# Characterization of the Complete Chloroplast Genome of Wild Populations of *Morina nepalensis* var. alba: A Postprint

**Authors:** Zhang Qian, Zhang Dequan

**Date:** 2022-02-14T21:21:03Z

## Abstract

Acanthocalyx alba is a commonly used medicinal plant in Tibetan regions, yet studies on its complete chloroplast genome remain scarce. To elucidate the fundamental characteristics of this species' chloroplast genome and investigate its lineage genetic structure, this study employed the Illumina sequencing platform to conduct second-generation sequencing on 10 individuals from five wild populations of Acanthocalyx alba. Following assembly and annotation, 10 complete chloroplast genome sequences were obtained, and their genomic features and inter-population phylogenetic relationships were preliminarily analyzed. The findings indicate that: (1) The complete chloroplast genome size of Acanthocalyx alba ranges from 155,335 to 156,266 bp, encoding a total of 113 genes, including 72 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. The chloroplast genome exhibits high conservation within the species regarding size, structure, GC content, and gene composition; (2) Comparative genomic analysis demonstrates that the most variable regions in Acanthocalyx alba are located within the single-copy regions, with no significant expansion or contraction observed at the IR boundaries; (3) Population genetic analysis reveals that wild populations of Acanthocalyx alba possess a distinct geographic genetic structure, with a certain correlation between genetic distance and geographic distance among different populations. This study demonstrates that the chloroplast genome of Acanthocalyx alba is relatively conserved at the intraspecific population level and can effectively reveal species' geographic genetic structure at the population level. These findings establish a foundation for future population genetics and phylogenomic studies of Acanthocalyx species.

# Full Text

## Preamble

### Analysis of Complete Chloroplast Genomes from Wild Populations of *Acanthocalyx alba*

**ZHANG Qian[1], ZHANG Dequan[1,2,*]**
[1]College of Pharmacy, Dali University, Dali 671000, Yunnan, China
[2]Yunnan Key Laboratory of Screening and Research on Anti-pathogenic Plant Resources from Western Yunnan, Dali 671000, Yunnan, China

**Abstract:** *Acanthocalyx alba* is a commonly used medicinal plant in Tibetan regions, yet studies on its complete chloroplast genome remain scarce. To elucidate the fundamental characteristics of this species' chloroplast genome and explore its phylogeographic structure, we performed next-generation sequencing on ten individuals from five wild populations using the Illumina platform. Following assembly and annotation, we obtained ten complete chloroplast genome sequences and conducted preliminary analyses of their genomic features and inter-population phylogenetic relationships. Our findings reveal: (1) The complete chloroplast genomes of *A. alba* range from 155,335 to 156,266 bp, encoding 113 genes including 72 protein-coding genes, 30 tRNA genes, and four rRNA genes. The chloroplast genome exhibits high conservation within the species regarding size, structure, GC content, and gene composition. (2) Comparative genomic analysis indicates that highly variable fragments are located in the single-copy regions, with no significant expansion or contraction at IR boundaries. (3) Population genetic analysis demonstrates distinct geographic genetic structure among wild populations, with genetic distances correlating with geographic distances. This study demonstrates that the chloroplast genome of *A. alba* is relatively conserved at the intraspecific population level and can reveal geographic genetic structure at the population level, laying a foundation for future population genetics and phylogenomic studies of *Acanthocalyx* species.

**Keywords:** *Acanthocalyx alba*, complete chloroplast genome, medicinal plant, gene, phylogenomics

---

## Introduction

*Acanthocalyx alba* (Hand.-Mazz.) M.J.Cannon, also known as "Baihua Cishen" (white-flowered thistle), belongs to the family Dipsacaceae and genus *Acanthocalyx*. This genus comprises four species and two varieties in China, primarily distributed across Yunnan, Sichuan, and Tibet (Hong et al., 2011). *A. alba* is a traditional Tibetan medicinal herb, known in Tibetan medicine as "Jiangcaigabao" and first documented in the *Four Medical Tantras* (State Administration of Traditional Chinese Medicine Chinese Herbal Editorial Board, 2002). It is one of three "Cishen" species included in the Ministry of Health's Tibetan

medicine standards (Qinghai Institute for Drug Control and Qinghai Institute of Tibetan Medicine, 1996), with therapeutic effects including stomachic and emetic properties. Internally, it is used to treat joint pain, urinary incontinence, lower back pain, dizziness, and facial paralysis, while external application treats sores and purulent wounds, and it exhibits anti-tumor activity (State Administration of Traditional Chinese Medicine Chinese Herbal Editorial Board, 2002; Yang, 1989). Recent research on *A. alba* has primarily focused on its active compounds, content determination, and extraction methods (Wu et al., 2011; Zhang et al., 2013; Yang et al., 2014; Zhang et al., 2015). For instance, Zhang et al. (2018) identified saponins, alkaloids, and sterols in *A. alba*, with saponins being the main active constituents. However, molecular biology research remains limited, with only Wang et al. (2020) reporting the chloroplast genome sequence of this species. This raises the question: what variations exist in the chloroplast genome sequences at the intraspecific population level?

Chloroplast genomes in angiosperms are typically maternally inherited and evolve more slowly than nuclear and mitochondrial genomes, with relatively conserved gene composition and structure (Smith, 2015; Szymon et al., 2016; Du et al., 2020). These characteristics make chloroplast genomes valuable for plant species identification and phylogenetic studies. Cui et al. (2019) compared chloroplast genomes of 32 *Amomum* species, demonstrating that complete chloroplast genomes can accurately identify species within the genus. Li et al. (2020) used chloroplast genomes to reveal phylogenetic relationships in the ethnic medicinal plant complex *Gaultheria leucocarpa* var. *yunnanensis*. Zhang et al. (2021) reconstructed species divergence times and phylogenetic relationships in Myrtiflorae based on chloroplast genomes. However, structural variations such as IR region contraction, inversion, and gene/intron loss occur during long-term evolution (Zhang et al., 2014; Liao et al., 2020; Jiang et al., 2020), providing genetic information for understanding species phylogeny and evolutionary relationships. Thus, plant chloroplast genome sequences offer rich genetic information significant for taxonomy, phylogeny, and evolution. The question remains whether chloroplast whole genomes are suitable for population genetics studies at the intraspecific level. Due to high sequencing costs and immature data analysis methods at the population level, such research remains limited.

This study utilized wild population individuals of *A. alba* as research material, employing next-generation sequencing technology for high-throughput sequencing, followed by assembly, annotation, and evolutionary analysis of the complete chloroplast genome. We addressed two scientific questions: (1) What are the characteristics of the *A. alba* chloroplast genome sequence? (2) Can the chloroplast whole genome be used to resolve genetic structure at the intraspecific population level? This research establishes a foundation for molecular genetic studies of *Acanthocalyx* species and provides preliminary insights into using chloroplast whole genomes for population genetics research.

### 1.1 Experimental Materials

Plant material of *A. alba* was collected from five wild populations in Garzê Prefecture, Sichuan Province, comprising ten samples (Table 1). All specimens were identified as *Acanthocalyx alba* (Hand.-Mazz.) M.J.Cannon by Professor Zhang Dequan of Dali University, with voucher specimens deposited in the Herbarium of Medicinal Plants and Crude Drugs, College of Pharmacy, Dali University.

### 1.2 Genomic DNA Extraction and Sequencing

Total genomic DNA was extracted from silica-dried leaf material using a modified CTAB method. The genomic DNA was fragmented using a Covaris ultrasonic disruptor, followed by end repair, A-tailing, adapter ligation, purification, and PCR amplification to construct sequencing libraries. After quality assessment, libraries were sequenced on the Illumina HiSeq 2500 platform with paired-end reads (2×300 bp) at Beijing Novogene Bioinformatics Technology Co., Ltd.

### 1.3 Chloroplast Genome Assembly and Annotation

Approximately 4 Gb of raw data were generated from next-generation sequencing. After filtering with Trimmomatic v.0.32, assembly was performed using GetOrganelle.py, with subsequent data processing following our group' s previous work (Hu and Zhang, 2021). Using the *Acanthocalyx alba* reference genome (Accession: NC_045055), annotation was conducted in Geneious (https://www.ncbi.nlm.nih.gov/). The physical map of the chloroplast genome was drawn using the online tool Organellar Genome Draw (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

### 1.5 Sequence Variation Analysis

The ten *A. alba* chloroplast genome sequences were aligned using MAFFT v.7.129 and manually adjusted with BioEdit. Nucleotide variability (Pi) was analyzed using DnaSP v.7.0.26 with a sliding window approach (window length: 600 bp, step size: 200 bp). P-distances were calculated using MEGA v.7.0.26. Additionally, annotated chloroplast genome sequences were converted and compared using the Shuffle-LAGAN mode in the online software mVISTA (http://genome.lbl.gov/vista/mvista/submit.shtml), with *Acanthocalyx alba* (NC_045055) as the reference.

### 1.6 Phylogenetic Analysis

Based on annotation information from the *A. alba* chloroplast genome, we downloaded complete chloroplast genome sequences of 11 Dipsacaceae and Caprifoliaceae species from the NCBI database for phylogenetic analysis. *Coffea arabica* L. and *C. arabica* Pierre ex Froehn. were selected as outgroups. Multiple sequence alignment was performed using MAFFT V.7.1. Phylogenetic trees were

constructed using three methods: Maximum Likelihood (ML), Maximum Parsimony (MP), and Bayesian Inference (BI). The nucleotide substitution model was selected as GTR+G using jModelTest v2.1.7. The ML tree was built with RAxML v.8.2.4 using rapid bootstrap algorithm with 1,000 replicates. The MP tree was constructed with MEGA v.7.0.26 with 1,000 bootstrap replicates. The BI tree was generated with MrBayes v.3.2.6 using MCMC algorithm for 1 million generations, sampling every 1,000 generations, discarding the first 25% of trees, and constructing a consensus tree from remaining samples.

## 2.1 Genome Structure and Basic Characteristics

The *A. alba* chloroplast genome exhibits the typical quadripartite structure, comprising two inverted repeat regions (IRs), a large single-copy region (LSC), and a small single-copy region (SSC) (Table 2, Figure 1). The assembled chloroplast genomes range from 155,335 to 156,266 bp in length, with GC content of 38.1–38.2%. The lengths of each region are: 89,027–89,076 bp (LSC), 17,689–17,842 bp (SSC), and 24,253–24,666 bp (IRs). The IR region shows the highest GC content (42.8–43.2%), followed by LSC (36.5%) and SSC (32.9%). Annotation revealed 113 genes, including 72 protein-coding genes, 30 tRNA genes, four rRNA genes, and seven pseudogenes (*clpP*, *accD*, *ycf2*, *ycf1*, *rps18*, *rps3*, and *ycf3*). Additionally, 16 genes contain introns, each with a single intron (Table 3).

## 2.3 Sequence Variation Analysis

Sliding window analysis of the five regional *A. alba* chloroplast genomes revealed that the SSC region exhibits the highest variation, while the IR region shows the lowest (Figure 4). Three highly variable sequences were identified in the LSC region (*rpoC1*) and SSC region (*ndhF* and *rpl32-trnL-UAG*), with *rpl32-trnL-UAG* showing the highest variability, followed by *ndhF* and *rpoC1*. Pairwise comparisons using SD01 as a reference against the other nine genomes showed that non-coding regions exhibit higher variation than protein-coding regions, and single-copy regions (LSC & SSC) show significantly greater variation than inverted repeat regions (IR). Overall, the chloroplast genomes from five regions are highly similar, with the most variable genes being *rpoC2*, *psbC*, *rrn23*, and *ycf1*, while other genes remain highly conserved. Intergenic regions show greater variation than gene regions, including *atpF-atpH*, *psaB-psaA*, *psaA-ycf3*, *trnM-CAU-atpE*, *psbF-psbE*, *psbE-petL*, *rrn5-trnN-ACG*, *trnR-ACG-trnN-GUU*, and *trnL-UAG-ccsA* (Figure 5). These regions can be developed as specific markers for phylogenetic and evolutionary studies at both interspecific and intraspecific levels.

## 2.4 Phylogenetic Analysis

P-distance analysis revealed interspecific genetic variation and nucleotide substitution patterns among the ten *A. alba* chloroplast genomes, with P-distances

ranging from 0.0000 to 0.0007 and nucleotide differences from 0 to 1,515. Generally, sequences from more distant geographic locations showed larger P-distances and nucleotide differences (Table 4). Phylogenetic analysis showed that evolutionary relationships among wild populations were consistent across the three methods (Figure 6) and corresponded with genetic distance results. In the phylogenetic trees, four individuals from Kangding (KD) and Daofu (DF) diverged earliest, followed by Yading (YD) and Sangdui (SD), with two individuals from Litang (LT) diverging most recently. However, the four individuals from YD and SD did not form distinct separate clades, likely due to relatively frequent gene flow between these geographically close populations.

## 3 Discussion and Conclusion

This study reports the chloroplast genome sequence characteristics of *Acanthocalyx alba* and reveals its geographic genetic structure at the population level. Chloroplast genomes from different wild populations show highly consistent gene categories, numbers, and arrangement orders, with similar GC content among individuals and the highest GC content in IR regions. The *A. alba* chloroplast genome contains seven pseudogenes, five of which are shared across Dipsacaceae (*clpP*, *accD*, *ycf2*, *ycf1*, *rps18*), suggesting that pseudogenization may be common in this family (Wang et al., 2020). Chloroplast SSR loci represent efficient molecular markers. In this study, *A. alba* chloroplast SSRs are primarily composed of A/T bases, particularly poly-A/T mononucleotide repeats, with fewer C or G tandem repeats, consistent with patterns in other angiosperms (Guo et al., 2017; Na et al., 2018; Chen et al., 2019). These SSRs are mainly distributed in the two single-copy regions, suggesting that high A/T content in SSRs and rRNA sequences in IR regions may contribute to the relatively low GC content and base composition differences across chloroplast genome regions (Zhang et al., 2020).

Expansion and contraction of IR and SC regions are considered important factors affecting chloroplast genome size in angiosperms (Wang et al., 2017; Song et al., 2019). This study found no significant expansion or contraction at the four boundaries among individuals from different wild populations, indicating that IR region size is highly conserved in *A. alba*, consistent with Wang et al. (2020). Highly variable regions identified in chloroplast genomes can provide rich genetic information for both species-level phylogenetic and identification studies and population-level dynamics and evolutionary history. Fatemeh et al. (2018) used *rpl32-trnL-UAG* for phylogenetic analysis and divergence time estimation in *Onosma*; Nahla et al. (2020) used *rpoC1* for phylogenetic analysis in *Medicago*; Chen et al. (2020) identified *ycf1* and *psbM-psbD* as specific barcodes for *Fritillaria* species identification. These studies confirm the special role of hypervariable regions in species evolution and identification. We identified three such regions (*rpoC1*, *ndhF*, and *rpl32-trnL-UAG*) suitable for phylogenetic and population genetic studies within *Acanthocalyx*.

Traditionally, chloroplast gene fragments have been used to study population

genetic structure and phylogeographic relationships, but their utility is limited by insufficient polymorphic sites (Zhang et al., 2019; Zhang et al., 2020; Liu et al., 2021). In contrast, complete chloroplast genomes provide abundant genetic variation for studying complex plant groups. Wang et al. (2020) revealed population-level genetic diversity in *Carya illinoinensis* based on nucleotide differences between two populations. Our study shows clear genetic structure among five wild populations of *A. alba*, with genetic distances and nucleotide differences correlating with geographic distances, consistent with phylogenetic tree topology. Notably, individuals from Sangdui (SD) and Yading (YD) did not form separate clades, likely due to relatively frequent gene flow between these geographically proximate populations. This result also illustrates the limitations of chloroplast genomes, which evolve more slowly and are uniparentally inherited compared to nuclear genomes.

In summary, complete chloroplast genome sequences provide rich genetic information for population genetics and phylogenetic studies in complex plant groups at both species and subspecies levels. However, due to the still relatively high cost of next-generation sequencing, our study had limited population sample sizes, which may affect the robustness of our conclusions. Therefore, whether chloroplast genomes can serve as a technical supplement to traditional molecular markers requires further validation. Additionally, data analysis methods for population genetics using complete chloroplast genomes need further refinement.

*Note: Figure translations are in progress. See original paper for figures.*