Gene losses and partial deletion of small single-copy regions of the chloroplast genomes of two hemiparasitic Taxillus species

Ying Li¹, Jian-guo Zhou¹, Xin-lian Chen¹, Ying-xian Cui², Zhi-chao Xu¹, Yong-hua Li², Jing-yuan Song¹, Bao-zhong Duan³ & Hui Yao¹

Numerous variations are known to occur in the chloroplast genomes of parasitic plants. We determined the complete chloroplast genome sequences of two hemiparasitic species, Taxillus chinesis and T. sutchuenensis, using Illumina and PacBio sequencing technologies. These species are the first members of the family Loranthaceae to be sequenced. The complete chloroplast genomes of T. chinesis and T. sutchuenensis comprise circular 121,363 and 122,562 bp-long molecules with quadripartite structures, respectively. Compared with the chloroplast genomes of Nicotiana tabacum and Osyris alba, all ndh genes as well as three ribosomal protein genes, seven tRNA genes, four ycf genes, and the infA gene of these two species have been lost. The results of the maximum likelihood and neighbor-joining phylogenetic trees strongly support the theory that Loranthaceae and Viscoaceae are monophyletic clades. This research reveals the effect of a parasitic lifestyle on the chloroplast structure and genome content of T. chinesis and T. sutchuenensis, and enhances our understanding of the discrepancies in terms of assembly results between Illumina and PacBio.

The chloroplast is a key plant cell organelle that carries out photosynthesis¹. The chloroplast genome is highly conserved and has multiple copies, which means that target genes are expressed at high levels²,³. In recent years, the chloroplast genome has increasingly been used as a source of molecular markers⁴,⁵ and barcoding identification⁶,⁷. Given that parasitic plants have either lost some or all their photosynthetic capacity, they absorb organic and inorganic nutrients as well as water from their hosts by maintaining a much higher transpiration rate and using specialized parasitic organs called haustoria⁸,⁹. Despite their large known diversity, only a few chloroplast genomes from parasitic plants have been obtained. The first complete parasitic plant chloroplast genome to be sequenced was from Epifagus virginiana¹⁰. All photosynthesis and energy producing genes in this species have been lost, although a few fragments remain as pseudogenes, and the entire chloroplast genome no longer performs photosynthesis¹¹. Subsequently sequenced chloroplast genomes include four species from the holoparasitic genus Cuscuta, including C. reflexa, C. gronovii, C. exaltata and C. obtusiflora¹²,¹³. Previous studies showed that the chloroplast genome of Rafflesia lagascae is completely lost¹⁴. The complete chloroplast genomes of several species within the parasitic family Orobanchaceae have been sequenced and analyzed in recent years, including the completely non-photosynthetic plants, Cistanche deserticola¹, Phelipanche ramosa⁴, Orobanche austrohispanica⁵, and Lathraea squamaria¹⁶. More recently, Petersen et al. sequenced and analyzed the complete chloroplast genome of one species of the genus Osyris and three species of the genus Visco⁶. A number of photosynthetic

¹The Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicine, Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, 100193, China. ²Department of Pharmacy, Guangxi Traditional Chinese Medicine University, Nanning, 530200, Guangxi, China. ³College of Pharmaceutical Science, Dali University, Dali, 671000, Yunnan, China. Ying Li and Jian-guo Zhou contributed equally to this work. Correspondence and requests for materials should be addressed to H.Y. (email: scauyaoh@sina.com)
and photorespiratory genes, some protein-coding genes, ribosomal protein genes, transfer RNA (tRNA) genes from some parasitic plants have either been completely lost or pseudogenized\textsuperscript{[23–25]}. Horizontal gene transfer also occurs between donor and recipient in some parasitic plants\textsuperscript{[26]}. Plants within Loranthaceae comprise hemiparasitic species that have retained photosynthesis and have seeds which are widely propagates by birds\textsuperscript{[27]}. The taxonomy of plants within Loranthaceae is controversial, particularly regarding the branching point between these taxa and Viscaceae. To date, the plants in China classified within Loranthaceae have been studied and the results demonstrated that, apart from the hemiparasitic characteristics, significant differences exist in pollen morphology\textsuperscript{[28]}, chemical composition\textsuperscript{[29]}, and DNA molecules\textsuperscript{[30]}, which nevertheless support the theory that Loranthaceae and Viscaceae are branched independently. However, one medicinal plant within Viscaceae, namely, Viscum coloratum (Kom.) Nakai, is assigned to Loranthaceae in the Chinese Pharmacopoeia\textsuperscript{[31]}. Approximately 70 genera comprising more than 900 species are classified within Loranthaceae\textsuperscript{[32]}. Most of these plant species primarily live in tropical and subtropical regions, with 8 genera and 51 species (18 endemic) found in China\textsuperscript{[33]}. Of these genera, the hemiparasitic plant genus Taxillus consists of species with degenerated chloroplasts and restricted photosynthetic capacity. Specifically, T. chinensis is used in traditional Chinese herbal medicine and is recorded in the Chinese Pharmacopoeia\textsuperscript{[34]}. Another species, namely, T. sutchuenensis, is used in folk medicine. These two medicinal plants are commonly used to treat diseases, such as rheumatism, hypertension, and fetal irritability\textsuperscript{[35–37]}. The recorded hosts of T. chinensis and T. sutchuenensis include species within Moraceae, Rutaceae, Acanthaceae, Anacardiaceae, Euphorbiaceae, Rosaceae, Theaceae and rarely Taxodiaceae\textsuperscript{[38]}. The third-generation sequencing platform PacBio is based on single-molecule real-time (SMRT) sequencing technology. The main advantage of this sequencing approach is the long read length, generating read lengths of over 10 kb on average, with some reads possibly reaching up to 60 kb\textsuperscript{[39–41]}. Previous studies have demonstrated that the long read lengths provide many benefits in genome assembly, including generating longer contigs and fewer unresolved gaps\textsuperscript{[39]}. PacBio has been successfully applied in a number of chloroplast genome sequencing projects involving three species of Fritillaria\textsuperscript{[42]}, Aconitum barbatum var. puberulum\textsuperscript{[43]}, and Swertia mussotii\textsuperscript{[44]}. However, PacBio has high rates of random error in single-pass reads\textsuperscript{[37]}. In this study, the chloroplast genome sequence of T. chinensis was sequenced using second-generation Illumina platform and third-generation PacBio system to verify the accuracy of the genome sequence. We report the complete chloroplast genome of T. chinensis and T. sutchuenensis, which are the first two sequences completed within Loranthaceae. We also present a comparative analysis of the genetic changes together with chloroplast genomes of five other species, including the previously reported sequence of Viscum minimum, to determine the effect of a parasitic lifestyle on chloroplast structure and the genome content. We also analyzed the phylogenetic relationships of T. chinensis and T. sutchuenensis within Dicotyledoneae based on the complete chloroplast genomes to provide baseline data for systematic classification of Loranthaceae.

Results

Chloroplast Genome Structures of T. chinensis and T. sutchuenensis. Results show that the chloroplast genome sequence of T. chinensis is a circular molecule that is 121,363 bp in length, which can be divided into a large single-copy (LSC) region of 70,357 bp and a small single-copy (SSC) region of 6,082 bp, and separated by a pair of inverted repeats (IRa and IRb) each 22,462 bp in length (Fig. 1). This sequence, which was assembled using the reads obtained by the Illumina sequencing platform, is 121,363 bp in length. By contrast, the sequence assembled using the reads obtained by the PacBio system is 12 bp shorter than that assembled from the reads obtained by the Illumina platform. After verification using PCR, we found that the complete chloroplast genome of T. chinensis is consistent with the assembly results obtained using the reads from second-generation sequencing. The chloroplast genome of T. sutchuenensis is extremely similar to that of T. chinensis in size and genomic structure; it is 122,562 bp in length and retains a typical structure comprising a LSC (70,630 bp), a SSC (6,102 bp), and two IRs, each having 22,915 bp (Fig. 2). The complete and correct chloroplast genome sequences of T. chinensis and T. sutchuenensis were deposited in GenBank under accession numbers KY996492 and KY996493, respectively.

Data reveal that both species have a GC content of 37.3%, which is unevenly distributed across the whole chloroplast genome. In both cases, the GC content of the IR regions exhibits the highest values across the complete chloroplast genome, 43.0% in T. chinensis and 42.8% in T. sutchuenensis, respectively. This high GC content in IR regions is the result of four rRNA genes (rrn16, rrn23, rrn4.5 and rrn5) that occur in this region\textsuperscript{[45]}. In addition, after the LSC, which has a GC content of 34.7%, lowest values of 26.2% are seen in SSC regions.

A total of 106 genes were identified in each genome, which include 66 protein-coding genes, 28 tRNAs, 8 rRNAs, and 4 pseudogenes. Simultaneously, we compared these two Taxillus species with autotrophic plants, including Nicotiana tabacum and Osyris alba. Genes encoding subunits of the NAD(P)H dehydrogenase complex (ndh genes) were missing from the chloroplast genome of the two species, whereas three genes for ribosomal proteins (rpl2, rps15, and rps16), seven tRNA genes (trnA-UGC, trnG-UGC, trnH-GUG, trnL-GAU, trnK-UUU, trnL-UAA, and trnV-UCG), four ycf genes (ycf1, ycf5, ycf9, and ycf10), and initiation factor gene (infA) were also lost (Table S1). Two ribosomal protein genes (rpl16 and rpl2) and the duplicate gene ycf15 have also been pseudogenized because their gene-coding regions are interrupted by deletion, insertions or internal stop codons, while the pseudogene rpl2 is located in IRb region. We designed the primers to perform PCR to verify the accuracy of the pseudogenes rpl16, ycf15 and rpl2. The primer sequences are listed in Supplementary Table S2.

The basic information and gene contents of the chloroplast genomes of T. chinensis and T. sutchuenensis compared to other five species are presented in Table 1 and Supplementary Table S1.

Introns play an important role in the regulation of gene expression. Introns enhance exogenous gene expression at specific sites within plants at particular times, resulting in desirable agronomic traits\textsuperscript{[46]}. Introns within these two species are similar to other angiosperms\textsuperscript{[44,45]}. Results reveal the presence of nine genes containing introns in each chloroplast genome, including atpF, rpoC1, ycf3, rps12, rpl2, ycf16, clpP, petB, and petD. In addition, the
ycf3 gene and rps12 gene each contain two introns and three exons. The ycf3 gene is located within the LSC, as seen in *Metasequoia glyptostroboides*[^56], *Aquilaria sinensis*[^57], while the rps12 gene is specialized for trans-splicing. The 5' exon is located in the LSC, and the 3' exon is located in the IR, as is the case in *Panax ginseng*[^44], *C. deserticola*[^1], and *L. squamaria*[^23]. Relevant lengths of exons and introns are listed in Table 2.

**Comparative genome analyses.** Data plotted using mVISTA (Fig. S1) reveal that non-coding regions of the chloroplast genomes of the two Taxillus species are more divergent than their coding counterparts. Moreover, the two IR regions have lower sequence divergence than the LSC and SSC regions. Similar results were obtained in previous research on the complete chloroplast genomes of five Lamiales species[^42] as well as in a comparative study of five *Epimedium* chloroplast genomes[^45]. In the present study, *rpl16* gene is the most divergent of the coding regions, probably because of pseudogenization. Thus, we conducted a series of linear rearrangement comparisons across the complete chloroplast genome sequences of six species (*T. chinensis, T. sutchuenensis, S. jasminodora, V. minimum, O. alba* and *N. tabacum*) aligned in Geneious using the Mauve algorithm (Fig. 3). The comparisons reveal the presence of two structural variants, including an approximately 24-kb-long inversion within the LSC region of the *V. minimum* chloroplast genome and an approximately 3-kb-long inversion in the SSC region of the *O. alba* chloroplast genome, which is consistent with a previous report[^44]. The lengths of the IR regions in our two species are also similar to that of other plants, with the exception of *S. jasminodora*[^48] where they are much shorter (at least 10 kb) than the length of the IR regions of the five species considered here, including *T. chinensis* and *T. sutchuenensis*.

**Codon Usage.** The calculations for the codon usage of protein-coding genes within *T. chinensis* and *T. sutchuenensis* chloroplast genomes are summarized in Fig. 4 and Supplementary Table S3. Results reveal the presence of 63 codons encoding 20 amino acids within the chloroplast protein-coding genes of these two species; of these,
Figure 2. Gene map of the complete chloroplast genome of *T. sutchuenensis*. Genes on the inside of the circle are transcribed clockwise, while those outside are transcribed counter clockwise. The darker gray in the inner circle corresponds to GC content, whereas the lighter gray corresponds to AT content.

Table 1. Comparisons among the chloroplast genome characteristics of *T. chinensis*, *T. sutchuenensis*, and other five species.
1711 encode leucine and 191 encode cysteine, which are respectively the most and least prevalent amino acids in *T. chinensis* chloroplast genome. Results also reveal that most of the amino acid codons have preferences, with the exception of methionine and tryptophan. Moreover, usage is generally biased toward A or T with high relative synonymous codon usage (RSCU) values, including TTA (2.12) in leucine, TAT (1.62) in tyrosine, and the stop-codon TAA (1.84) in the *T. sutchuenensis* chloroplast genome (Supplementary Table S3). The data presented in Fig. 4 illustrates that the RSCU value increases with the quantity of codons that code for a specific amino acid. High codon preference, especially a strong AT bias in codon usage, is very common in other land plant chloroplast genomes\(^42,44\). The present results are similar to the chloroplast genomes of *A. sinensis\(^46* and species within the genus *Ulmus\(^49* in terms of codon usage.

**Simple Sequence Repeats (SSRs) Analyses.** SSRs are ubiquitous throughout genomes and are also known as microsatellites. SSRs comprise tandem repeated DNA sequences that consist of between one and six repeat nucleotide units\(^50\). As such, SSRs are widely used as molecular markers in species identification, population genetics, and phylogenetic investigations because they exhibit high levels of polymorphism\(^51–53\). In total, 195 and 198 SSRs are identified within the chloroplast genomes of *T. chinensis* and *T. sutchuenensis*, respectively (Table 3; Supplementary Tables S4–S5), which mainly comprise mononucleotide repeats encountered 146 and 195 times in each case. In addition, A/T mononucleotide repeats (93.9% and 96.4%, respectively; Table 3) are the most common, while the majority of dinucleotide repeat sequences comprise AT/TA repeats (59.5% and 67.3%, respectively; Table 3). Results show that SSRs within the chloroplast genomes of *T. chinensis* and *T. sutchuenensis* are dominated by AT-rich repetitive motifs, which is consistent with the fact that AT content is also very high (62.7%) in these species. This result is also in agreement with previous studies showing higher proportions of polyadenine (polyA) and polythymine (polyT) relative to polycytosine (polyC) and polyguanine (polyG) within the chloroplast SSRs in many plants\(^4\).

**Phylogenetic Analyses.** Phylogenetic trees were constructed using two methods based on two datasets from different species (Fig. 5). Results revealed extremely similar tree topologies from each dataset irrespective of the method used, as supported by high bootstrap values. All nodes in our maximum likelihood (ML) and neighbor-joining (NJ) trees based on 54 protein-coding genes have 100% bootstrap support values, whereas four out of six nodes that received bootstrap values of ≥99% were recovered in both sets of trees when *matK* genes were used for analyses. All nodes in all phylogenetic trees received higher than 50% bootstrap support. All four phylogenetic trees showed that *T. chinensis* and *T. sutchuenensis* are sister taxa with respect to *S. jasminodora* (Olacaceae), whereas the three species within genus *Viscum* group with *Osyris alba* (Santalaceae) and all Santalales species are clustered within a lineage distinct from the outgroup.

**Discussion**

Numerous variations occur in the chloroplast genomes of parasitic plants. To date, however, most investigations on these genomes in parasitic and heterotrophic plants focused on nonphotosynthetic species\(^54\). For instance, some complete chloroplast genomes of holoparasitic plants from Orobanchaceae were reported\(^2,21,22\). A small number of hemiparasitic plants within Santalales and other groups have been studied\(^24,44\). In this study, the complete chloroplast genomes of *T. chinensis* and *T. sutchuenensis* from Santalales were assembled, annotated, and analyzed.

| Species | Gene | Location | Exon1(bp) | Intron1(bp) | Exon2(bp) | Intron2(bp) | Exon3(bp) |
|---------|------|----------|-----------|-------------|-----------|-------------|-----------|
| *T. chinensis* | *atpF* | LSC | 150 | 779 | 375 |
| | *clpP* | LSC | 335 | 621 | 229 |
| | *petB* | LSC | 6 | 755 | 642 |
| | *petD* | LSC | 9 | 718 | 483 |
| | *rpl2* | LSC, IR | 394 | 645 | 437 |
| | *ω/ρpl16* | LSC | 10 | 924 | 397 |
| | *ρps12* | LSC, IR | 114 | — | 232 | 543 | 26 |
| | *ρpsC1* | LSC | 456 | 752 | 1617 |
| | *ρef3* | LSC | 127 | 730 | 230 | 771 | 153 |

Table 2. Genes with introns in the chloroplast genomes of *T. chinensis* and *T. sutchuenensis* as well as the lengths of the exons and introns.
Figure 3. Comparison of the complete chloroplast genomes of six species using the MAUVE algorithm. Local collinear blocks are colored in this figure to indicate syntenic regions, while histograms within each block represent the degree of sequence similarity.

Figure 4. Codon content of 20 amino acid and stop codons in all protein-coding genes of the chloroplast genomes of two species. The histogram on the left-hand side of each amino acid shows codon usage within the T. chinensis chloroplast genome, while the right-hand side illustrates the genome of T. sutchuenensis.

Table 3. Types and amounts of SSRs in the T. chinensis and T. sutchuenensis chloroplast genomes.
Gene loss events have occurred within the chloroplast genomes of most parasitic plants and in a handful of autotrophic species. Previous work has shown that genes of *chlB*, *chlL*, *chlN*, and *trnP-GGG* have been lost from the chloroplast genomes of most flowering plants, whereas gene *infA*, which codes for a translation initiation factor, is either missing or has been transferred in many plants, including the two species observed in this study. All *ndh* genes have been lost from the chloroplast genomes of *T. chinensis* and *T. suchuenensis*, similar to the case of *Cuscuta gronovii* and *C. obtusiflora*. Similarly, nine out of eleven *ndh* genes have been pseudogenized within the chloroplast genome of *L. squamarilla*. This degree of *ndh* gene degradation is not only observed in heterotrophic organisms, but also in many autotrophic plants, including *Orchidaceae* and *Cactaceae*. Kim et al. reported that losses of *ndh* genes in angiosperms are usually associated with nutritional status and/or extensive rearrangements of chloroplast structures. Also, *ndh* genes loss events and pseudogenization have occurred in reported chloroplast genomes of parasitic plants, regardless of the degree of degradation in photosynthetic capacity. As a result, studies suggested that *ndh* genes were first lost in the transformation from autotrophy to heterotrophy.

In this study, seven transfer RNA genes, including *trnK-UUU*, have been lost from the chloroplast genomes of both species. Although similar tRNA losses have commonly occurred in most plants (Supplementary Table S1), the *trnK-UUU* gene, which is generally absent from most parasitic plants, is completely preserved (including its intron *matK* gene) within the chloroplast genome of *Gistanche deserticola*. Li et al. suggested that tRNA genes from the chloroplast genome were lost later than photosynthesis genes.

A pseudogene, which is a defective copy of the functional gene, is widespread in the chloroplast genome of plants and has lost the normal protein coding function. Loss of genic normal activity is generally caused by mutations inhibiting gene expression. Pseudogenes not only demonstrate gene mutation accumulation but are also associated with gene expression and regulation. Four pseudogenes exist in the chloroplast genomes of *T. chinensis* and *T. suchuenensis*; these pseudogenes include *rpl16*, *rpl2* and *ycf15* (duplicate gene). The gene of *ycf15* has been pseudogenized in many plants, including *S. jasminodora*, *C. reflexa*, and *C. exaltata*. Genes of *rpl16* and *rpl2* exist in most plants as functional genes, whereas they have been pseudogenized in the current study. A previous study pointed out that one early response of the chloroplast genomes to the evolution of a parasitic lifestyle was condensation via losses in numerous non-coding and unimportant regions; this event resulted in reduction of chloroplast genome size. Although gene loss can be regarded as a terminal evolutionary step, accumulation of point mutations leading to pseudogenization nevertheless occurred at previous steps.
region are correlated with degeneration of ndhF.

Discrepancies in assembly results obtained using Illumina and PacBio.

Table 4. Discrepancies in assembly results obtained using Illumina and PacBio.

| No. | Sites (bp) | Repeat unit | Number of repeat unit | Illumina | PacBio | Location |
|-----|------------|-------------|-----------------------|----------|--------|----------|
| 1   | 4615       | C           | 10                    | 9        |        | intergenic region |
| 2   | 17326      | C           | 9                     | 8        |        | introns |
| 3   | 17952      | C           | 7                     | 6        |        | introns |
| 4   | 27768      | G           | 7                     | 6        | psiC   |          |
| 5   | 32700      | C           | 6                     | 5        | psiA   |          |
| 6   | 41040      | C           | 10                    | 9        |        | intergenic region |
| 7   | 45020      | G           | 8                     | 7        |        | trnL     |
| 8   | 51046      | T           | 6                     | 7        |        | intergenic region |
| 9   | 55514      | C           | 10                    | 9        |        | intergenic region |
| 10  | 58056      | A           | 9                     | 8        |        | intergenic region |
| 11  | 87056      | G           | 7                     | 6        |        | intergenic region |
| 12  | 90138      | C           | 8                     | 7        |        | intergenic region |
| 13  | 101590     | G           | 8                     | 7        |        | intergenic region |
| 14  | 104671     | C           | 7                     | 6        |        | intergenic region |

Second-generation sequencing technology provides an efficient, novel, and rapid method for whole-genome sequencing. SMRT sequencing, which is combined with circular consensus sequencing (CCS), provides multiple reads of individual templates. Wu et al. have compared three generations of sequencing technologies (Sanger, Illumina and PacBio) on chloroplast genome assembly. Results demonstrated that long reads from PacBio showed potential for highly accurate "finished" genomes. However, the accuracy between second-generation and third-generation sequencing platforms was not compared thoroughly. In the present study, the complete chloroplast genome sequence of T. chinensis was sequenced using Illumina and PacBio platforms. Discrepancies in terms of assembly results between Illumina and PacBio were detected using PCR-based conventional Sanger sequencing, and the quality is very high. Results revealed that in PacBio platform, the error rate is high in homopolymers when the number of repeat units of a mononucleotide is higher than or equal to six. In the chloroplast genome of T. chinensis, polyA/T and polyC/G (repeat higher than or equal to six) included 509 and 72 sites, respectively. Although A/T mononucleotide repeats are the most common types (Table 3), these errors are mainly present in structures of polyC and polyG (Table 4). Among the 14 errors, 12 were G/C deletions, 1 was A/T deletion, and 1 was A/T insertion. All errors differed in terms of only one base.

As a result of multiple comparisons (Table 1 and Fig. 3), we observed that complete lengths of the chloroplast genomes of T. chinensis and T. sachuenensis are similar to those of S. jasminodora and V. minimum, whereas the lengths of SSC regions are much smaller (at least 3 kb). These regions, which contain most ndh genes, also encapsulate the largest variation within the chloroplast genome and have undergone dramatic reductions in some parasitic plants, including L. clandestine. Previous studies have demonstrated that positions of IR junction and SSC region are correlated with degeneration of ndhF and ycf1 genes. Loss of ycf1 and all ndh genes (including ndhF), as revealed by this study may explain why SSC chloroplast genome regions of the two considered species are shorter than those of others.

Chloroplast genomes have provided significant data for evolutionary, taxonomic, and phylogenetic studies. Specifically, the chloroplast gene of matK has been widely utilized in plant phylogenetic analyses. In this study, we constructed phylogenetic trees using ML and NJ methods based on matK and 54 protein-coding genes common to all present in the chloroplast genomes of ten species, including two medicinal hemiparasites in the current study. Phylogenetic results are extremely consistent, irrespective of method and dataset. All phylogenetic results strongly support the theory that Loranthaceae and Viscaceae diverged independently from one another. Phylogenetic results discussed in the present study are broadly consistent with those of a previous research, which utilized chloroplast trnL intron sequences to investigate inter-familial relationships within Santalales.

Conclusions

The complete chloroplast genome sequences of traditional medicinal hemiparasites T. chinensis and T. sachuenensis were obtained and analyzed. Results of this study revealed effects of parasitic lifestyle on chloroplast structure and genome content in these species and enhanced understanding of phylogenetic positions and relationships of T. chinensis and T. sachuenensis. This research also showed that sequences assembled using reads obtained by the Illumina platform is more accurate than those from PacBio.

Materials and Methods

Plant Material, DNA Extraction, and Sequencing. Fresh leaves of T. chinensis and T. sachuenensis were collected from Qinzhou City in Guangxi Province and from Lichuan City in Hubei Province, respectively. All samples were identified by Professor Yulin Lin, who is based at the Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College. The voucher specimens were deposited in the herbarium of IMPLAD. Approximately 100 g of samples frozen in −80 °C were used to extract total genomic DNA using DNeasy Plant Mini Kit (Qiagen Co., Germany). DNA quality was assessed based
on electrophoresis and optical density results. DNA of two species was used to generate libraries with average insert size of 500 bp and sequenced using Illumina Hiseq X in accordance with standard protocol. Approximately 4.4 Gb of raw data from *T. chinensis* and 3.7 Gb from *T. sutchuenensis* using Illumina sequencing platform were generated with 150 bp paired-end read lengths. To compare PacBio with Illumina sequencing technology when employed in chloroplast genome study, we sequenced a PacBio shotgun library of *T. chinensis* with an insert size of 3 kb on PacBio RS II platform using P6-C4 chemistry (Pacific Biosciences, Menlo Park, CA, USA). A total of 24,590 CCS reads with a length of 67,512,059 bp were obtained from one SMRT cell, and these reads were used for assembly. Assembly results showed that 13.6% of chloroplast sequences were detected in total data, revealing percentage of chloroplast DNA in total DNA during DNA extraction experiment.

**Chloroplast Genome Assembly.** Low-quality reads resulting from all samples were trimmed using the software Trimmomatic. The trimmed reads included a mixture of data from nuclear and organelle genomes. We used the chloroplast genome sequence of *Viscum minimum*, which was downloaded from GenBank to establish a Basic Local Alignment Search Tool (BLASTn) database. Then all trimmed reads were mapped onto this database, and the mapped reads were extracted from raw data based on coverage and similarity. Extracted reads were assembled to contigs using SOAPdenovo2. SSPACE was used to construct the scaffold of the chloroplast genome, and GapCloser was used to fill gaps. Reads sequenced using PacBio system were used to assemble the chloroplast genome according to the strategy described by Xiang et al. Assembly results obtained using Illumina and PacBio (Table 4) differed in terms of 14 sites, which are all homopolymers and mainly located at intergenic regions. To detect these discrepancies, we performed PCR-based conventional Sanger sequencing. The primer sequences are listed in Supplementary Table S6.

**Genome Annotation and Structural Analyses.** To verify accuracy, including boundaries of single copy and IR regions of assembled sequences, we designed a series of PCR primers (Supplementary Table S7). Annotations of genome sequences of two *Taxillus* species were performed using the online software Dual Organellar GenoMe Annotator (DOGMA, http://dogma.ccbb.utexas.edu/) and CPGAVAS with default settings and checked manually. We then used the software tRNAscan-SE to annotate tRNA genes. Boundaries of genes, introns/exons and coding regions were verified using BLAST versus reference sequences. Circular chloroplast genome map was constructed using an online program Organellar Genome DRAW (OGDRAW) and subsequently modified manually. GC content was analyzed using the software MEGA 6.0. Genome comparisons between *T. chinensis* and *T. sutchuenensis* were performed and plotted using the mVISTA program. The whole-genome alignment for chloroplast genomes of six species, including *T. chinensis*, *T. sutchuenensis*, *S. jasminodora*, *V. minimum*, *O. alba* and *N. tabacum*, was performed using the algorithm MAUVE in the software Geneious v10.1.2 (Biomatters Ltd., http://www.geneious.com/).

**Codon Usage and SSRs Analyses.** RSCU value, the ratio between frequency of use and expected frequency of a particular codon, is a simple method for detecting non-uniform synonymous codon usage (SCU) within a coding sequence. In the present study, utilizing the RSCU ratio, we performed statistical analyses to investigate the distribution of codon usage with the software CodonW (http://codonw.sourceforge.net/), applying a 1.00 value for no preference. In addition, a value less than 1.00 refers to a frequency of use that is less than expected, whereas a value higher than 1.00 indicates codons that are more frequently used than expected. Potential SSRs were exploited using the software MISA, with parameters set to encompass the number of repeat units of a mononucleotide SSR higher than or equal to eight; followed by higher than or equal to four repeat units for di- and tri-nucleotide SSRs; and higher than or equal to three repeat units for tetra-, penta- and hexa-nucleotides, respectively. In this study, we mainly searched for complete repetitive SSR loci, treating cycled or reverse complementary SSRs as the same type.

**Phylogenetic Analyses.** To determine phylogenetic positions of *T. chinensis* and *T. sutchuenensis* within Santalales, we analyzed the chloroplast genomes of ten species, encompassing five other taxa within this lineage, *V. album* (accession number: KT003925), *V. crassula* (KT070881), *V. minimum* (KJ512176), *O. alba* (KT070882), and *S. jasminodora* (KX775962). We also used the chloroplast genomes of *P. ginseng* (AY582139), *N. tabacum* (Z00044), and *Astragalus mongholicus* (KU665554) as outgroups, and constructed phylogenetic trees using ML and NJ methods in the software MEGA 6.0 with 1000 bootstrap replicates employing 54 protein-coding genes commonly present in the ten species and *matK* genes. ML analysis was conducted based on the Tamura-Nei model using a heuristic search for initial trees. This most appropriate model was determined by Modeltest 3.7. NJ trees were performed with NJ method, and evolutionary distances were computed using the Kimura 2-parameter method.

**References**

1. Li, X. *et al.* Complete chloroplast genome sequence of holoparasite *Cistanche deserticola* (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host *Haloxylon ammodendron* (Chenopodiaceae). *PloS One* 8, e58747 (2013).
2. Raubeson, L. A. & Jansen, R. K. In: *Diversity and Evolution of Plants: Genotypic and Phenotypic Variation in Higher Plants* (ed R. Henry) 45–68 (CABI Publishing, 2005).
3. Verma, D. & Daniell, H. Chloroplast vector systems for biotechnology applications. *Plant Physiol.* 145, 1129–1143 (2007).
4. Wu, F. H. *et al.* Complete chloroplast genome of *OncidiumGower Ramsey* and evaluation of molecular markers for identification and breeding in Orchidaceae. *BMC Plant Biol.* 10, 68 (2010).
5. Cheng, C. F. *et al.* The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis* orchids. *Plant Sci.* 190, 62–73 (2012).
6. Kuan, D. Y. *et al.* Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* 54, 663–673 (2011).
7. Nock, C. J. *et al.* Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol.* 9, 328–333 (2011).
8. Tangphatsornruang, S. et al. The chloroplast genome sequence of mungbean (Vigna radiata) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res. 17, 11–22 (2009).

9. Carbonell, I. C. et al. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus. Citrus. Mol. Biol. Evol. 32, 2015–2035 (2015).

10. Sasaki, C. et al. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol. Biol. 59, 309 (2005).

11. Sasaki, C. et al. Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes. Theor. Appl. Genet. 115, 571–590 (2007).

12. Li, X. W. et al. High-throughput pyrosequencing of the complete chloroplast genome of Magnolia officinalis and its application in species identification. Acta Pharm. Sinica 47, 124–130 (2012).

13. Zhang, Y. et al. The complete chloroplast genome sequence of Taxus chinensis var. mairii (Taxaceae): loss of an inverted repeat region and comparative analysis with related species. Gene 540, 201–209 (2014).

14. Lei, W. et al. Intraspecific and heteroplasmonic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. Sci. Rep. 6, 21669 (2016).

15. Huang, X. Y., Guan, K. Y. & Ai-Rong, L. I. Biological traits and their ecological significances of parasitic plants: A review. Chinese J. of Botol. 30, 1838–1844 (2011).

16. Wolfe, K. H., Morden, C. W. & Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA 89, 10648–10652 (1992).

17. Wolfe, K. H., Morden, C. W., Ems, S. C. & Palmer, J. D. Rapid evolution of the plastid translational apparatus in a non-photosynthetic plant: loss or accelerated sequence evolution of RNAI and ribosomal protein genes. J. Mol. Evol. 35, 304–317 (1992).

18. Mcneal, J. R., Kuehl, J. V., Boore, J. L. & de Pamphilis, C. W. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus Cuscuta. BMC Plant Biol. 7, 57 (2007).

19. Maier, U. G., Karin, K., Sabine, B., Funk, H. T. & Kirsten, K. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, Cuscuta reflexa and Cuscuta grossoni. BMC Plant Biol. 7, 1–12 (2007).

20. Molina, J. et al. Possible Loss of the Chloroplast Genome in the Parasitic Flowering Plant Rafflesia lagascae (Rafflesiacae). Mol. Biol. Evol. 31, 793 (2014).

21. Wicke, S. et al. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Plant Cell 25, 3711 (2013).

22. Cusimano, N. & Wicke, S. Massive intracellular gene transfer during plastid genome reduction in nongreen Orobancheae. New Phytol. 210, 680–693 (2015).

23. Samigullin, T. H., Logacheva, M. D., Penin, A. A. & Vallejeroman, C. M. Complete plastid genome of the recent holoparasite Lathraea squamaria reveals earliest stages of plastome reduction in Orobancheae. PloS One 11, e0150718 (2016).

24. Petersen, G., Cuenca, A. & Seberg, O. Plastome evolution in hemiparasitic mistletoes. Mol. Phylogenet. Evol. 43, 974–985 (2007).

25. Li, Y. et al. Study on medicinal plants of Loranthaceae resources in china. World Science & Technology: Modernization of Traditional Chinese Medicine and Materia Medica 11, 665–669 (2009).

26. Liu, L. F. & Qiu, H. X. Pollen morphology of Loranthaceae in China. Guihaia 13, 235–245 (1993).

27. Gong, Z. et al. A chemotaxonomic study of 27 species of the Loranthaceae plant from China. Guihaia 24, 493–487 (2004).

28. Han, R., Hao, G. & Zhang, D. Interfamilial relationships of Santalales as revealed by chloroplast trnl intron sequences. J. Tro. Subtr. Botany 12, 393–398 (2004).

29. Commission, C. P. The Chinese Pharmacopoeia. 51–52 (Beijing: Chemical Industry Press, 2015).

30. Vidal-Russell, R. & Nickrent, D. L. Evolutionary relationships in the showy mistletoe family (Loranthaceae). Am. J. Bot. 95, 1015–1029 (2008).

31. The Editorial Committee of Flora of China. Flora of China. Vol. 5, 246–269 (Beijing: Science Press, China; Missouri: Missouri Botanical Garden Press, USA, 2003).

32. Wang, Y., Zhang, S. Y., Ma, X. F. & Tian, W. X. Potent inhibition of fatty acid synthase by parasitic loranthus [Taxillus chinensis (DC.) Danser] and its constituent avicularin. J. Enzym. Inhib. Med. Ch. 21, 87–93 (2006).

33. Liu, C. Y. et al. Antioxidant, anti-inflammatory, and antiproliferative activities of Taxillus euchuenensis. Am. J. Chinese Med. 40, 335–348 (2012).

34. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138 (2009).

35. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. Genome Biology 14, 405 (2013).

36. Qi, L. et al. High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. New Phytol. 204, 1041–1049 (2014).

37. Hanrari, M. et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14, 670 (2013).

38. Chen, X., Li, Q., Li, Y., Qian, J. & Han, J. Chloroplast genome of Aconitum barbatum var. puberulum (Ranunculaceae) derived from CCS reads using the PacBio RS platform. Front. Plant Sci. 6, 42 (2015).

39. Xiang, B. et al. The complete chloroplast genome sequence of the medicinal plant Swertia mussotii using the PacBio RS II platform. Molecules 21, 1029 (2016).

40. Qian, J. et al. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PloS One 8, e57607 (2013).

41. Xu, J. et al. The first intron of rice EPSP synthase enhances expression of foreign gene. Sci. China C. Life Sci. 46, 561–569 (2003).

42. Kim, K. J. & Lee, H. L. Complete chloroplast genome sequences from Korean ginseng (Panax schinseng) and comparative analysis of sequence evolution among 17 vascular plants. DNA Research 11, 247–261 (2004).

43. Cucu, F. L., De, P. D., Danzi, D., Vendramin, G. G. & Sonnante, G. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. PloS One 10, e0120589 (2015).

44. Wang, Y. et al. Complete chloroplast genome sequence of Aquilaria sinensis (Lour.) Gilg and comparative analyses with other Malvales order. Front. Plant Sci. 7, 280 (2016).

45. Zhang, Y. et al. The complete chloroplast genome sequences of five Epimedium species: lights into phylogenetic and taxonomic analyses. Front. Plant Sci. 7, 306 (2016).

46. Su, H. J. & Hu, J. M. The complete chloroplast genome of hemiparasitic flowering plant Schoepfia jasminodora. Mitochondr. DNA Part B 1, 767–769 (2016).

47. Zuo, L. H. et al. The first complete chloroplast genome sequence of Ulmus species by de novo sequencing: genome comparative and taxonomic position analysis. PloS One 12, e0171264 (2017).

48. Powell, W., Morgante, M., Mcdevitt, R., Vendramin, G. G. & Rafalski, J. A. Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc. Natl. Acad. Sci. USA 92, 7759–7763 (1995).

49. Yang, A. H., Zhang, J. J., Yao, X. H. & Huang, H. W. Chloroplast microsatellite markers in Liriopendron tulipifera (Magnoliaceae) and cross-species amplification in L. chinense. Am. J. Bot. 98, e123 (2011).
52. 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).

53. Xue, J., Wang, S. & Zhou, S. L. Polymorphic chloroplast microsatellite loci in Nelumbo (Nelumbonaceae). Am. J. Bot. 99, 240–244 (2012).

54. Bryant, N., Lloyd, J., Sweeney, C., Myouga, F. & Meinke, D. Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in Arabidopsis. Plant Physiol. 155, 1678–1689 (2011).

55. Gao, L., Ying, H., S. U. & Wang, T. Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. J. Syst. Evol. 48, 77–93 (2010).

56. Millen, R. S. et al. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13, 645–658 (2001).

57. Kim, H. T. et al. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. PLoS One 10, e0142215 (2015).

58. Lin, C. S. S. et al. Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some orchids. Plant J. 90, 994–1006 (2017).

59. Chris, B. J., Guisinger, M. M. & Jens, R. K. Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). Plant Mol. Biol. 76, 263–272 (2011).

60. Sanderson, M. J. et al. Exceptional reduction of the plastid genome of saguaro cactus (Carnegiea gigantia): Loss of the ndh gene suite and inverted repeat. Am. J. Bot. 102, 1115–1127 (2015).

61. Martin, M. & Sabater, B. Plastid ndh genes in plant evolution. Plant Physiol. Bioch. 48, 636–645 (2010).

62. Li, W. H., Gojobori, T. & Nei, M. Pseudogenes as a paradigm of neutral evolution. Nature 292, 237–239 (1981).

63. Balakirev, E. S. & Ayala, F. J. Pseudogenes: are they “junk” or functional DNA? Annu. Rev. Genet. 37, 123–151 (2003).

64. Mardis, E. R. & Next-generation D. N. A. sequencing methods.

65. Metzker, M. L. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46 (2010).

66. Wu, Z. et al. A precise chloroplast genome of Nelumbo nuclifera (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. BMC Plant Biol. 14, 289 (2014).

67. Walker, J. F., Jansen, R. K., Zanis, M. J. & Emery, N. C. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. Am. J. Bot. 102, 1751–1752 (2015).

68. Delavault, P. M., Russo, N. M., Lusson, N. A. & Thalouarn, P. A. Organization of the reduced plastid genome of Lathraea clandestina, an achronophyllous parasitic plant. Physiol. Plantarum 96, 674–682 (1996).

69. Logacheva, M. D., Schelkunov, M. I., Nuraliev, M. S., Samigullin, T. H. & Penin, A. A. The plastid genome of mycoheterotrophic species of Parnassia. Curr. Genet. 50, W273 (2004).

70. Luo, R., Martin, M. & Sabater, B. CpGAV AS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13, 715 (2012).

71. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. Comput. Biol. Chem. 36, 2725 (2013).

72. Frazier, K. A., Pachtler, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 33, W273–W274 (2005).

73. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147 (2010).

74. Sharp, P. M. & Li, W. H. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295 (1987).

75. Posada, D. & Crandall, K. A. MODESTEST: testing the model of DNA substitution. Bioinformatics 14, 817 (1998).

76. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406 (1987).

77. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120 (1980).

Acknowledgements

This work was supported by CAMS Innovation Fund for Medical Sciences (CIFMS) (NO. 2016-21M-3–016), Major Scientific and Technological Special Project for “Significant New Drugs Creation” (No. 2014ZX09304307001) and Guangxi Natural Science Foundation (NO. 2013GXNSFAA019120).

Author Contributions

J.Z., X.C., and Z.X., performed the experiments; Y.L., J.Z., and Y.C., assembled sequences and analyzed the data; Y.L. and J.Z. wrote the manuscript; Y.H.L., B.D., and J.S. collected plant material; H.Y. conceived the research and revised the manuscript. All authors have read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-13401-4.

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

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