 doi:10.20944/preprints202110.0246.v1

6Article

The complete chloroplast genome sequence of four plant species, their SSR identification and phylogenetic analysis

Yueyi Zhu [2,†], Xianwen Zhang [3,†], Guopeng Li [4,†], Jiqian Xiang [1,5], Jinghua Su [6], Liwen Wu [1], Muhammad Khan Daud [7], Lei Mei [1,2,5*]

[1] Enshi Tujia & Miao Autonomous Prefecture Academy of Agricultural Sciences, Enshi 445000, China; 2416861638@qq.com (J. X.); 1907789166@qq.com (L. W.)
[2] Institution of Crop Science, Zhejiang University, Hangzhou 310058, China; 21916018@zju.edu.cn (Y.Z.);
[3] Institute of Virology and Biotechnology, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China; bestzxw@zju.edu.cn (X. Z.)
[4] College of Coastal Agricultural Science, Guangdong Ocean University, Zhanjiang 524088, China; gdszlgp@163.com (G. L.)
[5] Hubei Selenium Industrial Technology Research Institute, Enshi 445000, China
[6] Novogene Bioinformatics Institute, Beijing 100083, China; ywbiotech@yeah.net (J. S.)
[7] Department of Biotechnology and Genetic Engineering, Kohat University of Science and Technology, Kohat 26000, Pakistan; mkdaud@kust.edu.pk (M. K. D)
[†] Co-first authors: These authors contributed this research equally
[*] Correspondence: Lei Mei (meileihzruk@gmail.com or leimei@zuaa.zju.edu.cn)

Abstract: The chloroplast genome is conservative and stable, which can be employed to resolve genotypes. Currently, published nuclear sequences and molecular markers failed to differentiate the species from taxa robustly, including *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi*. In this study, the four chloroplast genomes were characterized, and then their simple sequence repeats (SSRs) and phylogenetic positions were analyzed. The results demonstrated the four chloroplast genomes consisted of 152.624 kb, 153.296kb, 156.309 kb, and 158.953 kb in length, involving 124, 130, 129, and 131 genes, respectively. Moreover, the chloroplast genomes contained typical four regions. Six classes of SSR were identified from the four chloroplast genomes, in which mononucleotide was the class with the most members. The types of the repeats were various within individual classes of SSR. Phylogenetic trees indicated that *M. leptophylla* was clustered with *M. yunnanensis*, and *H. exserta* was confirmed under family *Ocimeae*. Additionally, *R. bambusarum* and *R. henryi* were clustered together, whereas they did not belong to the same species due to the differing SSR features. This research would provide evidence for resolving the species and contributed new genetic information for further study.

Keywords: Chloroplast genome; *Machilus leptophylla*; *Hanceola exserta*; *Rubus bambusarum*; *Rubus henryi*; Simple sequence repeat; Phylogenetic analysis

1. Introduction

Machilus leptophylla is an evergreen broad-leaved tree in the family *Lauraceae* mainly distributed in most districts of China. Zhejiang, Jiangxi, Hunan, Fujian, and other regions of China [1]. Because of its fast growth, beautiful appearance, and high-quality wood, *M. leptophylla* has attracted more and more attention from commercial markets and related scholars. The genus *Machilus* includes nearly 100 species distributed in tropical and subtropical East and South Asia[2]. The reported nuclear sequences and genomic markers failed to resolve species in the genus [3]. To date, nine species in genus *Hanceola* were distributed in south China and identified out, based on the morphological features [4]. However, unlike most species of *Hanceola* are perennial herbs, *H. suffruticosa,* as a species newly discovered, is woody and robust stems. Hence, it is challenging to identify the species of *Hanceola* via morphology solely. There was no report on nuclear sequences and chloroplast genomic markers in this genus at the species level.

 doi:10.20944/preprints202110.0246.v1

Rubus bambusarum Focke (1891) was known as 'bamboo-leaved raspberry, presenting semi-evergreen, three-lobed foliage with white underside and olive green on the upper surface. *R. bambusarum* is a climbing bramble that produces long, prickled stems (canes), growing up to 20 feet tall if supported [5]. *Rubus henryi* is an evergreen shrub forming scrambling stems, and the slender stems can be up to 600 cm tall [6]. The plants are employed as wood for local, and the young leaves are used for making tea in some regions of China, such as Hubei, Guizhou Province, etc. There were plenty of complex taxonomic problems in the genus Rubus. For example, the blackberries were often mistakenly considered as *Rubus fruticosus* L. sp. agg.

The chloroplast is a cellular organelle that absorbs carbon dioxide and releases oxygen. Meanwhile, it converts light energy into chemical energy in organisms, including green plants, algae, and phototrophic bacterias [7-9]. Although photosynthesis is considered the prominent role of chloroplast, which also plays crucial roles in many biological processes, involving the synthesis of nucleotides, amino acids, fatty acids, vitamins, phytohormones, and plenty of metabolites and metabolism of sulfur and nitrogen [10]. The metabolites synthesized in chloroplasts are vital for plant survival since the chloroplast genome encodes substantial unique proteins involved in these metabolic processes, such as photosynthesis [10,11]. The development of high-throughput sequencing technologies has promoted and improved the study in the kingdom of chloroplast genetics and genomics. After the first complete chloroplast genome, tobacco (*Nicotiana tabacum*), was published, over 2000 complete chloroplast genomes were retrieved from the National Center for Biotechnology Information (NCBI) organelle genome database. Variable regions and multiple DNA fragments were employed for phylogenetic analysis. However, these sequences have insufficient information to differentiate the closely related taxa, especially some without knowing their taxonomic relationships. At higher taxonomic levels (family), conserved sequences and protein-coding regions of the chloroplast genome can be employed for phylogenetic analysis and domestication studies in plants [10,12,13].

DNA molecule markers allow broad use for genetic identification of parents, assessment of genetic variation, developing genetic linkage groups, and improving plants' genetic structure[14]. To date, more and more molecular markers are explored and available since the technologies are enhanced and updated, such as sequencing. These markers are classified according to their purposes, e.g., PCR-based versus non-PCR-based. Common molecular markers such as simple sequence repeats (SSR), sequence-characterized amplified regions (SCAR), and single nucleotide polymorphisms (SNP) are PCR-based. Also known as microsatellites or short tandem repeats, simple sequence repeats are tandemly repeating units of DNA 1 or 2-6 bp in length, and distributed in the whole genomes in plants[15]. Due to the high polymorphic, SSRs were employed as genetic markers in evolutionary analysis, parentage identification, genetic mapping, population genetics, and conservation [15,16].

To date, the chloroplast genome of *M. leptophylla*, *H. exserta*, *R. bambusarum*, and *R. henryi* had not been reported and existed nuclear sequences and molecular markers failed to resolve them from the genus. In this study, we characterized the chloroplast genome of the four species. After that, simple sequence repeats were identified globally, and the species' phylogenetic position was assessed, baing on the circle genome information. This research would provide evidence for resolving the species and contributed new genetic information for further study.

 doi:10.20944/preprints202110.0246.v1

2. Results

2.1. Sequencing profiles and Quality control

The morphology of *M. leptophylla*, *H. exserta*, *R. bambusarum,* and *R. henryi* is shown in Figure 1. The correct identification of the species would ensure the samples for fine sequencing. Leaves of *M. leptophylla* were grown with independent petioles, and several leaves share a common node. *H. exserta* grows soft, complex serrated leaves. *R. bambusarum* and *R. henryi* usually were considered as the same species owning to the highly similar plant morphology. There were certain variations on leaf sharps, even though both these two species exhibited okra leaves. The three-lobed foliage was much longer and shorter in length and width within R. bambusarum, respectively, comparing with those within *R. henryi*.

13,128,417, 11,399,665, 11,851,040 and 10,197,459 raw reads were generated from *M. leptophylla*, *H. exserta*, *R. bambusarum* and *R. henryi*, respectively. Consequently, 12,876,579, 11,348,725, 11,696,892, and 10,062,246 clean reads yielded via data filter, correspondingly. Within *M. leptophylla*, 3.94 and 3.86 Giga raw and clean bases were obtained separately, as the effective rate was 98.08%. Likewise, the 3.40, 3.51, and 3.02 G clean bases were produced from *H. exserta*, *R. bambusarum,* and *R. henryi*, respectively, and the effective rates were 99.55%, 98.70%, and 98.67% correspondingly. Overall, the sequencing error rates involved in four species were 0.03%, which kept a quite low level. In detail, the values Q_{20} and Q_{30} in terms of four species bellowed 97.78% and 93.65%, separately. It's evidence that the sequencing quality was fine. The GC content involved in four species was presented from 40.02% to 40.81%, and it illustrated the sequencing composition is proper in the experiment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

Figur+

+9-e 1 Profiles of the four species. *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi* were showed as panel (A), (B), (C), and (D), respectively. The small patches placed onto each board at the right top were the leaf morphology, correspondingly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

Table 1 The profiles of sequencing on DNA from four species.

Species	Raw Reads	Clean Reads	Raw Base (G)	Clean Base (G)	Effective Rate (%)	Error Rate (%)	Q_{20} (%)	Q_{30} (%)	GC Content (%)
M. leptophylla	13,128,417	12,876,579	3.94	3.86	98.08	0.03	97.56	93.41	40.81
H. exserta	11,399,665	11,348,725	3.42	3.40	99.55	0.03	97.78	93.65	40.38
R. bambusarum	11,851,040	11,696,892	3.56	3.51	98.70	0.03	97.42	93.14	40.02
R. henryi	10,197,459	10,062,246	3.06	3.02	98.67	0.03	97.41	93.00	40.67

Note: The quality scores are logarithmically linked to error probabilities, and Q_{20} and Q_{30} denote accuracy of a base call was 99% and 99.9% separately.

2.2. Assembly and annotation of four species

The circle chloroplast genome of *M. leptophylla*, *H. exserta*, *R. bambusarum*, and *R. henryi* were successfully assembled and annotated (Figure 2). The complete chloroplast genome of *M. leptophylla* was 152, 624 base pairs (bp) in length, and those of *H. exserta*, *R. bambusarum*, and *R. henryi* were 153,296 bp, 156,309 bp, and 158, 953 bp, respectively. The four chloroplast genomes presented a quite close length. The GC content of four species were 39.16%, 37.79%, 37.17% and 37.09% separately (Figure 3 (A)). The value of the GC content waved slightly. Regarding the genome structures, four specific regions, including large single copy (LSC), small single copy (SSC), and double inverted repeats, i.e., IRa and IRb, were identified from all four species. In the chloroplast genome of *M. leptophylla*, the lengths of LSC, SSC, IRa, and IRb were 93,670 bp, 18,806 bp, 20,074 bp, and 20,074 bp orderly. Correspondingly, the percentages were 61.37%, 12.32%, 13.53% and 13.53%. Regarding *H. exserta*, the four regions were 84,079 bp, 17,703 bp, 25,757 bp and 25, 757 bp, as the percentage were 54.85%, 11.55%, 16.80% and 16.80% respectively. In addition, *R. bambusarum* and *henryi* showed pretty similar representative chloroplast genome structures. In *R. bambusarum*, LSC, SSC, IRa, and IRb were 85, 880 bp, 18, 841 bp, 25, 749 bp, and 25, 749 bp in length, and they accounted for 54.95%, 12.05%, 16.50, and 16.50%. Similarly, those values were 88, 586 bp (55.74%), 18, 827 bp (11.84%), 25, 770 bp (16.21%) and 25, 770 bp (16.21%) respectively. Overall, the length showed trends as LSC> IRa= IRb> SSC in four species (Figure 3 (B)).

124, 130, 129, and 131 genes were characterized from the chloroplast genome of *M. leptophylla*, *H. exserta*, *R. bambusarum*, and *R. henryi*, respectively. Within *M. leptophylla*, the three classes of genes involving coding sequence, tRNA, and rRNA were 80, 36, and 8, respectively. Regarding the chloroplast genome of *H. exserta*, *R. bambusarum*, and *R. henryi*, there were 85, 84, and 84 genes involving in the coding sequence, separately. Correspondingly, the numbers of genes in terms of tRNA were 37, 37, and 39 in the above three species, whereas those regarding rRNA were 8 for each species. In four species, the value of the genes contained introns were 22, 23, 22, and 22, respectively. What's more, they all processed two genes that had more than two introns (Figure 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

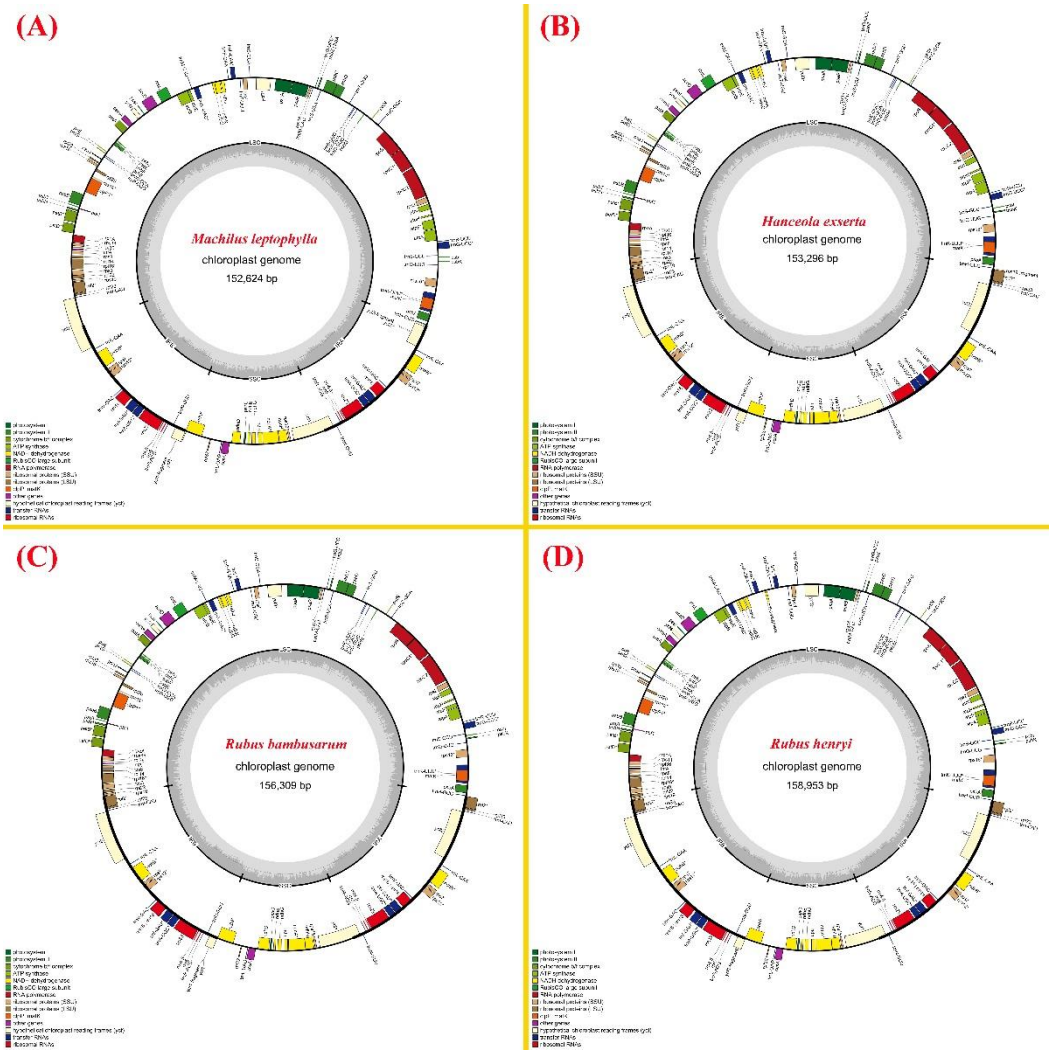

Figure 2 Assembly and annotation on complete chloroplast genome of four species. (A), (B), (C), and (D) represent species *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi*, respectively.

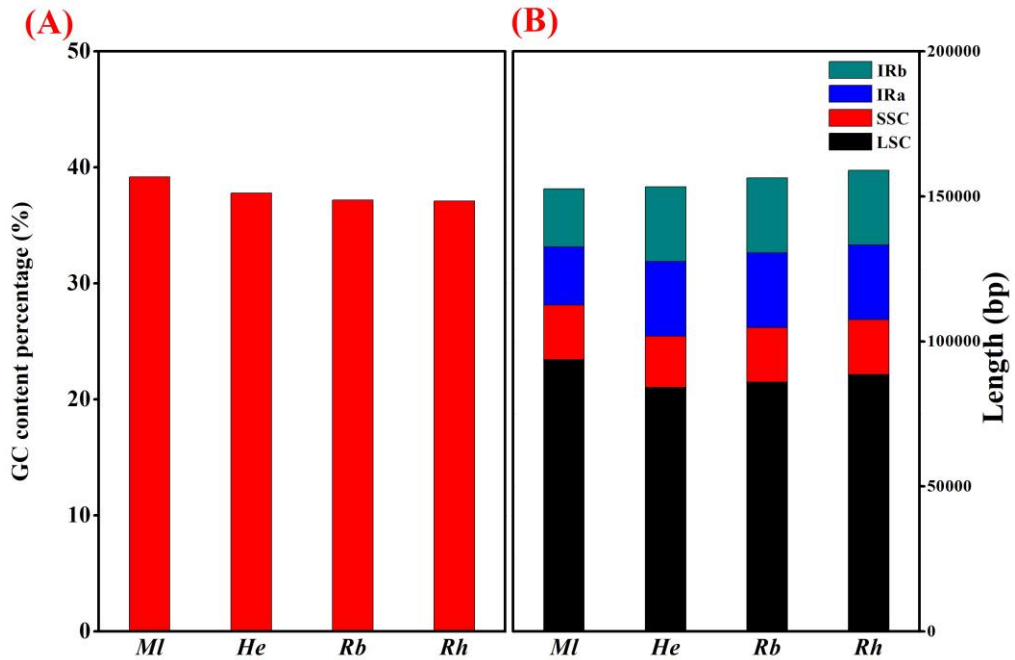

Figure 3 GC content and length of DNA structures of complete chloroplast genome among four species. (A) presented GC content percentage within the whole chloroplast genomes, and (B) shown

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

the genome structures including LSC, SSC, IRa, and IRb, respectively. LSC, SSC, IRa, IRb, *Ml*, *He*, *Rb*, and *Rh* were abbreviated from large single copy, small single copy, inverted repeats a, inverted repeats b, *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi*, orderly.

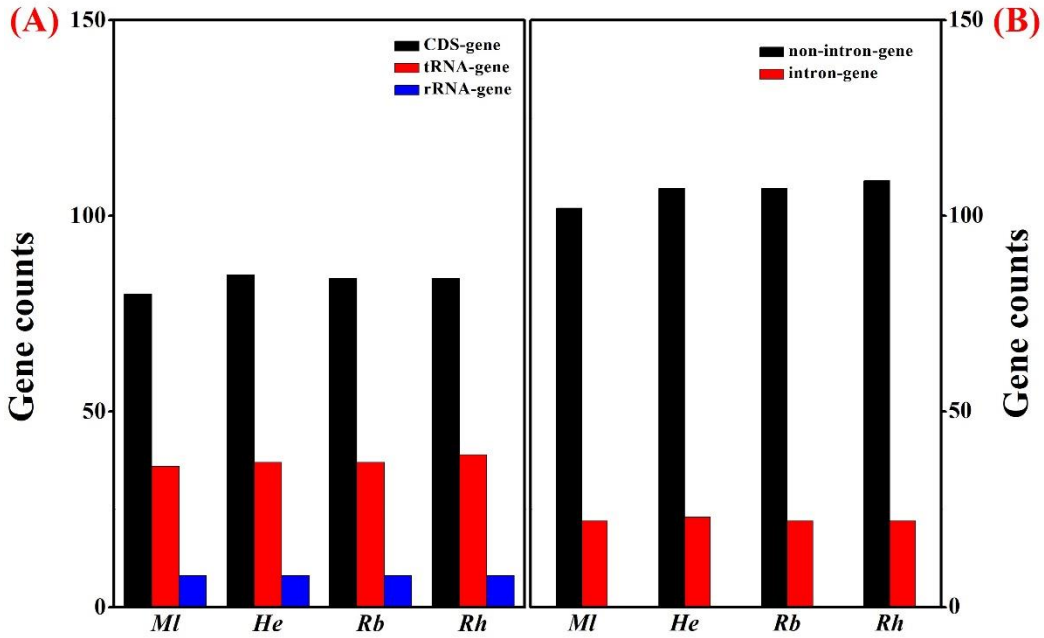

Figure 4 Classes of genes distributed in four species. *Ml*, *He*, *Rb*, and *Rh* were abbreviated from *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi*, separately.

2.3. Identification of simple sequence repeats

Total 82 SSRs were identified from the chloroplast genome of *M. leptophylla* (Table 2). Less SSRs were discovered from *H. exserta*, *R. bambusarum*, and *R. henryi*, with 56, 58, and 62, respectively. In all four chloroplast genomes, the number of SSR containing sequences was one, and the number of sequences containing more than 1 SSR was one, too. Additionally, the number of SSRs present in compound formation from *M. leptophylla* chloroplast genome was 8, whereas those in *H. exserta*, R. *bambusarum*, and *R. henryi* were 6, 6, and 7, respectively. Six different repeat classes were identified from the species. In the chloroplast genome of *M. leptophylla*, the counts of mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides were 55, 12, 3, 9, 2, and 1, orderly. There were 35, 7, 2, 11, 0, and 1 SSRs belonging to mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide within the chloroplast genome from *H. exserta*. Regarding those within *R. bambusarum* and *R. henryi*, 42 and 46 mononucleotide SSRs were identified separately. Correspondingly, 8 and 9 dinucleotide SSRs were discovered. Moreover, the number of trinucleotide SSRs were 2 and 1 within those two species. Additionally, both six tetranucleotides were found from the chloroplast genomes of *R. bambusarum* and *R.henryi*. However, not any pentanucleotide and hexanucleotide SSRs were identified from those two chloroplast genomes. Overall, the numbers of mononucleotide SSRs in the four species were much more than those of other SSR classes.

The types of repeats were various in the same SSR class (Figure 5). In mononucleotide SSRs, the number of type A/T was much more than those of type C/G. The SSRs involving 10~ 17 consecutive A/T were found in the chloroplast genome of *M. leptophylla*. The mononucleotide SSRs including 10~14 repeated A/T were discovered in that of *H. exserta*. In the chloroplast genome of *R. bambusarum* and *R. henryi*, SSRs related to A/T are mainly presented as 10~12 nucleotide repeats. Regarding mononucleotide SSRs of type C/G, only one consecutive 10, 11, 11-nucleotides SSR could be identified from *M. leptophylla*, *H. exserta*, and *R henryi*. For dinucleotide SSRs, both five repeated AG/CT and AT/AT were found in all four chloroplast genomes. SSRs of both 6 and 7-repeated dinucleotide can be discovered from that of *M. leptophylla*. Three trinucleotide SSR types can be identified from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

all four species: AAG/CTT, AAT/ATT, and ATC/ATG. What's more, 9 types involved in tetranucleotide SSR were discovered among species, i.e. AAAC/GTTT, AAAG/CTTT, AAAT/ATTT, AACT/AGTT, AATG/ATTC, AATT/AATT, ACAG/CTGT, ACAT/ATGT, and AGAT/ATCT. Additionally, the chloroplast genome of *M. leptophylla* comprised two pentanucleotide and one hexanucleotide SSR type. They were AAATC/ATTTG, AAATT/AATTT, and AAATAG/TTTCTC. The only hexanucleotide type in that of *H. exserta* was AAGATC/ATCTTG.

Table 2 Numbers of SSR among four species

Species		*M. leptophylla*	*H. exserta*	*R. bambusarum*	*R. henryi*
Total number of identified SSRs		82	56	58	62
Number of SSR containing sequences		1	1	1	1
Number of sequences containing more than 1 SSR		1	1	1	1
Number of SSRs present in compound formation		8	6	6	7
Number of different repeat classes	mononucleotide	55	35	42	46
	dinucleotide	12	7	8	9
	trinucleotide	3	2	2	1
	tetranucleotide	9	11	6	6
	pentanucleotide	2	0	0	0
	hexanucleotide	1	1	0	0

Note: SSR was abbreviated from simple sequence repeats. The quality scores are logarithmically linked to error probabilities, and Q_{20} and Q_{30} denote accuracy of a base call was 99% and 99.9% separately.

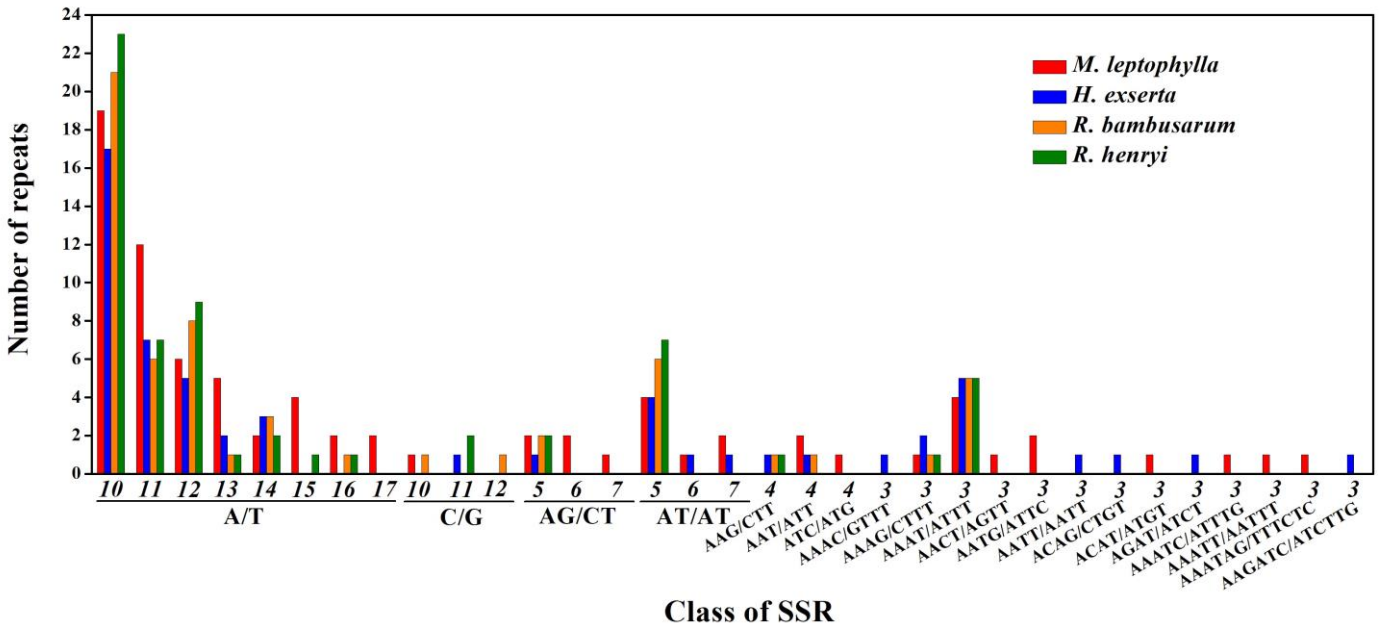

Figure 5 Different repeat types of SSR motifs among species. The number under the X-axis denoted the consecutive repeated sequences (nucleotides).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

2.4. Phylogenetic analysis basing on chloroplast genome

The topological structure of the phylogenetic tree of 22 species, including *Machilus leptophylla*, is illustrated in Figure 6-1. *M. leptophylla* was clustered with *M. yunnanensis* under genus *Machilus*. They showed closer phylogenetic relationships with the other two *Machilus* species *balansae* and *robusta*, with strong bootstrap support. Outgroup *Chimonanthus praecox* established distant phylogenetic relationships with the other 21 species. It's clear that *C. praecox*, chosen as an out-group, played the right roles to differentiate the species. The family *Calycanthaceae* containing out-group *C. praecox*, and *Lauraceae* involving the other 21 species, including *M. leptophylla*, were belonged to order *Laurales*. For identifying the phylogenetic position of *H. exserta*, 20 species involved in 15 genera were analyzed. From Figure 6-2, *H. exserta* was clustered with genus *Isodon*, *Nepeta*, *Ocimum*, *Platostoma*, and *Lavandula*, which were belonged to the family *Ocimeae*. The other species were involved in the family *Mentheae*. Out-group *Scutellaria kingiana* was distant from the other 19 species in phylogenetic relationships. For confirming the phylogenetic position of *R. bambusarum* and *R. henryi*, 22 species were added, and the phylogenetic tree was presented in Figure 6-3. *R. bambusarum* and *henryi* were clustered together, indicating a much closer phylogenetic relationship and similar genetic backgrounds. The species *R. chingiigi*, *R. crataegifoliusgi*, and *R. leucanthusgi* clustered together and showed the class of parallel branches. *Euonymus schensianus*, as the out-group, was far from other species in the phylogenetic tree.

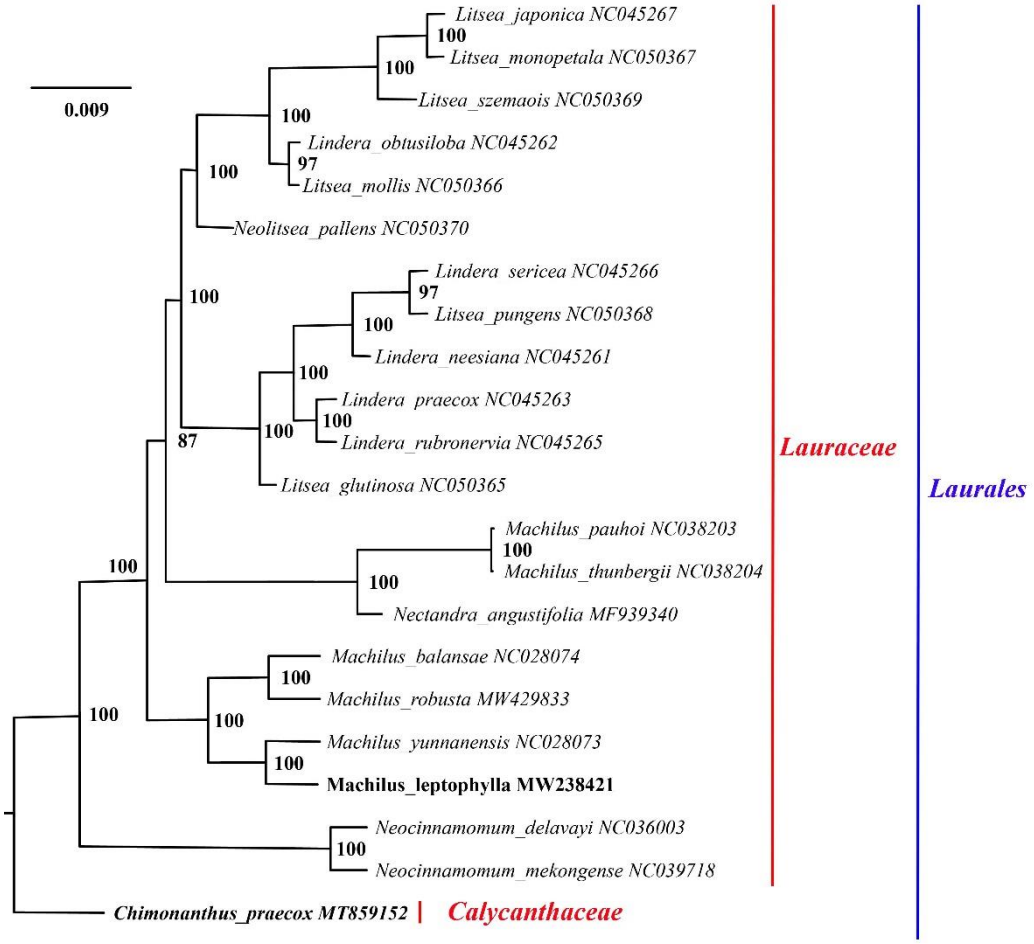

Figure 6-1 The Maximum-Likelihood (ML) phylogenetic tree shows the relationships among *Machilus leptophylla* (showed as bold texts) and other 21 species, and *Chimonanthus praecox* (showed as bold and italic texts) is presented as the out-group. Bootstrap support values from 1000 replicates are given close to the nodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

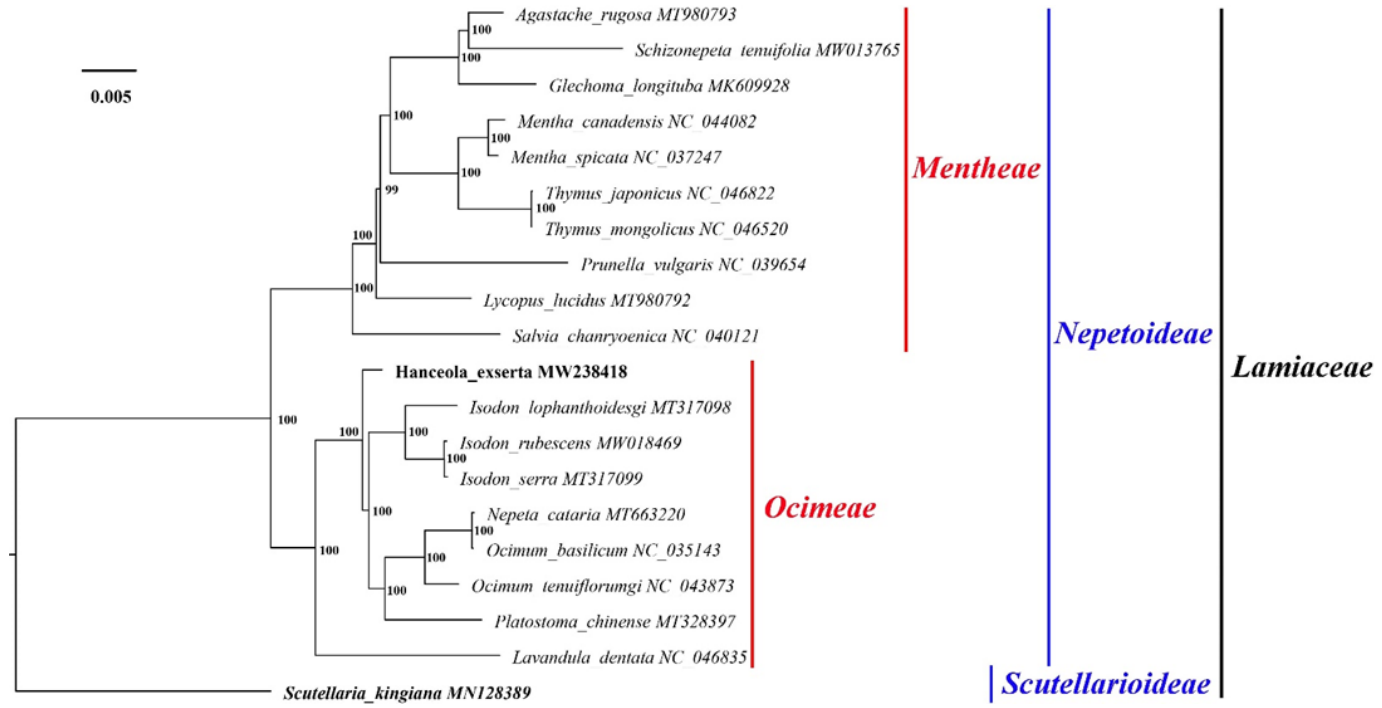

Figure 6-2 Total of 20 complete chloroplast genomes was showed as a phylogenetic tree via Maximum-Likelihood (ML) method. *Hanceola exserta* was denoted with bold text, and *Scutellaria kingiana* was used as an out-group species as bold and italic text showed. The numbers adjacent to the branch nodes, presented bootstrap support values from 1000 replicates.

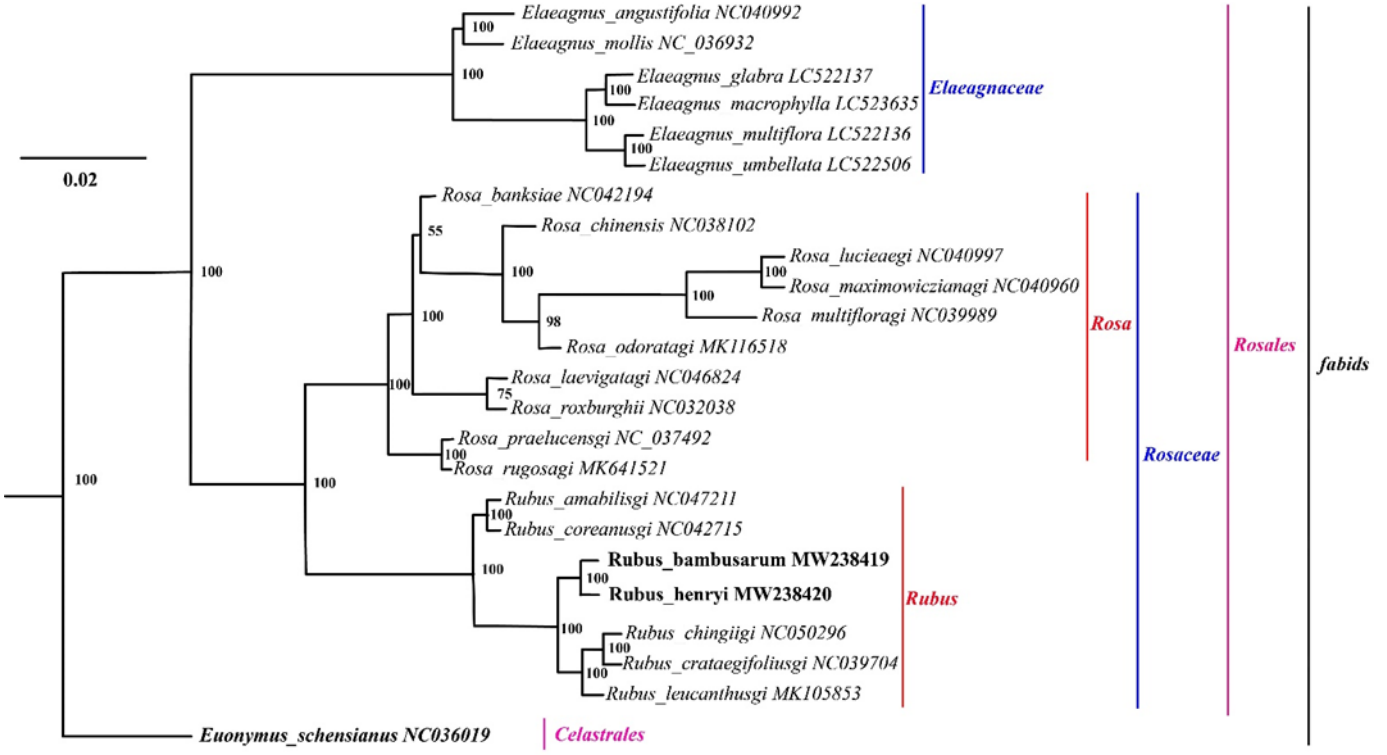

Figure 6-3 The Maximum-Likelihood (ML) phylogenetic trees of 23 complete chloroplast genomes: Rubus bambusarum showed with bold text, *Euonymus schensianus* used as an out-group species and showed with bold and italic text. The numbers adjacent to nodes show bootstrap support values from 1000 replicates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2021 doi:10.20944/preprints202110.0246.v1

3. Discussion

3.1 Chloroplast genome processing featured constructs

Most chloroplast genome comprises specific four regions, i.e., large and small single copies and two inverted repeats[12]. Complete four regions in chloroplast commonly mean related full biological functions in related species. Our results showed that these four regions could be found in *M. leptophylla*, *H. exserta*, *R. bambusarum*, and R. henryi. What's more, the specific regions showed regular length individually, which indicated there was no loss of significant long fragments. Usually, the chloroplast genome contained 120~130 genes, and that length fall ranged from 107-218 kb[12]. Within these four chloroplast genomes, 124, 130, 129, and 131 genes were identified, and the whole genomes were 152, 624 bp, 153,296 bp, 156,309 bp, and 158, 953 bp in length, correspondingly. Our results were consistent with the report. It's documented one copy of the IR was missed in some species, such as family *Papilionoideae*, as formed IR lacing clade[17,18]. It's evident there were some exceptions, even though the sub-structures and gene counts were relatively conserved and stable in the chloroplasts. Our results indicate most of the identified genes are without introns. The number of genes that contained introns was 22-23 over the four species, and all four chloroplasts had two genes that processed more than two introns. Obviously, the number of introns involved in chloroplast genes kept at an average level. However, the loss of introns chloroplast genes had been reported in many plants, such as *Cicer arietinum*, *Manihot esculenta*, *Bambusa sp*., and *Hordeum vulgare* [19-22]. In chloroplast genomes, intron loss tended to happen in diverse plants such as *Poaceae*, *Onagraceae*, *Oleaceae*, and *Pinus* [23]. Our data probably inferred the genus *Machilus*, Hanceola, and Rubus may be more stable involving introns.

3.2 SSR from chloroplast genome provide essential genetic information

Simple sequence repeats are tandem repeats, which comprise 1-6 nucleotides in the genomes of organisms [24]. Among species, even genotypes, the number of repeats units may change as the tandem arrays of varies on SSR motifs. A substantial number of SSRs distributed all over the genome, including organellar DNA [25]. In model plants rice and *Arabidopsis thaliana*, it was reported that SSRs presented to be organized and altered in regions of the genes [26]. Generally, SSRs showed properties of high mutation rate in the locus of generations, locus specificity, intraspecific polymorphism, reproducibility, and multiallelic across taxa [27]. In our studies, mononucleotide SSRs were the richest SSRs in the chloroplast genomes involved in species *M. leptophylla*, *H. exserta*, R. bambusarum, and *R. henryi*. In the chloroplast genome of single-petal (SP) and double-petal (DP) *Jasminum sambac* L. (*Oleaceae*), the mononucleotides SSRs accounting for 62.71% (74/118) and 62.39% (73/117), respectively [28]. However, in the chloroplast genome of *Rhus chinensis*, mononucleotide SSRs account for 28.74%, which was less than dinucleotide SSRs with 60% [29]. It is thus clear evidence that the diverse classes of SSRs in the plant chloroplast genome probably depend on the categories of plants. In nuclear genomes, the mononucleotide SSRs take higher portions in all six classes (mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide). In the nuclear genome of *Zanthoxylum bungeanum*, mononucleotide repeats were the most abundant class, with the value of 19706, which is the four times of dinucleotide repeats (5154) [30]. Similarly, in the nuclear genome, the mononucleotide presents the highest proportion in *Chinese jujube* (Ziziphus jujuba) [31]. Moreover, the SSRsclass mononucleotide was the most abundant expressed sequence tag, such as tobaccos (*Nicotiana tabacum* L.). In mononucleotide SSRs of the four chloroplasts, they were determined to be rich in A/T and rare in tandem G or C repeats, and this was consistent with reported [32-34].

3.3 Comparision on chloroplast genome offered a robust tool to study phylogenetic relationship and evolution among plant species

Complete chloroplast genome sequences provide insights into the understanding of plants' biology and diversity[10]. Within phylogenetic clades, chloroplast genomes contributed significantly in phylogenetic studies of several plant families and resolving evolutionary relationships[10]. Furthermore, as within and between plant species involving

 doi:10.20944/preprints202110.0246.v1

both sequence and structural variation, considerable variation was revealed by chloroplast genome sequences. The information from chloroplast genomes was precious to understand the environments, promoting the breeding of closely related species [35,36]. The phylogenetic tree (Figure 6-1, Figure 6-2, and Figure 6-3) were constructed by three groups of complete chloroplast genome sequences. Overall, the topological structure of the species in this study demonstrated highly consistent with taxa relationship from the database of Taxonomy under National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/taxonomy/?term=). Exceptionally, in Figure 6-2, the species *Nectandra angustifolia* was clustered with *Machilus pauhoi* and *thunbergii*, which form a new clade. It inferred that some taxa problems probably existed in the current genus *Machilus* or*Nectanra*. Within Figure 6-3, species R. bambusarum and R henryi were clustered together with robust bootstrap support. Hence, the information from the phylogenetic tree did not successfully answer the questions of whether the R. bambusarum and R. henryi were the same species or not. However, the specific information from the chloroplast genomes provides the evidence to differentiate these two species, such as differences in the length of the complete circle genome and the distribution and classes of SSRs. The chloroplast genome sequence offered a robust approach to resolve the close species.

4. Materials and Methods

4.1. Sampling and DNA extraction

The fresh leaves of *M. leptophylla* were sampled from Zijingang Campus, Zhejiang University (120°51′32″ E, 30°18′08″ N). Those of *H. exserta*, *R. bambusarum*, and *R. henryi* were collected from Hangzhou Botany Garden (120°07′36″ E, 30°15′15″ N). Consequently, the specimens were deposited in Institute of Crop Sciences, Zhejiang University at Specimen code: *LM001*, *LM002*, *LM003*, and *LM004*, orderly. The DNA extraction was performed as follows: 1) Weight 80-150 mg fresh samples and mixed them with 800μl of CTAB buffer. 2) Grind the mixture to homogenate, and then vortex them for 3 minutes. 3) Place the tube containing the mixtures in a water bath for 35 minutes at 65 °C. 4) Centrifuge the homogenate for 10 minutes at 13 000 rpm. After that, transfer the supernatant into a new centrifuge tube. 5) aliquot 4μl of RNase A working solution and add them into each tube for incubating at 37°C for 15 minutes. 6) Add phenol/chloroform/isoamyl alcohol (25:24:1) into the tubes, make the final volume were folded. 7) Vortex for mixing and then centrifuge the tubes at 13 000 rpm for 2 minutes. 8) transfer the upper layer of liquid into a new centrifuge tube. 9) Add half-volume pre-cold isopropanol and incubate at the frozen fridge at -20°C for 20 minutes. 10) Centrifuge the tubes at 13000 rpm for 8 minutes, and then discard the supernatant at the condition of ensuring peace of the pellet. 11) Wash it with pre-cold 70% ethanol and dry the pellet at the laminar flow cabinet. 12) Add 50 μl TE buffer to dissolve the DNA. The total DNA quality was detected by *NanoDrop Microvolume Spectrophotometers and Fluorometer* (ThermoFisher Scientific, USA). The values of *OD260/OD280* fall into the range from 1.7 to 1.9 would be kept for further study.

4.2. DNA sequence and raw data processing

According to the manufacturer's instructions, the TruSeq Library Construction Kit (Illumina, San Diego, CA, USA) was employed to construct the sequencing libraries. The total DNA samples were fragmented by g-TUBE, followed by centrifuging at 4000 rpm for 3 min and processed orderly via end-repair, adapter, ligation, and exonuclease. The sequencing was conducted by the *Illumina HiSeq 2000 platform* referring to the standard protocols at *Tianjin Sequencing Center*, *Tianjin Novogene Technology Co., Ltd., China*. A genomic shotgun library with an injection size of 150 bp was constructed, and more than three Giga base clean data was obtained. Adapter sequences, potential contamination, and low-quality bases of the raw data were removed by Adapter Removal. The *CLC-quality trim tool* was employed to filtered fine reads.

4.3. Chloroplast genome assembling and annotating

For identifying the chloroplast sequences of *M. leptophylla*, the Illumina reads were mapped to the reference chloroplast sequence of *M balansae* (KT348517) in the NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genom e/organ /) by Bwa (version 0.7.17)[37]. Similarly, the reference chloroplast for *H. exserta* was used by *Ocimum basilicum* (KT348517), and those of *R. bambusarum* and *R. henryi* shared the same reference in terms of *Rubus crataegifolius (NC_039704).* The reads were assembled and finally polished by SPAdes[38] and Pilon[39] separately. The order of contigs was evaluated based on the collinearity analysis by the tool *Mummer* [40]. Consequently, the initiation and termination sites of the two inverted repeat sequences were identified by aligning the targeting and reference chloroplast genome with the tool *Blast* [41]. All four chloroplast genomes were annotated by Dual Organellar GenoMe Annotator (*DOGMA*) under manual corrections [42]. *BLASTX*, *BLASTN*, and tRNAscan-SE1.21 were employed to identify putative gene types involving protein-coding，rRNA and tRNA[43,44]. The circular chloroplast genomes were drawn and illustrated by *Organellar Genome DRAW* [45].

4.4. Identification of simple sequence repeat among chloroplast genomes

Small sequence repeats (SSR) of the chloroplast genome were identified by tool *MIcroSAtellite* (MISA²)[46]. The parameter set as followed: 1) Definition (unit_size, min_repeats): 1-10 2-5 3-4 4-3 5-3 6-3; 2) interruptions (max_difference_between_2_SSRs): 100 bp.

4.5. Phylogenetic analysis

For phylogenetic analysis, 22 chloroplast genomes of representative species, including *M. leptophylla,* were selected, in which that of Chimonanthus praecox (MT859152) served as the out-group. Similarly, to determine the phylogenetic positions of *H. exserta*, a total of 20 chloroplast genomes were employed to analyze, and *Scutellaria kingiana* (MN128389.1) was selected as the out-group. For *R. bambusarum* and henryi, a total of 24 chloroplast genomes was employed. In this group, *Euonymus schensianus* (NC036019) was used as an out-group. The chloroplast genomes were aligned using MAFFT (V7.407)[47], and after that, the phylogeny trees were constructed via the maximum likelihood (ML) method by IQtree (*Version 1.7*) [48]. The internal branching support was estimated through 1000 bootstrap replicates.

Conclusions: The main findings were concluded as follows:

1) The four chloroplast genomes, involved in *Machilus leptophylla*, *Hanceola exserta*, *Rubus bambusarum*, and *Rubus henryi*, comprised 152.624 kb, 153.296kb, 156.309 kb, and 158.953 kb in length, as well as 124, 130, 129, and 131 genes, respectively. Moreover, they presented the typical four regions in chloroplast genome structures.

2) Six classes of SSR were identified from the four chloroplast genomes, in which mononucleotide was the class with the highest numbers. However, SSR classes regarding trinucleotide, pentanucleotide, and hexanucleotide processed a few numbers. The types of repeats were various within individual classes of SSR.

3) Phylogenetic trees indicated that *M. leptophylla* was clustered with *M. yunnanensis* under genus *Machilus*, *H. exserta* was confirmed under family *Ocimeae*. Additionally, *R. bambusarum* and *R. henryi* were clustered together, whereas they did not belong to one species due to the differing SSR features.

Author Contributions: Conceptualization, L.M. and Y.Z.; methodology, G. L and J. X.; software, J. S. and L. W.; formal analysis, L. M. and M. K. D.; resources, X. Z.; data curation, G. L. and L. W.; writing—original draft preparation, L. M., Y. Z. and X. Z.; writing—review and editing, M. K. D.; visualization, Y. Z. and X. Z.; supervision, L.M.; project administration, L.M.; funding acquisition, L.M. and G. L. all authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by China Postdoctoral Science Foundation, grant number 2021M690633, Natural Science Foundation of Hubei Province, China, grant number 2021, and Natural Science Foundation of Guangdong Province, China, grant number 2016A030307002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The genome sequence data that support the findings are openly available in GenBank of NCBI at https://www.ncbi.nlm.nih.gov/. The accessions involved in species are *M. leptophylla*, *H. exserta*, *R. bambusarum*, and *R. henryi*, which are *MW238421*, *MW238418*, *MW238419*, and *MW238420*, respectively. The associated related BioProject number is *PRJNA722038*.

Acknowledgments: We appreciated Dr. Chao Feng and Dr. Chen Feng at South-China Botany Garden, who offered comprehensive favors involving data analysis consults on data analysis and professional proofreading.

Conflicts of Interest: The authors declare no conflict of interest.

References:

1. Zhong, Q.; Cheng, D.; Hu, S.; He, L.; Tang, C.; Wen, Y.; Qiu, J.; Li, X. Chlorophyll content and net photosynthetic rate of *Machilus pauhoi* and *M. leptophylla*. Chinese Journal of Applied Ecology, 2009, 20, 271-276.
2. Tang, S.; Xu, W.; Wei, F. *Machilus parapauhoi* sp. nov. and a new synonym of Machilus (Lauraceae) from east Asia. Nord J Bot 2010, 28, 503-505, doi:10.1111/j.1756-1051.2010.00748.x.
3. Song, Y.; Dong, W.; Liu, B.; Xu, C.; Yao, X.; Gao, J.; Corlett, R.T. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. Front Plant Sci 2015, 6, doi:10.3389/fpls.2015.00662.
4. Harley, R.M.; Atkins, S.; Budantsev, A.L.; Cantino, P.D.; Conn, B.J.; Grayer, R.; Harley, M.M.; de Kok, R.; Krestovskaja, T.; Morales, R., et al. Labiatae. In The Families and Genera of Vascular Plants, Springer: Berlin and Heidelberg, 2004; Vol. 7, pp 167-275.
5. Focke, W.O. Rubus bambusarum Focke. Hooker's Icones Plantarum. In 1891; Vol. 30, p 1952.
6. http://flora.huh.harvard.edu/china/ (accessed on.
7. Yin, D.; Wang, Y.; Zhang, X.; Ma, X.; He, X.; Zhang, J. Development of chloroplast genome resources for peanut (Arachis hypogaea L.) and other species of Arachis. Sci Rep 2017, 7, 11649, doi:10.1038/s41598-017-12026-x.
8. Mauriello, E. How bacteria arrange their organelles. Elife 2019, 8, doi:10.7554/eLife.43777.
9. Melis, A.; Chen, H.C. Chloroplast sulfate transport in green algae--genes, proteins and effects. Photosynth Res 2005, 86, 299-307, doi:10.1007/s11120-005-7382-z.
10. Daniell, H.; Lin, C.; Yu, M.; Chang, W. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 2016, 17, doi:10.1186/s13059-016-1004-2.
11. Bobik, K.; Burch-Smith, T.M. Chloroplast signaling within, between and beyond cells. Front Plant Sci 2015, 6, 781, doi:10.3389/fpls.2015.00781.
12. Tian, C.; Li, X.; Wu, Z.; Li, Z.; Hou, X.; Li, F.Y. Characterization and Comparative Analysis of Complete Chloroplast Genomes of Three Species From the Genus Astragalus (Leguminosae). Front Genet 2021, 12, doi:10.3389/fgene.2021.705482.
13. Jansen, R.K.; Cai, Z.; Raubeson, L.A.; Daniell, H.; Depamphilis, C.W.; Leebens-Mack, J.; Müller, K.F.; Guisinger-Bellian, M.; Haberle, R.C.; Hansen, A.K., et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 2007, 104, 19369-19374, doi:10.1073/pnas.0709121104.
14. Younis, A.; Ramzan, F.; Ramzan, Y.; Zulfiqar, F.; Ahsan, M.; Lim, K.B. Molecular Markers Improve Abiotic Stress Tolerance in Crops: A Review. Plants 2020, 9, 1374, doi:10.3390/plants9101374.
15. Bhargava, A.; Fuentes, F.F. Mutational dynamics of microsatellites. Mol Biotechnol 2010, 44, 250-266, doi:10.1007/s12033-009-9230-4.
16. Buschiazzo, E.; Gemmell, N.J. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 2006, 28, 1040-1050, doi:10.1002/bies.20470.
17. Xiong, Y.; Xiong, Y.; He, J.; Yu, Q.; Zhao, J.; Lei, X.; Dong, Z.; Yang, J.; Peng, Y.; Zhang, X., et al. The Complete Chloroplast Genome of Two Important Annual Clover Species, Trifolium alexandrinum and T. resupinatum: Genome Structure, Comparative Analyses and Phylogenetic Relationships with Relatives in Leguminosae. Plants (Basel) 2020, 9, doi:10.3390/plants9040478.
18. Martin, G.E.; Rousseau-Gueutin, M.; Cordonnier, S.; Lima, O.; Michon-Coudouel, S.; Naquin, D.; de Carvalho, J.F.; Aïnouche, M.; Salmon, A.; Aïnouche, A. The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann Bot 2014, 113, 1197-1210, doi:10.1093/aob/mcu050.
19. Saski, C.; Lee, S.B.; Fjellheim, S.; Guda, C.; Jansen, R.K.; Luo, H.; Tomkins, J.; Rognli, O.A.; Daniell, H.; Clarke, J.L. Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes. Theor Appl Genet 2007, 115, 571-590, doi:10.1007/s00122-007-0567-4.
20. Daniell, H.; Wurdack, K.J.; Kanagaraj, A.; Lee, S.B.; Saski, C.; Jansen, R.K. The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. Theor Appl Genet 2008, 116, 723-737, doi:10.1007/s00122-007-0706-y.
21. Jansen, R.K.; Wojciechowski, M.F.; Sanniyasi, E.; Lee, S.B.; Daniell, H. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol 2008, 48, 1204-1217, doi:10.1016/j.ympev.2008.06.013.

22. Wu, F.H.; Kan, D.P.; Lee, S.B.; Daniell, H.; Lee, Y.W.; Lin, C.C.; Lin, N.S.; Lin, C.S. Complete nucleotide sequence of Dendrocalamus latiflorus and Bambusa oldhamii chloroplast genomes. Tree Physiol 2009, 29, 847-856, doi:10.1093/treephys/tpp015.
23. Jansen, R.K.; Cai, Z.; Raubeson, L.A.; Daniell, H.; Depamphilis, C.W.; Leebens-Mack, J.; Müller, K.F.; Guisinger-Bellian, M.; Haberle, R.C.; Hansen, A.K., et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 2007, 104, 19369-19374, doi:10.1073/pnas.0709121104.
24. Taheri, S.; Abdullah, T.L.; Ahmad, Z.; Abdullah, N.A. Effect of acute gamma irradiation on Curcuma alismatifolia varieties and detection of DNA polymorphism through SSR marker. Biomed Res Int 2014, 2014, 631813, doi:10.1155/2014/631813.
25. Phumichai, C.; Phumichai, T.; Wongkaew, A. Novel Chloroplast Microsatellite (cpSSR) Markers for Genetic Diversity Assessment of Cultivated and Wild Hevea Rubber. Plant Mol Biol Rep 2015, 33, 1486-1498, doi:10.1007/s11105-014-0850-x.
26. Lawson, M.J.; Zhang, L. Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biol 2006, 7, R14, doi:10.1186/gb-2006-7-2-r14.
27. Taheri, S.; Lee Abdullah, T.; Yusop, M.; Hanafi, M.; Sahebi, M.; Azizi, P.; Shamshiri, R. Mining and Development of Novel SSR Markers Using Next Generation Sequencing (NGS) Data in Plants. Molecules 2018, 23, 399, doi:10.3390/molecules23020399.
28. Qi, X.; Chen, S.; Wang, Y.; Feng, J.; Wang, H.; Deng, Y. Complete chloroplast genome of Jasminum sambac L. (Oleaceae). Braz J Bot 2020, 43, 855-867, doi:10.1007/s40415-020-00638-z.
29. Zuo, R.; Jiang, P.; Sun, C.; Chen, C.; Lou, X. Analysis of the chloroplast genome characteristics of Rhus chinensis by de novo sequencing. Sheng Wu Gong Cheng Xue Bao 2020, 36, 772-781, doi:10.13345/j.cjb.190354.
30. Li, J.; Li, S.; Kong, L.; Wang, L.; Wei, A.; Liu, Y. Genome survey of Zanthoxylum bungeanum and development of genomic-SSR markers in congeneric species. Bioscience Rep 2020, 40, doi:10.1042/BSR20201101.
31. Xiao, J.; Zhao, J.; Liu, M.; Liu, P.; Dai, L.; Zhao, Z. Genome-Wide Characterization of Simple Sequence Repeat (SSR) Loci in Chinese Jujube and Jujube SSR Primer Transferability. Plos One 2015, 10, e127812, doi:10.1371/journal.pone.0127812.
32. Hong, S.; Cheon, K.; Yoo, K.; Lee, H.; Cho, K.; Suh, J.; Kim, S.; Nam, J.; Sohn, H.; Kim, Y. Complete Chloroplast Genome Sequences and Comparative Analysis of Chenopodium quinoa and C. album. Front Plant Sci 2017, 8, doi:10.3389/fpls.2017.01696.
33. Liu, W.; Kong, H.; Zhou, J.; Fritsch, P.; Hao, G.; Gong, W. Complete Chloroplast Genome of Cercis chuniana (Fabaceae) with Structural and Genetic Comparison to Six Species in Caesalpinioideae. Int J Mol Sci 2018, 19, 1286, doi:10.3390/ijms19051286.
34. Ni, L.; Zhao, Z.; Xu, H.; Chen, S.; Dorje, G. The complete chloroplast genome of Gentiana straminea (Gentianaceae), an endemic species to the Sino-Himalayan subregion. Gene 2016, 577, 281-288, doi:10.1016/j.gene.2015.12.005.
35. Brozynska, M.; Furtado, A.; Henry, R.J. Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol J 2016, 14, 1070-1085, doi:10.1111/pbi.12454.
36. Wambugu, P.W.; Brozynska, M.; Furtado, A.; Waters, D.L.; Henry, R.J. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep 2015, 5, 13957, doi:10.1038/srep13957.
37. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25, 1754-1760, doi:10.1093/bioinformatics/btp324.
38. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D., et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012, 19, 455-477, doi:10.1089/cmb.2012.0021.
39. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K., et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. Plos One 2014, 9, e112963, doi:10.1371/journal.pone.0112963.
40. Delcher, A.L.; Salzberg, S.L.; Phillippy, A.M. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics 2003, Chapter 10, 10-13, doi:10.1002/0471250953.bi1003s00.
41. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T.L. NCBI BLAST: a better web interface. Nucleic Acids Res 2008, 36, W5-W9, doi:10.1093/nar/gkn201.
42. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 2004, 20, 3252-3255, doi:10.1093/bioinformatics/bth352.
43. Chen, Y.; Ye, W.; Zhang, Y.; Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 2015, 43, 7762-7768, doi:10.1093/nar/gkv784.
44. Chan, P.P.; Lowe, T.M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. In Springer New York: New York, NY, 2019; Vol. 1962, pp 1-14.
45. Lohse, M.; Drechsel, O.; Kahlau, S.; Bock, R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 2013, 41, W575-W581, doi:10.1093/nar/gkt289.
46. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: a web server for microsatellite prediction. Bioinformatics 2017, 33, 2583-2585, doi:10.1093/bioinformatics/btx198.
47. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 2013, 30, 772-780, doi:10.1093/molbev/mst010.
48. Nguyen, L.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 2015, 32, 268-274, doi:10.1093/molbev/msu300.