**Plastid Phylogenomics and Adaptive Evolution of *Gaultheria* series *Trichophyllae*
(Ericaceae), a Clade from Sky Islands of the Himalaya-Hengduan Mountains**

Ming-Ying Zhang^{a, b}, Peter W. Fritsch^c, Peng-Fei Ma^{a, b}, Hong Wang^{a, b}, Lu Lu^{a, b*},
De-Zhu Li^{a, b*}

^a Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming
Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201,
China.

^b Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species,
Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan
650201, China.

^c Botanical Research Institute of Texas, 1700 University Drive, Fort Worth, TX
76017, U.S.A.

* Corresponding authors.

Email addresses: dzl@mail.kib.ac.cn (D.-Z. Li), lulu@mail.kib.ac.cn (L. Lu)

Abstract

Gaultheria series *Trichophyllae* Airy Shaw is an angiosperm clade of high-alpine shrublets endemic to the Himalaya-Hengduan Mountains and characterized by recent species divergence and convergent character evolution that has until recently caused much confusion in species circumscription. Although multiple DNA sequence regions have been employed previously, phylogenetic relationships among species in the group have remained largely unresolved. Here we examined the effectiveness of plastid genome for improving phylogenetic resolution within the *G.* series *Trichophyllae* clade. Plastid genomes of 31 samples representing all 19 recognized species of the series and three outgroup species were sequenced with Illumina Sequencing. Maximum likelihood (ML), maximum parsimony (MP) and Bayesian inference (BI) phylogenetic analyses were performed with various datasets, i.e., that from the whole plastid genome, coding regions, noncoding regions, large single-copy region (LSC) and inverted-repeat region a (IRa). The partitioned whole plastid genome with inverted-repeat region b (IRb) excluded was also analyzed with ML and BI. Tree topologies based on the whole plastid genome, noncoding regions, and LSC region datasets across all analyses, and that based on the partitioned dataset with ML and BI analyses, are identical and generally strongly supported. *Gaultheria* series *Trichophyllae* form a clade with three species and one variety that is sister to a clade of the remaining 16 species; the latter comprises seven main subclades. Interspecific relationships within the series are strongly supported except for those based on the coding-region and IRa-region datasets. Eight divergence hotspot

regions, each possessing > 5% percent variable sites, were screened across the whole plastid genome of the 28 individuals sampled in the series. Results of morphological character evolution reconstruction diagnose several clades, and a hypothesis of adaptive evolution for plant habit is postulated.

Keywords: Plastid genome, Character evolution, *Gaultheria*, Phylogenetic relationship

1. Introduction

Gaultheria series *Trichophyllae* Airy Shaw (Ericaceae: tribe Gaultherieae) form an angiosperm clade of 19 species endemic to the Himalaya-Hengduan Mountains of East Asia. The species grow as low shrublets, typically prostrate or ascending, and are characterized by small leaves (generally < 1 cm long), paired apical bracteoles, solitary axillary pentamerous flowers, and capsules surrounded by an expanded fleshy calyx (Fritsch et al., 2015a; Lu et al., 2010; Middleton, 1991). They are narrowly restricted to isolated montane regions, mainly between 2400 and 4500 m elevation (Fritsch et al., 2015a; Fritsch et al., 2015b), and thus form part of “sky island” (Heald, 1951) vegetation zones.

The tribe was erected by Airy Shaw (1941) with seven species and several ambiguously defined varieties, and later recognized in the morphologically based global classification of *Gaultheria* by Middleton (1991). The group has been taxonomically challenging because of the relatively few collections available for study (at least until recently), the fact that fruit characters, recently understood to harbor important taxonomic characters, are poorly preserved on herbarium specimens, and the difficulty in field accessibility of these plants due to the challenging environmental conditions and remote locations in which these species occur. Fritsch et al. (2008) postulated that the considerably high morphological variation within some species that appear to intergrade with other sympatric species (e.g., *G. sinensis*, *G. trichophylla*) suggests hybridization and reticulate evolution, leading to confusion in identification and classification. More recent molecular phylogenetic evidence,

however, suggests that such variation results from the combination of convergent character evolution and the presence of a much higher species number than was previously recognized (Lu et al., 2010). Subsequent taxonomic work has revealed that taxonomic characters such as leaf shape, leaf size, and corolla shape employed in the previous treatments (e.g., Fang and Stevens, 2005) are of little significance in species delimitation, whereas other previously neglected characters, including the number of leaf marginal teeth (setae), calyx lobe shape and size, and various characters of the fruit are valuable (Fritsch et al., 2015a; Lu et al., 2010). On this basis, a number of new species have now been described, either as completely new taxa or as former varieties raised to the species level, and the circumscriptions of many species have been clarified (Fritsch et al., 2008; Fritsch et al., 2015a; Fritsch et al., 2015b; Fritsch et al., 2016). Other plants still known only from herbarium specimens, or only in flower or fruit, likely represent further undescribed species in the group.

Phylogenetic studies of *Gaultheria* ser. *Trichophyllae* based on plastid genic regions (*matK*, *ndhF*, *rpl16*, *trnL-trnF* and *trnS-trnG*) plus the ITS region (Fritsch et al., 2011; Lu et al., 2010) failed to clarify most interspecific relationships in this group, and an analysis of this series in the context of the new species taxonomy has not yet been conducted. In prior analyses, *G. dolichopoda* and *G. trichophylla* were well supported as forming the sister group to the rest of the clade, and in the latter, *G. cardiosepala* + *G. thymifolia* form the sister to the clade of the remaining members. The relationships among the remaining species were generally poorly resolved, irrespective of species taxonomy.

The poor phylogenetic resolution of *Gaultheria* ser. *Trichophyllae* in previous studies resulted from limited amount of DNA sequence data variation available, likely because of the group's recent origin and rapid radiation (Lu et al., 2010; Lu et al., unpublished data). Whole plastid genome data from next-generation sequencing technology offer promise for generating adequate levels of variation for phylogenetic reconstruction among species (Yang et al., 2014). Plastid genomes have many advantages for phylogenetic reconstruction, including haploidy, small size, high copy number, conserved gene content and arrangement, uniparental inheritance, and general absence of recombination (Jansen and Ruhlman, 2012; Palmer et al., 1988; Raubeson and Jansen, 2005; Rousseau-Gueutin et al., 2015; Wu et al., 2010). They have thus been shown to be highly useful in the generation of robustly supported phylogenetic trees of taxa at low taxonomic levels which otherwise show poor clade support with traditional Sanger-based methods (e.g., Cai et al., 2015; Carbonell-Caballero et al., 2015; Huang et al., 2014; Parks et al., 2009; Yang et al., 2013).

A robust phylogeny provides the foundation for investigating the evolution of morphological traits and the potential underlying patterns in their transformations (Luo et al., 2015). However, the poor phylogenetic resolution of *Gaultheria* ser. *Trichophyllae*, together with the previously unclear taxonomic significance of morphological character variation, has precluded an attempt to trace the morphology of the series onto phylogenetic trees. With the new data available from morphological characters for species delimitation, the generation of whole plastid

genome data can address this, with the ultimate goal of better understanding the diversification, character evolution and “sky island” distribution of this group. Here we use whole plastid sequence data to establish a well resolved phylogenetic estimate of *Gaultheria* ser. *Trichophyllae* with the aid of next-generation sequencing. We reconstruct evolutionary patterns of morphological characters and assess their significance for classification. Finally, we compare the variation of various genic regions across the plastid genome among the species of the series and screen those with the highest sequence divergence for future evolutionary studies.

2. Materials and methods

2.1. Sampling strategy

Twenty-eight individuals representing all 19 species, one presumed hybrid (*L. Lu et al. LL-2013-54*), and one variety of *Gaultheria* ser. *Trichophyllae* were sampled. Thirteen species are represented by a single sample, and six species, those exhibiting extensive morphological variation, a relatively large geographic range, or non-monophyly as recovered in prior studies, are represented by two or more samples with different individuals of the same species denoted as a, b, c (Table 1). *Gaultheria hookeri*, *G. nummularioides* and *G. praticola* from *G. ser. Leucothoides* s.l., the sister clade of *G. ser. Trichophyllae* (Fritsch et al., 2011; Lu et al., 2010), were selected as outgroup representatives.

2.2. DNA sequence generation

Total genomic DNA of each sample with fresh green leaves (ca. 100 mg), either from living plants collected from the field or dried in silica gel, was extracted with a modified CTAB method (Doyle and Doyle, 1987; Table 1). Two sequencing methods were used to obtain the plastid genome DNA sequences. For the samples with fresh leaves, amplification through long-range PCR was performed with the nine primer pairs designed by Yang et al. (2014). Amplicon DNA concentrations were detected by visual approximation with gel electrophoresis and the amplicons of each sample were pooled together in roughly equal mass mixtures for subsequent sequencing. PCR amplicons were fragmented (~500 bp) and used for constructing short-insert libraries in accordance with the manufacturer's manual (Illumina). Paired-end sequencing was performed on the Illumina Miseq 2000 analyzer at the molecular biology experiment platform in the Southwest China Germplasm Bank of Wild Species, Kunming Institute of Botany. For the samples with dried leaves, paired-end sequencing was performed on an Illumina Hiseq 2500 instrument at the Beijing Genomics Institute (BGI) in Shenzhen, China.

The raw sequence reads were filtered for primer/adaptor sequences and low-quality reads with the NGS QC Tool Kit (Patel and Jain, 2012). The cut-off value for percentage of read length was 80, and that for PHRED quality score was 30. The filtered high-quality short reads were assembled into contigs by using the CLC Genomics Workbench version 8.0.1 (CLCBio, Aarhus, Denmark) with a word size of 60 or 64 (based on the length of the Contig N50 value and the count of contigs), and a

minimum contig length of 1000 bp. For plastid genome assembly, the complete plastid genome of *Vaccinium macrocarpon* Ait., the taxon phylogenetically closest to *Gaultheria*, was downloaded from GenBank (NC_019616) and used as a reference genome. The assembled plastid genomes were annotated with the program DOGMA [Wyman et al. (2004); <http://phylocluster.biosci.utexas.edu/dogma/>]. Start and stop codons, and boundaries of introns and exons, were edited manually.

2.3. Sequences alignment and data partitioning

The assembled whole plastid genomes displayed the typical quadripartite structure of most angiosperms, with one large single-copy (LSC) region, one small single-copy (SSC) region (sequencing of this region failed for 12 samples), and two inverted-repeat regions (IRa and IRb). The 31 whole plastid genome sequences recovered in this study were aligned initially by using MAFFT version 7.017b (Katoh and Toh, 2010) with default settings and then through subsequent manually adjustment in Geneious version 4.8.5 (Drummond et al., 2009); this dataset was used for phylogenetic reconstruction. To test the phylogenetic utility of different plastid genome regions, the LSC and IRa regions, which were obtained for all samples, were each extracted from the whole dataset and analyzed as two separate datasets. In addition, all coding (including protein coding genes, tRNAs and rRNAs) and noncoding (intergenic regions and introns) regions were extracted and concatenated to form two other distinct datasets, because coding and noncoding regions have been shown to exhibit distinct rates of nucleotide substitution (Jansen and Ruhlman, 2012;

Wolfe et al., 1987). One additional analysis was conducted by using the whole plastid genome excluding inverted repeat region b (IRb), divided into partitions (hereafter defined as the partitioned dataset). The aligned whole plastid genome dataset (without IRb) was first divided into 356 subsets (all protein coding regions, divided into 81 subsets by gene, and then by codon position; all noncoding regions, divided into 111 subsets; all tRNAs, combined into a subset; and all rRNAs, combined into a subset). These subsets were then analyzed with the heuristic search (search setting to rcluster) option in PartitionFinder version 1.1.1 (Lanfear et al., 2012) to select an optimal partitioning scheme in accordance with the Bayesian information criterion (BIC). Results yielded 17 partitions. In total, we constructed six datasets for phylogenetic analyses, i.e., sequence data from the whole plastid genome, LSC region, IRa region, coding regions, noncoding regions, and the partitioned dataset.

2.4. Phylogenetic analyses

For phylogenetic reconstruction, each dataset except for the partitioned dataset was analyzed with maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference (BI). The partitioned dataset was analyzed with only ML and BI. The MP analyses were performed with PAUP* version 4.0b10 (Swofford, 2002). All characters were assumed to be independent, unordered, and equally weighted, and gaps were treated as missing data (the latter of which also applied to the ML and BI analyses). The MP strict consensus trees were obtained by using the heuristic search option with 1000 random stepwise-addition replicates and tree bisection-reconnection

(TBR) branch swapping, and MULTREE option in effect. The extent of homoplasy was estimated by the consistency index (CI) and retention index (RI). Bootstrap support values for each node were calculated by running 1000 bootstrap replicates with 100 random stepwise-addition replicates, 10 trees held at each step, and TBR branch swapping. The ML analyses were run with the program RAxML version 7.2.6 (Stamatakis, 2006). Bootstrap analyses (1000 replicates) were performed simultaneously with the ML analysis (option “-f a”) under the GTR + G model as suggested in the RAxML manual.

The BI analyses were conducted with MrBayes version 3.2 (Ronquist et al., 2012). The substitution model and parameter settings [the gamma distribution of rate variation (shape), the proportion of invariable sites (pinvar), the stationary nucleotide frequencies of the GTR rate matrix (statefreq) and the six substitution rates of the GTR rate matrix (revmat)] for each of the five datasets were selected with the Akaike Information Criterion (AIC) from JModelTest version 2.14 (Darriba D et al., 2012). Two independent Markov chain Monte Carlo (MCMC) runs with four chains each (one cold chain and three incrementally heated chains) were run for 10 million generations [with the average standard deviation of split frequencies (ASDSF) achieving values lower than 0.01] with random starting trees and trees sampled every 1000 generations. Tracer version 1.6 (<http://tree.bio.ed.ac.uk/software/tracer/>) was also used to check for stabilization of overall likelihood within and convergence between runs for each input file. Twenty-five percent of the trees were discarded as

burn-in and Bayesian posterior probabilities were estimated from the consensus of remaining trees.

For Bayesian inference of the partitioned dataset, the substitution model for each subset was defined simultaneously in the process of selecting partitioning schemes by PartitionFinder. In addition, all parameters (shape, pinvar, statefreq and revmat) were unlinked across partitions except those for branch lengths and topology. Each partition was set to evolve under a specific rate, with “prset applyto” set to “all” and “ratepr” set to “variable”. Four simultaneous MCMC chains were each run for 2 million generations with random starting trees and trees sampled every 100 generations. Twenty-five percent of the trees were discarded as burn-in. Bayesian posterior probabilities were estimated from the consensus of remaining trees.

2.5. Divergence hotspot regions

All protein-coding genes and noncoding (introns and intergenic spacers) regions (> 200 bp) were extracted from each of the 28 plastid genomes from *Gaultheria* ser. *Trichophyllae* and homologous regions were aligned. The percentage of variable sites was calculated for each region as in Kress et al. (2005) to screen sequence divergence hotspot regions. The average percentages of variable sites in the combinations of protein coding genes and noncoding regions (both including those less than 200 bp) located among variously circumscribed regions of the plastid genome (i.e., across the whole plastid genome, LSC region, IRa region) were calculated as average percentage = (percentage₁ + ... percentage_n)/n. The average

percentages of variable sites among all intergenic regions and introns across the whole plastid genome were also calculated, with the same method. To examine the phylogenetic efficiency of the screened divergence hotspot regions, ML trees were reconstructed with each of the selected regions and their concatenation. Bootstrap analyses (1000 replicates) were performed simultaneously with the ML analysis (option “-f a”) under the GTR + G model, as suggested in the RAxML manual.

2.6. *Reconstruction of morphological character evolution*

Phylogenetic estimates inferred with plastid DNA sequences have been employed as frameworks for morphological character evolution reconstruction in numerous evolutionary studies (e.g., Chaowasku et al., 2014; Clement et al., 2014; Ding et al., 2014; Lagomarsino et al., 2014). To assess the systematic significance of morphological characters in *Gaultheria* ser. *Trichophyllae* and reveal the extent of convergent evolution, 10 characters previously suggested having diagnostic importance and taxonomic utility in the series were selected for analysis, i.e., habit, presence of puberulence on the lamina midvein adaxial-basally, presence of setose trichomes on the abaxial surface of the leaf blade, calyx lobe trichome presence, corolla color, corolla shape, anther theca awn number, fruiting calyx openness, fruiting calyx color, and capsule color (Fang and Stevens, 2005; Fritsch et al., 2008; Lu et al., 2010; Middleton, 1991; Xu, 1981; Table 2). Characters were scored from literature sources, herbarium specimens, and field observations (Table 1). Each character was treated as unordered, binary or multistate, with unknown data scored as

missing (Table 3), and then optimized onto the ML tree derived from the whole plastid genome sequences. Fitch Parsimony optimization was conducted with Mesquite 2.75 (Maddison and Maddison, 2011).

3. Results

3.1. Genome assembly

Eighteen individuals sequenced with long-range PCR amplicon DNA by the Illumina Miseq 2000 produced 711,286 to 2,142,744 paired-end reads (250 bp average read length), and the other 13 individuals sequenced with total genomic DNA by the Illumina Hiseq 2500 produced 17,232,330 to 22,474,868 paired-end reads (90 bp average read length). After reference-guided assembly, the obtained plastid genome sequences ranged from 62505 bp to 163526 bp in length (Table 1).

3.2. Phylogenetic resolution

The LSC region possesses the most variable and parsimony-informative sites. As expected, the noncoding regions possess the highest percentages of both variable and parsimony-informative sites, whereas the coding regions possess the lowest percentages of variable and parsimony-informative sites. With the shortest aligned sequence length, the IRa region possesses the fewest variable sites and parsimony-informative sites. The MP tree based on the LSC region dataset has the highest CI and RI values, followed by those of the noncoding region and coding

region dataset, and that of IRa region dataset has the lowest CI and RI values (Table 4). The whole plastid genome dataset contains 2.6% missing data.

Phylogenetic trees estimated from the whole plastid genome were strongly supported with ML BS > 98 and MP BS > 95 (usually 100), and PP = 1.00 (Figure 1). *Gaultheria* ser. *Trichophyllae* is divided into a clade (Clade 1) comprising *G. dolichopoda*, *G. minuta*, *G. trichophylla* var. *trichophylla*, and *G. trichophylla* var. *tetracme*, and another (Clade 2) comprising the remaining species. Clade 2 comprises several other subclades, here referred to as Subclades 2A–G.

Analyses based on the noncoding regions, LSC region, and the partitioned dataset, irrespective of reconstruction method, all yielded identical topologies, which are also identical to that estimated from the whole plastid genome dataset. The clades from these analyses generally exhibit high support values, i.e., ML and MP BS > 95 (usually 100) and PP = 1, except for two clades from the trees based on the noncoding regions dataset and one clade from the trees based on the LSC region, respectively (Figures S1 – S3). Although the coding-region analysis also yielded the same tree across all reconstruction methods (Figure S4), the support values are somewhat lower for several nodes as compared to the whole plastid genome tree, and two conflicts were found between the topologies based on these two datasets (cf. Figures 1 and S4), i.e., Subclades 2A and 2B form a sister group (ML BS = 91, MP BS = 86, PP = 1), and *G. obovata* nests within Subclade 2F and is sister to *G. ciliisepala_c* (ML BS = 100, MP BS = 97, PP = 1). The analyses based on IRa region dataset generated the most unresolved topologies, with low to moderate support,

i.e., > 10 nodes with ML and MP values < 65 and PP values generally < 0.90 (Figure S5).

3.3. Divergence hotspot regions

Among the 65 protein-coding genes with sequence length > 200 bp, only *rpl23b* (a pseudogene) and *rps3b* possess > 5% variable sites (Figure 2). Forty-two genes possess < 1% variable sites, and nine of these are invariant genes. Among the 85 noncoding regions with sequence length > 200 bp, six possess > 5% variable sites. Twenty-four regions (28.2% of the total) possess > 3% variable sites, 16 regions have < 1% variable sites, and only one region is invariant. Therefore in total eight regions, i.e., *rpl36_infA*, *trnF(GAA)_ndhJ*, *trnD(GUC)_psbM*, *trnL(UAA)_trnF(GAA)*, *trnT(UGU)_trnL(UAA)*, *rpl23b*, *rps3b* and *petN_trnC(GCA)* (listed in order of decreasing variation) possess > 5% variable sites. None of these regions alone effectively resolve the phylogenetic relationships of *Gaultheria* ser. *Trichophyllae* (trees not shown), but the dataset of these combined regions partially resolve the inter- and infra-specific relationships of this series (Figure S6).

3.4. Reconstruction of morphological character evolution

Each of the 10 morphological characters included in this study shows a distinct pattern of evolutionary change (Figure 3) with the number of state changes per character ranging from two to 12 (Figure 4).

4. Discussion

4.1. Phylogenetic relationships

Our study has clarified previously unresolved phylogenetic relationships among species of *Gaultheria* ser. *Trichophyllae* with plastid phylogenomic analyses. Two clades and seven subclades are recovered, and most of these can be characterized by at least one most recent common ancestral morphological character state.

Phylogenetic relationships of this series are discussed as follows.

Clade 1. Results of morphological character reconstruction indicate that a pink-colored corolla and blue-colored fruiting calyx are the most recent common ancestral states of Clade 1 (Figure 3). The phylogenetic relationships among species in this clade are well resolved. Consistent with the result of Lu et al. (2010) and Fritsch et al. (2011), *Gaultheria dolichopoda* is inferred as sister to the rest of the clade, and this species can be identified by the unique elongate fruiting pedicel.

Gaultheria minuta, *G. trichophylla* var. *trichophylla*, and *G. trichophylla* var. *tetracme* all bear notably long leaf marginal setae, in contrast to the short setae of *G. dolichopoda*. Habit, leaf size and shape, number of anther awns, and other characters are variable and appear to be continuous within *G. trichophylla*, with no correlated morphological gaps (Fritsch et al., 2008), making it easy to confuse *G. trichophylla* with other morphologically similar species. *G. trichophylla* was revealed as non-monophyletic by Lu et al. (2010) with one sample (6N) diverging early and forming the basal clade with *G. dolichopoda* and two other samples (2L and 4F)

grouping together and nesting in a distinct clade. We could not obtain sequences of the species in the latter clade for this study, and so further studies with more sampling should be done to clarify this issue. Although the sister relationship of *G. trichophylla* var. *trichophylla* and *G. trichophylla* var. *tetracme* is strongly supported by phylogenomic analyses in this study, these taxa are distinct in haplotype due to 121 variable sites throughout the plastid genome sequences, with a p-distance of 0.002. In addition, *G. trichophylla* var. *trichophylla* is characterized by one-awned anther thecae (rarely with two awns) and is distributed in the western Himalaya-Hengduan region mainly along the Himalaya, whereas *G. trichophylla* var. *tetracme* always possesses two-awned thecae and is distributed in the Hengduan Shan. These data may warrant the recognition of *G. trichophylla* var. *tetracme* at the species level.

Clade 2. In Clade 2, *Gaultheria cardiosepala*, *G. nivea* and *G. thymifolia* of Subclade 2A grow at relatively low elevations (mainly 2700–3000 m) and are characterized by a white closed fruiting calyx and lack of setose trichomes on the abaxial surface of the leaf blade. Within this subclade, *G. cardiosepala* is distinguished by the taller habit (10–30 cm), urceolate corolla, and anther awns 0.3–0.6 mm long (versus 2.5–10 cm, campanulate corolla, and anther awns < 0.3 mm). *Gaultheria nivea* was treated as a species in Airy Shaw (1941) and a variety of *G. sinensis* in Fang and Stevens (2005). Following Airy Shaw's treatment and based on morphological characters, Fritsch et al. (2008) proposed its affinity to *G. cardiosepala* and *G. thymifolia* rather than *G. sinensis*. Our plastid phylogenomic

analyses support the treatment of Fritsch et al. (2008), although the two individuals of *G. nivea* sampled successively diverge instead of forming a clade. Our results yielded 211 variable sites between *G. cardiosepala* and *G. nivea_a* across the whole genome, with a p-distance of 0.001. Only 33 variable sites were recovered between *G. nivea_a* and *G. thymifolia*, and 51 between the two samples of *G. nivea* with both p-distances equal to zero. Both *G. nivea* and *G. thymifolia* were found growing in the same Chukuai Lake basin in the Gaoligong Mountains (Fritsch et al., 2008). This similarity between the two species may have arisen from reticulate evolution, lineage sorting, or cryptic species. Composed of *G. bryoides* and *G. marronina*, Subclade 2B can be characterized by one-awned anther theca, a white fruiting calyx, and a maroon capsule. Maroon capsules are unique to this subclade in *G. ser. Trichophyllae*, except for their sporadic and rare occurrence in *G. eciliata* and *G. sinensis*.

Subclade 2C comprises *Gaultheria albiflora*, *G. obovata* and *G. sinensis*. *Gaultheria albiflora* has been confused morphologically with *G. eciliata* but is clearly distinct by several reliable morphological characters (Fritsch et al., 2015a) and also with molecular data (Lu et al., 2010). Our plastid phylogenomic analyses recovered these two species as occurring in two separate subclades (C and E), corroborating their distinctness. *Gaultheria obovata* was traditionally treated as a variety of *G. trichophylla* (Airy Shaw, 1941) due to the presence of notably long stem and leaf marginal setae but has been shown to be distinct with molecular data (Lu et al., 2010). Fritsch et al. (2015a) distinguished *G. obovata* from *G.*

trichophylla based on floral and fruit characters, and recognized this variety as a species, a treatment which is also supported by the present study. The phylogenetic position of both *G. albiflora* and *G. obovata* were poorly resolved in Lu et al. (2010) but our analyses strongly support the sister relationship between *G. obovata* and *G. sinensis*, and support *G. albiflora* is their sister. In addition, these three species all possess an open fruiting calyx and grow at elevations > 3200 m.

Subclade 2D comprises *Gaultheria hypochlora* and the three samples of *G. stenophylla*. An erect or ascending-erect habit and a ciliolate calyx lobe are shared by these two species. They also share the character states of 8 to 15 leaf marginal teeth per side, a campanulate white corolla, and a prolate to ellipsoid fruiting calyx (Fritsch et al., 2015a). In addition, both blue- and white-colored fruiting calyces are observed occurring in these two species in the field. It often requires care to distinguish between them in the field, especially when they are growing intermixed (Fritsch et al., 2008; Fritsch et al., 2015a). In the present study, the sister relationship between *G. stenophylla_a* and *G. stenophylla_b* is strongly supported with ML and MP BS = 100, and PP = 1. However, *G. stenophylla_c* groups with *G. hypochlora*, also with high support values (ML and MP BS > 98, PP = 1), suggesting reticulate evolution or lineage sorting among these two species. However, further more work at population level of these two species is needed.

Gaultheria eciliata and *G. gonggashanensis*, distributed in the Gaoligong Mountains mainly in Yunnan Province and Gongga Mountains in Sichuan Province, respectively, are distinct in plant size (smaller in *G. eciliata*), both growing generally

at > 3000 m elevation. Fritsch et al. (2015b) suggested that *G. gonggashanensis* is most similar to *G. sinensis* in terms of its general leaf features, flower bud, bud bracteole size and shape, and eciliolate calyx lobes. However, our analyses support a sister relationship between *G. eciliata* and *G. gonggashanensis*. In addition, the characters in *G. eciliata* are generally consistent with those of *G. gonggashanensis* except for three differences, i.e., puberulence present on the lamina midvein adaxial-basally (versus absent), ciliolate calyx lobes (versus eciliolate), and one-awned anther theca (versus two-awned), albeit within-species polymorphism is exhibited in several other characters.

Subclade 2F consists of *Gaultheria ciliisepala* and *G. jingdongensis*. Fritsch et al. (2015a) took up the unpublished name *Gaultheria sinensis* var. *ciliisepala* written on herbarium specimens by Airy Shaw but recognized it as a species, and suggested that it is closely related to *G. eciliata*, *G. hypochlora*, *G. stenophylla*, and *G. jingdongensis* because all these species share ciliolate calyx lobes, a character unique to these species in *G. ser. Trichophyllae*. The analyses of Lu et al. (2010) also found a sister relationship of *G. ciliisepala* and *G. jingdongensis* but failed to resolve their relationship with the other species in *G. ser. Trichophyllae*. From our data, Subclade 2F is sister to Subclade 2G with ML and MP BS = 98, and PP = 1, although further morphological evidence is needed to support this relationship.

Within Subclade 2G, *Gaultheria crassifolia* and *L. Lu et al. LL-2013-54* form a strongly supported sister group (ML and MP BS = 100, PP = 1). These entities can be distinguished from *G. major* by an ascending-erect habit and a closed fruiting calyx.

On the basis of morphology (e.g., intermediate habit and leaf length), phylogenetic placement, and occurrence in the same area as the two putative parents, we hypothesize that *L. Lu et al. LL-2013-54* is a hybrid between *G. major* and *G. crassifolia*. In this regard, the plastid genomes of *L. Lu et al. LL-2013-54* and *G. crassifolia* are identical, and so the latter could be the maternal parent. Nevertheless, evidence from nuclear data is needed to test this.

Two conflicts were found between the topologies estimated from the dataset of the coding regions and that of the whole plastid genome. The sister relationship between Subclades 2A and 2B found in the former could be attributed to homoplasy, because there are only 409 parsimony-informative sites detected in the dataset of the coding regions, which is less than one-fourth of that (1778) from the whole plastid genome dataset; 17 parsimony-informative sites are detected as shared by *Gaultheria obovata* and *G. ciliisepala_c* in the *rps3b* gene, which is also one of the eight screened sequence divergence hotspot regions, and located in the IRa region. Thus, the inferred sister relationship of *G. obovata* and *G. ciliisepala_c* could be caused by homoplasious mutations in the *rps3b* gene.

4.2. Utility of divergence hotspot regions for phylogenetic resolution

Our results indicate that the proportion and distribution patterns of variable sites among protein-coding genes and noncoding regions are strikingly different. As a whole, the percentage of variable sites in the noncoding regions is higher than that in the protein coding genes, and the LSC region possesses a higher proportion of

variable site than the IRa region. These results are generally consistent with those of most other studies on plastid genomes (Clegg et al., 1994; Huang et al., 2014; Jansen and Ruhlman, 2012; Perry and Wolfe, 2002; Wolfe et al., 1987). Noncoding regions from the LSC region show a higher average variation ratio than those from IRa region (2.5% and 1.1%, more than twice, Table 5), and the average variation ratio of the protein-coding genes from the LSC region is less than that from the IRa region (0.85% and 1.0%), indicating that the overall higher variation observed in the LSC region is chiefly contributed by the noncoding regions. The average variation ratio of the introns (1.4%) is lower than that of the intergenic regions (2.3%) but higher than that of protein-coding genes (0.89%). This has been suggested as a general pattern occurring across different angiosperm families (Jansen and Ruhlman, 2012).

Phylogenetic relationships among species in *Gaultheria* ser. *Trichophyllae* were not well resolved in previous studies with multiple plastid sequence analyses (Lu et al., 2010). A limited number of DNA regions with insufficient phylogenetically informative sites, as well as low rate of nucleotide substitutions in the selected regions, might account for the low resolution (Shaw et al., 2007). For this group, which may have undergone recent rapid radiation, it is necessary to explore more molecular markers to facilitate future evolutionary studies. Among the eight screened sequence divergence hotspot regions with variable sites exceeding 5%, *trnF*(GAA)*_ndhJ* and *trnT*(UGU)*_trnL*(UAA) have also been identified as the most variable regions for the taxa of Araliaceae (Li et al., 2013). In addition, the *rpl16* intron and *ndhC-trnV*(UAC) intergenic spacer, which have been suggested to be

among the most variable regions across angiosperm lineages (Shaw et al., 2014), are also detected as possessing relatively high character variation (4.16% and 3.18%, respectively, though $< 5\%$) in *G. ser. Trichophyllae*. However, these eight regions still failed to sufficiently resolve the phylogenetic relationships of *G. ser.*

Trichophyllae. Further studies might be conducted to test the efficiency of these regions in evolutionary studies of *Gaultheria* and related taxa which also occur in the Himalaya-Hengduan Mountains.

4.3. Evolution of morphological characters

Ancestral state reconstructions suggest that the 10 studied characters have been subjected to various degrees of selection pressure during the evolutionary history of this rapidly radiating group. With increasing elevation, plant habit in *Gaultheria ser. Trichophyllae* evolves from erect or ascending-erect to prostrate or ascending/creeping prostrate. The low dense prostrate habit of the high-alpine species may create a warmer microclimate and moderate the effect of wind (Körner, 2012; Yang and Sun, 2006). Thus, this evolutionary pattern might be explained by morphological adaptation of plants to the harsh alpine environments of the Himalaya-Hengduan Mountains.

Acknowledgements

The authors thank Jie Cai and Ting Zhang for the sample collection of *Gaultheria dolichopoda*; Lei-Lei Zhang and Yu-Cun Sun for assistance with field work; Jun-Bo

Yang, Xiang Zhao and Jing Yang for laboratory support; and Hong-Tao Li, Hai-Dong Li and Xiao-Juan Li for their help with data analysis. This study was supported by grants from the National Natural Science Foundation of China (41671052, 31270272 to LL), the National Geographic Society (4-03-9329 to PWF and LL), National Key Basic Research Program of China (2014CB954100 to DZL), Key Programs of the Chinese Academy of Sciences (KSZD-EW-Z-011-01 to HW), and U.S. National Science Foundation (DEB-0717711 to PWF).

References

- Airy Shaw, H., 1941. Studies in the Ericales: IV. Classification of the Asiatic species of *Gaultheria*. Bull. Misc. Inf. (R. Bot. Gard., Kew). 1940, 306–330.
<http://dx.doi.org/10.2307/4120216>.
- Bush, C.M., Lu, L., Fritsch, P.W., Li, D.-Z., Kron, K.A., 2009. Phylogeny of Gaultherieae (Ericaceae: Vaccinioideae) based on DNA sequence data from *matK*, *ndhF*, and nrITS. Int. J. Plant Sci. 170, 355–364.
<http://dx.doi.org/10.1086/596330>.
- Cai, J., Ma, P.-F., Li, H.-T., Li, D.-Z., 2015. Complete plastid genome sequencing of four *Tilia* species (Malvaceae): a comparative analysis and phylogenetic implications. PloS one 10, e0142705.
<http://dx.doi.org/10.1371/journal.pone.0142705>.
- Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., Dopazo, J., 2015. 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. <http://dx.doi.org/10.1093/molbev/msv082>.
- Chaowasku, T., Thomas, D.C., van der Ham, R.W., Smets, E.F., Mols, J.B., Chatrou, L.W., 2014. A plastid DNA phylogeny of tribe Miliuseae: insights into relationships and character evolution in one of the most recalcitrant major clades of Annonaceae. Am. J. Bot. 101, 691–709.
<http://dx.doi.org/10.3732/ajb.1300403>.

- Clegg, M.T., Gaut, B.S., Learn, G.H., Jr., Morton, B.R., 1994. Rates and patterns of chloroplast DNA evolution. *Proc. Natl. Acad. Sci. USA* 91, 6795–6801.
- Clement, W.L., Arakaki, M., Sweeney, P.W., Edwards, E.J., Donoghue, M.J., 2014. A chloroplast tree for *Viburnum* (Adoxaceae) and its implications for phylogenetic classification and character evolution. *Am. J. Bot.* 101, 1029–1049.
<http://dx.doi.org/10.3732/ajb.1400015>.
- Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772.
<http://dx.doi.org/10.1038/nmeth.2109>.
- Ding, H.-H., Chao, Y.-S., Callado, J.R., Dong, S.-Y., 2014. Phylogeny and character evolution of the fern genus *Tectaria* (Tectariaceae) in the Old World inferred from chloroplast DNA sequences. *Mol. Phylogenet. Evol.* 80, 66–78.
<http://dx.doi.org/10.1016/j.ympev.2014.06.004>.
- Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.
- Drummond, A., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R.,
Stones-Havas, S., Thierer, T., Wilson, A., 2009. Geneious v4.8.5. Biomatters
Ltd. Auckland, New Zealand.
- Fang, R.-C., 1999. New taxa of Ericaceae from China. *Novon* 9, 162–178.
<http://dx.doi.org/10.2307/3391793>.

- Fang, R.-C., Stevens, P.F., 2005. *Gaultheria*. In: Wu, Z.-Y., Raven, P.H., Hong, D.-Y. (Eds.), Flora of China, Apiaceae through Ericaceae, Vol. 14. Science Press/Missouri Botanical Garden Press, Beijing/St. Louis, pp. 464–475.
- Fritsch, P.W., Zhou, L.-H., Lu, L., Bartholomew, B., 2008. The flowering plant genus *Gaultheria* (Ericaceae) in the Gaoligong Shan, along the border region of China and Myanmar. Proc. Calif. Acad. Sci., Ser. 4, 147–214.
- Fritsch, P.W., Lu, L., Bush, C.M., Cruz, B.C., Kron, K.A., Li, D.-Z., 2011. Phylogenetic analysis of the wintergreen group (Ericaceae) based on six genic regions. Syst. Bot. 36, 990–1003.
<http://dx.doi.org/10.1600/036364411X604994>.
- Fritsch, P.W., Lu, L., Hong, W., Li, D.-Z., 2015a. New species, taxonomic renovations, and typifications in *Gaultheria* series *Trichophyllae* (Ericaceae). Phytotaxa 201, 1–26. <http://dx.doi.org/10.11646/phytotaxa.201.1.1>.
- Fritsch, P.W., Lu, L., Wang, H., 2015b. *Gaultheria gonggashanensis* sp. nov. (Ericaceae) from Sichuan, China. Nord. J. Bot. 33, 582–585.
<http://dx.doi.org/10.1111/njb.00837>.
- Fritsch, P.W., Lu, L., Xu, B., Wang, H., Li, D.-Z., 2016. *Gaultheria marronina* sp. nov. (Ericaceae) from Sichuan, China. Nord. J. Bot. 000, 001–005.
<http://dx.doi.org/10.1111/njb.01129>.
- Heald, W.F., 1951. Sky islands of Arizona. Nat. Hist. 60, 56–63, 95–96. URI:
<http://hdl.handle.net/2246/6381>.

- Huang, H., Shi, C., Liu, Y., Mao, S.-Y., Gao, L.-Z., 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol. Biol. 14, 151. <http://dx.doi.org/10.1186/1471-2148-14-151>.
- Jansen, R.K., Ruhlman, T.A., 2012. Plastid genomes of seed plants. In: Bock, R., Knook, V. (Eds), Genomics of chloroplasts and mitochondria. Springer, Netherlands , pp. 103–126. http://dx.doi.org/10.1007/978-94-007-2920-9_5.
- Körner, C., 2012. Alpine treelines: functional ecology of the global high elevation tree limits. Springer, New York.
- Katoh, K., Toh, H., 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26, 1899–1900. <http://dx.doi.org/10.1093/bioinformatics/btq224>.
- Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H., 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. USA 102, 8369–8374. <http://dx.doi.org/10.1093/10.1073/pnas.0503123102>.
- Lagomarsino, L.P., Antonelli, A., Muchhala, N., Timmermann, A., Mathews, S., Davis, C.C., 2014. Phylogeny, classification, and fruit evolution of the species-rich Neotropical bellflowers (Campanulaceae: Lobelioideae). Am. J. Bot. 101, 2097–2112. <http://dx.doi.org/10.3732/ajb.1400339>.
- Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic

- analyses. *Mol. Biol. Evol.* 29, 1695–1701.
<http://dx.doi.org/10.1093/molbev/mss020>.
- Li, R., Ma, P.-F., Wen, J., Yi, T.-S., 2013. Complete sequencing of five Araliaceae chloroplast genomes and the phylogenetic implications. *PloS one* 8, e78568.
<http://dx.doi.org/10.1371/journal.pone.0078568>.
- Lu, L., Fritsch, P.W., Cruz, B.C., Wang, H., Li, D.-Z., 2010. Reticulate evolution, cryptic species, and character convergence in the core East Asian clade of *Gaultheria* (Ericaceae). *Mol. Phylogenet. Evol.* 57, 364–379.
<http://dx.doi.org/doi:10.1016/j.ympev.2010.06.002>.
- Luo, Y., Lu, L., Wortley, A.H., Li, D.-Z., Wang, H., Blackmore, S., 2015. Evolution of angiosperm pollen. 3. Monocots. *Ann. Mo. Bot. Gard.* 101, 406–455.
<http://dx.doi.org/10.3417/2014014>.
- Maddison, W., Maddison, D., 2011. Mesquite: a modular system for evolutionary analysis. 2011. Version 2.75. <http://mesquiteproject.org>.
- Merrill, E.D., 1941. The upper Burma plants collected by captain F. Kingdon Ward on the Vernay-Cutting expedition, 1938–39. *Brittonia* 4, 20–188.
<http://dx.doi.org/10.2307/2804985>.
- Middleton, D.J., 1991. Infrageneric classification of the genus *Gaultheria* L. (Ericaceae). *Bot. J. Linn. Soc.* 106, 229–258.
<http://dx.doi.org/10.1111/j.1095-8339.1991.tb02293.x>.

- Palmer, J.D., Jansen, R.K., Michaels, H.J., Chase, M.W., Manhart, J.R., 1988. Chloroplast DNA variation and plant phylogeny. *Ann. Mo. Bot. Gard.* 75, 1180–1206. <http://dx.doi.org/10.2307/2399279>.
- Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7, 84. <http://dx.doi.org/10.1186/1741-7007-7-84>.
- Patel, R.K., Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One* 7, e30619. <http://dx.doi.org/10.1371/journal.pone.0030619>.
- Perry, A.S., Wolfe, K.H., 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J. Mol. Evol.* 55, 501–508.
- Raubeson, L.A., Jansen, R.K., 2005. Chloroplast genomes of plants. In: Henry, R.J. (Ed.), *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*. CABI Publishing, UK, pp. 45–68. <http://dx.doi.org/10.1079/9780851999043.0045>.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. <http://dx.doi.org/10.1093/sysbio/sys029>.
- Rousseau-Gueutin, M., Bellot, S., Martin, G.E., Boutte, J., Chelaifa, H., Lima, O., Michon-Coudouel, S., Naquin, D., Salmon, A., Ainouche, K., Ainouche, M., 2015. The chloroplast genome of the hexaploid *Spartina maritima* (Poaceae,

- Chloridoideae): comparative analyses and molecular dating. *Mol. Phylogenet. Evol.* 93, 5–16. <http://dx.doi.org/10.1016/j.ympev.2015.06.013>.
- Shaw, J., Lickey, E.B., Schilling, E.E., Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.* 94, 275–288. <http://dx.doi.org/10.3732/ajb.94.3.275>.
- Shaw, J., Shafer, H.L., Leonard, O.R., Kovach, M.J., Schorr, M., Morris, A.B., 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *Am. J. Bot.* 101, 1987–2004. <http://dx.doi.org/10.3732/ajb.1400398>.
- Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. <http://dx.doi.org/10.1093/bioinformatics/btl446>.
- Swofford, D., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other Methods). Version 4.0 b10. Sinauer Associates, Sunderland, MA.
- Wolfe, K.H., Li, W.-H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* 84, 9054–9058.
- Wu, F.-H., Chan, M.-T., Liao, D.-C., Hsu, C.-T., Lee, Y.-W., Daniell, H., Duvall, M.R., Lin, C.-S., 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in

- Oncidiinae. BMC Plant Biol. 10, 68.
<http://dx.doi.org/10.1186/1471-2229-10-68>.
- Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252–3255.
<http://dx.doi.org/10.1093/bioinformatics/bth352>.
- Xu, T.-Z., 1981. Preliminary study of classification on Chinese *Gaultheria*. Acta Bot. Yunnanica 3, 417–434.
- Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R., Li, D.-Z., 2013. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol. Biol. 13, 84. <http://dx.doi.org/10.1186/1471-2148-13-84>.
- Yang, J.-B., Li, D.-Z., Li, H.-T., 2014. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Mol. Ecol. Resour. 14, 1024–1031. <http://dx.doi.org/10.1111/1755-0998.12251>.
- Yang, Y., Sun, H., 2006. Advances in the functional ecology of alpine and arctic plants. Acta Bot. Yunnanica 28, 43–53.

Figure captions

Figure 1. Maximum likelihood (ML) phylogenetic tree of *Gaultheria* series *Trichophyllae* based on the whole plastid genome dataset. Numbers along branches are bootstrap values (BS) from ML, maximum parsimony (MP), and Bayesian inference (posterior probabilities, PP), in that order. Asterisks indicate nodes with BS (ML and MP) = 100 and PP = 1. Major clades are denoted for reference in the text.

Figure 2. Percentage of variable sites in each homologous region and their distribution patterns across the plastid genome of taxa in *Gaultheria* series *Trichophyllae*. A: protein coding genes; B: noncoding regions. LSC = large single-copy region; IRa = inverted-repeat region a.

Figure 3. State changes of 10 morphological characters traced with Fitch parsimony over the maximum likelihood tree from the whole plastid genome dataset. Numbers on bars indicate state transformations. Numbers before the colon indicate characters (see legend for details); numbers after the colon indicate state transformations from lower to upper node with states represented by numbers before and after the greater than sign. Characters and states follow Table 2. Major clades are denoted for reference in the text.

Figure 4. Bar diagram of the number of state changes of the 10 characters

throughout the phylogeny of *Gaultheria* series *Trichophyllae* traced with Fitch parsimony over the ML tree from the whole plastid genome dataset.

Figure 5. Habit, leaf morphology, flowers and fruit in selected species of *Gaultheria*

series *Trichophyllae*. A: *Gaultheria marronina* (L. Lu et al. LL-2011-37) showing prostrate habit and white slightly open fruit with maroon capsule; B: *G. dolichopoda* (L. Lu et al. LL-2014-25) showing erect habit, blue fruit and long pedicel; C: *G. trichophylla* var. *trichophylla* (L. Lu et al. LL-2013-47) showing leaves with long marginal setae; D: *G. cardiosepala* (L. Lu et al. LL-2014-44) showing urceolate white corolla; E: *G. stenophylla* (L. Lu et al. LL-2014-8) showing campanulate white corolla with two-awned anther thecae; F: *G. tetracme* (L. Lu & M. Y. Zhang LL-2015-14) showing campanulate pink corolla and leaves with notable marginal setae; G: *G. ciliisepala* (L. Lu et al. LL-2013-20) showing open fruit with green capsule; H: *G. cardiosepala* (L. Lu et al. 6-0516) showing closed white fruit; I: *G. stenophylla* (L. Lu et al. LL-2013-7) showing nearly closed blue fruit; J: *G. ciliisepala* (L. Lu et al. LL-2013-20) showing open blue fruit; K: *G. crassifolia* (L. Lu et al. LL-2013-13) showing closed blue fruit; L: *G. stenophylla* (L. Lu et al. LL-2013-7) showing nearly closed blue fruit. Photographs: A, C, E, H, J, K, L by L. Lu; B by J. Cai; D, F, G, I by M. Y. Zhang.

Table 1

Sample voucher information, with assembled plastid genome size, GenBank accession number, and references for morphological data. All vouchers are deposited in the herbarium of the Kunming Institute of Botany (KUN).

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
<i>G. albiflora</i> (Ching ex T.Z. Hsu) P.W. Fritsch & Lu Lu	<i>L. Lu et al. LL-2013-50</i>	Yanmen Township, Deqin, Yunnan, China	88312	XXXXXXX	Fritsch et al., 2015a
<i>G. bryoides</i> P.W. Fritsch & L.H. Zhou	<i>L. Lu et al. LL-2014-24</i>	Dulongjiang Township, Gongshan, Yunnan , China	89063	XXXXXXX	Fritsch et al., 2008; field observation
<i>G. cardiosepala</i> Hand.-Mazz.	<i>L. Lu et al. LL-2014-38</i>	Pianma Township, Lushui, Yunnan, China	141330	XXXXXXX	Fritsch et al., 2008; Fang and Stevens, 2005; field observation
<i>G. ciliisepala</i> Airy Shaw ex P.W. Fritsch & Lu Lu (a)	<i>L. Lu & P. W. Fritsch LL-2011-8</i>	Cang Shan, Dali, Yunnan, China	161398	XXXXXXX	Fritsch et al., 2015a; field observation

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
<i>G. ciliisepala</i> (b)	<i>L. Lu et al. LL-2013-34</i>	Cikai Township, Gongshan, Yunnan, China	161025	XXXXXXX	Fritsch et al., 2015a; field observation
<i>G. ciliisepala</i> (c)	<i>L. Lu et al. LL-2014-40</i>	Pianma Township, Lushui, Yunnan, China	143440	XXXXXXX	Fritsch et al., 2015a; field observation
<i>G. crassifolia</i> (Airy Shaw) P.W. Fritsch & Lu Lu	<i>L. Lu et al. LL-2014-30</i>	Lishadi Township, Fugong, Yunnan, China	163526	XXXXXXX	Fritsch et al., 2015a; field observation
<i>G. dolichopoda</i> Airy Shaw	<i>L. Lu et al. LL-2014-25</i>	Dulongjiang Township, Gongshan, Yunnan, China	161696	XXXXXXX	Fritsch et al., 2008; field observation; herbarium specimens
<i>G. eciliata</i> (Rae & D.G. Long) P.W. Fritsch & L.H. Zhou	<i>L. Lu et al. LL-2014-13</i>	Danzhu village, Gongshan, Yunnan, China	108595	XXXXXXX	Fritsch et al., 2015a; field observation

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
<i>G. gonggashanensis</i> P.W.	<i>L. Lu & P.W. Fritsch</i>	Luding, Sichuan, China	159979	XXXXXXX	Fritsch et al., 2015b
Fritsch & Lu Lu	<i>LL-2011-33</i>				
<i>G. hookeri</i> C.B. Clarke	<i>L. Lu et al. LL-2013-18</i>	Cikai Township, Gongshan, Yunnan, China	137404	XXXXXXX	Fritsch et al., 2008, herbarium specimens
<i>G. hypochlora</i> Airy Shaw	<i>L. Lu et al. LL-2014-9</i>	Dulongjiang township, Gongshan, Yunnan, China	141850	XXXXXXX	Fritsch et al., 2008, 2015a; field observation; herbarium specimens
<i>G. jingdongensis</i> R.C. Fang	<i>L. Lu et al. LL-2013-58</i>	Linjie Township, Jingdong, Yunnan, China	160015	XXXXXXX	Fang, 1999; Fang and Stevens, 2005; field observation; herbarium specimens
<i>G. major</i> (Airy Shaw) P.W.	<i>L. Lu et al. LL-2014-60</i>	Yezhi Township, Weixi,	141665	XXXXXXX	Fritsch et al., 2015a; field

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
Fritsch & Lu Lu (a)		Yunnan, China			observation
<i>G. major</i> (b)	<i>L. Lu et al.</i>	Xinzhu Village, Yulong,	138605	XXXXXXX	Fritsch et al., 2015a; field
	<i>LL-2014-64A</i>	Yunnan, China			observation
<i>G. marronina</i> P.W. Fritsch	<i>L. Lu & P.W. Fritsch</i>	Tianquan, Sichuan, China	161292	XXXXXXX	Fritsch et al., 2016; Field
& Lu Lu	<i>LL-2011-37</i>				observation
<i>G. minuta</i> Merr.	<i>L. Lu & P.W. Fritsch</i>	Yubeng Village, Deqin,	80318	XXXXXXX	Merrill, 1941; field
	<i>LL-2011-23</i>	Yunnan, China			observation; herbarium specimens
<i>G. nivea</i> (J. Anthony) Airy	<i>L. Lu et al. LL-2013-16</i>	Cikai Township,	158000	XXXXXXX	Fritsch et al., 2008; field
Shaw (a)		Gongshan, Yunnan, China			observation; herbarium specimens
<i>G. nivea</i> (b)	<i>L. Lu et al. LL-2013-56</i>	Yanmen Township, Deqin,	161202	XXXXXXX	Fritsch et al., 2008; field

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
		Yunnan, China			observation; herbarium specimens
<i>G. nummularioides</i> D. Don	<i>L. Lu et al. LL-2014-14</i>	Cikai Township, Gongshan, Yunnan, China	147187	XXXXXXX	Fritsch et al., 2008; herbarium specimens
<i>G. obovata</i> (Airy Shaw)	<i>L. Lu et al. LL-2013-49</i>	Yanmen Township, Deqin, Yunnan, China	103634	XXXXXXX	Fritsch et al., 2015a; field observation
P.W. Fritsch & Lu Lu					
<i>G. praticola</i> C.Y. Wu	<i>L. Lu et al. LL-2014-23</i>	Dulongjiang Township, Gongshan, Yunnan, China	136819	XXXXXXX	Fritsch et al., 2008; field observation; herbarium specimens
<i>G. sinensis</i> J. Anthony (a)	<i>N. Yang et al. NZ031</i>	Bomi, Tibet, China	92287	XXXXXXX	Field observation
<i>G. sinensis</i> (b)	<i>L. Lu et al. LL-2014-28</i>	Yaduo village, Fugong, Yunnan, China	79502	XXXXXXX	Fritsch et al., 2015a; field observation

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
<i>G. stenophylla</i> P.W. Fritsch & Lu Lu (a)	<i>L. Lu et al. LL-2013-7</i>	Dulongjiang township, Gongshan, Yunnan, China	160321	XXXXXXX	Fritsch et al., 2015a; field observation
<i>G. stenophylla</i> (b)	<i>L. Lu et al. LL-2013-9</i>	Dulongjiang township, Gongshan, Yunnan, China	160939	XXXXXXX	Fritsch et al., 2015a; field observation
<i>G. stenophylla</i> (c)	<i>L. Lu et al. LL-07122</i>	Duoxiongla, Mêdog, Tibet, China	82534	XXXXXXX	Field observation
<i>G. thymifolia</i> Stapf ex Airy Shaw	No voucher	Cultivated, RSF	78821	XXXXXXX	Fritsch et al., 2008; field observation
<i>G. trichophylla</i> var. <i>trichophylla</i> Royle	<i>L. Lu et al. LL-2013-47</i>	Bangdang township, Gongshan, Yunnan, China	161139	XXXXXXX	Fritsch et al., 2008; field observation; herbarium specimens
<i>G. trichophylla</i> var.	<i>L. Lu & P.W. Fritsch</i>	Luding, Sichuan, China	62505	XXXXXXX	Field observation;

Taxon	Collection number	Locality	Assembled length (bp)	GenBank accession number	Sources for morphological data
<i>tetracme</i> Airy Shaw	<i>LL-2011-31</i>				herbarium specimen measurement
/	<i>L. Lu et al. LL-2013-54</i>	Yanmen Township, Deqin, Yunnan , China	163526	XXXXXX	Field observation

Note: Samples in bold represent those with fresh leaves available for DNA extraction.

RSF = Rhododendron Species Foundation, USA.

/ = indicates a presumed hybrid.

Table 2

The 10 morphological characters and their states analyzed in this study. For detailed description and illustration of micro-features (character 2, 3, and 4), see Fritsch et al. (2015a).

Characters	Character states
1 Habit	0: prostrate (ascending, creeping); 1: erect (ascending-erect)
2 Presence of puberulence on lamina midvein adaxial-basally	0: present; 1: absent
3 presence of setose trichomes on the abaxial surface of the leaf blade	0: present; 1: absent
4 Calyx lobe trichome presence	0: eciliolate; 1: ciliolate
5 Corolla color	0: white; 1: pink; 2: red
6 Corolla shape	0: campanulate; 1: urceolate; 2: globose-urceolate
7 Anther theca awn number	0: 1; 1: 2
8 Fruiting calyx openness	0: open; 1: closed
9 Fruiting calyx color	0: white; 1: blue (including light blue, deep blue and bluish purple); 2: black

Characters	Character states
10 Capsule color	0: green; 1: maroon

Table 3

Coding matrix of morphological characters for the species of *Gaultheria* series

Trichophyllae and three outgroup species. Characters ordered as in Table 2.

	1	2	3	4	5	6	7	8	9	10
<i>Gaultheria albiflora</i>	0	1	1	0	0	0	01	0	1	0
<i>G. bryoides</i>	0	0	1	0	0	0	0	1	0	1
<i>G. cardiosepala</i>	1	0	1	0	01	1	1	1	0	0
<i>G. ciliisepala</i>	0	0	0	1	01	0	1	0	1	0
<i>G. crassifolia</i>	1	0	01	0	0	1	01	1	1	0
<i>G. dolichopoda</i>	1	0	01	0	1	1	1	1	1	0
<i>G. eciliata</i>	0	0	01	1	0	0	0	0	1	0
<i>G. gonggashanensis</i>	0	1	01	0	0	0	1	0	1	0
<i>G. hookeri</i>	1	0	0	?	01	1	?	1	1	0
<i>G. hypochlora</i>	1	0	0	1	01	0	01	0	01	0
<i>G. jingdongensis</i>	1	1	0	1	0	0	0	?	?	?
<i>G. major</i>	0	0	01	0	0	0	1	0	1	0
<i>G. marronina</i>	0	0	1	0	0	0	0	0	0	1
<i>G. minuta</i>	0	1	1	0	1	?	0	0	1	0
<i>G. nivea</i>	1	0	1	0	0	0	1	1	0	0
<i>G. nummularioides</i>	0	0	0	1	01	0	1	1	2	0
<i>G. obovata</i>	0	0	01	0	0	0	1	0	1	0
<i>G. praticola</i>	1	0	0	1	012	02	1	1	2	1

	1	2	3	4	5	6	7	8	9	10
<i>G. sinensis</i>	0	0	01	0	01	0	1	0	1	0
<i>G. stenophylla</i>	1	0	01	1	0	0	01	0	01	0
<i>G. thymifolia</i>	1	0	1	0	0	0	1	1	0	0
<i>G. trichophylla</i> var. <i>trichophylla</i>	1	0	01	0	012	0	01	1	1	0
<i>G. trichophylla</i> var. <i>tetracme</i>	1	1	1	0	1	0	1	1	1	0
<i>L. Lu et al. LL-2013-54</i>	1	0	01	0	?	?	01	1	1	0

Table 4

Sequence variation and tree statistics for the different datasets used in phylogenetic analyses.

		Total sequences	Coding regions	Noncoding regions	LSC region	IR region	Partitioned dataset
	Aligned length (bp)	169160	76105	93055	109261	28626	140532
	Variable sites (%)	3808 (2.25)	850 (1.12)	2958 (3.18)	3015 (2.76)	348 (1.22)	3456 (2.46)
	Parsimony-informative sites (%)	1778 (1.05)	409 (0.54)	1369 (1.47)	1371 (1.25)	179 (0.63)	1589 (1.13)
ML	Tree length	0.035018	0.016851	0.050583	0.043773	0.018582	0.038972
	Optimization likelihood	-268952.832245	-113885.439747	-153723.380594	-177202.958932	-43254.966047	-222944.761373
MP	Tree length	4770	1071	3761	3658	474	/
	Consistency index (CI)	0.827	0.815	0.837	0.855	0.755	/
	Retention index (RI)	0.841	0.831	0.851	0.868	0.783	/

		Total sequences	Coding regions	Noncoding regions	LSC region	IR region	Partitioned dataset
BI	Model	GTR + I + G	GTR + I + G	GTR + I + G	GTR + I + G	GTR + I + G	GTR + G & GTR + I + G
	AIC	536655.5849	227324.0204	307038.1257	353978.6974	86276.9673	/
	BIC	/	/	/	/	/	447223.4859
	-lnL	268257.7925	113592.0102	153449.0629	176919.3487	43068.4837	222183.4335
	Proportion of invariable sites	0.883	0.945	0.818	0.838	0.941	/
	Gamma distribution	0.769	0.748	0.793	0.972	0.383	/

Table 5

Percentage of sequence variation within different plastid regions of species in *Gaultheria* series *Trichophyllae*, including sequences < 200 bp.

		Highest (%)	Average (%)
Protein coding genes	Total	6.1	0.89
	LSC	3.5	0.85
	IRa	6.1	1.0
Noncoding sequences	Total	8.5	2.2
	LSC	8.5	2.5
	IRa	3.9	1.1
	Intergenic regions	8.5	2.3
	Introns	4.2	1.4

Numbers along branches: BS ML/BS MP/PP BI

Asterisks: BS ML = 100, BS MP = 100, PP BI = 1