Genome Survey Indicated Complex Evolutionary History of Garuga Roxb. Species

Dongbo Zhu
Southwest Forestry University
Rui Rao
Southwest Forestry University
Yu Du
Technology Center of Kunming Customs
Chunmin Mao
Southwest Forestry University
Rong Chen
Southwest Forestry University
Sun Hang
Kunming Institute of Botony, Chinese Academy of Science
Liangliang Yue
yueliangliang@swfu.edu.cn

Yunnan Key Laboratory of Plateau Wetland Conservation, Restoration and Ecological Services, Southwest Forestry University

Research Article

Keywords: Genome survey, K-mer, Flow cytometry, SNP, Phylogenetic, Garuga forrestii, Nuclear-cytoplasmic conflict

Posted Date: February 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3905007/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomics on October 23rd, 2024. See the published version at https://doi.org/10.1186/s12864-024-10917-8.

Abstract

Background

Garuga Roxb. is a genus endemic to southwest China and other tropical regions in Southeast Asia facing risk of extinction due to the loss of tropical forests and changes in land use. Conducting a genome survey of *G. forrestii* contribute to a deeper understanding and conservation of the genus.

Results

This study utilized genome survey of *G. forrestii* generated approximately 54.56 GB of sequence data, with approximately 112 × coverage. K-mer analysis indicated a genome size of approximately 0.48 GB, smaller than 0.52GB estimated by flow cytometry. The heterozygosity is of about 0.54%, and a repeat rate of around 51.54%. All the shotgun data were assembled into 339,729 scaffolds, with an N50 of 17,344 bp. The average content of guanine and cytosine was approximately 35.16%. A total of 330,999 SSRs were detected, with mononucleotide repeats being the most abundant at 70.16%, followed by dinucleotide repeats at 20.40%. A pseudo chromosome of *G. forrestii* and a gemone of *Boswellia sacra* were used as reference genome to perform a primer population resequencing analysis within three *Garuga* species. PCA indicated three distinct groups, but genome wide phylogenetics represented conflicting both between the dataset of different reference genomes and between maternal and nuclear genome.

Conclusion

In summary, the genome of *G. forrestii* is small, and the phylogenetic relationships within the *Garuga* genus are complex. The genetic data presented in this study holds significant value for comprehensive whole-genome analyses, the evaluation of population genetic diversity, investigations into adaptive evolution, the advancement of artificial breeding efforts, and the support of species conservation and restoration initiatives. Ultimately, this research contributes to reinforcing the conservation and management of natural ecosystems, promoting biodiversity conservation, and advancing sustainable development.

1 Background

Garuga Roxb. is an endemic genus to southwest China and other tropical regions in Southeast Asia. These deciduous trees bloom between March and April, prior to the onset of the rainy season, produce fruits maturing from May to November, with the main maturation phase occurring in July and August. There are approximately five species or varieties within this genus, including *G. forrestii* W. W. Smith, *G. floribunda* var. *gamblei* (King ex Smith) Kalkm., *G. floribunda* var. *floribunda* Decne., *G. pierrei* Guill., and *G. pinnata* Roxb.. Four of these species are found in China, except *G. floribunda* var. *floribunda*. *G. forrestii* is an endemic species of China.

Garuga forrestii, is a Chinese endemic species of this genus, distributed in the arid and warm river valleys of Yunnan, S.W. China, such as Jinshajiang River (JR), Lancangjiang River (LR), Red River (RR), Nanpanjiang

River (NPR) and their tributaries. It is also the only species distributed beyond to the Tropic of Cancer, in the JR, extra north to the tropical regions. It extends to Leibo, Sichuan, making it the highest latitude distribution within the Burseraceae family. It distributed upper to the altitudes nearing 1600 meters at many points in the JR. It demonstrates an extraordinary ability to flourish under arid conditions. *G. floribunda* var. *gamblei* and *G. pinnata* grow in dense tropical forests of South Yunnan, further to Guangxi, Guangdong and Hainan. Our survey along the middle LR unveiled the coexistence of three species: *G. forrestii*, *G. floribunda* var. *gamblei* and *G. pinnata*. But individuals of this genus are all in danger of extinction due to the loss of tropical forests and changes of land uses.

The taxonomy of *Garuga* has experienced substantial reorganization, resulting in alterations to their taxonomic classifications. *G. pierrei* previously considered a variety of *G. pinnata*, exhibits leaf characteristics similar to *G. pinnata*. However, distinctions arise in the short and soft hairs present on the axis and leaflets of *G. pierrei*, contrasting with the shorter hairs found on the latter. *G. pierrei* bears spherical fruits, *G. pinnata*'s fruits are nearly spherical and occasionally exhibit soft hairs. Nonetheless, typical *G. pierrei* specimens eluded discovery during our fieldwork. The fruiting attributes of *G. floribunda* var. *gamblei* share similarities with those of *G. forrestii*, yet notable differences manifest in terms of hair coverage, pedicel, fruit dimensions and shape. Within the Lancang-Mekong River region, these three species coexist, displaying a continuum of variations in tree structure and leaf morphology. Consequently, the current taxonomic classification needs further tests by phylogenetic and phylogenomic evidence.

The Sino-Himalaya region is in the southeast margin of the Qinghai-Tibetan Plateau (QTP) hosting several Asian rivers, including the Mekong, RRB and JRB, along with their tributaries. Presently, these rivers flow separately to the ocean [1, 2]. However, they once flowed southward into the Paleo-Red River (PRR), creating an extensive drainage network that ultimately emptied into the South China Sea millions of years before [3, 4]. The PRR system disassembled subsequently due to uplift of QTP and river capture events since the late Miocene [4]. This reorganization disrupted the previously continuous distribution pattern, resulting in distinctive genetic and biogeographical attributes, facilitating species' genetic differentiation giving rise to new taxa [5, 6]. The river captures also made separate ranges unique promoting lineage fusion and diversity accumulation [7].

Genome sequencing technology enables the thorough investigation of species evolution [8]. Next-generation sequencing, NGS, technique has greatly reduced the cost of DNA sequencing [9], and was widely used in most fields of genetics and genomics, providing researchers with a more convenient, refined, and comprehensive investigation method [10]. A genome survey is the most convenient strategy to provide a rough reference genome for those species without whole genome data [11–13].

This study employed genome survey to explore the genome size and characteristics of *G. forrestii*. The gathered genomic information from this research will serve to 1) learn deeper in our comprehension of the *G. forrestii* genome; 2) strive to construct a preliminary long contig or even a draft scaffold, offering a foundation for future evolutionary surveys of this genus.

2 Materials and methods

2.1 Sample collection and DNA extractions

Fresh leaf from a mature *G. forrestii* individual were collected for genome survey during the summer of 2023 near to the Gangou Bridge in the Red River valley, Yunnan, China. Fresh leaves from other 25 individuals of three species (*G. floribunda* var. *gamblei*, *G. pinnata*, *G. forrestii*) in the *Garuga* genus. Among these, 6 individuals of *G. floribunda* var. *gamblei*, 8 of *G. pinnata*, and 11 of *G. forrestii* (Fig. 1; Table 1). No specific permissions were required for the collection of specimens for this study which were neither privately owned nor protected and the field study did not involve endangered or protected species. We complied with the IUCN Policy Statement on Research Involving Species at Risk of Extinction and the Convention on the Trade in Endangered Species of Wild Fauna and Flora. After the formal identification of the plant material was carried out by Liangliang Yue, voucher specimens were prepared and deposited at the herbarium of College of Wetlands, Southwest Forestry University (Yue, accession number Yue2023051032-1).

A total of 26 samples were sequenced. The genomic DNA was isolated using the CTAB method [14]. DNA was detected by 0.8% agarose gel electrophoresis, while DNA was quantified by UV spectrophotometer.

Table 1
Sampling Information

Species	Number of Individuals	Population	Province	Small Scale	Longitude	Latitude
G. pinnata	3	Pi_BB	Yunnan	Bubeng	101.585	21.60944
	3	Pi_YXG	Yunnan	Yexianggu	100.8648	22.18184
	1	Pi_SHZWY	Yunnan	Shanghaizhiwuyuan	121.4429	31.14649
	1	Pi_DTS	Yunnan	Datianshan	99.87337	24.7832
G. floribunda var. gamblei	2	Fl_SMC	Yunnan	Sanmaicun	100.6193	22.00044
	2	Fl_YXG	Yunnan	Yexianggu	100.8648	22.18184
	1	Fl_ML	Yunnan	Menglong	100.7445	21.7359
	1	Fl_DTBC	Yunnan	Datianbacun	99.88069	24.73952
G. forrestii	1	Fo_XP	Yunnan	Xinping	101.5044	24.14715
	2	Fo_DC	Yunnan	Dacun	100.4178	25.01894
	1	Fo_LJZ	Yunnan	Luzhijiang	101.9574	24.6514
	2	Fo_HN	Yunnan	Huaning	102.9734	24.0748
	2	Fo_KYC	Yunnan	Kaiyuancheng	103.2985	23.89117
	1	Fo_NXH	Yunnan	Nanxihe	101.8507	23.64461
	1	Fo_YJZ	Sichuan	Yinjiangzhen	101.7854	26.59825
	1	Fo_FF	Sichuan	Fenfang	101.9786	26.76728

2.2 Genome size estimation by flow cytometry

For the experimental material, we utilized tender leaves from tomato plants (genome size: 900 Mb) that were one month old following seed germination. The samples were carefully placed within 0.8 mL of pre-chilled MG^b dissociation solution, composed of 45 mM $MgCl_2 \cdot 6H_2O$, 20 mM MOPS, 30 mM sodium citrate, 1% (W/V) PVP 40, 0.2% (V/V) Tritonx-100, 10 mM Na_2EDTA, 20 μL/mL β-mercaptoethanol, and adjusted to pH 7.5. The tissues were swiftly sectioned vertically using a sharp blade and allowed to rest in the dissociation solution on ice for a duration of 10 minutes. Following this, the mixture was filtered through a 40-micron mesh to obtain a nuclear cell suspension. This suspension was then combined with a suitable volume of pre-chilled propidium iodide (PI) solution, having a stock concentration of 1 mg/mL, along with a fitting amount of RNAase solution at a stock concentration of 1 mg/mL. The combined mixture was subsequently subjected to a dark ice-cold staining process for a period of 0.5-1 hour. The effective concentration of both the PI staining solution and the RNAase solution was maintained at 50 μg/mL [15, 16]. The stained suspension of nuclear cells underwent detection using a BD FACScalibur flow cytometer. This involved utilizing a 488 nm blue light excitation to measure the emitted fluorescence intensity of propidium iodide. Each detection cycle involved the collection of 10,000 particles. The coefficient of variation (CV%) was maintained at a level below 5%. Modifit3.0 software was employed to conduct graphing and analysis. The genome size was calculated using the subsequent formula [17]:

$$Samplegenomesize = standardgenomesize \times \frac{sampleG_0/G_1peakmean}{standardG_0/G_1peakmean}$$

2.3 Genome Sequencing and Data Quality Control

Dried leaves from the mature tree of *G. forrestii* was sent to Personalbio company for paired-end sequencing. We followed the standard protocol using Illumina's TruSeq DNA PCR-free prep kit reagents for sequencing library preparation. Extracted DNA sample underwent random shearing through ultrasonication, followed by end repair to eliminate overhanging bases at the 5' end and to add a phosphate group while filling in missing bases at the 3' end. To prevent self-ligation of DNA fragments and ensure compatibility with sequencing adaptors, an A base was added to the 3' end of the DNA sequence. Sequencing adaptors with library-specific tags were ligated to the 5' end of the DNA sequence, facilitating the immobilization of DNA molecules onto the Flow Cell. We used AMPureXP beads (Beckman Coulter, Brea, CA) for a selective removal of adaptor-ligated fragments and purification of the resulting library system. Subsequently, PCR amplification was performed on DNA fragments ligated with adaptors to enrich the sequencing library templates. A second purification step with BECKMAN AMPure XP Beads was carried out to purify the enriched library products. Finally, we conducted 2% agarose gel electrophoresis to select and purify the library's final fragments. The resulting library insert fragments were approximately 400 bp in size. We performed paired-end sequencing with 2 × 150 bp reads using an Illumina NovaSeq instrument (Table 2). Furthermore, we also subjected the remaining 25 samples to sequencing processing, including chloroplast genome sequencing.

Table 2
Sequencing Overview

Sample	Insert Size	Sequencing platform	Sequencing Mode
G. forrestii	400 bp	Illumina Novaseq	Paired-end, 2×150bp

2.4 High-quality data acquisition

Raw sequence was filtered using fastp [18] and using the sliding window method to generate high quality sequence (high quality data). The size of the slide window is set to 5 bp. Slide the window from the 3' end to the 5' end, calculate the average Q value of the bases in the window, if the Q value is < 20, delete the bases in the window; if the Q value is ≥ 20, stop sliding. Length filtering, if the length of any one reads in the bipartite end ≤ 50 bp, then remove the bipartite end reads. fuzzy base N filtering, if the number of N bases in the bipartite end ≥ 5, then remove the bipartite end reads.

2.5 Ploidy analysis and k-mer analysis

Smudgeplot software [19] was used to analyze the genome structure and to count the number of heterozygous k-mer pairs by comparing the total number of k-mer pair coverage (CovA + CovB) and relative coverage (CovB / (CovA + CovB)) and to obtain the genomic ploidy. The distribution of k-mer was calculated using jellyfish software [20], which gives information about the heterozygosity of *G. forrestii*, the proportion of repetitive sequences, and then genome size estimation based on the distribution of k-mer frequencies and the number of 19-mer.

2.6 De novo assembly and GC content analysis

The pair-end DNA sequencing data was de novo assembled with standard parameters using the MEGAHIT [21]. Contigs were further spliced into scaffold levels using SOAPdenovo [22] software. Sliding window calculations were performed using a window size of 10 kb. The average depth and GC content of each window were computed. The quality of genome assembly was assessed using the Assembly-stats utility, which calculated metrics such as N50, scaffold count, scaffold size, scaffold length, and genome length.

Before chloroplast assembly, low-quality sequences were filtered using SOAPnuke [23], and de novo assembly was performed following the pipeline of GetOrganelle [24].

2.7 Identification of microsatellite motifs

Microsatellite Identification Software (MISA) [25] was utilized to detect microsatellite patterns within the scaffolds generated above. The minimum number of repeats required for identifying mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide repeat sequences is set as follows: mononucleotide repeats with less than 10 repetitions, dinucleotide repeats with less than 6 repetitions, and all other repeat types with less than 5 repetitions [26].

2.8 Population resequencing variation detection and filtering of SNP data

The Picard [27], BWA [28], and Samtools [29] software is used to build the genome index and perform processes such as mapping, sorting, and deduplication, all aimed at subsequent variation detection. We used *G. forrestii* (we sampled, sequenced, and assembled the genome to ensure its usability as a reference genome.) and *Boswellia sacra* Flück. (downloaded from NCBI, accession: SNVD00000000) as reference genomes for mapping the 25 individuals. The GATK [30] software was employed to perform variant calling using HaplotypeCaller for each sample. This study involved the assembly of the genome of an individual of *G. forrestii* at the scaffold level, as well as the genomes of the remaining 25 individuals, which included 6 *G. floribunda* var. *gamblei*, 8 *G. pinnata*, and 11 *G. forrestii*. Each genome of the mentioned individuals was divided into 10 pseudo-chromosomes. The GATK software was applied to filter and extract SNPs and INDELs.

2.9 Phylogenetic reconstruction and PCA

We applied Plink [31] software to perform LD (Linkage Disequilibrium) filtering on the SNP data, resulting in a refined VCF file to generate an alignment for phylogenomic reconstruction and PCA analysis. The dataset encompassed nucleotide polymorphisms from the whole genomes of 25 individuals of the three species. We employed iqtree [32] to construct two Maximum Likelihood (ML) trees for the nuclear genome respectively using *B. sacra* Flück. and *Garuga forrestii* as reference genomes. Phylogenomic trees were visualized using Figtree [33]. The PCA based on the SNP data extracted from the 25 samples was conducted using Plink [31] and GCTA [34] software, followed by visualization of the results using an R script. We also constructed ML trees using iqtree [32] for the chloroplast genomes of 25 individuals.

3 Results

3.1 Genome size estimation by flow cytometry

The utilization of flow cytometry analysis produced a detailed high-resolution histogram (Fig. 2; Table 3), revealing a fluorescence intensity of 17.04 for the sample and 29.67 for the internal reference. The ratio between these values is 0.57, leading to an estimated genome size of approximately 0.52 GB.

Table 3
Flow Cytometry Information

Sample	Reference Selection	Reference Fluorescence Intensity	Fluorescence Intensity of the Test Sample	Ratio	Genome Size (Gb)
G. forrestii	tomato	29.67	17.04	0.57	0.52

3.2 Sequence filtering and data quality control

Insertion of paired-end libraries with a size of 400 bp (Table 2) yielded approximately ~ 54.56 GB of raw sequence (Table 4). The percentage of bases with quality score ≥ Q20 was 97.46% and ≥ Q30 was 92.86% (Table 4). Approximately 1.1% of the low-quality data was filtered, and the remaining ~ 53.97 GB was used for downstream analysis (Table 5).

Table 4
Sequencing Data Statistics

Sample	Read Num	Total base	N_rate	GC_Content (%)	Q20_rate	Q30_rate
G. forrestii	361299934	54556290034	0	35.1	97.46	92.86

Table 5
High-Quality Data Statistics

Sample	HQ Reads	HQ Reads (%)	HQ Data (bp)	HQ Data (%)
G. forrestii	358261400	99.16	53967212246	98.92

The GC content is 35.10% (Table 4). The base distribution showed no separation of AT and GC either at the R1 end or the R2 end (Fig. 3), which laid the foundation for the subsequent quantitative analysis. The average sequencing error rates for single-base loci were all less than 0.1% (Fig. 4)

3.3 K-mer analysis and ploidy estimation

K-mer analysis provides an estimate of the genome size based on the substrings of length k contained in the biological sequence. The data volume was multiplied to achieve a coverage of 112x. Besides, this it also indicates low quality or contamination in the sequences. 19-mer frequency analysis identified the 90X depth as the dominant peak based on the k-mer count (Fig. 5). After dividing the k-mer number by 90, the genome size was predicted to be ~ 483 Mb (483,508,243 bp). The other peak at 1/2 the depth of the main peak (45X) is most likely due to heterozygosity (~ 0.54%), While the peak at twice the depth of the main peak (180X) is caused by the repetitive sequence (~ 51.54%).

The results of ploidy assessment yielded 83% AB, 12% AAB, and 5% AABB (Fig. 6), from which *G. forrestii* was predicted to be diploid, and this ploidy assessment is for reference only.

3.4 De-novo assembly and gene prediction

A total of 360 million paired-end reads were utilized to establish the initial genome assembly of *G. forrestii*. The genome, assembled using SOAPdenovo software, comprises 339,729 scaffolds (509.9 Mb) with scaffolds exceeding 200 bp chosen to exclude low-quality sequences. The genome size, excluding Ns, accounts for 499,986,768 bp, and the final assembly exhibits a low value of Ns content of roughly 0.01%. The largest scaffold spans 345,990 bp, with N50 = 17,344 bp, N90 = 293 bp (Table 6).

Table 6
Statistics of Contigs and Scaffolds

Assembly Level	Number	Sum (bp)	Longest Seq	N50	N90	N count	Gaps	Software
contig	514298	516660543	179938	6972	280			MEGAHIT
scaffold	339729	509934743	345990	17344	293	6439439	146457	SOAPdenovo

3.5 Analysis of GC content

The resultant scaffolds longer than 200 bases in length were chosen. A window size of 10 kb was used for non-repetitive advancement in the sequence and calculation of the mean depth and GC content of every window to generate a GC depth plot. Most windows displayed GC content ranging from 25–50%, resulting in an estimated GC ratio of ~ 35.16% (Fig. 7). We did not identify any significant regions with abnormal accumulation. This suggests that the DNA samples used for sequencing were not mixed with DNA from other species.

3.6 Identification of SSRs

The assembled scaffolds were searched for the presence of SSR markers by using the MISA software. It yielded 330,999 SSR motifs (Table 7). Among these, mononucleotides were the largest in number (232,225; 70.16%), dinucleotides were the second in number (67,518; 20.40%), which was followed by trinucleotide (23,506; 7.10%), tetranucleotide (5,103; 1.54%), pentanucleotide (1,472; 0.44%), and hexanucleotide (1,175; 0.35%) SSR markers (Fig. 8a).

Table 7
SSR Information

Total number of sequences examined:	**339729**
Total size of examined sequences (bp):	506426207
Total number of identified SSRs:	330999
Number of SSR containing sequences:	6914
Number of sequences containing more than 1 SSR:	34647
Number of SSRs present in compound formation:	51666

Within mononucleotide repeat sequences, the highest content is represented by A/T (230,708; 99.35%) (Fig. 8b). Among dinucleotide repeat sequences, the predominant composition is AT/AT (48,604; 71.99%), followed by AG/CT (12,133; 17.97%), AC/GT (6,723; 9.96%), and CG/CG (58; 0.09%) (Fig. 8c). Regarding trinucleotide repeat sequences, the principal repetitive motifs include AAT/ATT, AAG/CTT, ATC/ATG, and AAC/GTT, accounting for contents of 16,427 (69.88%), 3,005 (12.78%), 1,337 (5.69%), and 1,062 (4.52%), respectively (Fig. 8d).

3.7 Phylogenetic reconstruction and PCA

Both two nuclear genomic reconstruction suggested that the *G. forrestii* located at the basal position (Fig. 9ab). However, plastid phylogenomic reconstruction indicated conflict structure. *G. floribunda* var. *gamblei* and *G. pinnata* are clustered at basal position, (Fig. 9c). The Maximum Likelihood (ML) reconstruction of the nuclear genome with *G. forrestii* as the reference genome, reveals that *G. forrestii* constitutes a monophyletic lineage in the Clade-A. *G. pinata* and *G. floribunda* var. *gamblei* individuals clustered into mixed clades. In the Clade-B, 2 individuals of *G. floribunda* var. *gamblei* (from populations of FL_DTBC and FL_YXG) were found mixed with 6 individuals of *G. pinata* (from Pi_YXG, Pi_SHZWY and Pi_DTS). In the Clade-C, 2 individuals of *G. pinata* (from of Pi_YXG) were mixed with 4 individuals of *G. floribunda* var. *gamblei* (from Fl_YXG, Fl_ML and Fl_SMC) (Fig. 9a). Phylogenomic reconstruction based on dataset using *B. sacra* as the reference genome revealed similar patterns mentioned above, with differences lie in the internal topology of

the branches (Fig. 9ab). But *G. forrestii* does not form a monophyletic lineage (Clade-F, Clade-G and Clade-H), mixing free with the other two species (Fig. 9b). The plastid ML tree shows that *G. forrestii* and *G. pinata* are mixed with each other in the Clade-D ((Fig. 9c), *G. pinata* and *G. floribunda* var. *gamblei* are mixed in Clade-E (Fig. 9c).

25 individuals clustered into three distinct groups in the PCA (Fig. 10). In the first group, two individuals of *G. floribunda* var. *gamblei* (from Fl_DTBC and Fl_YXG) were found mixed within the *G. pinata* group, whereas within the predominantly *G. floribunda* var. *gamblei* group, two individuals of *G. pinata* (from Pi_YXG) were also mixed.

4 Discussion

4.1 Genome Characterization

We observed variations in genome size when measuring it using both k-mer analysis and flow cytometry for the same individual. In this study, for the first time, we conducted genome survey sequencing on *G. forrestii* and obtained ~ 53.97 GB of clean data. The 19 K-mer analyses showed that the *G. forrestii* genome was approximately 483.51 Mb, slightly smaller than that by flow cytometry (~ 520.00 Mb) (Fig. 2, 5). We observed similar variations in other published studies and statistically analyzed the differences among these 15 species (Table 8). The genomic size differences obtained by these two methods ranged from 16.00 Mb to 138.09 Mb, with *G. forrestii* showing a difference of 37 Mb, falling within the range of differences we calculated and being lower than the average difference of 82.93 Mb. Different principles of the two methods determined the differences. Flow cytometry is a technique based on staining undamaged nuclei with a fluorescent dye that adheres quantitatively to the DNA to calculate the amount of DNA. Processes are varying in different laboratories in sample preparation, staining/staining strategies. Random drift of the instruments may lead to significant deviation in genome size estimation [16, 35]. K-mer approach emerged as a computational technique to generate a k-mer frequency distribution (similar to a Poisson distribution) by plotting the coverage distribution of all k-mers in a sequence, where the peak of the distribution would be centered on the average sequencing depth of the genome. The genome size would be better inferred by directly sequencing reads and analyzing the frequency of k-mers [36].

Table 8
Genome Size Difference between Flow Cytometry and K-mer

Species	Genome Size of K-mer	Genome Size of Flow Cytometry	Absolute Value (Low Flow Cytometry - K-mer)	Absolute Value (High Flow Cytometry - K-mer)	Average	Article
Carex cristatella	317.3	255.2 ± 13.1	75.2	49	62.1	Genome Survey Sequencing for the Characterization of the Genetic Background of Rosa roxburghii Tratt and Leaf Ascorbate Metabolism Genes
Carex scoparia	294.5	268.8 ± 9.7	35.4	16	25.7	
Juncus effusus	225.5	198.9 ± 4.6	31.2	22	26.6	
Juncus inflexus	196	286.4 ± 4.7	85.7	95.1	90.4	
Raddia distichophylla	608	589	19		19	The draft genome sequence of herbaceous diploid bamboo Raddia distichophylla
Psammochloa villosa	1564	1503.27 ± 3.41	64.14	57.32	60.73	Estimation of genome size for Psammochloa villosa by flow cytometry and K-mer analysis
Reseda lutea	934	867.7	66.3		66.3	Estimation of Genome Size in the Endemic Species Reseda pentagyna and the Locally Rare Species Reseda lutea Using comparative Analyses of Flow Cytometry and K-Mer Approaches
R. pentagyna	1022	896	126		126	
Aspalathus linearis	1070	1.24 ± 0.01	180	160	170	Rooibos (Aspalathus linearis) Genome Size Estimation Using Flow Cytometry and K-Mer Analyses
Acer henryi	561.72	691.12 ± 8.69	120.71	138.09	129.4	Estimation of genome sizes of six Acer species by flow cytometry and K-mer analysis
A. buergerianum	743	863.90 ± 8.69	112.21	129.59	120.9	
A. elegantulum	777.87	896.50 ± 4.35	114.28	122.98	118.63	

Species	Genome Size of K-mer	Genome Size of Flow Cytometry	Absolute Value (Low Flow Cytometry - K-mer)	Absolute Value (High Flow Cytometry - K-mer)	Average	Article
A. griseum	771.51	893.24 ± 8.69	113.04	130.42	121.73	
A. pentaphyllum	650.64	766.10 ± 8.69	106.77	124.15	115.46	
A. tegmentosum	1103.46	1 154.04 ± 13.04	37.9	63.98	50.94	

GC content of *G. forrestii* is ~ 35.16% (Fig. 7), based on most of the plant genetic data summarized by previous authors, it was found that most of the GC content ranged from 30–47% [37]. This value can serve as an indicator of genome stability, as DNA sequences with higher GC content tend to exhibit tolerance to extreme temperatures and drought [38, 39]. We assessed the genomic GC content of three species at shallow sequencing depths, assembled to the contig level. It was observed that *G. forrestii* has the lowest GC content (~ 33.60%), *G. floribunda* var. *gamblei* has the highest (~ 33.94%), while *G. pinnata* falls in between (~ 33.81%) (Table 9). This GC content may be due to the species' adaptation to environmental stress [40], and it may be similar. As a report found that the highest GC content was found in gramineous plants [41]. During the Tertiary period's global cooling, Poaceae plants underwent differentiation and adaptation to this stress, consequently establishing their dominance in thriving under today's extreme climatic conditions [42–44]. Two species of *G. floribunda* var. *gamblei* and *G. pinnata* did suffer more interspecific competition from other higher tree species from rainforests and from longer, hotter and drier situations in South Yunnan area.

Table 9
Three Species GC Content

Species	Individual	GC Content (%)
G. pinnata	Pi_BB	33.7
	Pi_BB	33.67
	Pi_BB	33.87
	Pi_YXG	33.59
	Pi_YXG	33.63
	Pi_YXG	34.29
	Pi_SHZWY	33.97
	Pi_DTS	33.5
Average	33.7775	
G. floribunda var. gamblei	Fl_SMC	34.52
	Fl_SMC	34.51
	Fl_YXG	33.63
	Fl_ML	33.56
	Fl_DTBC	33.64
Average	33.972	
G. forrestii	Fo_XP	33.59
	Fo_DC	33.69
	Fo_DC	33.59
	Fo_LJZ	33.62
	Fo_HN	33.81
	Fo_HN	33.6
	Fo_KYC	33.91
	Fo_KYC	33.7
	Fo_NXH	33.7
	Fo_YJZ	33.52
	Fo_FF	33.57
Average	33.66363636	

The individual Fl_YXG of *G. floribunda* var. *gamblei* was not included in the calculation.

Plant genomes present the most challenging task for sequencing and assembly due to their high levels of heterozygosity, complex polyploidy, and abundant repeat content [45]. A genome with a heterozygosity exceeding 1% becomes highly challenging for de novo assembly [46]. However, the K-mer analysis in this study indicated a relatively low heterozygosity of approximately 0.54% in the *G. forrestii* genome (Fig. 6), making de novo assembly in this aspect more accurate. The repeat rate of the genome is 51.54%, This high repeat rate likely due to gene family expansion [47], gene transposition activity [48] and genome recombination [49]. Chromosome-level assembly for *G. forrestii* is needed to gain deeper insights into its genome and facilitate a comprehensive understanding of speciation and population dynamics of this species.

SSR has consistently been one of the most preferred molecular markers for plant genotyping due to its high level of polymorphism, wide distribution in the majority of plant genomes, user-friendly, and anticipated significance as a valuable tool for numerous species in the future [50–52]. Mononucleotide repeats are the predominant type, in accordance with previous reports [53]. The high abundance of AT/AT (48,604; 71.99%) motifs (Fig. 8c) in our study was consistent with some earlier genomic surveys. As an illustration, *Akebia trifoliata* exhibits an AT/AT content of 50.21% [54], while *Cunninghamia lanceolata* (Lamb.) Hook. showcases a content of 59% [55], and *Acer truncatum* Bunge boasts a significantly higher AT/AT content at 71.31% [56], confirming their representation as the most typical dinucleotide motifs in higher plants.

We investigated several species exited also in warm and hot valleys, *Terminalia franchetii* Gagnep. [57], *Osteomeles schwerinae* Schneid. [58], *Nouelia insignis* Franch. [59], *Buddleja crispa* Benth. [60] and *Excoecaria acerifolia* Didr. [61] for a comparison analysis. No raw reads of four species mentioned can be found in the GenBank for genome survey. Our data would provide a first view for evolutionary pattern of this area.

4.2 Nuclear-Cytoplasmic Conflict

In this study, the topography of the basal clades is different among phylogenomic trees. The *G. forrestii* referred tree indicates that *G. forrestii* is monophyletic, while the *B. sacra* referred tree suggests polyphyletic. For if we used a reference genome from other genus could have overestimated more ancient characters in the dataset. Consequently, a significantly higher number of SNP loci were generated through the alignment, three times more than those produced in the latter (Table 10). The disparity in SNP loci counts results in the observed differences in the topological structures of the two nuclear gene phylogenetic trees [62].

Table 10
SNP Calling Information

Reference	Number of Individual	Number of SNP Sites	VCF Size (Mb)
G. forrestii	26	244936	43
B. scara	26	79691	15.5

The topological structure distinctly shows a significant increase in the branch length of Fo_HN relative to other taxonomic units, designating it as the earliest diverging species. This could indicate a result of long-branch attraction, explaining the anomalous position in the phylogenetic tree [63, 64]. The basal clade of the

chloroplast genome phylogenetic tree is composed of *G. pinnata* and *G. floribunda* var. *gamblei*, showing a nuclear-cytoplasmic conflict with the two nuclear gene trees. Hybridization could properly explain [65, 66]. The rapid elevation of the QTP at the beginning of the Paleogene, and the formation and intensification of the monsoon climate zone brought abundant precipitation leading to river erosion, and that the rapid downcutting and uptake events led to changes in the spatial pattern of the PRR drainage system, which in turn affected the geographic distribution pattern of plants [1, 2, 67, 68]. Their sympatric distribution in river valleys may produce interspecific hybridization as otherwise disconnected river systems are linked together, making it easier for species to migrate and spread along river valleys. A previous report on the study of three river valley species of the genus *Ostryopsis* Decne indicates that the inconsistent interspecific relationship of *Ostryopsis intermedia* B. Tian & J. Q. Liu with the other two species (*Ostryopsis davidiana* Decaisne and *Ostryopsis nobilis* I. B. Balfour et W. W. Smith) recovered from two different sets of molecular markers strongly indicates its origin through hybrid speciation in the southeast Qinghai-Tibet Plateau through hybrid speciation [69]. To address this issue in the *Garuga* genus, we plan to conduct further Hi-C or Hi-Fi sequencing in the future, studying at the chromosomal level to acquire additional genetic information. By employing sophisticated models, we aim to consider this situation comprehensively and reconstruct the evolutionary history between species more accurately.

5 Conclusion

This study conducted a preliminary investigation into the genome size and characteristics of *G. forrestii* for the first time and reconstructed the phylogenetic relationships of three species within the *Garuga* genus, namely *G. floribunda* var. *gamblei*, *G. pinnata*, and *G. forrestii*. It provided valuable genomic resources for the further exploration and utilization of *G. forrestii* and offered initial insights into the phylogenetic relationships within the *Garuga* genus. However, there is further work to be done as the construction of two nuclear gene phylogenetic trees for these three species resulted in differences in the topological structure at the base clade (*G. forrestii*) with one being monophyletic and the other polyphyletic. Additionally, conflicts arose between the two nuclear gene phylogenetic trees and the chloroplast genome phylogenetic tree (where the base clade in the nuclear gene trees is *G. forrestii*, while in the chloroplast genome tree, it includes *G. floribunda* var. *gamblei*, *G. pinnata*, and *G. forrestii*). To clearly explain the differences in the topology of nuclear trees and the nuclear-cytoplasmic conflict with the chloroplast tree, we plan to supplement the study with Hi-C or Hi-Fi sequencing at the chromosomal level in future whole-genome sequencing. This will lead to better assembly results, and we will use complex models to study the aforementioned issues.

The genetic data presented in this study holds significant value for comprehensive whole-genome analyses, the evaluation of population genetic diversity, investigations into adaptive evolution, the advancement of artificial breeding efforts, and the support of species conservation and restoration initiatives. Ultimately, this research contributes to reinforcing the conservation and management of natural ecosystems, promoting biodiversity conservation, and advancing sustainable development.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

A raw sequence data of *Garuga floribunda* var. *gamblei* were deposited in NCBI under the BioProject ID: PRJNA783803. The other raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2022), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA014668, CRA014684, CRA014694, CRA014677) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number GWHERCJ00000000 that is publicly accessible at https://ngdc.cncb.ac.cn/gwh. The chloroplast assembly and annotation data for this study are deposited in NCBI with accession numbers PP337695-PP337718.

Competing interests

The authors declare that they have no competing interests

Funding

This research was funded by the National Natural Science Foundation of China (grant No. 42161015).

Authors' contribution

LLY and HS conceived the study. CMM collected plant materials and drew the figure. RC organized the plant materials. RR performed data analysis for SNP calling and generated plots. DBZ carried out genome assembly, data analysis, figure generation, and contributed to the paper writing. DBZ and RR contribute equally in this study. All authors read and approved the final manuscript.

Authors' information

(1) Authors and Affiliations

Southwest Forestry University, Kunming, PR, 650224, China

Dongbo Zhu (zdb1186070789@outlook.com)

Southwest Forestry University, Kunming, PR, 650224, China

Rui Rao (raorui0825@gmail.com)

Technology Center of Kunming Customs, Kunming, PR, 650228, China

Yu Du (dyciq1@163.com)

Southwest Forestry University, Kunming, PR, 650224, China

Chunmin Mao (mcm18213835817@126.com)

Southwest Forestry University, Kunming, PR, 650224, China

Rong Chen (cc18687745468@163.com)

Kunming Institute of Botony, Chinese Academy of Science, Kunming, PR, 650201, China

Hang Sun (hsun@mail.kib.ac.cn)

Yunnan Key Laboratory of Plateau Wetland Conservation, Restoration and Ecological Services, Southwest Forestry University, Kunming, PR, 650224, China. National Plateau Wetland Research Center, Kunming, PR, 650224, China

Liang-Liang Yue (yueliangliang@swfu.edu.cn)

(2) Contributions

LLY and HS conceived the study. CMM collected plant materials and drew the figure. RC organized the plant materials. RR performed data analysis for SNP calling and generated plots. DBZ carried out genome assembly, data analysis, figure generation, and contributed to the paper writing. DBZ and RR contribute equally in this study. All authors read and approved the final manuscript.

(3) **Corresponding authors**

Correspondence to Liangliang Yue and Hang Sun

(4) Acknowledgements

We appreciate the assistance of Yanchun Liu from the Shanghai Botanical Garden and Jinlong Dong from the Xishuangbanna Tropical Botanical Garden for their support in sample collection for this study.

References

1. Spicer RA, Farnsworth A, Su T: **Cenozoic topography, monsoons and biodiversity conservation within the Tibetan Region: An evolving story**. *Plant Diversity* 2020, **42**(4):229-254.

2. Nie J, Ruetenik G, Gallagher K, Hoke G, Garzione CN, Wang W, Stockli D, Hu X, Wang Z, Wang Y *et al*: **Rapid incision of the Mekong River in the middle Miocene linked to monsoonal precipitation**. *Nature Geoscience* 2018, **11**(12):944-948.
3. Ming Q, Shi Z, Zhang H: **The evolution of the landform and environment in the region of the three parallel rivers**. *Tropical geography* 2006, **26**(2):122.
4. Clark MK, Schoenbohm LM, Royden LH, Whipple KX, Burchfiel BC, Zhang X, Tang W, Wang E, Chen L: **Surface uplift, tectonics, and erosion of eastern Tibet from large-scale drainage patterns**. *Tectonics* 2004, **23**(1).
5. Brookfield ME: **The evolution of the great river systems of southern Asia during the Cenozoic India-Asia collision: rivers draining southwards**. *Geomorphology* 1998, **22**(3-4):285-312.
6. Sun H, Li Z, Landis JB, Qian L, Zhang T, Deng T: **Effects of drainage reorganization on phytogeographic pattern in Sino-Himalaya**. *Alpine Botany* 2021, **132**(1):141-151.
7. Sun H, Zhang J, Deng T, Boufford DE: **Origins and evolution of plant diversity in the Hengduan Mountains, China**. *Plant Diversity* 2017, **39**(4):161-166.
8. Yu T, Hu Y, Zhang Y, Zhao R, Yan X, Dayananda B, Wang J, Jiao Y, Li J, Yi X *et al*: **Whole-Genome Sequencing ofAcer catalpifoliumReveals Evolutionary History of Endangered Species**. *Genome Biology and Evolution* 2021, **13**(12).
9. Hert DG, Fredlake CP, Barron AE: **Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods**. *Electrophoresis* 2008, **29**(23):4618-4626.
10. Aird D, Ross MG, Chen W, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries**. *Genome Biology* 2011, **12**(2):1-14.
11. Bi Q, Zhao Y, Cui Y, Wang L: **Genome survey sequencing and genetic background characterization of yellow horn based on next-generation sequencing**. *Molecular Biology Reports* 2019, **46**(4):4303-4312.
12. Huang G, Cao J, Chen C, Wang M, Liu Z, Gao F, Yi M, Chen G, Lu M: **Genome Survey of Misgurnus Anguillicaudatus to Identify Genomic Information, Simple Sequence Repeat (SSR) Markers and Mitochondrial Genome**. *Reaserch Square* 2021.
13. Liang X, Bai T, Wang J, Jiang W: **Genome survey and development of 13 SSR markers in Eucalyptus cloeziana by NGS**. *Journal of Genetics* 2022, **101**(2).
14. Doyle JJ: **A rapid DNA isolation procedure for small quantities of fresh leaf tissue**. *Phytochem Bull* 1987, **19**:11-15.
15. Dolezel J: **Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size**. *Annals of Botany* 2005, **95**(1):99-110.
16. Doležel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry**. *Nature Protocols* 2007, **2**(9):2233-2244.
17. Xinming T, Xiangyan Z, Na G: **Applications of Flow Cytometry in Plant Research—Analysis of Nuclear DNA Content and Ploidy Level in Plant Cells**. *Chinese Agricultural Science Bulletin* 2011, **27**(9):21-27.

18. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.
19. Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nature Communications* 2020, **11**(1):1432.
20. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**. *Bioinformatics* 2011, **27**(6):764-770.
21. Li D, Liu C, Luo R, Sadakane K, Lam T: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**. *Bioinformatics* 2015, **31**(10):1674-1676.
22. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**(1):2047-2217X-2041-2018.
23. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z *et al*: **SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data**. *GigaScience* 2018, **7**(1):gix120.
24. Jin J, Yu W, Yang J, Song Y, dePamphilis CW, Yi T, Li D: **GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes**. *Genome Biology* 2020, **21**(1):241.
25. Thiel T, Michalek W, Varshney R, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.)**. *Theoretical and Applied Genetics* 2003, **106**(3):411-422.
26. Beier S, Thiel T, Münch T, Scholz U, Mascher M, Valencia A: **MISA-web: a web server for microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583-2585.
27. Wysokar A, Tibbetts K, McCown M, Homer N, Fennell T: **Picard: A set of tools for working with next generation sequencing data in BAM format**. *Retrieved Aug* 2014.
28. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.
31. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly M: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *The American journal of human genetics* 2007, **81**(3):559-575.
32. Nguyen L, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies**. *Molecular Biology and Evolution* 2015, **32**(1):268-274.
33. Rambaut A: **FigTree v1. 4.2, a graphical viewer of phylogenetic trees**. In.; 2014.
34. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: A Tool for Genome-wide Complex Trait Analysis**. *The American Journal of Human Genetics* 2011, **88**(1):76-82.

35. Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R: **Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison**. *Annals of Botany* 1998, **82**(suppl_1):17-26.
36. Li X, Waterman MS: **Estimating the repeat structure and length of DNA sequences using ℓ-tuples**. *Genome Res* 2003, **13**(8):1916-1922.
37. Pellegrini M, Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X, Fang J: **Evaluation of Genome Sequencing Quality in Selected Plant Species Using Expressed Sequence Tags**. *PLOS ONE* 2013, **8**(7):e69890.
38. Šmarda P, Knápek O, Březinová A, Horová L, Grulich V, Danihelka J, Veselý P, Šmerda J, Rotreklová O, Bureš P: **Genome sizes and genomic guanine+cytosine (GC) contents of the Czech vascular flora with new estimates for 1700 species**. *Preslia* 2019, **91**(2):117-142.
39. Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová O: **Ecological and evolutionary significance of genomic GC content diversity in monocots**. *Proceedings of the National Academy of Sciences* 2014, **111**(39):E4096-E4102.
40. Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB: **Three keys to the radiation of angiosperms into freezing environments**. *Nature* 2014, **506**(7486):89-92.
41. Singh R, Ming R, Yu Q: **Comparative Analysis of GC Content Variations in Plant Genomes**. *Tropical Plant Biology* 2016, **9**(3):136-149.
42. Strömberg CA: **Evolution of grasses and grassland ecosystems**. *Annual review of Earth and planetary sciences* 2011, **39**:517-544.
43. Edwards EJ, Osborne CP, Strömberg CA, Smith SA, Consortium CG, Bond WJ, Christin P-A, Cousins AB, Duvall MR, Fox DL: **The origins of C4 grasslands: integrating evolutionary and ecosystem science**. *science* 2010, **328**(5978):587-591.
44. Zachos J, Pagani M, Sloan L, Thomas E, Billups K: **Trends, rhythms, and aberrations in global climate 65 Ma to present**. *science* 2001, **292**(5517):686-693.
45. Michael TP, VanBuren R: **Building near-complete plant genomes**. *Current Opinion in Plant Biology* 2020, **54**:26-33.
46. Zhou P, Zhang Q, Li J, Li F, Huang J, Zhang M: **A first insight into the genomic background of Ilex pubescens (Aquifoliaceae) by flow cytometry and genome survey sequencing**. *BMC Genomics* 2023, **24**(1).
47. Han Y, Luthe D: **Identification and evolution analysis of the JAZ gene family in maize**. *BMC Genomics* 2021, **22**(1).
48. Zhao D, Ferguson AA, Jiang N: **What makes up plant genomes: The vanishing line between transposable elements and genes**. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2016, **1859**(2):366-380.
49. Hufton AL, Panopoulou G: **Polyploidy and genome restructuring: a variety of outcomes**. *Current Opinion in Genetics & Development* 2009, **19**(6):600-606.
50. Hayden MJ, Nguyen TM, Waterman A, McMichael GL, Chalmers KJ: **Application of multiplex-ready PCR for fluorescence-based SSR genotyping in barley and wheat**. *Molecular Breeding* 2007, **21**(3):271-281.

51. Gramazio P, Plesa IM, Truta AM, Sestras AF, Vilanova S, Plazas M, Vicente O, Boscaiu M, Prohens J, Sestras RE: **Highly informative SSR genotyping reveals large genetic diversity and limited differentiation in European larch (Larixdecidua) populations from Romania**. *Turkish Journal of Agriculture and Forestry* 2018, **42**(3):165-175.
52. Liu XB, Feng B, Li J, Yan C, Yang ZL: **Genetic diversity and breeding history of Winter Mushroom (Flammulina velutipes) in China uncovered by genomic SSR markers**. *Gene* 2016, **591**(1):227-235.
53. Manee MM, Al-Shomrani BM, Al-Fageeh MB: **Genome-wide characterization of simple sequence repeats in Palmae genomes**. *Genes & Genomics* 2020, **42**(5):597-608.
54. Zhang Z, Zhang J, Yang Q, Li B, Zhou W, Wang Z: **Genome survey sequencing and genetic diversity of cultivated Akebia trifoliata assessed via phenotypes and SSR markers**. *Molecular Biology Reports* 2021, **48**(1):241-250.
55. Lin E, Zhuang H, Yu J, Liu X, Huang H, Zhu M, Tong Z: **Genome survey of Chinese fir (Cunninghamia lanceolata): Identification of genomic SSRs and demonstration of their utility in genetic diversity analysis**. *Scientific Reports* 2020, **10**(1).
56. Wang R, Fan J, Chang P, Zhu L, Zhao M, Li L: **Genome Survey Sequencing of Acer truncatum Bunge to Identify Genomic Information, Simple Sequence Repeat (SSR) Markers and Complete Chloroplast Genome**. *Forests* 2019, **10**(2).
57. Zhang T, Comes HP, Sun H: **Chloroplast phylogeography of Terminalia franchetii (Combretaceae) from the eastern Sino-Himalayan region and its correlation with historical river capture events**. *Molecular Phylogenetics and Evolution* 2011, **60**(1):1-12.
58. Gomory D, Wang Z, Chen S, Nie Z, Zhang J, Zhou Z, Deng T, Sun H: **Climatic Factors Drive Population Divergence and Demography: Insights Based on the Phylogeography of a Riparian Plant Species Endemic to the Hengduan Mountains and Adjacent Regions**. *PLOS ONE* 2015, **10**(12).
59. Zhao Y, Gong X: **Genetic divergence and phylogeographic history of two closely related species (Leucomeris decora and Nouelia insignis) across the 'Tanaka Line' in Southwest China**. *BMC Evolutionary Biology* 2015, **15**(1):134.
60. Yue L, Chen G, Sun W, Sun H: **Phylogeography of Buddleja crispa (Buddlejaceae) and its correlation with drainage system evolution in southwestern China**. *American Journal of Botany* 2012, **99**(10):1726-1735.
61. Wang Z, Zhang T, Luo D, Sun W, Sun H: **Phylogeography of Excoecaria acerifolia (Euphorbiaceae) suggests combined effects of historical drainage reorganization events and climatic changes on riparian plants in the Sino–Himalayan region**. *Botanical Journal of the Linnean Society* 2019, **192**(2):350-368.
62. Leaché AD, Banbury BL, Felsenstein J, De Oca AN-M, Stamatakis A: **Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies**. *Systematic biology* 2015, **64**(6):1032-1047.
63. Philippe H: **Opinion: long branch attraction and protist phylogeny**. *Protist* 2000, **151**(4):307-316.
64. Bergsten J: **A review of long-branch attraction**. *Cladistics* 2005, **21**(2):163-193.
65. Degnan JH, Rosenberg NA: **Gene tree discordance, phylogenetic inference and the multispecies coalescent**. *Trends in ecology & evolution* 2009, **24**(6):332-340.

66. Pelser PB, Kennedy AH, Tepe EJ, Shidler JB, Nordenstam B, Kadereit JW, Watson LE: **Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies**. *American Journal of Botany* 2010, **97**(5):856-873.
67. Spicer RA: **Tibet, the Himalaya, Asian monsoons and biodiversity – In what ways are they related?** *Plant Diversity* 2017, **39**(5):233-244.
68. Tada R, Zheng H, Clift PD: **Evolution and variability of the Asian monsoon and its potential linkage with uplift of the Himalaya and Tibetan Plateau**. *Progress in Earth and Planetary Science* 2016, **3**(1):4.
69. Lu Z, Tian B, Liu B, YANG C, Liu J: **Origin of Ostryopsis intermedia (Betulaceae) in the southeast Qinghai–Tibet Plateau through hybrid speciation**. *Journal of Systematics and Evolution* 2014, **52**(3):250-259.

Figures

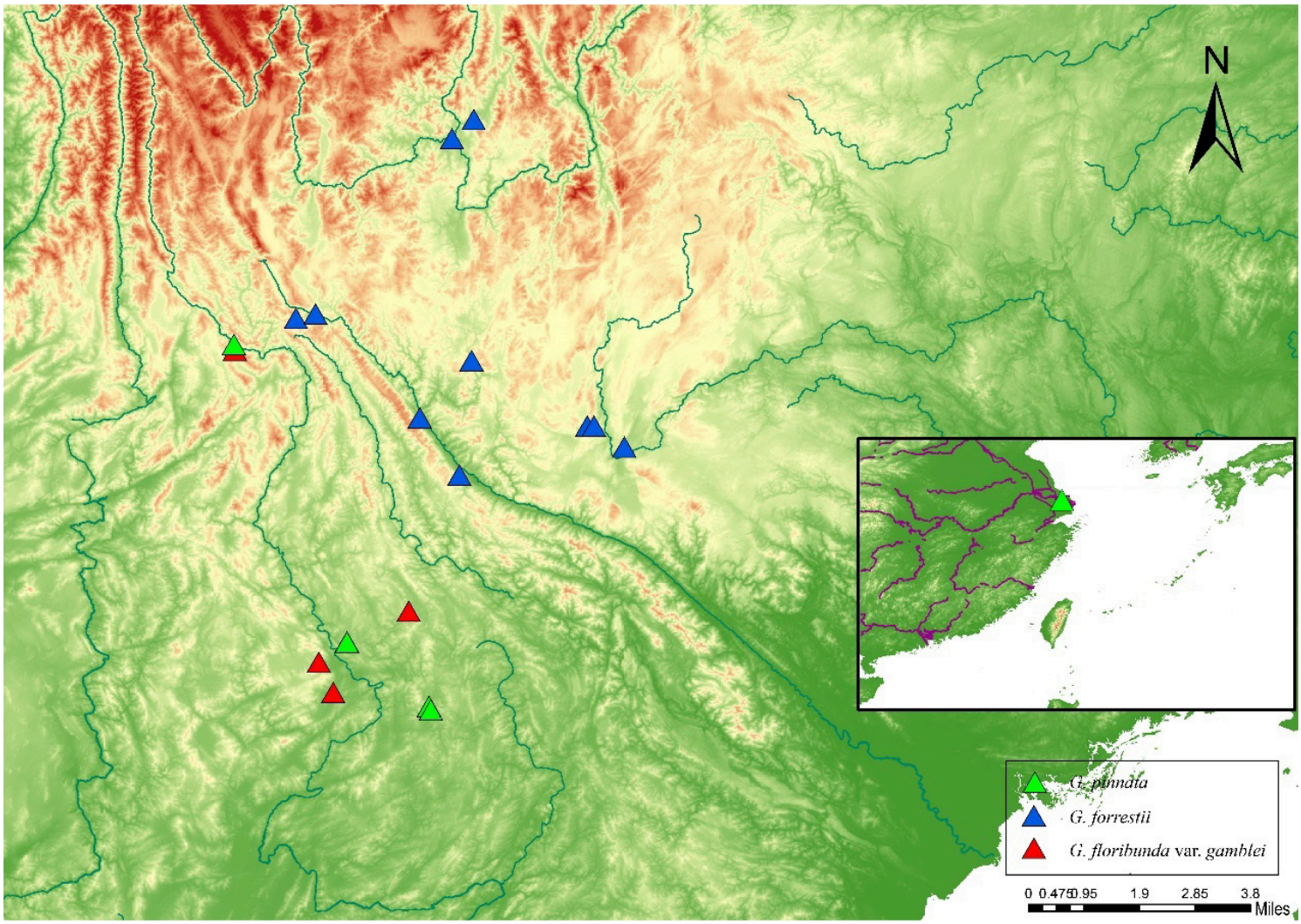

Figure 1

Sample location. Green triangles represent *G. pinnata*, blue triangles represent *G. forrestii*, and red triangles represent *G. floribunda* var. *gamblei*. One sample of *G. pinnata* was collected from Chenshan Botanical

Garden, Shanghai.

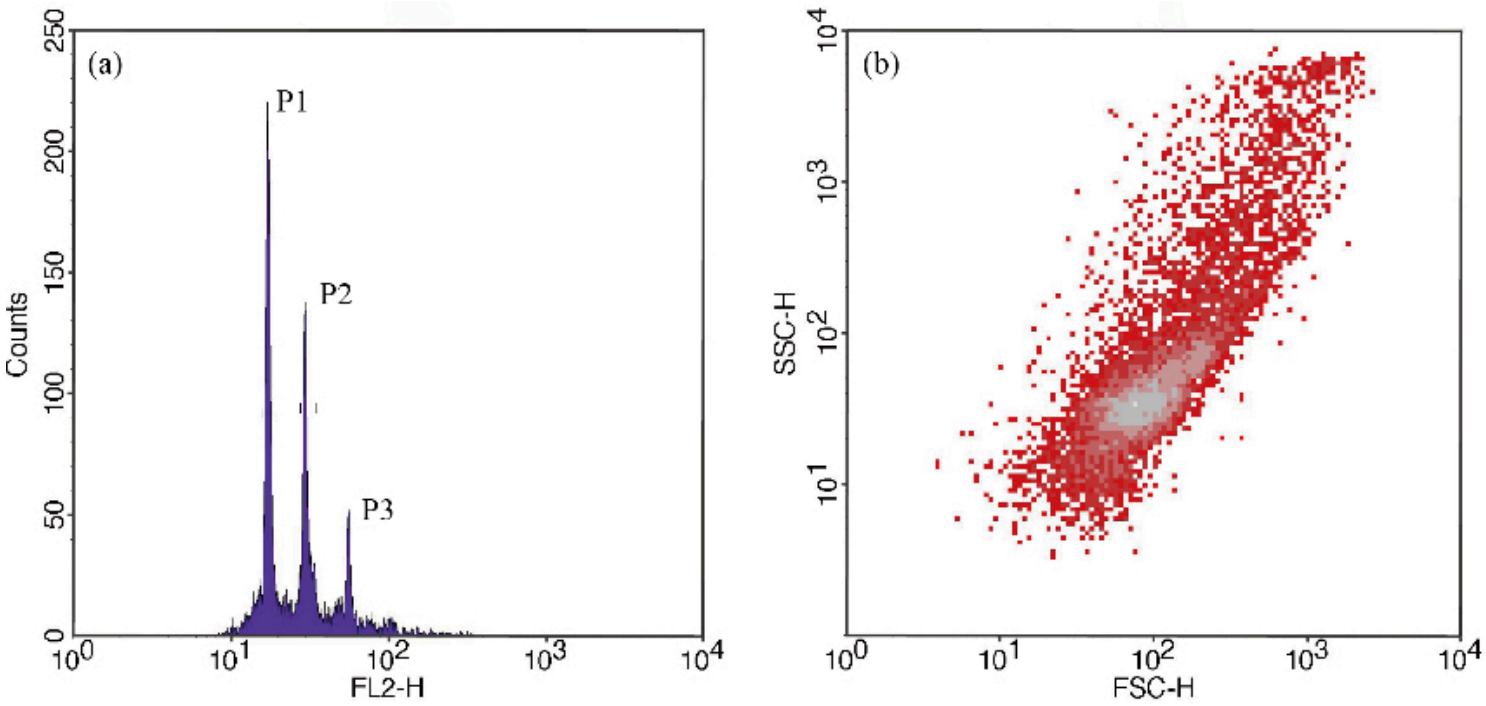

Figure 2

The flow cytometry (FCM) analysis results for the *G. forrestii*sample are as follows: (a) Represents the histogram of relative fluorescence intensity. Peak 1 corresponds to the G1 phase, Peak 2 corresponds to the S phase, and Peak 3 corresponds to the G2/M phase; (b) Represents the scatter plot of side scatter (SSC) versus PI fluorescence. The particle clusters exhibit a pronounced, lucid, and concentrated appearance, underscoring the effective differentiation of samples under the specific experimental conditions.

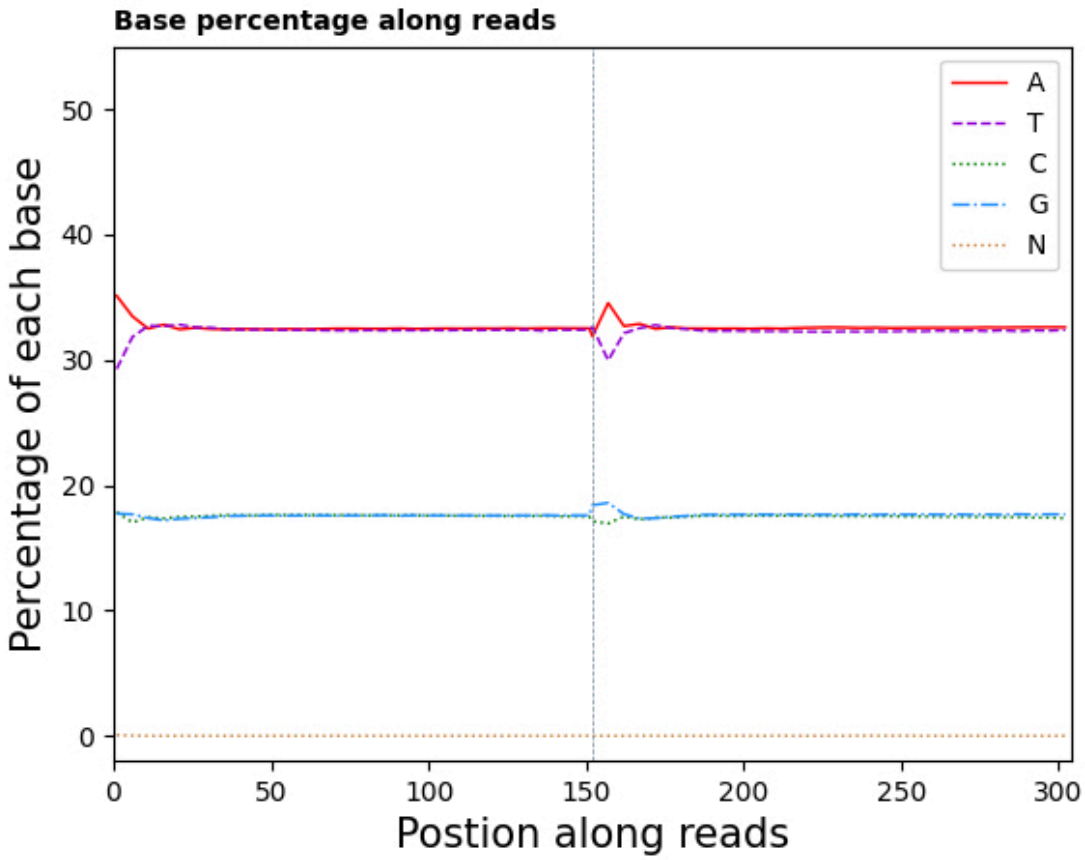

Figure 3

Reads sequencing base content distribution. The x-axis represents the position of bases within the reads, and the y-axis indicates the average percentage of base content at that position. Different colored lines represent different types of bases. The dashed line on the left shows the base content distribution for the R1 end of paired-end sequencing reads, while the dashed line on the right shows the base content distribution for R2 end sequencing reads.

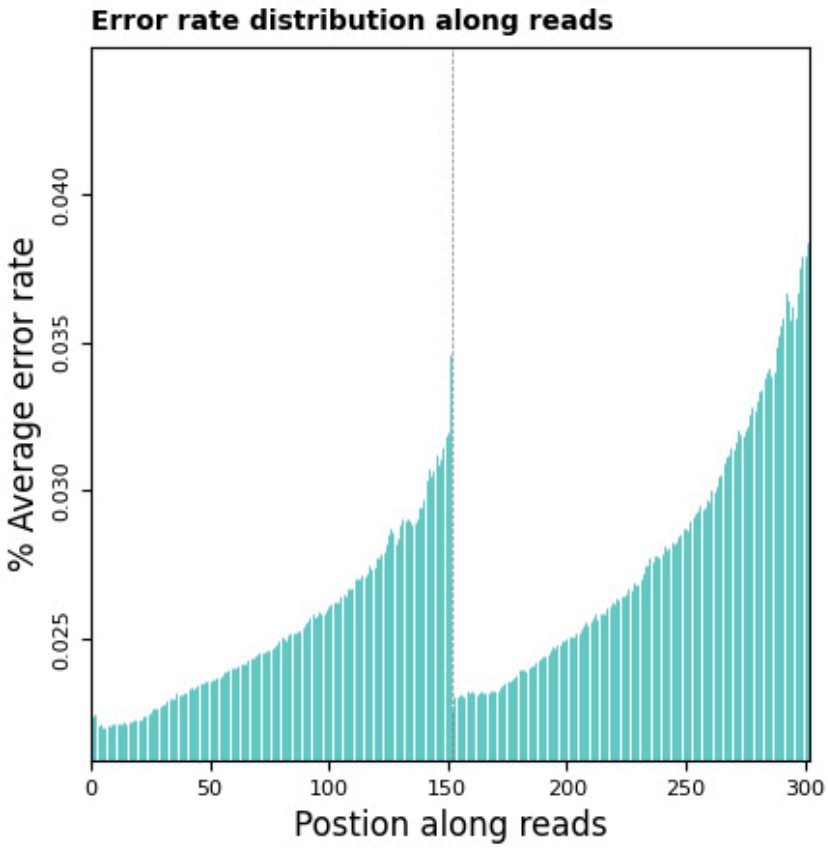

Figure 4

Average error rate distribution of reads. The x-axis represents the position of bases within the reads, and the y-axis indicates the average percentage of error rate at that position. The dashed line on the left shows the error rate distribution for the R1 end of paired-end sequencing reads, while the dashed line on the right shows the error rate distribution for R2 end sequencing reads.

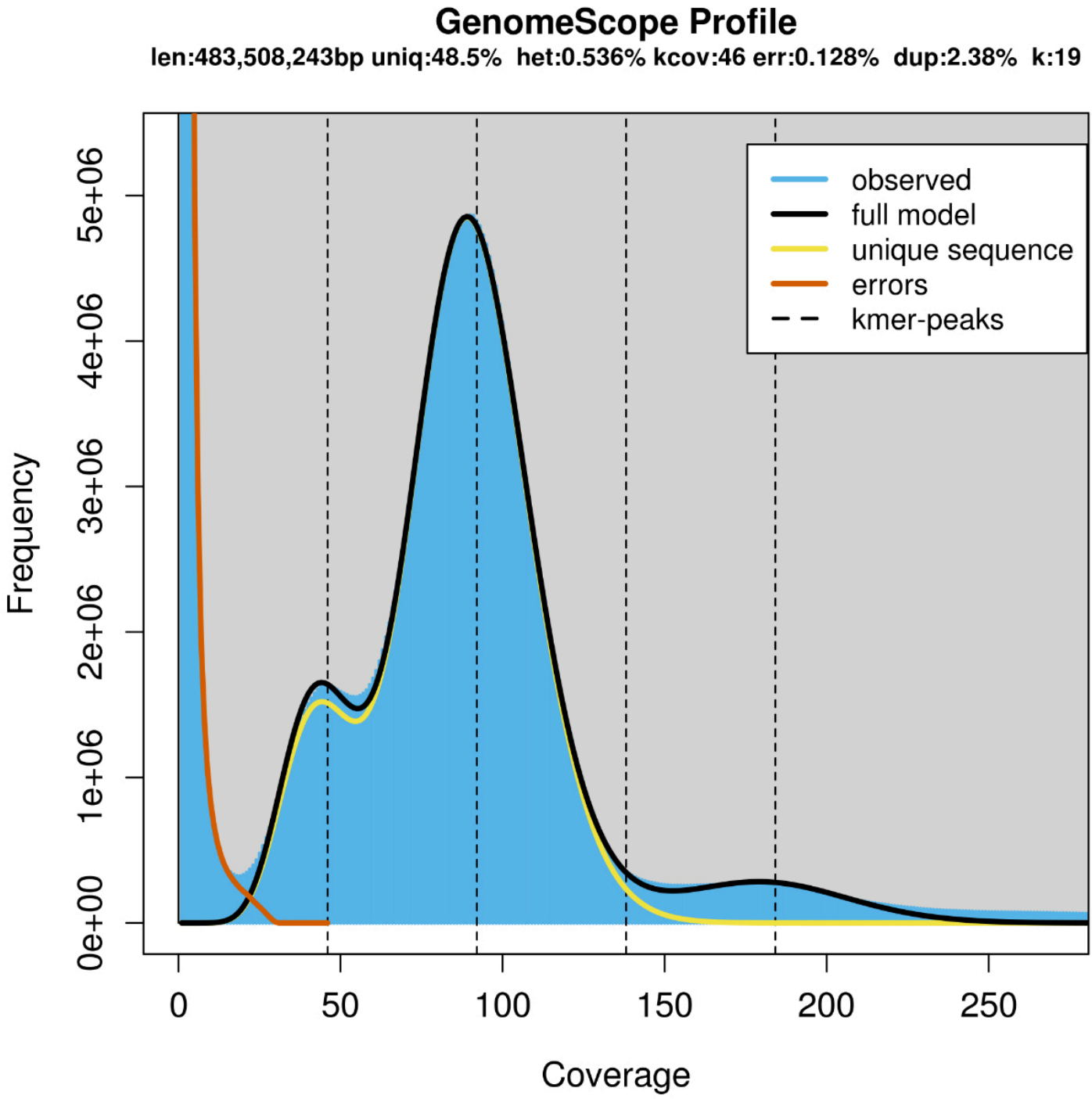

Figure 5

K-mer analysis. The x-axis represents the K-mer depth, and the y-axis indicates the number of K-mer types at that depth.

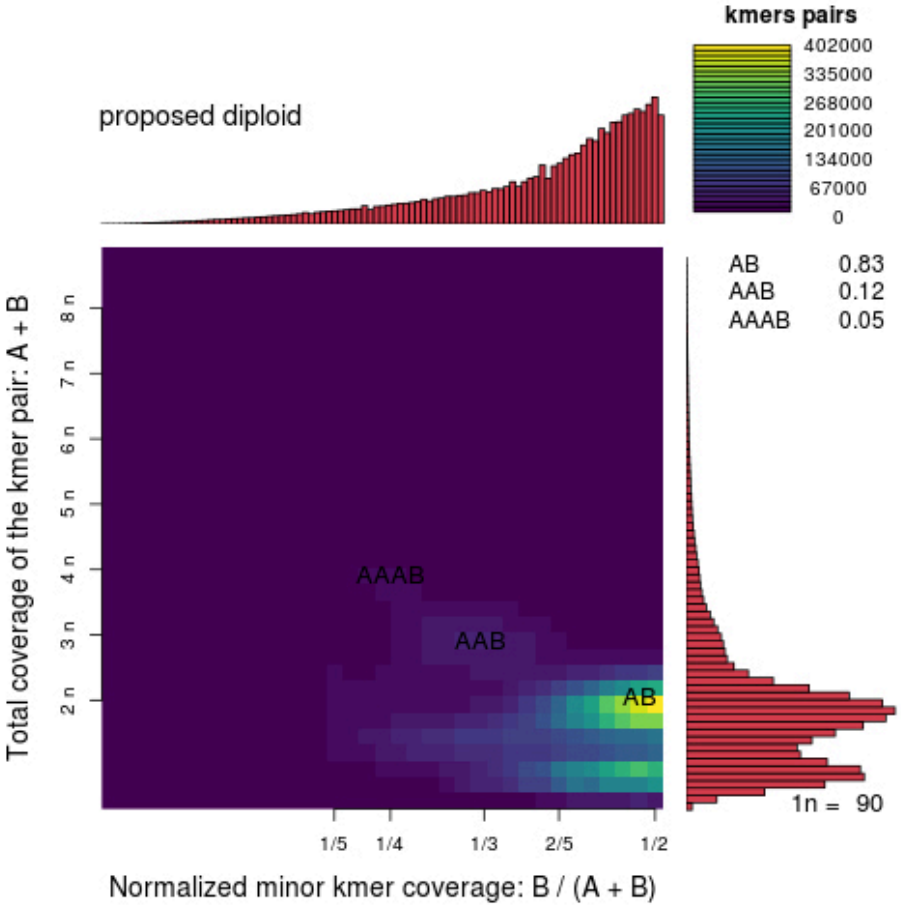

Figure 6

Polyploidy analysis of *G. forrestii*. The x-axis represents the relative coverage depth, the y-axis represents the total coverage depth, and the color intensity indicates the frequency of k-mer pairs. The single ploidy structure with the highest frequency is considered as the predicted species ploidy result.

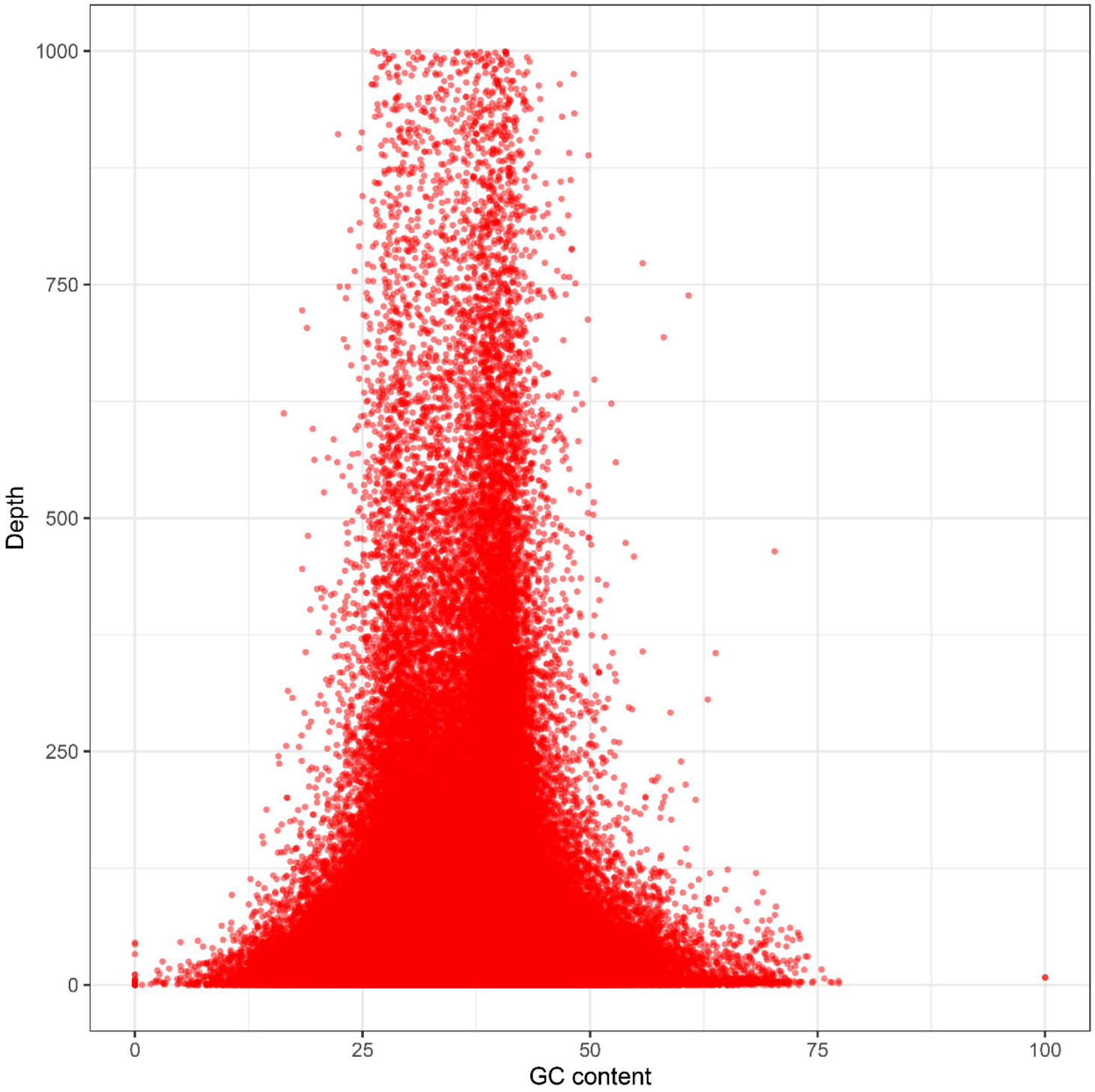

Figure 7

Analysis of the correlation between Guanine-Cytosine (GC) content and sequencing depth of the scaffold of the *G. forrestii* genome is obtained after assembly. The x-axis represents GC content, and the y-axis represents sequence depth. Estimation of GC% was performed utilizing a sliding window of 10 kb size with a 5 kb step.

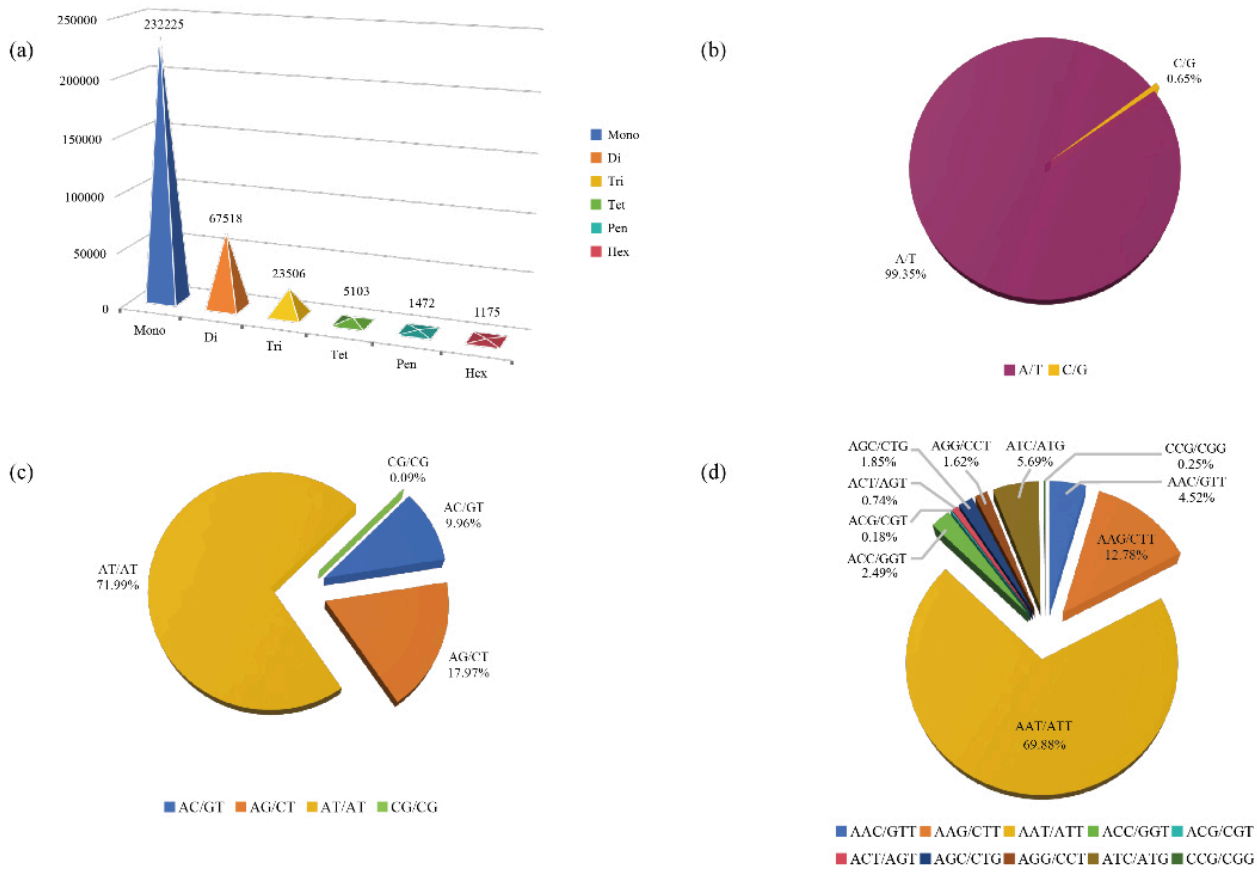

Figure 8

Characteristics of microsatellite motifs of *G. forrestii*after assembly. (a) Frequency of different microsatellite patterns; (b) Frequency of different mononucleotide patterns; (c) Frequency of different dinucleotide patterns; (d) Frequency of different trinucleotide patterns.

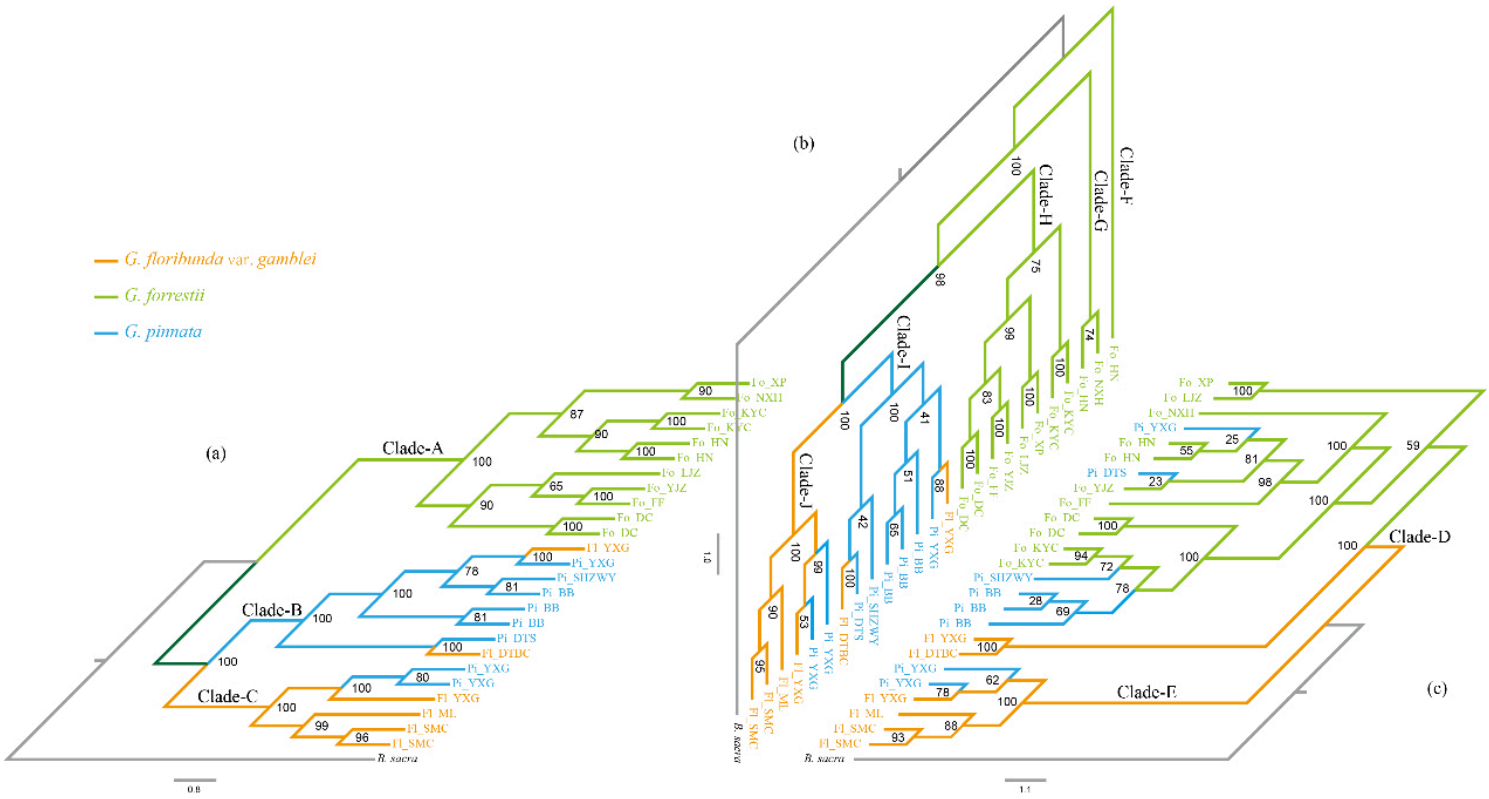

Figure 9

The phylogenetic trees of the nuclear-cytoplasmic gene systems for three species with a total of 25 individuals. (a) Nuclear gene phylogenetic tree constructed based on *Garuga forrestii* as the reference genome; (b) Nuclear gene phylogenetic tree constructed based on *Boswellia scara* as the reference genome; (c) Chloroplast gene phylogenetic tree. Orange lines represent *Garuga floribunda* var. *gamblei*, green lines represent *Garuga forrestii*, and blue lines represent *Garuga pinnata*. The text at the tips of the branches represents the species and population. In (a) and (c), *Garuga forrestii* clustered at the evolved clades, while in (b) clustered at the basal position.

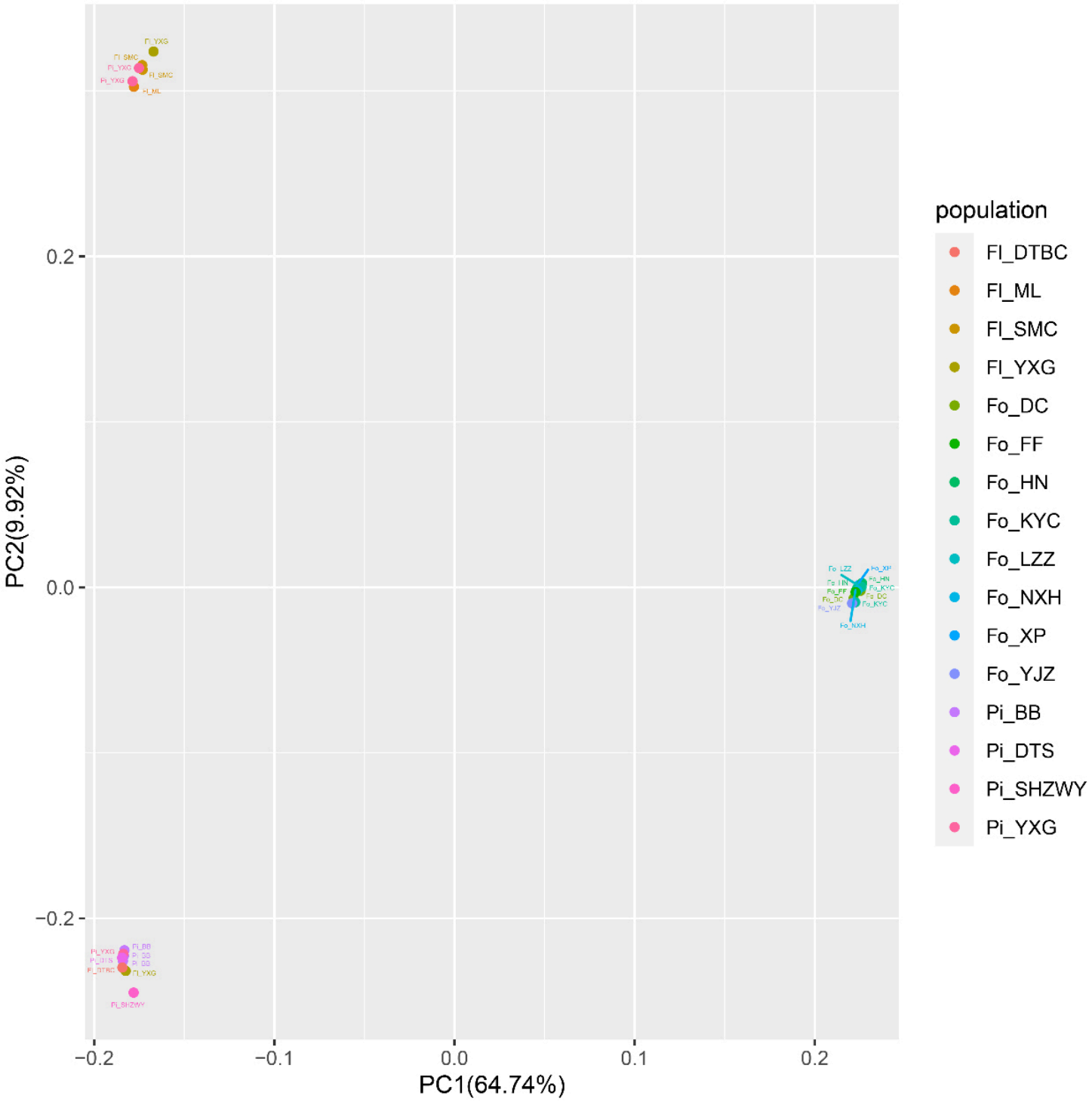

Figure 10

Principal Component Analysis (PCA). Based on the SNP similarity of 25 individuals from 16 populations of the three species using *G. forrestii*as the reference genome. All the individuals clustered into three distinct groups.