

1 **Comparative analysis of chloroplast genome between wild and**
 2 **cultivated *Rehmannia glutinosa* Libosch.**

3

4 Hao Yang¹, Yohei Sasaki², Conglong Lian¹, Liping Zhang¹, Lili Wang¹, Jingfan
 5 Yang¹, Shujuan Xue¹, Suiqing Chen^{1,3*}

6

7 1 School of Pharmacy, Henan Key Laboratory of Chinese Medicine Resources and
 8 Chinese Medicine Chemistry, Henan University of Traditional Chinese Medicine,
 9 Boxue Road, Jinshui District, Zhengzhou 450046, China.

10 2 School of Medicine and Health Care, Department of Molecular Pharmacognosy,
 11 Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan.

12 3 Collaborative Innovation Center for Respiratory Disease Diagnosis and Treatment &
 13 Chinese Medicine Development of Henan Province, Henan University of Chinese
 14 Medicine, No. 156, Jinshui East Road, Zhengzhou, 450046, China.

15

16 * Corresponding author: Suiqing Chen

17 E-mail: suiqingchen0371@163.com.

ABSTRACT

Rehmanniae radix, as a tonic Chinese medicine, has a good therapeutic effect on cardiovascular and cerebrovascular diseases and central nervous system diseases, and has a good development prospect in the international market. *R.glutinosa* Libosch has a large number of source varieties, and the wild and cultivated varieties are different types of *R.glutinosa* Libosch. There has been a lot of controversy in the formulation of scientific names for wild and cultivated types of *R.glutinosa* Libosch. In order to learn more about the genetic information of wild *R.glutinosa* Libosch and explore the phylogenetic relationship between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch. In the current study, we constructed and annotated a complete circular chloroplast genome of wild *R.glutinosa* Libosch. The chloroplast genome of wild *R.glutinosa* Libosch. is 153,678 bp in length, including two inverted repeat (IR) regions of 25,759 bp, separated by a large single copy (LSC) region of 84,544 bp and a small single copy (SSC) region of 17,616 bp. The GC content of whole chloroplast genome is 37.9%. The genome contains 133 different genes, including 88 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. Among them, 23 genes contained introns, 50 SSR loci and 49 long repeats were found by Microsatellite scan, and codon use showed obvious A , T base preference. Nine variation regions were screened by mVISTA analysis and combined with DnaSP analysis, and *psaJ-rpl33* was finally identified as a region with high variation, which could be used as a candidate DNA barcode for later interspecific identification of *R.glutinosa* Libosch.. Neighbor-joining method phylogenomic analysis showed that wild *R.glutinosa*

Libosch. formed a monophyletic group, and was sister to other groups of *R.glutinosa*
*Libosch.*The results of this study revealed the chloroplast genome information of wild
R.glutinosa Libosch in detail, and found the genetic information difference between
wild and cultivated *R.glutinosa Libosch.* which laid the foundation for the variety
identification, genetic engineering, biosynthesis and other studies.

Introduction

Rehmanniae radix is the fresh or dried tuber root of *Rehmannia glutinosa*
Libosch of Scrophulariaceae. With a long history of cultivation and medicine,
Rehmanniae radix has been widely used in the clinical treatment of human diseases.
Globally, *R.glutinosa Libosch* is mainly produced in China, Japan and other places.
Chinese scholars Lou Zhicen, Wen Xuesen, Zhou Yanqing et al. have carried out
researches on the chemical composition, cultivation and genetic diversity of
R.glutinosa Libosch, completed the mechanism analysis of the main chemical
composition and pharmacological effect of different varieties of *R.glutinosa Libosch*,
improved and constantly improved the quality of *R.glutinosa Libosch*, and optimized
the planting, cultivation and production and processing methods. Meanwhile, the
genetic structure and content differences among *R.glutinosa Libosch* varieties and
regions were analyzed[1]. Japanese scholars, represented by Hoon Kitagawa et al,
have carried out analysis of rhizome components of *R.glutinosa Libosch.*, studies on
components of *R.glutinosa Libosch* from China, Japan and South Korea and chemical
composition changes before and after processing. They have reported for the first time
that fresh *R.glutinosa Libosch* contains iridoid glycosides and the main component is

catalpa, and successfully isolated four new iridoid glycosides[2]. In order to increase yield and cultivate excellent varieties, scholars from various countries have done a lot of research on the germplasm resources of *R.glutinosa* Libosch, genetic breeding, cultivation and planting, chemical composition and efficacy, extraction and analysis of active ingredients, plant tissue culture, virus-free rapid propagation and molecular biology.

By summarizing previous studies on *R.glutinosa* Libosch, it was found that as early as one thousand years ago, *R.glutinosa* Libosch started to "Wild varieties to cultivated varieties". There were obvious differences between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch in terms of microscopic internal genetic information, content of active components and macroscopic external morphology and physiological characteristics[3, 4]. Both wild and cultivated *R.glutinosa* Libosch belong to the same plant with very similar morphological characteristics. However, there were significant differences in the size of underground tubers. Therefore, different scholars have different naming and classification of *R.glutinosa* Libosch. Both Chinese Flora and Chinese Pharmacopoeia unified the scientific name of *R.glutinosa* Libosch as *R.glutinosa* Libosch from the perspective of large species. According to Xie Zongwan, the quality of wild *R.glutinosa* Libosch is different from that of cultivated *R.glutinosa* Libosch because of the different active ingredient content of the underground parts. In order to ensure the correct supply according to the drug name, the two scientific names are still suitable for separate use[5], several papers reported that the genetic basis of the cultivated varieties was narrow, the

genetic diversity at the genome level was low, the genetic differentiation between varieties was not obvious[6], and the wild populations of *R.glutinosa* Libosch had higher genetic diversity than the cultivated varieties, which could be used as the screening material for the excellent germplasm resources of *R.glutinosa* Libosch.

With the continuous development of molecular biology technology and the continuous innovation of high-throughput sequencing technology, researchers can explore and study from the level of the whole genome of organisms[7]. Currently, the whole genome sequencing of cultivated *R.glutinosa* Libosch has been completed and the genetic diversity and function of some genes have been explored[8], but there are few background reports on the genetic information of wild *R.glutinosa* Libosch. Due to its high conservation and strong species resolution, chloroplast genome has gradually become a potential super barcode for near source inter-species and intra-species identification within the genus[9]. Therefore, this study assembled the complete chloroplast genome of wild *R.glutinosa* Libosch based on the next generation sequencing technology, and analyzed the basic structure of the chloroplast genome of wild *R.glutinosa* Libosch. The phylogenetic relationship between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch was revealed by constructing phylogenetic tree, which provided molecular basis for the classification of wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch, and also provided important reference for genetic engineering, biosynthesis and variety improvement of *R.glutinosa* Libosch.

Materials and methods

Plant materials, DNA extraction, and DNA sequencing

Fresh leaves of wild *R.glutinosa* Libosch were collected from Wanxianshan (113°61'78.22"E, 35°72'27.10"N) in Xinxiang City, Henan Province, China and stored in the herbarium of Henan University of Chinese Medicine with voucher specimen number: HZYYHC15. Total genomic DNA was extracted using a Rapid Plant Genomic DNA Isolation extraction kit. The concentration of total DNA was detected quantitatively by fluorescence meter Qubit and the quality of total DNA was detected by 1% agarose gel electrophoresis. This experiment adopts the Illumina HiSeq PE150 platform, a 300 bp (insertion size) double-ended library was constructed by splicing DNA.

Genome assembly, annotation, and analysis

Quality control software (fastqc) is used to conduct quality control on raw data and quality assessment before and after quality control. Trimmomatic software was used to cut the low-quality regions in Illumina sequencing data to obtain relatively accurate and effective data. SPAdes was used to assemble the obtained cleanreads after quality control to obtain contigs. Appropriate chloroplast contigs were selected according to the similarity and coverage between sequences, and then GapFiller software was used to fill the vacant sites of each chloroplast genome contigs. Finally, sequence corrections are used in a PriSeS-G, to correct the editing errors[10]. and insertion and deletion of small fragments in the splice process, to obtain the final complete chloroplast genome sequence. The Fasta file of the chloroplast genome obtained in the above step was uploaded to the online analysis software GeSeq

(<https://chlorobox.mpimp-golm.mpg.de/geseq.html#>), and the IR reverse repeat region and rps12 trans-shear gene were selected to annotate. tRNA gene was annotated by tRNAscan-SEv2.70 software[11]. and wild *R.glutinosa* Libosch sequence published in NCBI (MW013795) was selected as the reference sequence, and the whole chloroplast genome physical map was automatically generated from the output results. Meanwhile, software MEGA 7.0 (molecular evolutionary genetics analysis 7.0) was used to analyze the chloroplast genome GC content of the sequenced samples.

Repeat sequences, SSRs, and codon usage analysis

Random repeat analysis of chloroplast genome of wild *R.glutinosa* Libosch was conducted by using the functional module of REPuter of online program BiBiServ2 (<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>) Select Forward/direct, Reverse, complementary, and Palindromic to analyze the four types of repeats. The Hamming Distance is set to 3 and the program defaults for the remaining parameters, which are 60 for Maximum Computed Repeats and 30 for Minimal Repeat Size[12]. The online program MISA (<https://webblast.ipk-gatersleben.de/misa/>) was used for SSR analysis of the chloroplast genome of wild *R.glutinosa* Liboschi, and the parameters were set as: the number of single nucleotide repeats was not less than 10, the number of double nucleotide repeats was not less than 5, and the number of trinucleotide and above repeats was not less than 4. The four, five and six nucleotide repeats were no less than 3[13]. All the protein coding sequences (CDS) of chloroplast genomes of different wild *R.glutinosa* Libosch were extracted, CDS with repetition and length less than 300bp were eliminated and codon bias analysis was performed on them by

CodonW1.4.2 (<http://downloads.fyxm.net/CodonW-76666.html>). The fasta file containing CDS sequence information was input, and the out and blk files were output to obtain the number of each codon and RSCU value of all the encoded amino acids in CDS.

Comparative analysis of chloroplast genomes

Using mVISTA (<http://genome.lbl.gov/vista/index.shtml>) online analytical tools to the wild *R.glutinosa* Libosch as reference sequence and the remaining 5 cultivated *R.glutinosa* Libosch visualization genome-wide comparison analysis[14]. DnaSP v6.0 was used to determine the nucleic acid variation values among six *R.glutinosa* Libosch chloroplast genomes. Midpoint and K(JC-Total) were used for Excel mapping, and mVISTA output was compared to mark the variation areas. Using on-line analysis software IRscope (<https://irscope.shinyapps.io/irapp/>) to upload GenBank annotation format files, generate visual boundary analysis diagram, The boundary information of the LSC region, SSC region and reverse repeat IR region of the annotated wild *R.glutinosa* Libosch chloroplast genome was compared, and the contraction and expansion of the IRa and IRb regions were analyzed.

Phylogenetic Analysis

In order to further explore the genetic relationship between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch, We selected 16 chloroplast genome sequences of *Rehmannia* and *Lysionotus pauciflorus* Maxim and *Boea hygrometrica* as outgroups, and these chloroplast genome sequences were compared by Alignment function of online software MAFFT[15]. The compared files were uploaded to

MEGA6 software, and the system evolution tree was constructed using the maximum likelihood method[16].

Results and Discussion

Genome sequencing and assembly

In general, there are many incorrect or low-quality reads in the original data obtained by sequencing. In order to ensure the accuracy of subsequent analysis, quality control must be carried out on the original data. The total number of read sequences in this sequencing was 20.66 million, the total amount of data was 2.9Gb, and the average sequence length was 145bp. Bases with base detection quality above 30 accounted for 88.82% of the total number of bases, and the GC content of the original data was 38.78%. The above indicated that the error rate of the whole sequencing base was low, which met the experimental requirements. (S1 Table).The sequence was uploaded to NCBI, and the serial number was MW007380.

Features of the wild *R.glutinosa* Libosch chloroplast genome

Like other angiosperms, the chloroplast genome of wild *R.glutinosa* Libosch has a typical tetrad structure, consisting of two reverse repeat regions IRs, a large single-copy region LSC, and a small single-copy region SSC. The total length of the splice wild *R.glutinosa* Libosch genome was 153,678bp (Fig 1). and the length of each segment was as follows: 84,544bp(LSC), 17,616bp(SSC), 25,759bp(IRs). GC content is an important characteristic of nucleic acid sequence composition and can be used as an indicator of evolution[17].The total GC content was 37.9%, the skew of AT ranged from -0.0218-0.0058, the skew of GC ranged from -0.0471-0.0405.the use

frequency of A/T was higher than that of G/C, indicating that wild *R.glutinosa* Libosch has A/T base preference, and the highest GC content was in IR region (43.1%), followed by LSC region (35.95%) and SSC region (32.17%) (S2 Table). The results of this experiment echo those of other plants[18].

Fig 1. Circular gene map of the wild *R.glutinosa* Libosch chloroplast genome.

The map contains four rings. From the center going outward, the first circle shows the forward and reverse repeats connected with red and green arcs respectively. The next circle shows the tandem repeats marked with short bars. The third circle shows the microsatellite sequences identified using MISA. The fourth circle is drawn using drawgenemap and shows the gene structure on the plastome. The genes were colored based on their functional categories.

A total of 133 genes have been annotated in the chloroplast genome of wild *Rehmannia rehmanniae*, including 88 protein-coding genes, 37 tRNA genes and 8 rRNA genes. For most terrestrial green plants, the functions of genes in the chloroplast genome can be divided into three main categories: the first category contains genes related to self-replication, which are annotated to the The second group is the genes related to photosynthesis. The third category is other genes involved in chloroplast biosynthesis and some genes with unknown functions, including genes related to photosynthetic system, transcription, translation, cytochrome complex, amino acid, etc[19]. Genes of the above three categories have been annotated in the chloroplast genome of wild *R.glutinosa* Libosch (Table 1). respectively, among which

ycf1 and rps12 genes are special. There were 2 copies of ycf1 gene, but only 1 copy was properly expressed, and the other copy lost its coding function due to failure to replicate completely and became pseudogene ycf1. rps12 was the only trans spliced gene in chloroplast genome, and its 5' end was located in LSC region and its 3' end was located in IR region. This phenomenon is very similar to findings in other plants species[20].

Table 1. List of genes in the chloroplast genome of wild *R. glutinosa* Libosch

Category for genes	Group of genes	Name of genes
Photosynthesis	Subunits of photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>
	Subunits of photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
	Subunits of cytochrome b/f complex	<i>petA, petB^b, petD^b, petG, petL, petN</i>
	Large subunit of Rubisco	<i>rbcL</i>
	Subunits of ATP synthase	<i>atpA, atpB, atpE, atpF^b, atpH, atpI</i>
	Subunits of NADH-dehydrogenase	<i>ndhA^b, ndhB^{a,b}, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
	Ribosomal RNA genes	<i>rrn16^a, rrn23^a, rrn5^a, rrn4.5^a</i>
Self-replication	Transfer RNA genes	<i>trnH-GUG, trnK-UUU^b, trnQ-UUG, trnS-GCU, trnG-UCC^b, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-GCC, trnM-CAU, trnS-GGA, trnT-UGU, trnL-UA^b, trnF-GAA, trnV-UAC^b, trnM-CAU, trnW-CCA, trnP-UGG, trnL-UAG, trnI-CAU^a, trnL-CAA^a, trnV-GAC^a, trnI-GAU^{ab}, trnA-UGC^{ab}, trnR-ACG^a, trnN-GUU^a</i>
	Small subunit of ribosome	<i>rps2, rps3, rps4, rps7^a, rps8, rps11, rps12^{a,b,c}, rps14, rps15, rps16^b, rps18, rps19</i>
	Large subunit of ribosome	<i>rpl2^{a,b}, rpl14, rpl16^b, rpl20, rpl22, rpl23^a, rpl32, rpl33, rpl36</i>
	DNA-dependent RNA polymerase	<i>rpoA, rpoB, rpoC1^b, rpoC2</i>
Other genes	Maturase	<i>matK</i>
	Envelope membrane protein	<i>cemA</i>
	Subunit of acetyl-CoA	<i>accD</i>
	C-type cytochrome synthesis gene	<i>ccsA</i>
	Protease	<i>clpP</i>
	Translational initiation factor	<i>infA</i>
unknown function	Conserved Open reading frames	<i>ycf1^a, ycf2^a, ycf3^b, ycf4, ycf15^a</i>

- Two gene copies.
- Genes containing introns.
- Genes divided into two independent transcription units.

Introns of higher plant protein-coding genes can influence gene expression

activity by splicing[21]. According to the statistical analysis of chloroplast gene introns of wild *R.glutinosa* Libosch, 23 genes contained introns, among which rps12, ycf3 and clpP contained 2 introns, and the remaining 19 genes contained only 1 intron. (S3 Table). The statistical analysis of the length of introns and exons showed that, Most of the intron lengths in the chloroplast genome of wild *R.glutinosa* Libosch are between 600-900bp, and The trnK-UUU gene intron has the longest codon length of 2525, which is consistent with the results of other studies[22].At the same time, studies have shown that intron length plays an important role in gene regulation or expression[23]. Some scholars have confirmed in rice and Arabidopsis that high-expressed genes contain more and longer introns than low-expressed genes[24]. Therefore, it is reasonable to speculate that trnK-UUU may be a highly expressed gene in wild *R.glutinosa* Libosch.

Codon Usage

In the long-term biological evolution, the combined effect of gene selection and mutation leads to the difference in the use frequency of most biological codons[25]. Each species has a unique codon use preference[26]. In this experiment, protein-coding genes and transfer RNA coding genes after the preliminary data screening were selected as the analysis objects, with a length of 67,445 bp. Containing 22,472 codons, the synonymous codon frequencies of different amino acids in wild *R.glutinosa* Libosch. chloroplast genome are significantly different (Figure 2). RSCU value is the ratio between the usage frequencies of specific codons, which is used to detect the non-uniform use of synonymous codons within the coding sequence. The

codon with an RSCU value greater than 1.00 is set to 1.00, and the codon with an RSCU value less than 1.00 is lower than expected. The RSCU value of methionine (Met) and tryptophan (Trp) is 1, so there is no codon with a bias. Leucine (Leu) had the highest coding rate among the 20 amino acids, and the number of codons used was 2,353, accounting for 10.47% of the total, while tryptophan (Trp) had the lowest coding rate and the number of codons used was 419 (1.86%). The analysis of the use of four bases on the third codon of the chloroplast genome of wild *R.glutinosa* Libosch. showed that GC3 was 30.19%, which was similar to the GC content of the whole chloroplast genome in the previous paper[27]. further confirming that the codon of wild *R.glutinosa* Libosch chloroplast genome preferred to end in A and T.(Table 2).

Fig 2. Codon bias map of the wild *R.glutinosa* Libosch chloroplast genome. The x-coordinate is the abbreviation for amino acid and the y-coordinate is the relative synonymous codon usage.

Table 2. Codon-anticodon recognition patterns and codon usage in the chloroplast genome of wild *R.glutinosa* Libosch

abbr	Amino Acid	Codon	No.	RSCU	tRNA	abbr	Amino Acid	Codon	No.	RSCU	tRNA
*	Ter	UAA	39	1.39		M	Met	AUG	523	1.00	trnfM-CAU
										trnM-CAU	
										trnI-CAU	
		UAG	31	1.11				N	Asn	AAC	229
A	Ala	UGA	14	0.50		P	Pro	AAU	819	1.56	
		GCA	348	1.12	trnA-UGC			CCA	257	1.08	trnP-UGG
		GCC	205	0.66				CCC	185	0.78	
		GCG	150	0.48				CCG	142	0.60	

		GCU	542	1.74				CCU	364	1.54	
C	Cys	UGC	65	0.51	trnC-GCA	Q	Gln	CAA	600	1.49	trnQ-UUG
		UGU	190	1.49				CAG	208	0.51	
D	Asp	GAC	171	0.38	trnD-GUC	R	Arg	AGA	413	1.79	trnR-UCU
		GAU	732	1.62				AGG	147	0.64	
E	Glu	GAA	914	1.52	trnE-UUC			CGA	311	1.35	
		GAG	288	0.48				CGC	100	0.43	
F	Phe	UUC	413	0.66	trnF-GAA			CGG	119	0.52	
		UUU	843	1.34				CGU	291	1.26	trnR-ACG
G	Gly	GGA	612	1.57	trnG-UCC	S	Ser	AGC	98	0.34	trnS-GCU
		GGC	179	0.46	trnG-GCC			AGU	355	1.23	
		GGG	274	0.70				UCA	327	1.14	trnS-UGA
		GGU	493	1.27				UCC	285	0.99	trnS-GGA
H	His	CAC	120	0.45	trnH-GUG			UCG	191	0.66	
		CAU	413	1.55				UCU	471	1.64	
I	Ile	AUA	551	0.88		T	Thr	ACA	334	1.18	trnT-UGU
		AUC	394	0.63	trnI-GAU			ACC	217	0.77	trnT-GGU
		AUU	924	1.48				ACG	121	0.43	
K	Lys	AAA	871	1.50	trnK-UUU			ACU	460	1.63	
		AAG	287	0.50		V	Val	GUA	472	1.49	trnV-UAC
L	Leu	CUA	326	0.83	trnL-UAG			GUC	150	0.48	trnV-GAC
		CUC	155	0.40				GUG	164	0.52	
		CUG	154	0.39				GUU	477	1.51	
		CUU	494	1.26		W	Trp	UGG	419	1.00	trnW-CCA
		UUA	750	1.91	trnL-UAA	Y	Tyr	UAC	159	0.39	trnY-GUA
		UUG	474	1.21	trnL-CAA			UAU	648	1.61	

SSRs and Long-Repeat Analysis

Microsatellite sequences are special sequences with highly variable length that can be distributed at different positions in the whole genome, and are widely used in species identification, population genetics and phylogenetic polymorphism studies[28, 29]. In this study, a total of 50 SSR loci were identified, including 30 single base repeats, 3 double base repeats, 2 three-base repeats, 8 four-base repeats, 3 five-base repeats and 4 complex base repeats, which were not annotated to 6 base repeats. Overall, SSRs in chloroplast genome of wild *R.glutinosa* Libosch have the most single nucleotide repeats and are dominated by bases A and T accounting for 60% of

the total SSRs. that is, SSRs are usually composed of short A or T repeats, while few are composed of tandem G or C repeat. This is consistent with the A and T base bias in wild *R.glutinosa* Libosch base composition mentioned above, which may be related to the fact that A/T is more variable than G/C(Table 3).

Table 3. SSRs in the chloroplast genome of *wild R.glutinosa* Libosch

SSR nr.	SSR type	SSR	size	start	end
1	p1	(T) 10	10	7762	7771
2	p1	(T) 11	11	12862	12872
3	p1	(T) 10	10	14272	14281
4	p1	(T) 11	11	18331	18341
5	p1	(T) 12	12	35326	35337
6	p1	(A) 14	14	44269	44282
7	p1	(T) 10	10	46168	46177
8	p1	(T) 10	10	46825	46834
9	p1	(T) 10	10	52727	52736
10	p1	(T) 10	10	54726	54735
11	p1	(T) 12	12	59212	59223
12	p1	(A) 10	10	59743	59752
13	p1	(A) 10	10	61804	61813
14	p1	(A) 10	10	63746	63755
15	p1	(T) 12	12	65904	65915
16	p1	(T) 11	11	66081	66091
17	p1	(T) 10	10	66705	66714
18	p1	(A) 10	10	67066	67075
19	p1	(T) 11	11	69685	69695
20	p1	(A) 11	11	71821	71831
21	p1	(A) 10	10	74653	74662
22	p1	(A) 10	10	75672	75681
23	p1	(T) 10	10	78628	78637
24	p1	(T) 11	11	80528	80538
25	p1	(A) 10	10	84560	84569
26	p1	(A) 15	15	113420	113434
27	p1	(T) 11	11	125458	125468
28	p1	(T) 10	10	126116	126125
29	p1	(T) 12	12	126234	126245
30	p1	(T) 10	10	153654	153663
31	p2	(TA) 7	14	8192	8205
32	p2	(AT) 5	10	19704	19713
33	p2	(TA) 6	12	46328	46339

34	p3	(TTC) 4	12	35102	35113
35	p3	(TAA) 4	12	42436	42447
36	p4	(ATTG) 3	12	5798	5809
37	p4	(CAAT) 3	12	27717	27728
38	p4	(ATTT) 3	12	29197	29208
39	p4	(TCTA) 3	12	31971	31982
40	p4	(AAAC) 3	12	68626	68637
41	p4	(GAAA) 3	12	112830	112841
42	p4	(AATA) 3	12	115555	115566
43	p4	(TTCT) 3	12	119386	119397
44	p5	(TAAAA) 3	15	260	274
45	p5	(ATTAG) 3	15	99646	99660
46	p5	(TAATC) 3	15	138559	138573
47	c	(A) 11tg (A) 10	23	36134	36156
48	c	(AATT) 3actaattaacagagccaaacctacaatttttggataaaat (A) 11	63	36522	36584
49	c	(GAAA) 3tttc (T) 11	27	61165	61191
50	c	(TA) 5cgaaatacagaacccatcataaaaaaatgagtatttttggga aataactcgcaaagaggggatgta (T) 11	86	67424	67509

285

286 Repeated sequences are widespread and ubiquitous in plant genomes[30].

287 Studies have shown that environmental changes can cause heritable changes in the

288 number of repeats, which may imply that repeats are related to plant adaptability[31].

289 Long repeats include forward repeats, palindromic repeats, reverse repeats and

290 complementary repeats. The quantitative distribution of long repeats in wild

291 *R.glutinosa* Libosch showed that there were 49 repeats in total, and there were four

292 types of repeats, mainly palindromic repeats and forward repeats, accounting for

293 55.1% and 34.7% respectively, and only one complementary repeat accounted for 2%.

294 (Table 4). It was found that 3 repeats were located in the IRa region, 20 in the IRb

295 region, 21 in the LSC region, and 5 in the SSC region. Therefore, repeats were

296 distributed unevenly in the four regions, among which 71.4% were located in the gene

297 coding region and 28.6% in the gene spacer region. The DNA repeats screened in this

study can be used as molecular markers, genetic map construction and variety identification of wild *R. glutinosa* Libosch.

Table 4. Long repeat sequences identified in the chloroplast genome of wild *R. glutinosa* Libosch

No.	Repeat size	Type	Repeat 1	Repeat 2	E-Value	Gene	Region
1	60	F	91742	91760	4.62×10^{-21}	ycf2	IRb
2	60	P	91742	146404	4.62×10^{-21}	ycf2	IRb;IRa
3	60	P	91760	146422	4.62×10^{-21}	ycf2	IRb;IRa
4	60	F	146404	146422	4.62×10^{-21}	ycf2	IRa
5	58	P	115184	115184	8.00×10^{-26}	IGS(ccsA,ndhD)	SSC
6	47	F	91759	91777	3.26×10^{-15}	ycf2	IRb
7	47	P	91759	146400	3.26×10^{-15}	ycf2	IRb;IRa
8	47	P	91777	146418	3.26×10^{-15}	ycf2	IRb;IRa
9	46	P	78262	78262	1.25×10^{-14}	IGS(petD,rpoA)	LSC
10	45	P	113321	113321	2.06×10^{-12}	IGS(rpl32,trnL-UAG)	SSC
11	41	F	98745	120011	1.69×10^{-13}	IGS(rps12,trnV-GAC); ndhA-intron1	IRb;SSC
12	41	P	120011	139438	1.69×10^{-13}	ndhA-intron1; IGS(trnV-GAC-2,rps12)	SSC;IRa
13	39	F	43606	120013	2.57×10^{-12}	ycf3-intron1; ndhA-intron1	LSC;SSC
14	39	F	43606	98747	1.47×10^{-10}	ycf3-intron1; IGS(rps12,trnV-GAC)	LSC;IRb
15	39	P	43606	139438	1.47×10^{-10}	ycf3-intron1; IGS(trnV-GAC-2,rps12)	LSC;IRa
16	39	P	123158	123158	5.42×10^{-9}	IGS(rps15,ycf1);	SSC
17	38	P	74837	74843	2.00×10^{-8}	IGS(psbT,psbN)	LSC
18	36	P	44599	62162	1.52×10^{-10}	IGS(ycf3,trnS-GGA); cemA	LSC
19	35	F	95707	120016	9.94×10^{-7}	ndhB-intron1; ndhA-intron1	IRb;SSC
20	35	P	120016	142482	9.94×10^{-7}	ndhA-intron1; ndhB-2-intron1	SSC;IRa
21	32	R	46908	46908	1.61×10^{-6}	IGS(trnT-UGU, trnL-UAA)	LSC
22	32	F	8028	35475	4.82×10^{-5}	trnS-GCU;trnS-UGA	LSC
23	32	F	91756	91792	4.82×10^{-5}	ycf2	IRb
24	32	P	91756	146400	4.82×10^{-5}	ycf2	IRb;IRa

25	32	P	91792	146436	4.82×10^{-5}	ycf2	IRb;IRa
26	32	F	146400	146436	4.82×10^{-5}	ycf2	IRa
27	31	R	99635	99635	6.03×10^{-6}	IGS(rps12,trnV-GAC)	IRb
28	31	C	99635	138558	6.03×10^{-6}	IGS(rps12,trnV-GAC)	IRb;IRa
29	31	R	138558	138558	6.03×10^{-6}	IGS(trnV-GAC,rps12)	IRa;IRa
30	31	P	55006	66311	1.75×10^{-4}	IGS(atpB,rbcl); IGS(psbE,petL)	LSC
31	31	F	75672	84558	1.75×10^{-4}	petB-intron1; IGS(rps19,rpl2)	LSC;IRb
32	31	P	75672	153635	1.75×10^{-4}	petB-intron1; IGS(rpl2-2,trnH-GUG)	LSC;IRa
33	31	P	76129	120013	1.75×10^{-4}	petB-intron1; ndhA-intron1	LSC;SSC
34	30	P	8030	45203	5.76×10^{-9}	trnS-GCU;trnS-GGA	LSC
35	30	F	91772	91790	5.76×10^{-9}	ycf2	IRb
36	30	P	91772	146404	5.76×10^{-9}	ycf2	IRb;IRa
37	30	P	91790	146422	5.76×10^{-9}	ycf2	IRb;IRa
38	30	P	29557	29597	2.26×10^{-9}	IGS(petN,psbM)	LSC
39	30	P	75673	75673	2.26×10^{-9}	petB-intron1	LSC
40	30	R	5320	75672	6.32×10^{-4}	rps16-intron1; petB-intron1	LSC
41	30	F	9553	36383	6.32×10^{-4}	trnG-UCC-exon2;trnG-GCC	LSC
42	30	P	35477	45203	6.32×10^{-4}	trnS-UGA;trnS-GGA	LSC
43	30	F	43618	98759	6.32×10^{-4}	ycf3-intron1; IGS(rps12,trnV-GAC)	LSC;IRb
44	30	F	43618	120025	6.32×10^{-4}	ycf3-intron1; ndhA-intron1	LSC;SSC
45	30	P	43618	139435	6.32×10^{-4}	ycf3-intron1; IGS(trnV-GAC,rps12)	LSC;IRa
46	30	F	89336	89378	6.32×10^{-4}	ycf2	IRb
47	30	P	89336	148816	6.32×10^{-4}	ycf2	IRb;IRa
48	30	P	89378	148858	6.32×10^{-4}	ycf2	IRb;IRa
49	30	F	99630	99635	6.32×10^{-4}	IGS(rps12,trnV-GAC)	IRb

F: forward repeats, P: palindromic repeats, R: reverse repeats, C: complementary repeats

Intra-specific Comparative Analysis of Genomic Structure

Phenotypic variation of wild and cultivated *R.glutinosa* Libosch is obvious, and

phenotypic variation can be analyzed from the two perspectives of environment and gene. There are many environmental factors and it is difficult to control them. Therefore, most scholars have carried out genetic diversity analysis of *R.glutinosa* Libosch with different germplasm and between different regions. However, short fragments carry less genetic information and cannot fully reflect the differences between *R.glutinosa* Libosch species[32]. Therefore, the chloroplast genome of wild and cultivated *R.glutinosa* Libosch was analyzed in this paper. The results showed that the number of genes, protein coding genes, tRNA and rRNA of wild *R.glutinosa* Libosch and other cultivated *R.glutinosa* Libosch were consistent in terms of the basic structure and composition of the genome. However, the total length of chloroplast genome of wild *R.glutinosa* Libosch was longer than that of cultivated *R.glutinosa* Libosch, and the main difference was in the LSC region. (S4 Table). According to the functional classification of the above genes, it was found that the genes related to photosystem I, photosystem II and cytochrome synthesis were mainly distributed in LSC. Therefore, it was reasonable to presume that this phenomenon might be related to chloroplast photosynthesis.

In order to further explore the size of variation between wild and cultivated species of *R.glutinosa* Libosch, using wild *R.glutinosa* Libosch as the reference genome, the chloroplast genome sequences of 6 *R.glutinosa* Libosch were globally compared and analyzed by mVISTA software (Fig 3). The results showed that the six chloroplast genome sequences had high similarity and basically the same sequence, among which the variation in the non-coding region (red part) was higher than that in

the conserved protein coding region (blue part), the variation in the LSC region and the IR region was significantly higher than that in the SSC region and the IR region, and the rRNA gene was highly conserved without variation. It can be seen from the figure that *ycf3* and *ycf1* are highly variable genes, and other genes are highly conserved. Most of the genes were more than 90% similar, and the gene interregions with large variation included *rpoB-trnC-GCA*, *psbC-trnS-UGA*, *psaA-ycf3*, *ycf3-trnS-GGA*, *ycf4-cemA*, *trnP-psaJ*, *psaJ-rpl33*, and *rps12-trnV-GAC*, *rps15-ycf1*. These results are mostly consistent with those reported for the chloroplast genomes of other plant species[33]. these genes and regions provide strong support for further study of *R.glutinosa* Libosch phylogeny, genetic diversity and species identification. These regions can be used to distinguish the variation sites between wild and cultivated *R.glutinosa* Libosch and are candidates for barcode design.

Fig 3. Sequence alignment of five *R.glutinosa* Libosch chloroplast genomes using mVISTA. the annotation of wild *R.glutinosa* Libosch was used as the reference. Gray arrows above the alignment indicate the position and direction of transcription of each gene. The scale on the vertical axes represents the percent sequence identity between 50% and 100%.

DnaSP software was used to analyze and detect the chloroplast genome of wild and cultivated *R.glutinosa* Libosch. The degree of nucleic acid variation (K value) was calculated by pairings to show the degree of variation at the sequence level. The results showed that the variation degree of the chloroplast genome IR region was

significantly lower than that of the LSC region and the SSC region, and the K value was generally below 0.001. Generally, the K value greater than 0.04 was set as the highly variable gene, and the variation value within the *R.glutinosa* Libosch species was all lower than this value, which proved that the variation within the *R.glutinosa* Libosch species was small. At the same time, it can be found that there are certain nucleic acid variation between different varieties of cultivated *R.glutinosa* Libosch, but the variation is smaller than that between wild and cultivated *R.glutinosa* Libosch. Among them, five regions had nucleic acid variation values greater than 0.015, all of which were located in the LSC region, namely petN-psbM, psbD-trnT-GGU,rps14-psaB, ycf3, and psaJ-rpl33. (Fig 4). Combined with the above mVISTA analysis, Both of them were annotated to the psaJ-rpl33 region and the maximum nucleic acid variation value was 0.02384. Therefore, psaJ-rpl33 could be used as a candidate DNA barcode for later *R.glutinosa* Libosch variety identification.

Fig 4. Sliding window analysis of five complete *R.glutinosa* Libosch chloroplast genomes. Window length: 800 bp; step size: 800 bp. X-axis: Nucleotide positions in the chloroplast genomes. Y-axis: K values of nucleotide diversity. A: wild *R.glutinosa* Libosch from XinXiang, Henan province (MW007380) , B: cultivated *R.glutinosa* Libosch (MW007382) , C: cultivated *R.glutinosa* Libosch (MW007383) D:cultivated *R.glutinosa* Libosch (MW007384) E: cultivated *R.glutinosa* Libosch (MW007379) F: cultivated *R.glutinosa* Libosch (MW007381) Total: the total nucleotide diversity value of A vs B, A vs C, A vs D, A vs E and A vs F.

The different size of chloroplast genome is mainly caused by the expansion and contraction of the boundaries of the four structures to different degrees[34]. We conducted a comparative analysis of the IR boundaries of the whole genome of the 6

kinds of *R.glutinosa* Libosch in this study. The results showed that: The JLB(LSC/IRb), JSB(IRb/SSC), JSA(SSC/IRa) and JLA(LSC/IRa) boundary genes of 6 *R.glutinosa* Libosch were completely consistent. In this study, the JLB boundary was rps19 and rpl2 genes, the JSB boundary was ycf1 and ndhF genes, and the JLA boundary was rpl2 and trnH genes. The JSA boundary is two genes, ycf1 and trnN (Fig 5), among which ycf1 gene extends to overlap with IRa boundary, and the overlap length range is 1083bp. These overlapping fragments made ycf1 gene have a pseudogene fragment of ycf1 at the IRa/SSC boundary. Meanwhile, it was found that wild *R.glutinosa* Libosch ycf1 gene had gene contraction at the SSC/IRa boundary compared with other 5 cultivated *R.glutinosa* Libosch species. The length of IR in the SSC region was 6bp shorter than that in other genomes. Therefore, the overall IR length of the six *R.glutinosa* Libosch was highly conserved, but there was still a slight difference in the boundary region between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch.

Fig 5. Comparison of the LSC, SSC, and IR border regions among the six *R.glutinosa* Libosch chloroplast genomes of examined in this study.

Phylogenetic analysis

Chloroplast genome is of great significance for phylogenetic studies[35]. In this study, the chloroplast genome sequences of 10 *Rehmannia* genus and other groups downloaded from GenBank and the chloroplast genome sequences of 6 *R.glutinosa* Libosch species measured in this study were used as the features-based construction

method, The maximum likelihood method (ML method) was used to construct the phylogenetic tree, in which *Lysionotus pauciflorus* Maxim and *Boea hygrometrica* were exotaxa. In the resulting phylogenetic tree bootstrap (BS) analysis showed that 11 nodes were strongly supported with BS values $\geq 90\%$, and 8 nodes had BS values of 100% (Fig 6). As can be seen from the results, the wild *R.glutinosa* Libosch measured in this study was first clustered into a single branch with 100% support rate, and then clustered into a branch with 100% support rate with 5 cultivated *R.glutinosa* Libosch self-sequenced and NCBI downloaded *R.glutinosa* Libosch, and finally clustered into a branch with *Rehmannia* species downloaded from GenBank database. Therefore, it can be clearly seen that there are some differences between wild *R.glutinosa* Libosch and cultivated *R.glutinosa* Libosch, which may be due to the change of gene structure in the long-term evolution process, which is consistent with the phenomenon of base loss in PsbA-trnH of wild and cultivated *R.glutinosa* Libosch discovered by our previous research team, and the evolutionary relationship between different species in the genus *Rehmannia* can be found. *Rehmannia henryi* and *Rehmannia solanifolia* formed the basal group of the evolutionary system as sister groups, *Rehmannia piasezkii* Maxim. differentiated successively, and *Rehmannia elata*. and *Rehmannia chingii*. sister groups formed the evolutionary group. The interspecific relationship of *Rehmannia* in this study is consistent with the results of previous studies[36].

Fig 6. Phylogenetic tree based on the complete chloroplast genomes from

Rehmannia species constructed using the maximum likelihood method. The *Lysionotus pauciflorus* Maxim. and *Boea hygrometrica*. cpDNA sequence were used as the outgroup. Bootstrap values are shown at the nodes.

Conclusions

The comparative study of wild and cultivated products is one of the important contents of traditional Chinese medicine resources, and the accurate identification and differentiation of species is the premise and basis of biological research and resource conservation. By comparing the chloroplast genome differences between wild and cultivated *R. glutinosa* Libosch, this study proved that there were not only morphological differences but also significant differences in the gene level between wild and cultivated *R. glutinosa* Libosch. The phylogenetic tree constructed from the whole chloroplast genome can successfully distinguish wild *R. glutinosa* Libosch from other cultivated *R. glutinosa* Libosch, which provides molecular basis for the classification of *R. glutinosa* Libosch. At the same time, it can be seen that the chloroplast genome contains more genes and more information of loci than the previous short DNA barcode. It can make up for the lack of information loci in the combined analysis of a few chloroplast fragments, which leads to the low resolution and low support rate, and has a stronger ability of species differentiation at the level below *R. glutinosa* Libosch species. Therefore, the chloroplast genome has the potential as a super DNA barcode for interspecific identification of *R. glutinosa* Libosch.

The specificity of wild *R. glutinosa* Libosch should be paid more attention by

relevant researchers, which is not only conducive to the research on the germplasm protection of *R.glutinosa* Libosch, but also has important reference value for the improvement of the quality of *R.glutinosa* Libosch. The continuous development of molecular biotechnology provides new technologies and new methods for this research. This study described in detail the introns, codon, repeats and other important genetic information of wild *R.glutinosa* Libosch. At the same time, the high variation region of *psaJ-rpl33* was screened, and the DNA barcoding study of this segment could be carried out in the later stage, which not only laid a solid foundation for the smooth development of future genetic engineering and biosynthesis research of *R.glutinosa* Libosch. It also provided an important reference for the identification and improvement of *R.glutinosa* Libosch varieties.

Acknowledgments

We are grateful to the members of the lab for their assistance and helpful discussions.

References

1. Zhou YQ, Yao HL, Zhou CE, Duan HY, Lu SX. Reveiw on *Rehmannia glutinosa* breeding Guihaia. 2010;30(03):373-6. doi: 10.3969/j.issn.1000-3142.2010.03.018 (In Chinese)
2. Kitagawa isao, Nishimura sei, Kobayashi amiko, Yoshioka itiro. Rhizome components of *Rehmannia* Huaiqing. Journal of Pharmaceutical Sciences. 1971;91(5):593-6.3.(In Japanese)
3. Wang L, Liu Y, Sun H, Li XE. Comparison of Carbohydrate Content Variation Between Wild Species and Cultivars of *Rehmannia glutinosa*. Chinese Agricultural Science Bulletin. 2013;29(31):7. doi:10.3969/j.issn.1000-6850.2013.31.027 (In Chinese)
4. Li XE, Sun P, Qi JJ, Zhou LL, Wang SS. Changes of Hormones in Cultivars and Wild-Type Varieties of *Rehmannia glutinosa* Libosch. ACTA Agronomica Sinica. 2013;(07):1276-83.doi:10.3724/SP.J.1006.2013.01276 (In Chinese)
5. Xie ZW. genuine regional drug. Journal of Chinese Medicine. 1990;(10):43-6. doi: 10.13288/j.11-2166/r.1990.10.020. (In Chinese)

6. Li CC, Hu SW, Xia Z. Genetic diversity analysis on cultivar and wild population of *Rehmannia glutinosa* based on ISSR markers. *Chinese Traditional and Herbal Drugs*. 2020;51(23):6054-61.doi:10.7501/j.issn.0253-2670.2020.23.019 (In Chinese)
7. Chen XC, Dong LL, Yao H, Chen SL, Han JP. Study on metagenomics and genuine traditional Chinese medicinal materials. *Chinese Traditional and Herbal Drugs*. 2012;43(12):6. doi:CNKI:SUN:ZCYO.0.2012-12-002 (In Chinese)
8. Lma B, Cda B, Chi SC, Xwa B, Xza B, Yan ND, et al. De novo genome assembly of the potent medicinal plant *Rehmannia glutinosa* using nanopore technology. *Computational and Structural Biotechnology Journal*. 2021; doi:19:3954-63.10.1016/j.csbj.2021.07.006
9. Zhang Z, Zhang Y, Song M, Guan Y, Ma X. Species Identification of *Dracaena* Using the Complete Chloroplast Genome as a Super-Barcode. *Frontiers in Pharmacology*. 2019; doi:10.3389/fphar.2019.01441
10. Wang Q, Liu YD, Liu Q, Liu X, Yang F, Wang M, et al. Isolation and Complete Genome of the Marine *Pseudoalteromonas* Phage C7 from Coastal Seawater of Yellow Sea, China *Curr Microbiol*.2020;77: 279-285.doi:10.1007/s00284-019-01815-4
11. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*. 2005;33(suppl_2):W686-W9. doi:10.1093/nar/gki366.
12. Stefan K, Choudhuri JV, Enno O, Chris S, Jens S, Robert G. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*. 2001;29(22):4633-42. doi:10.1093/nar/29.22.4633
13. Sebastian B, Thomas T, Thomas M, Uwe S, Martin M. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. 2017;33(16):2583-5. doi:10.1093/bioinformatics/btx198
14. Frazer KA, Lior P, Alexander P, Rubin EM, Inna D. VISTA: computational tools for comparative genomics. *Nucleic Acids Research*. 2004;32(suppl_2):W273-W9. doi:10.1093/nar/gkh458
15. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*. 2019;20(4):1160-6. doi:10.1093/bib/bbx108
16. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology & Evolution*. 2013;30(12):2725-9. doi:10.1093/molbev/mst197
17. Smarda P, Bures P, Horova L, Leitch IJ, Mucina L, Pacini E, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(39):E4096-102. doi:10.1073/pnas.1321152111
18. Xu J, Shen X, Liao B, Xu J, Hou D. Comparing and phylogenetic analysis chloroplast genome of three *Achyranthes* species. *Scientific reports*. 2020;10(1):10818. doi:10.1038/s41598-020-67679-y
19. Jansen RK, Raubeson LA, Boore JL, Depamphilis CW, Cui L. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology*. 2005;395:348-84. doi:10.1016/S0076-6879(05)95020-9
20. Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, et al. The Complete Chloroplast Genome Sequence of the Medicinal Plant *Salvia miltiorrhiza*. *Plos One*. 2013;8(2):e57607. doi:10.1371/journal.pone.0057607
21. Shaul O. How introns enhance gene expression. *The International Journal of Biochemistry & Cell*

- 518 Biology. 2017;91:145-55. doi:10.1016/j.biocel.2017.06.016.
- 519 22. Liang C, Wang L, Lei J, Duan B, Ma W, Xiao S, et al. A Comparative Analysis of the Chloroplast
520 Genomes of Four *Salvia* Medicinal Plants. *Engineering*. 2019;5(5):907-15.
521 doi:10.1016/j.eng.2019.01.017
- 522 23. Kalari KR, Casavant M, Bair TB, Keen HL, Scheetz TE. First exons and introns--a survey of GC
523 content and gene structure in the human genome. *Silico Biology*. 2006;6(3):237.
524 doi:10.1017/S0038713412003089
- 525 24. Ren X-Y, Vorst O, Fiers MWEJ, Stiekema WJ, Nap J-P. In plants, highly expressed genes are the
526 least compact. *Trends in Genetics*. 2006;22(10):528-32. doi: 10.1016/j.tig.2006.08.008.
- 527 25. Liu H, Lu Y, Lan B, Xu J. Codon usage by chloroplast gene is bias in *Hemiptelea davidii*. *Journal*
528 *of Genetics*. 2020;99(1):1-11. doi: 10.1007/s12041-019-1167-1
- 529 26. Wei L, He J, Jia X, Qi Q, Liang Z, Zheng H, et al. Analysis of codon usage bias of mitochondrial
530 genome in *Bombyx mori* and its relation to evolution. *Bmc Evolutionary Biology*.
531 2014;14(1):1-12. doi:10.1186/s12862-014-0262-4
- 532 27. Wang WB, Yu H, Lei WJ, Gao JH, Qiu XP, Wang JS. The Complete Chloroplast Genome
533 Sequences of the Medicinal Plant *Forsythia suspensa* (Oleaceae). *INT J MOL SCI*.
534 2017;18(11):2288. doi:10.3390/ijms18112288
- 535 28. Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA. Polymorphic simple sequence
536 repeat regions in chloroplast genomes: applications to the population genetics of pines.
537 *Proceedings of the National Academy of Sciences*. 1995;92(17):7759-63.
538 doi:10.1073/pnas.92.17.7759
- 539 29. Meng LZ, Chao X, Liu HW, Li Y, Mai CY, Yu LQ, et al. The impact of modern plant breeding on
540 dominant Chinese wheat cultivars (*Triticum aestivum* L.) revealed by SSR and functional
541 markers. *Genetic resources and crop evolution*. 2018;65:55-65. doi:10.1007/s10722-017-0508-2
- 542 30. Maria ME, Kaul T, Deininger P. Long-Distance Relationships: Suppression of Repeat-Mediated
543 Deletions. *TRENDS IN GENETICS*. 2018;34(8):572-4. <https://doi.org/10.1016/j.tig.2018.05.003>
- 544 31. Lapitan N, Sears RG, Gill BS. Amplification of repeated DNA sequences in wheat X rye hybrids
545 regenerated from tissue culture. *Theoretical & Applied Genetics*. 1988;75(3):381-8.
546 doi:10.1007/BF00276739
- 547 32. Alzahrani D, Albokhari E, Yaradua S, Abba A. Complete Chloroplast Genome Sequence of
548 *Dipterygium glaucum* and *Cleome chrysantha* and Other Cleomaceae Species, Comparative
549 Analysis and Phylogenetic Relationships. *Saudi Journal of Biological Sciences*.
550 2021;28(4):2476-90.doi:10.1016/j.sjbs.2021.01.049
- 551 33. Li B, Lin F, Huang P, Guo W, Zheng Y. Complete Chloroplast Genome Sequence of *Decaisnea*
552 *insignis*: Genome Organization, Genomic Resources and Comparative Analysis. *Scientific reports*.
553 2017;7(1):10073. doi:10.1038/s41598-017-10409-8
- 554 34. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the
555 inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *Bmc*
556 *Evolutionary Biology*. 2008;8. doi:10.1186/1471-2148-8-36
- 557 35. Sun J, Wang Y, Liu Y, Xu C, Yuan Q, Guo L, et al. Evolutionary and phylogenetic aspects of the
558 chloroplast genome of *Chaenomeles* species. *Scientific reports*. 2020;10(1):11466.
559 doi:10.1038/s41598-020-67943-1
- 560 36. Xiao JJ, Wang XQ, Li CQ, Liu HZ, Chen F. Phylogeny of *Rehmannia* and Related Genera Based
561 on Chloroplast Genome. *Molecular Plant Breeding*. 2020;18(8):7.doi:10.13271/j.mpb.018.002527

562 (In Chinese)

563

564 **Supporting information**

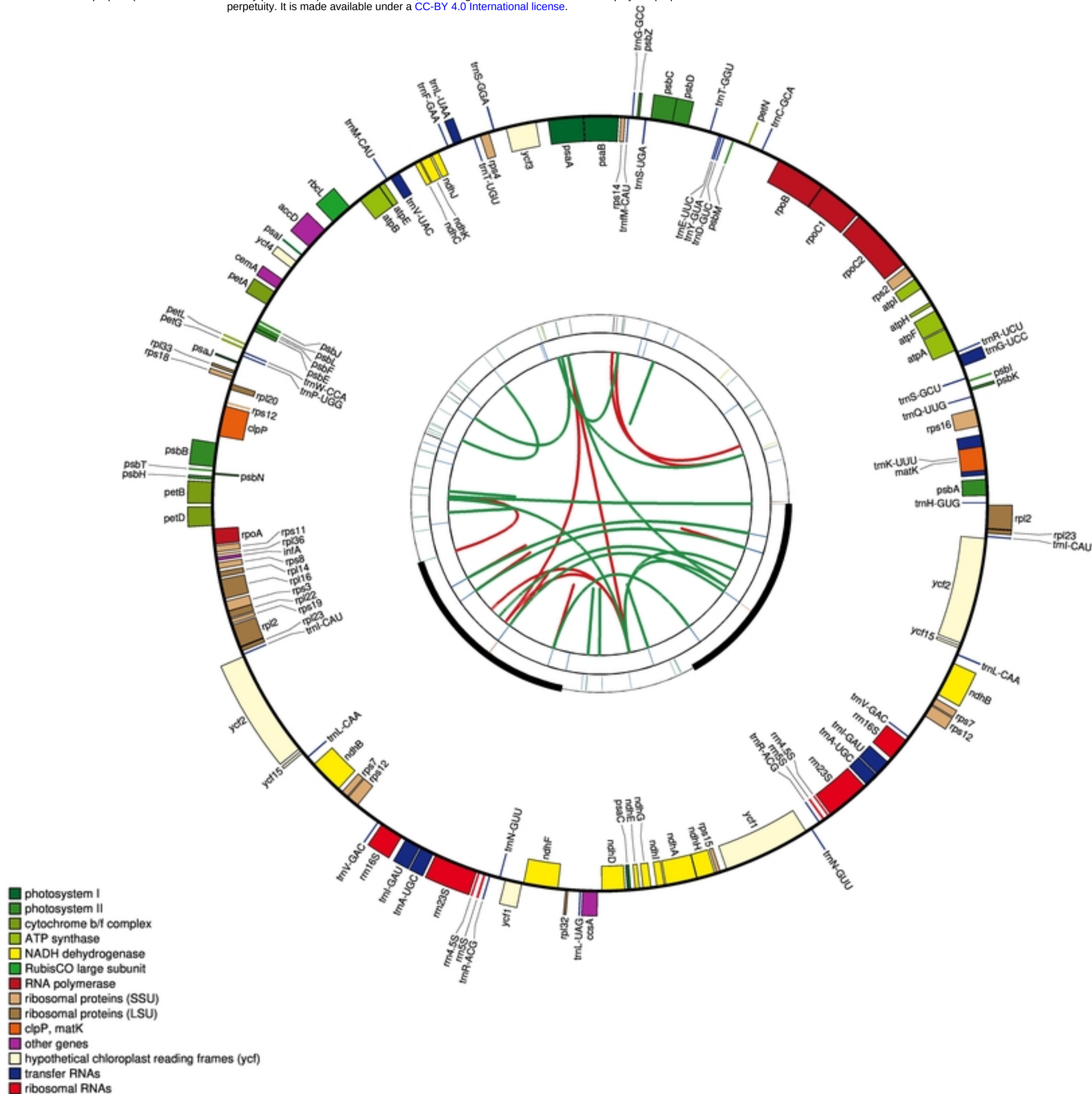
565 **S1 Table. Raw data quality information of wild *R.glutinosa* Libosch chloroplast genome**

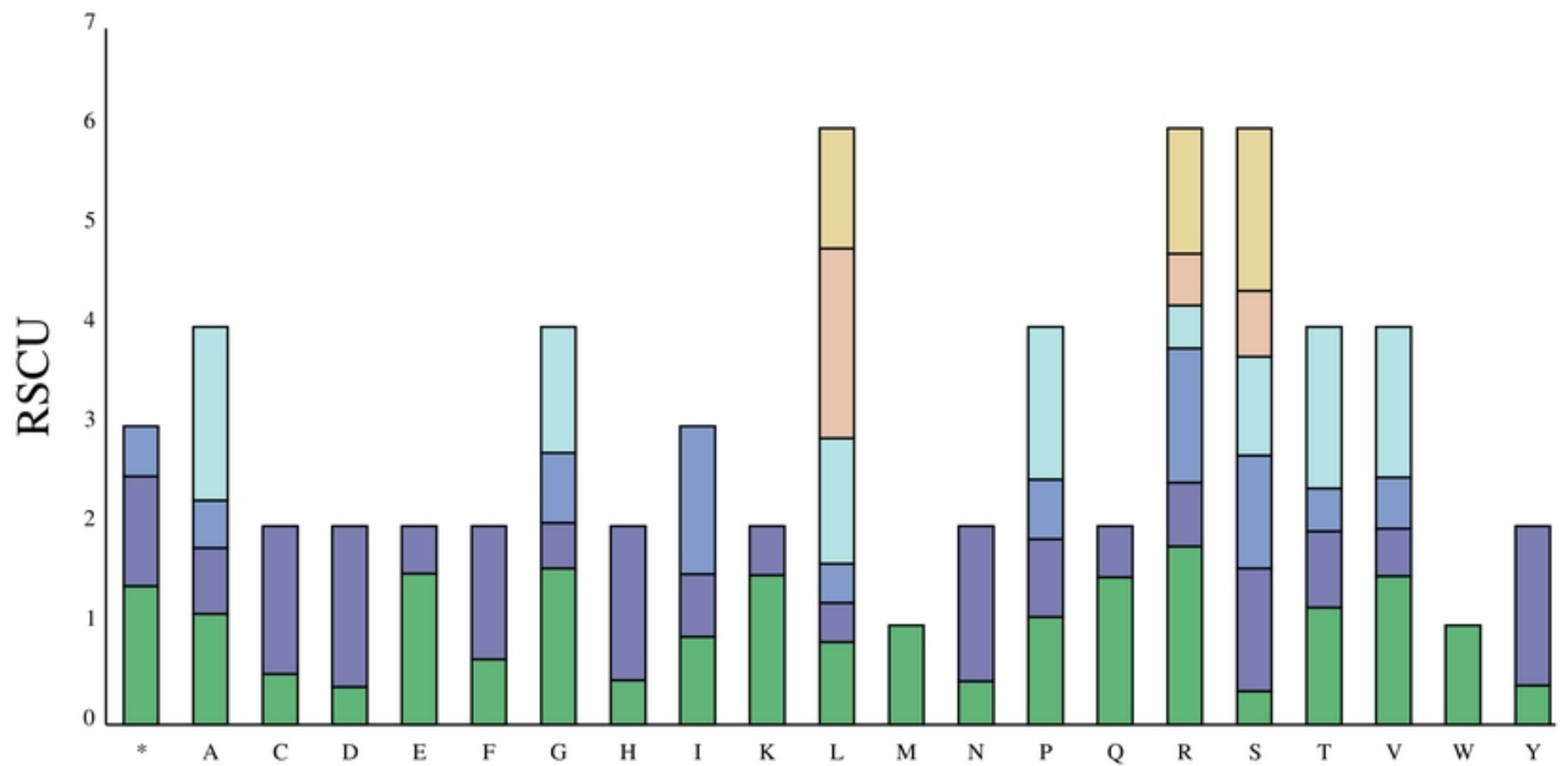
566 **S2 Table. Base composition of the wild *R.glutinosa* Libosch chloroplast genome.**

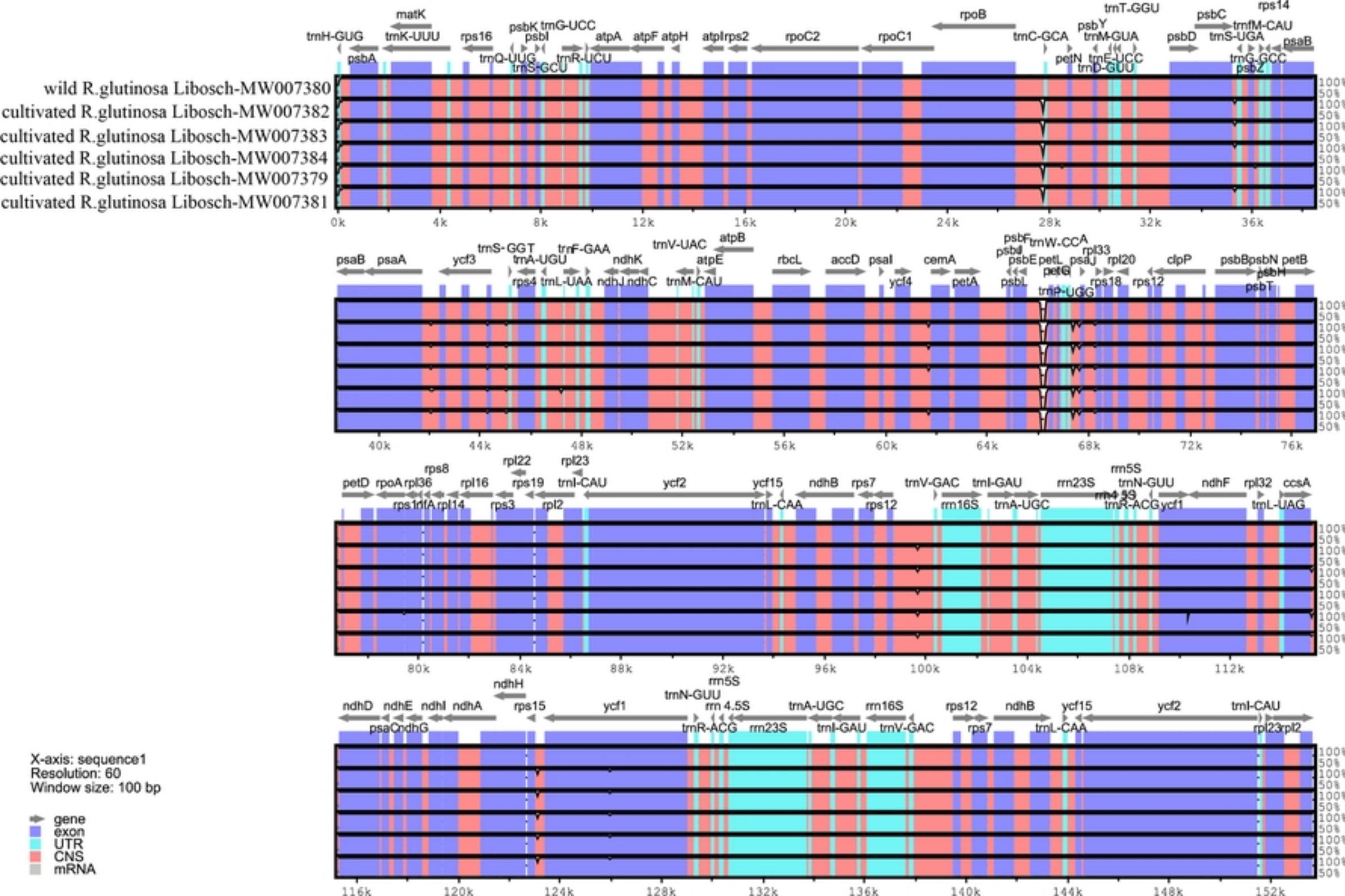
567 **S3 Table. The intron and exon analysis of intron contain genes in the chloroplast**

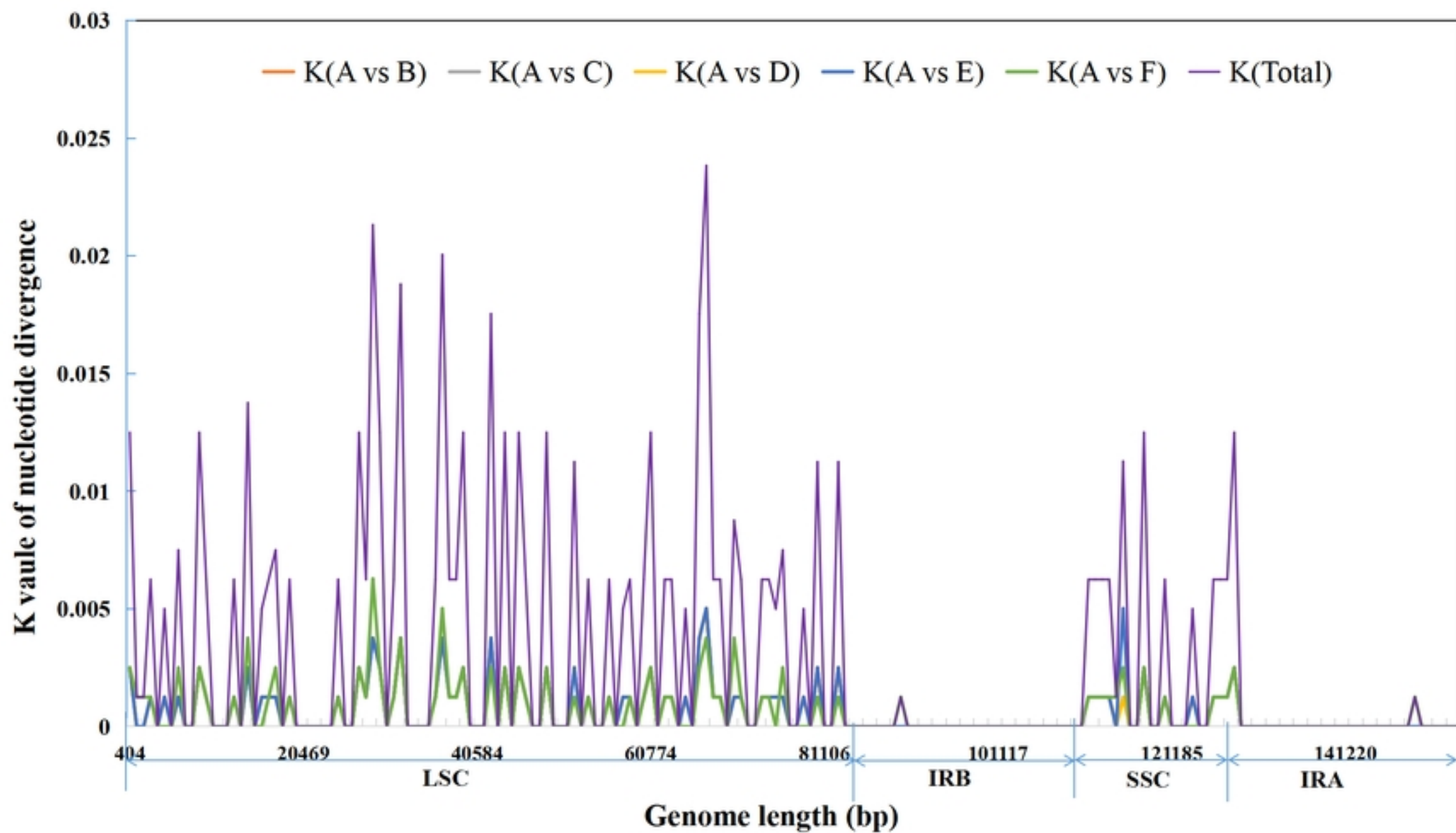
568 **genome of wild *R.glutinosa* Libosch**

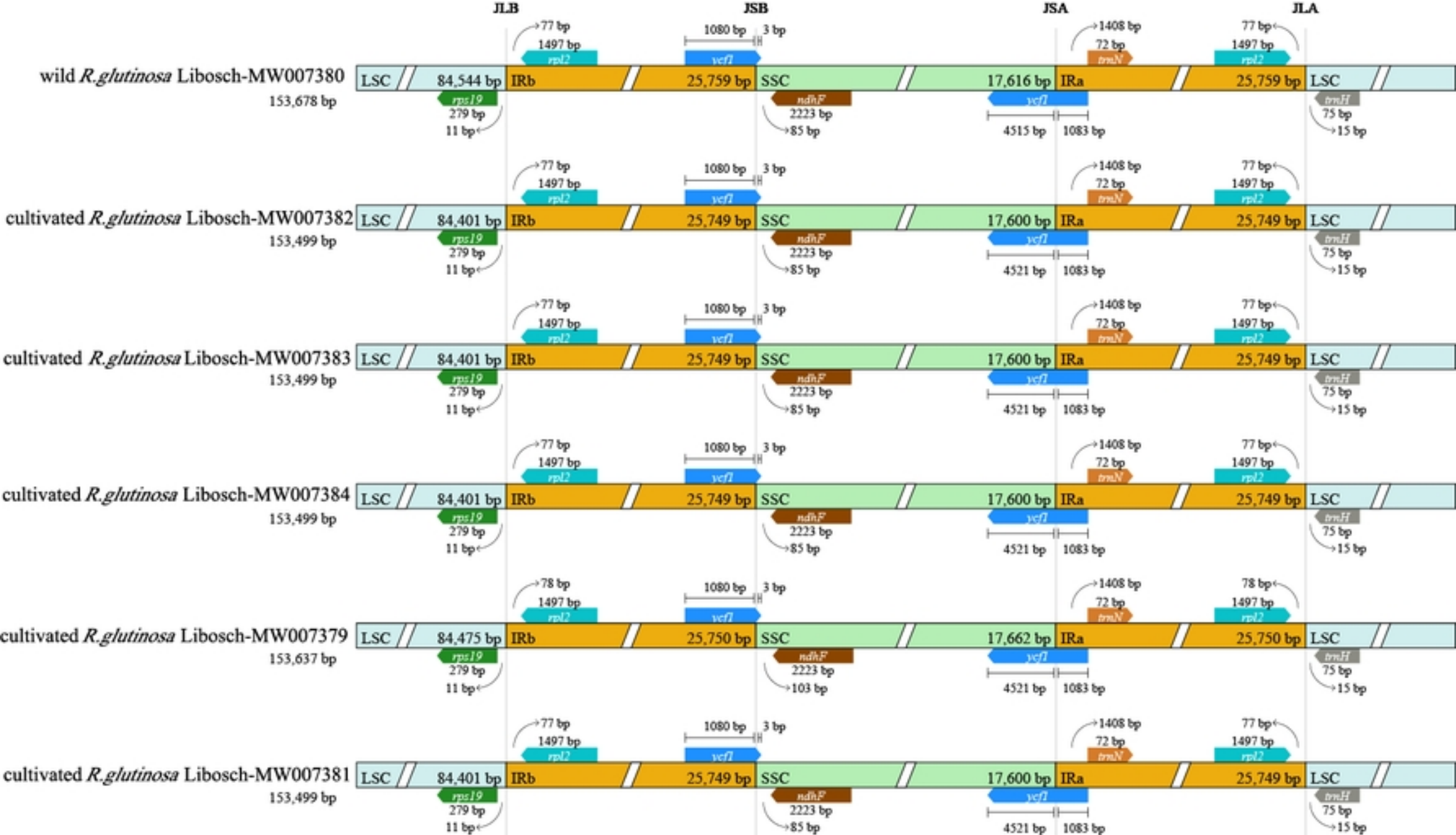
569 **S4 Table. Summary of six chloroplast genomes features of *R.glutinosa* Libosch**











LOREM IPSUM

