substitutions, and insertion–deletion mutations in chloroplast genomes of plant family Malvaceae. *J. Syst. Evol.* https://doi.org/10.1111/jse.12585 (2020).

30. Abdullah, et al. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. *Genomics* 112, 581–591 (2020).

31. Liu, E. 

32. Chumley, T. W. 

33. Ogihara, Y. 

34. Daniell, H., Lin, C. S., Yu, M. & Chang, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Plant Divers.* 40, 158–164 (2018).

35. Henriquez, C. L. et al. Evolutionary dynamics in chloroplast genomes of subfamily Aroideae (Araceae). *Genomics* 112, 2349–2360 (2018).

36. Wang, W. & Messing, J. High-throughput sequencing of three Lemnoidae (duckweeds) chloroplast genomes from total DNA. *PLoS ONE* 6, e24670 (2011).

37. Abdulch., et al. Complete chloroplast genome of *Anthurium haustatum* and *Pothos scandens* (Pothoideae, Araceae): unique inverted repeat expansion and contraction affect rate of evolution. *J. Mol. Evol.* 88, 562–574 (2020).

38. Oldenburg, D. J. & Bendich, A. J. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Curr. Genet.* 62, 431–442 (2016).

39. Sun, J. H. et al. Evolutionary and phylogenetic aspects of the chloroplast genome of *Chaenomeles* species. *Sci. Rep.* 10, 11466 (2020).

40. Liu, H. J. et al. Comparative analysis of complete chloroplast genomes of *Anemoclema, Anemone, Pulsatilla,* and *Hepatica* revealing structural variations among genera in tribe Anemoneae (Ranunculaceae). *Front. Plant Sci.* 9, 1097 (2018).

41. Meng, J. et al. Comparative analysis of the complete chloroplast genomes of four *Aconitum* medicinal species. *Molecules* 23, 1015 (2018).

42. Du, Y. P. et al. Complete chloroplast genome sequences of *Lilium:* insights into evolutionary dynamics and phylogenetic analyses. *Sci. Rep.* 7, 5751 (2017).

43. Bi, Y. et al. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on *Fritillaria.* *Sci. Rep.* 8, 1184 (2018).

44. Choi, K. S., Chung, M. G. & Park, S. The complete chloroplast genome sequences of three Veroniceae species (Plantaginaceae): comparative analysis and highly divergent regions. *Front. Plant Sci.* 7, 1–8 (2016).

45. Menezes, A. P. A. et al. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. *Sci. Rep.* 8, 1–12 (2018).

46. Li, Y., Zhang, Z., Yang, J. & Li, G. Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS ONE* 13, e0194613 (2018).

47. Grassi, F., Labra, M., Scienza, A. & Imazio, S. Chloroplast SSR markers to assess DNA diversity in wild and cultivated grapevines. *Vitis* 41, 157–158 (2002).

48. Chen, C., Zhou, P., Choi, Y. A., Huang, S. & Gmitter, F. G. Jr. Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.* 112, 1248–1257 (2006).

49. Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G. & Rafalski, J. A. Polymorphic simple sequence repeat regions in chloroplast DNA. *Curr. Genet.* 112, 2349–2360 (2018).

50. Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G. & Rafalski, J. A. Polymorphic simple sequence repeat regions in chloroplast DNA. *Curr. Genet.* 112, 2349–2360 (2018).

51. Kaundun, S. S. & Matsumoto, S. Heterologous nuclear and chloroplast microsatellite amplification and variation in tea, *Camellia sinensis.* *Genome Biol.* 45, 1041–1048 (2002).

52. Doorduin, L. et al. The complete chloroplast genome of 17 individuals of pest species *Jacobaee vulgaris:* SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Res.* 18, 93–105 (2011).

53. Jiao, Y. et al. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). *BMC Genomics* 13, 201 (2012).

54. He, S., Wang, Y., Yuan, S., Li, D. & Yi, T. Genetic diversity and population structure: implications for conservation of wild soybean (*Glycine soja* Sieb. et Zucc.) based on nuclear and chloroplast microsatellite variation. *Int. J. Mol. Sci.* 13, 12608–12628 (2012).

55. Zhang, N. et al. An analysis of *Echinacea* chloroplast genomes: implications for future botanical identification. *Sci. Rep.* 7, 216 (2017).

56. Li, H. T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470 (2019).

57. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649 (2012).

58. Qi, X. L., Moore, M., Li, D. D. & Yi, T. S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15, 12 (2019).

59. Tillich, M. et al. GeoSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11 (2017).

60. Lowe, T. M. & Chen, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44, W54–W57 (2016).

61. Lobse, M., Drexchel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274 (2007).

62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).

63. Amirouyesei, A., Hyvönen, J. & Poczi, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031 (2018).

64. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279 (2004).

65. Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403 (2004).

66. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452 (2009).

67. Kurtz, S. et al. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4653–4662 (2001).
68. Thiel, T., Michealek, W., Varshney, R. & Graner. A. Exploiting EST databases for the development and characterization of gene-derived SSR markers in barley (Hordeum vulgare L.). Theor. Appl. Genet. 106, 411–422 (2003).
69. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
70. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES science gateway for inference of large phylogenetic trees. New Orleans, LA. Proceedings of the Gateway Computing Environments Workshop (GCE), pp. 1–8 (2010).
71. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772 (2012).
72. Swofford, D. L. PAUP*: phylogenetic analysis using Parsimony (*and other methods), version 4.0b10 (Sinauer Associates, Sunderland, 2003).

Acknowledgements
We sincerely thank Prof. Zhong-hu Li, Dr. Jian-jun Jin, De-long Guan and Molecular Biology Experiment Center, Germplasm Bank of Wild Species in Southwest China for help with data analysis. This project was supported by the National Nature Science Foundation of China (No. 31770200, 31300158, 31770203, 31660048) and the Fundamental Research for the Central Universities (No. 2452020179).

Author contributions
L.Zhao designed the research. Q.J.L., L.Zhang, X.H.Z., Z.Y.C., and L.Zhao collected the samples. Q.J.L., N.S., L.Zhang, R.C.T., and J.R.W conceived the experiments. Q.J.L., N.S., L.Zhang, X.H.Z., Z.Y.C., L.Zhao and D.P. did computational analysis and deposited sequences. Q.J.L., N.S., L.Zhang, X.H.Z., Z.Y.C., L.Zhao and D.P. wrote the manuscript. All authors have read and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-76699-7.

Correspondence and requests for materials should be addressed to L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. This article and the version of record therein are made available under the terms of the Creative Commons Attribution 4.0 International License, which permits any use, distribution and reproduction in any medium or format, provided the original work is properly cited. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020