

¹Key Laboratory of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China
²School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China

The complete chloroplast genome sequence of *Hyssopus cuspidatus* Boriss. and analysis of phylogenetic relationships

Zhi Zhang^{1,2}, Yujing Miao¹, Guoshai Zhang¹, Xinke Zhang¹, Huihui Zhang¹, Guoxu Ma¹, Junbo Xie¹, Zhaocui Sun^{1,*}, Linfang Huang^{1,*}

(Submitted: May 11, 2023; Accepted: November 18, 2023)

Summary

Hyssopus cuspidatus is a member of the Lamiaceae family, members of which are often used to treat cough and asthma by the Uigurs. However, the *Hyssopus* genus has a limited number of known chloroplast genomes, making it difficult to compare species within the genus and to classify species within and outside the genus accurately. The introduction of the chloroplast genome method would therefore help improve the classification of the *Hyssopus* genus. This report presents the complete chloroplast sequences of *Hyssopus cuspidatus*. The chloroplast genome of *H. cuspidatus* is 149,678 bp long and contains 129 genes, including 85 protein-coding genes, 36 tRNA genes, and 8 rRNA genes. We identified 46 single sequence repeats (SSRs), most of which were mononucleotide adenine–thymine. The analysis of the repeat sequences, codon usage, and comparison of chloroplast genomes showed a high degree of conservation. The plastid genomes exhibited a typical quartile structure. Four hypervariable regions were identified: *accD-psal*, *psbZ-trnG-GCC*, *trnH-GUG-psbA*, and *atpH-atpI*. Phylogenetic analysis revealed that the *Hyssopus* genus was closely related to the adjacent genus *Dracocephalum*. Our research conducted a comprehensive analysis of the characteristics of the *Hyssopus* genus and provided a detailed comparison of the differences between species within and outside of this genus. Through IR comparison, phylogenetic analysis, and variation region analysis, we discovered a close relationship between the genera *Hyssopus* and *Dracocephalum* and propose a new perspective on the phylogenetic classification of *H. cuspidatus*. These findings will support the continued identification, classification, and evolutionary analysis of this genus.

Keywords: *Hyssopus cuspidatus*, chloroplast genome, phylogenetic relationship, comparative analysis.

Introduction

The *Hyssopus* genus is an important group of plants within the Labiatae family. *Hyssopus cuspidatus* Boriss. is mainly distributed from the East Mediterranean to central Asia and Mongolia and is used as a condiment and spice for its culinary and medicinal properties (AHMADI et al., 2020). The name *Hyssopus* originates from Hebrew *ezob*, which means “sacred herb” (SHARIFI-RAD et al., 2022). *Hyssopus cuspidatus* has been a commonly used herbal medicine for Chinese ethnic Uighur medicine and contains chemical components, including volatile oils, flavonoids, alkaloids, organic acids, and lipids (SHOMIRZOEVA et al., 2020; SHOMIRZOEVA et al., 2019). Modern pharmacological studies found that the water extract of *H. cuspidatus* has a significant effect on allergic asthma in mice (YUAN et al., 2019). The overground portion of *H. cuspidatus* has

been identified as having medicinal properties, and both ancient medicine and modern pharmacology show that this plant has significant efficacy in treating asthma. The antibacterial and anti-inflammatory capabilities have been shown to be due to its abundant essential oil and flavonoids (ZHAO et al., 2020). The dry weight yield of *H. cuspidatus* essential oils was 0.6% (w/w) with four main components of oxyterpene (66.33%), monoterpene (26.14%), octane (1.85%), and oxysesquiterpene (1.25%) (ZHOU et al., 2010).

The chloroplast (CP) is a vital organelle of green plants and an important place for photosynthesis. The complete CP genome of *Cyanobacteria* and *Arabidopsis* show that plant CPs originate from *Cyanobacteria* through endosymbiogenesis (RAVEN and ALLEN, 2003). CP DNA (cpDNA) is usually a double-stranded structure in higher plants, and part of the cpDNA has been shown to be circular. cpDNA consists of four typical regions: two reverse repeats (IRA and IRB) separated by a large and small copy region (LCS and SSC, respectively). An in-depth understanding of the CP genome will help us understand and protect important medicinal plants. Changes in the sequence and structure of CP genomes also indicate significant differences among species. Research on plants at the genus level can help improve understanding of phylogenetic role of each plant. The variation of cpDNA sequences provides substantial evidence for understanding the relationship between different plants (DANIELL et al., 2016).

Herein, we sequenced the CP genome of *H. cuspidatus*, and compared this with that of two species of the genus *Dracocephalum* and of *H. officinalis* for phylogenetic analysis. This study aimed to provide a new insight into the species classification in the *Hyssopus* genus from the perspective of phylogeny through the CP genome and further deepen the research on *Hyssopus* species.

Materials and methods

Plant material and DNA extraction

Hyssopus cuspidatus plant material was collected from Xinjiang, China. Samples were identified by Professor Linfang Huang and stored at the Herbarium of the Chinese Academy of Medical Science and Peking Union Medicinal College (CMPB13402). *Hyssopus cuspidatus* was stored at -80°C for DNA extraction. The CTAB method (ARSENEAU et al., 2017) 2017 was performed to extract the total genomic DNA. The Illumina HiSeq 2500 platform (Novogene Technologies, Inc., Beijing, China) was applied to sequence the genomic DNA. PRINSEQ lite Ver0.20.4 was performed to filter the raw data reads to get clean reads (SCHMIEDER and EDWARDS, 2011). The CP genomes were assembled from the highest quality clean reads by using NOVOPlasty (v.2.7.2) with kmer 39 using the CP genome of *Cucumis melo* var. *makuwa* isolate M1-15 chloroplast (MF_536700) as reference.

* Corresponding authors

Sequencing, assembly, and annotation

Genomic DNA was sequenced using the Illumina Hiseq 2500 platform (Novogene Technologies, Inc., Beijing, China). Raw reads were filtered using PRINSEQ lite Ver0.20.4 to obtain clean reads. High-quality reads were then assembled using NOVOPlasty v.3.8.3 and annotated through the CPGAVAS2 web server. tRNA gene annotation was confirmed using tRNAscan-SE v.2.03, and genome maps were generated using OGDRAW.

Sequence data was deposited at the National Center for Biotechnology Information (NCBI) under GenBank Accession No. OQ590030.1 (URL: <https://www.ncbi.nlm.nih.gov/nucleotide/OQ590030.1>).

Repeat sequences, codon usage, and IR analysis in *H. cuspidatus* CP genome

First, the protein-coding sequence was extracted with PhyloSuite v1.22, and then the codon usage distribution was investigated by using CodonW1.4.2 through relative synonymous codon usage (RSCU) values, which indicates the difference between actual use and prediction (ARELLA et al., 2021). An RSCU value >1 indicates that the frequency of codon use is higher than predicted. VMATCH was used to find dispersed repeat sequences, and PREPuter was used to predict the four types of repeat sequences: forward (F), palindromic (P), reverse (R), and complement (C).

Gb files of *H. officinalis*, *D. psammophilum*, and *D. rupestre* were downloaded from National Center for Biotechnology Information (NCBI) and imported with *H. cuspidatus* into the online IR analysis website (<https://irscope.shinyapps.io/irapp/>) to obtain the IR analysis diagrams for these four species.

Genome comparison

To determine the interspecific differences for phylogenetic analysis, we selected *H. officinalis*, a plant of the same genus as *H. cuspidatus*, and *D. psammophilum* and *D. rupestre*, two species of

Dracocephalum genus, adjacent to the phylogenetic tree, for plastid comparative analysis. Gene sequences of the four species were compared through Geneious 9.0.2 e and output in fas format (MASTERS et al., 2011). Then, DnaSP v5.0 (Window Length set to 600 sites, Step size set to 200 sites) was used for sliding window analysis to calculate nucleotide diversity (ROZAS et al., 2017).

Phylogenetic analysis and the nucleotide substitution rate

A total of 282 CP genomes were used for cluster analysis, including from two species belonging to the genus *Hyssopus*. *Arabidopsis thaliana* (NC_000932) was selected as the outgroup. In addition to the CP genome of *H. cuspidatus* obtained by testing, the complete CP genome of the remaining 281 species was downloaded from NCBI (nih.gov). Then, 92 CDSs were extracted, aligned, and connected using PhyloSuite and MAFFT (v 7.450) (ROZEWICKI et al., 2019) 2019. The phylogenetic tree was constructed using the maximum likelihood (ML) method (NGUYEN et al., 2015) method of RAXML v8.2.8. One thousand repetitions of bootstrap analysis were performed. Nucleotide polymorphisms were analyzed with DNASPv5.0 (KHAN et al., 2008).

Nonsynonymous and synonymous substitution rate analysis

A total of 80 common protein-coding genes were extracted from nine *Hyssop* and *Dracocephalum* species with PhyloSuite. After removing specific protein-coding genes, nonsynonymous (Ka) and synonymous (Ks) ratios for each protein-coding gene from these species were obtained with PAML4.9 (JIANG et al., 2022).

Results

The CP genomes features of *H. cuspidatus*

The whole CP genome length of *H. cuspidatus* is 149,678 bp. The CP genome of *H. cuspidatus* comprises four typical regions, including

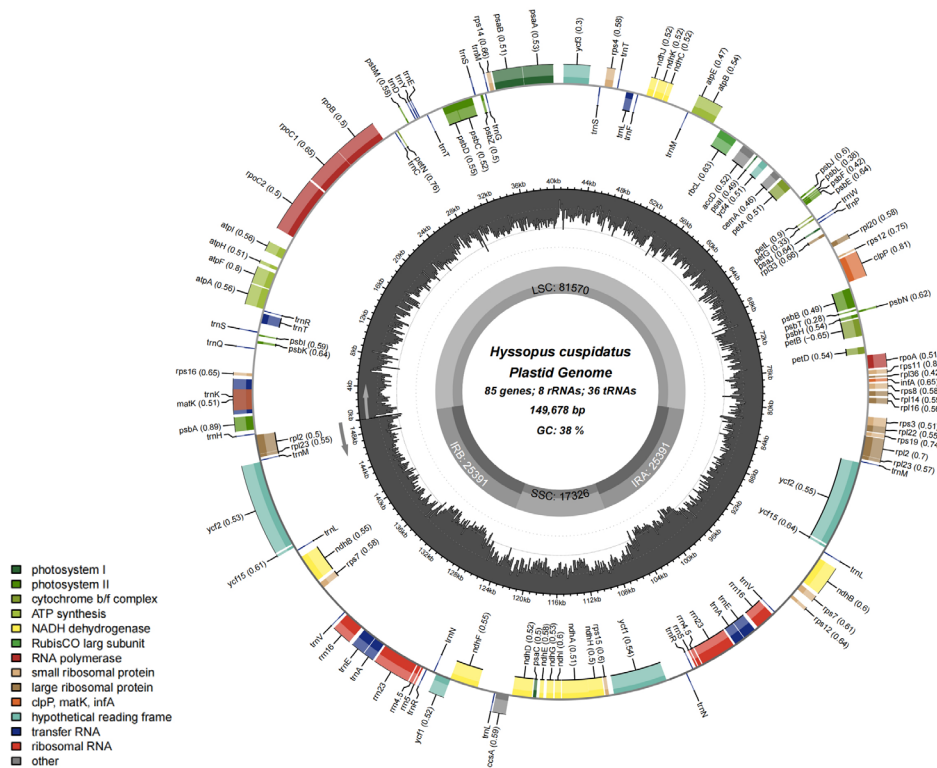


Fig 1: Chloroplast genome maps of *Hyssopus cuspidatus*. The genes in the circle of the chloroplast genome map are transcribed clockwise, while those outside the circle are transcribed counterclockwise. Genes with different functions are color coded. The darker grey in the inner circle shows GC content, while the lighter grey shows AT content.

two inverted repeats (IRa and IRb) and LSC and SSC regions. The LSC region is 81,570 bp, and the SSC region is 17,326 bp. The LSC and SSC regions are separated by a pair of IR regions (25,391 bp) (Fig. 1). Of the four regions, the IR region has the highest GC content, accounting for 43.11%, followed by the LSC region and the SSC region, which had the lowest GC content, with a GC content of 35.9% and 31.6%, respectively (Tab. 1). We also identified the direct and palindromic repeat sequences (Tab. S1).

The CP genome of *H. cuspidatus* contains 129 different genes, including 85 protein-coding genes, 36 tRNA genes, and 8 rRNA genes. Among them, four rRNAs (*rrn16s*, *rrn23s*, *rrn4.5s*, and *rrn5s*), five tRNAs (*trnL-CAA*, *trnV-GAC*, *trnA-UGC*, *trnR-ACG*, *trnN-GUU*), and ten protein encoding-proteins (*ndhB*, *psaC*, *rpl2*, *rpl22*, *rps12*, *rps7*, *ycf1*, *ycf2*, and *ycf15*) contain two repeating units while the tRNA (*trnE-UUC*, *trnM-CAU*) contains three repeating units (Tab. 2). Introns were present in 19 genes, including 8 tRNAs (*trnK-UUU*, *trnT-CGU*, *trnL-UAA*, *trnC-ACA*, *trnE-UUC* × 2, *trnA-UGC* × 2) and 10 protein-coding genes (*atpF*, *rpoC1*, *rps19*, *petB*, *rpl2* × 3, *ndhB* × 3, *NdhA*) and two protein-coding genes (*ycf3*, *clpP*) (Tab. 3).

Repetitive sequence and codon usage analysis

We detected 46 SSRs in the CP genome of *H. cuspidatus* (Tab. S2). Among all SSRs, the single nucleotide is the most abundant. Among them, we detected four repetitive sequences: the P sequence had the largest number of repeats, with 21 occurrences, followed by the F sequence, with 17 occurrences; the R and C sequences have fewer occurrences, with 8 and 4, respectively (Fig. 2).

Tab. 1: Basic features of the chloroplast genome of *Hyssopus cuspidatus*

Species	Name	Length(bp)	GC(%)
<i>Hyssopus cuspidatus</i>	IR	25,391	43.11
<i>Hyssopus cuspidatus</i>	IR	25,391	43.11
<i>Hyssopus cuspidatus</i>	SSC	17,326	31.6
<i>Hyssopus cuspidatus</i>	LSC	81,570	35.9

Tab. 2: Gene composition of *Hyssopus cuspidatus* chloroplast genome

Category of genes	Group of genes	Name of genes
RNA	rRNA	<i>rrn16s</i> ×2, <i>rrn23s</i> ×2, <i>rrn4.5s</i> ×2, <i>rrn5s</i> ×2
	tRNA	<i>trnL-CAA</i> ×2, <i>trnV-GAC</i> ×2, <i>trnA-UGC</i> ×2, <i>trnR-ACG</i> ×2, <i>trnN-GUU</i> ×2, <i>trnE-UUC</i> ×3, <i>trnM-CAU</i> ×3, <i>trnH-GUG</i> , <i>trnK-UUU</i> , <i>trnQ-UUG</i> , <i>trnS-GCU</i> , <i>trnT-CGU</i> , <i>trnR-UCU</i> , <i>trnC-GCA</i> , <i>trnD-GUC</i> , <i>trnY-GUA</i> , <i>trnT-GGU</i> , <i>trnS-UGA</i> , <i>trnG-GCC</i> , <i>trnS-GGA</i> , <i>trnT-UGU</i> , <i>trnF-GAA</i> , <i>trnW-CCA</i> , <i>trnP-UGG</i> , <i>trnL-CAA</i> , <i>trnL-UAG</i> , <i>trnL-UAA</i>
photosynthesis	Subunits of ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF</i> , <i>atpH</i> , <i>atpI</i>
	Subunits of photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i> , <i>ycf3</i>
	Subunits of NADH-dehydrogenase	<i>ndhA</i> , <i>ndhB</i> ×2, <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Subunits of cytochrome b/f complex	<i>petA</i> , <i>petB</i> , <i>petD</i> , <i>petG</i> , <i>petL</i> , <i>petN</i>
	Subunits of photosystem I	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> ×2, <i>psaI</i> , <i>psaJ</i>
	Subunit of rubisco	<i>rbcL</i>
Self-replication	Large subunit of ribosome	<i>rpl14</i> , <i>rpl16</i> , <i>rpl2</i> ×2, <i>rpl20</i> , <i>rpl22</i> ×2, <i>rpl33</i> , <i>rpl36</i> × <i>rps19</i>
	DNA dependent RNA polymerase	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1</i> , <i>rpoC2</i>
	Small subunit of ribosome	<i>rps11</i> , <i>rps12</i> ×2, <i>rps14</i> , <i>rps15</i> , <i>rps16</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> ×2, <i>rps8</i>
Other genes	Subunit of Acetyl-CoA-carboxylase	<i>accD</i>
	c-type cytochrom synthesis gene	<i>ccsA</i>
	Envelop membrane protein	<i>cemA</i>
	Protease	<i>clpP</i>
	Translational initiation factor	<i>infA</i>
	Maturase	<i>matK</i>
Unkown	Conserved open reading frames	<i>ycf1</i> ×2, <i>ycf15</i> ×2, <i>ycf2</i> ×2, <i>ycf4</i>

Codon usage of the CP genome of *H. cuspidatus* (Tab. S3) was performed to gain a deeper understanding of the protein-coding genes (HU et al., 2023). *H. cuspidatus* has 49,892 common codons that encode protein genes that include all 20 amino acids. Since the codon occurrence frequency and expectation of various amino acids are different, we chose the RSCU value instead of the number of codons to describe the actual occurrence frequency of codons encoding each amino acid. RSCU, as a measure of the uneven use of synonymous codons in the coding sequence, is the ratio of the frequency of a given codon to the expected frequency (WANG et al., 2023). At an RSCU ratio >1 the frequency of specific codons is higher than that of other synonymous codons, and the frequency of codons is higher than expected (SHARP and LI, 1987) 1987. We found that the total RSCU ratios of three amino acids (Leu, Ser, Arg) were significantly higher than those of other amino acids. Leu contained 5,670, Ser contained 4,395, and Arg contained 3,229 (Fig. 3).

Nucleotide polymorphism

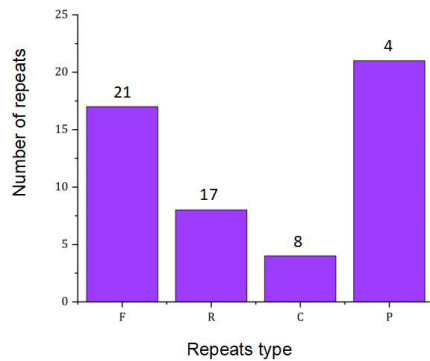
We conducted an in-depth study of nucleotide polymorphism in the *Hyssopus* genus. We used Geneious 9.0.2 to compare cpDNA sequences of two species and analyzed nucleotide polymorphism with DnaSPv5.0. Interestingly, we found a large genetic difference in the 106,797–127,350-bp sequence. The difference was also large when compared with a species of the same genus of *H. officinalis*. We used mVISTA to identify regions with high mutation rates in this sequence, namely *ycf1-ndhF*, *ccsA-ndhD*, *ndhE-ndhI*, and *ndhA-ycf1*. In these four regions, nucleotide polymorphism occurred at the highest frequency (Fig. 4). The discovery of this sequence will contribute to a more detailed analysis of the *Hyssopus* genus.

IR expansion and contraction

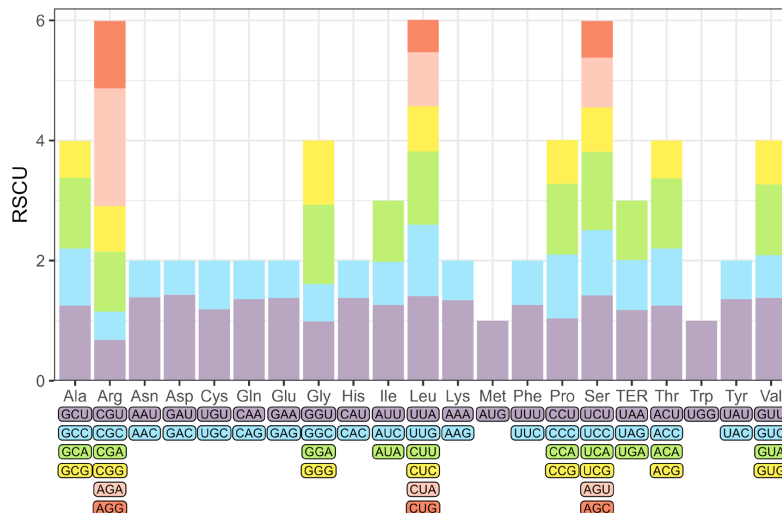
The contraction and expansion of IR can change the length of the complete plastid DNA (cpDNA). Studying the difference in the IR region can help clarify the phylogenetic relationship between species. Here, we selected four species for comparison and analyzed the differences between their genome size and the LSC, SSC, IRa, and

Tab. 3: Genes with intron in the *Hyssopus cuspidatus* chloroplast genome and length of exons and introns.

Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII	ExonIII
<i>trnK-UUU</i>	-	1,672	4,270	37	2,526	36		
<i>trnT-CGU</i>	+	8,684	9,438	35	677	43		
<i>atpF</i>	-	11,420	12,675	147	716	393		
<i>rpoC1</i>	-	20,138	22,943	433	748	1,625		
<i>ycf3</i>	-	40,981	42,929	126	709	228	733	153
<i>trnL-UAA</i>	+	45,914	46,481	35	483	50		
<i>trnC-ACA</i>	-	50,091	50,734	38	550	56		
<i>clpP</i>	-	67,617	69,531	71	696	291	631	226
<i>petB</i>	+	72,433	73,800	6	720	642		
<i>rps19</i>	-	81,290	81,568	391	661	434		
<i>rpl2</i>	-	81,625	83,110	391	661	434		
<i>ndhB</i>	-	91,779	939,88	775	677	758		
<i>rpl2</i>	+	99,266	100,284	32	947	40		
<i>ndhB</i>	+	100,339	101,138	37	727	36		
<i>trnE-UUC</i>	+	113,346	115,459	553	1,019	542		
<i>trnA-UGC</i>	-	130,111	130,910	37	727	36		
<i>ndhA</i>	-	130,965	131,983	32	947	40		
<i>trnA-UGC</i>	+	137,261	139,470	775	677	758		
<i>trnE-UUC</i>	+	148,139	149,624	391	661	434		
<i>ndhB</i>	+	148,139	149,624	391	661	434		
<i>rpl2</i>	+	8,684	9,438	35	677	43		

**Fig 2:** Types and numbers of interspersed repeats in *Hyssopus cuspidatus*. F = forward, P = palindromic, R = reverse and C = complement

IRb regions (Fig. 5). By comparing the length of the genome, we found that the length of the cpDNA of the four species was similar with no significant difference. However, *H. cuspidatus* significantly differed in gene connection compared with the other three species in that *ndhF* was absent at the SSC and IRb junction. Among the other three species, the length of *ycf1* was 61–1,075 bp across the junction of SSC and IRb, whereas the length of *ycf1* in *H. cuspidatus* in the SSC region was 226 bp. At the junction of SSC and IRa, *H. cuspidatus* had one more copy of *ndhF* than the other three species, and *ycf1* at the junction of SSC and IRa was mainly located in IRa. Although the cpDNA of *H. cuspidatus* was quite different from the other three species, the cpDNA of the *Hyssopus* genus was similar to that of the other two species. *H. officinalis*, *D. psammophilum*, and *D. rupestre* had similar cpDNA. *NdhB* and *ycf1* spanned the junction of SSC and IRb while *ycf1* also spanned the junction of SSC and IRa.

**Fig 3:** Codon content and stop codon of 20 amino acids in protein-coding genes.

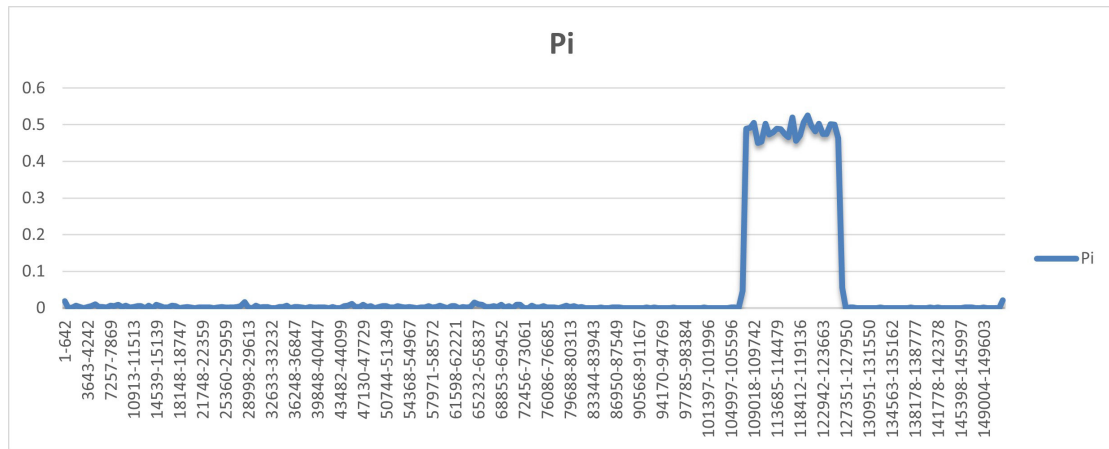


Fig 4: Nucleotide polymorphism distribution of *Hyssopus cuspidatus*

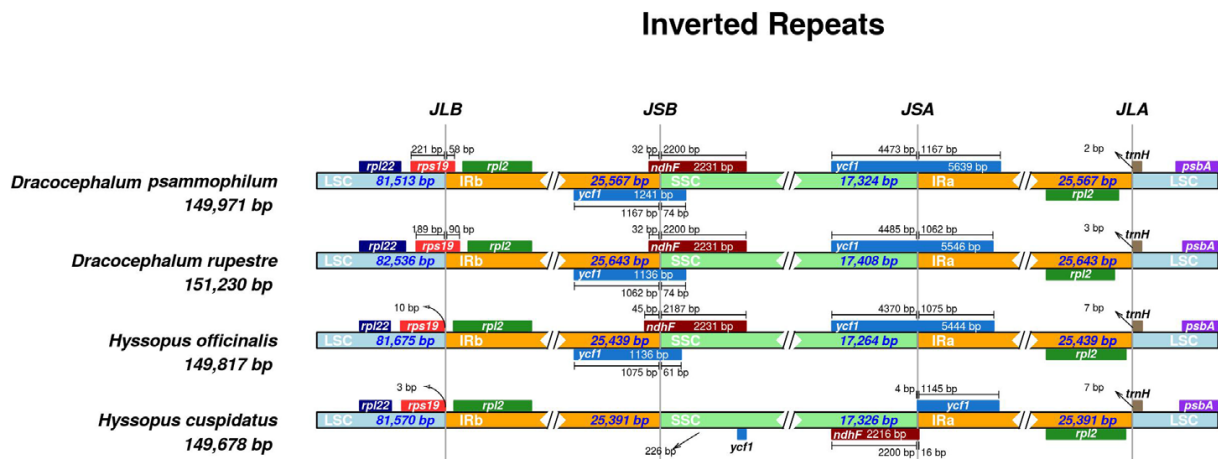


Fig 5: Analysis of gene linkage sites in four species. Genes transcribed clockwise are presented below the track, whereas those transcribed counterclockwise are presented on top of the track.

Phylogenetic analysis

To explore the plant characteristics of Labiatae, we systematically studied the *Hyssopus* genus found in this family. Using 92 CDS, all species data of Lamiaceae classification were obtained from the NCBI. A total of 282 species (Fig. 6), 281 species in the Labiatae family and *A. thaliana*, were selected to construct the maximum likelihood tree (ML). All nodes have high bootstrap support. Through phylogenetic analysis, we found that the *Hyssopus* and *Dracocephalum* species are closely related. Among them, the *Dracocephalum* species can also be classified into two categories according to their close relationship with the *Hyssopus* genus. We found that *D. psammophilum* and *D. rupestre*, which are both found in the Lamioideae family, clustered together with the *Hyssopus* genus.

Kimura's two-parameter (K2P) analysis

We selected four species to participate in the K2P model because of the close relationship identified between the *Hyssopus* genus and two species in genus *Dracocephalum* genus. Highly variable regions can be used to analyze and distinguish species with close phylogenetic relationships. We analyzed the genetic distance of the four species and found that specific regions could be used as potential molecular markers (Figure 7). We selected 69 intergenic regions in the four species and calculated the genetic distance using the K2P model. After the screening, we identified 36 gene regions, including *accD-psal*,

psbZ-trnG-GCC, *trnH-GUG-psbA*, and *atpH-atpI*, with the highest genetic distance. These regions can be considered potential markers for future molecular marker development.

Comparative analysis of genome

To conduct a comprehensive analysis of the genome of *H. cuspidatus*, we selected three other species for comparison: *H. officinalis*, a member of the same genus as *H. cuspidatus*, and the two species that are most closely related in the phylogenetic tree, *D. rupestre* and *D. psammophilum*. Using the phylogenetic tree as a guide, we analyzed the gene sequences of these four species. To do this, we used the online platform mVISTA and selected *H. officinalis* on NCBI as the reference sequence and then identified sequence differences among the four species (Fig. 8). Notably, we found that *H. cuspidatus* and *H. officinalis* had a high degree of similarity in their cpDNA sequence arrangements. Overall, the protein-coding genes were relatively conserved, with the greatest variation concentrated in regions such as *trnH-GUG-psbA*, *atpH-rps2*, *trnG-GCC-psaB*, *petA-psbJ*, and *ycf2-ycf15*.

Selective pressure analysis

Changes in DNA sequence depends on natural selection, and the possibility of natural selection can be represented by nonsynonymous

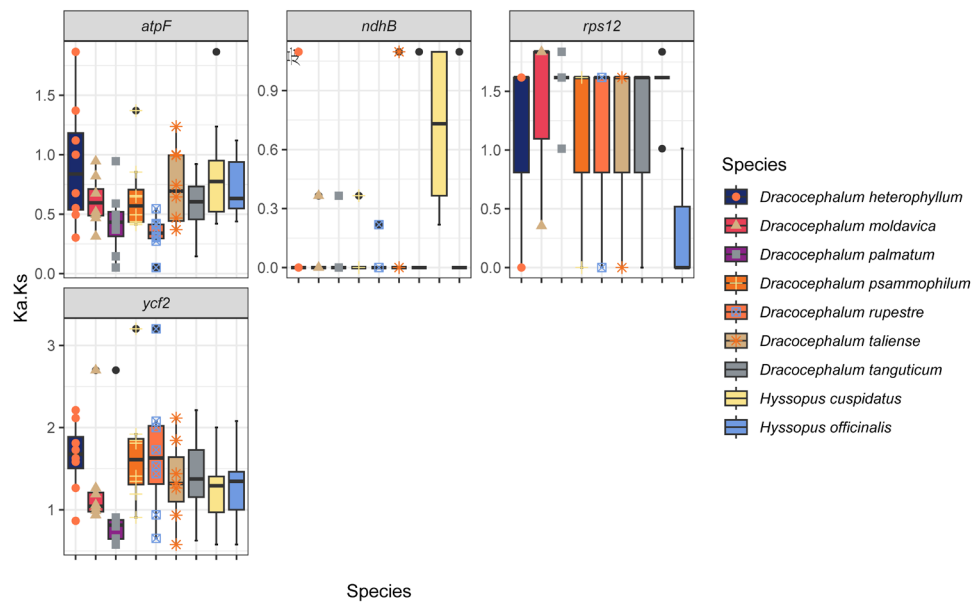


Fig 9: The Ka/Ks ratios of all genes between the *Hyssopus* and the *Dracocephalum* genus. The 65 coding genes with ka/ks value were sequenced, and the top four showed *atpF*, *ndhB*, *rps12* and *ycf2*, as shown in the figure.

(amino acid substitution, K_a) and synonymous (K_s) substitution and their ratio (K_a/K_s) (Fig. 9) (MAQSOOD et al., 2022). A K_a/K_s ratio <1 means that the synonymous substitution rate is high, that the amino acid at this site remains unchanged, and that the mutation possibility of this segment is not high; a K_a/K_s ratio >1 indicates that the rate of nonsynonymous substitution is high, that the amino acid changes, and the mutation possibility of this segment is high. Most of the K_a/K_s ratios obtained were <1 because during natural selection synonymous nucleotide substitution is more common in the protein-coding region. Two protein-coding genes with $0.6 \leq K_a/K_s < 1$ were *atpF* and *ndhB*, respectively. Although the K_a/K_s ratios were <1 , they were still at the forefront of 65 protein-coding genes with K_a/K_s with high mutagenicity. Only two protein-coding genes *rps12* and *ycf2*, had K_a/K_s ratios >1 , indicating that these two genes had the highest nonsynonymous substitution rate among the 65 protein-coding genes (Supplementary Fig. 1). In the future, mutations at these sites may accelerate the development of this species.

Discussion

The genome structure of *H. cuspidatus* CPs consists of four basic regions (LSC, SSC, IRa, and IRb). The length of these four regions varies widely, with the LSC region the longest at 81,570 bp, the IR region the second longest at 25,391 bp, and the SSC region the shortest at 17,326 bp. The GC content (43.11%) in the IR region was significantly higher than that in the LSC and SSC regions because the DNA encoding RNA in the IR region contained a high GC content. The analysis of repetitive sequences helps us understand the composition of repetitive sequences in *H. cuspidatus*. The palindromic (P) sequences have the highest number of repeats (21). SSR analysis showed that most of the CP genomes had a high proportion of single nucleotide repeats. SSR are usually located in noncoding gene regions and rarely in protein-coding gene regions and are often used as molecular markers for species identification (JAKOBSSON et al., 2007; PROVAN, 2000). Analysis of codon usage allows us to understand the specific characteristics of *H. cuspidatus*, and the frequent use of codons for Leu, Ser, and Arg indicated the species characteristics. By constructing a phylogenetic tree, we found that *Hyssopus* and *Dracocephalum* genera are closely related and belong to the same

clade, which is divided into upper and lower clades. Fig. 6 shows that the *Hyssopus* genus is most closely related to the upper clade and to *D. rupestre* and *D. psammophilum*. To observe the associations and differences between species with a high degree of intimacy between two genera, we selected these four species for IR, K2-P, and mVISTA analysis. In the final selection pressure, to show the difference of pressure between the *Hyssopus* and *Dracocephalum* genera as much as possible, the whole branches of these genera were included in the analysis to explore the relationship.

Through IR analysis, we explained the similarities and differences between these genera from the perspective of the CP genome. Although *H. cuspidatus* and *H. officinalis* belonged to same genus, significant differences were present in their gene sequences. In the junction between the LSC and IRb, *H. cuspidatus* lacked the *ndhF* gene and had a shorter *ycf1* gene than *H. officinalis*, and in the junction between the SSC and IRa, *H. cuspidatus* had an additional *ndhF* gene and most of the *ycf1* gene at this site was located in the IRa region, which is clearly different from these regions in the other three species. These differences within the same genus demonstrated the diversity of the *Hyssopus* genus and suggested that the species within this genus may have significant differences. At the genus level, we found that although *H. officinalis* and two species of the genus *Dracocephalum* belong to different genera, their CP genomes were highly similar, and the phylogenetic analysis placed the four species together, indicating a strong relationship between the two genera. The IR analysis also indirectly confirmed the reasonable division of the phylogenetic classification.

Our previous analysis grouped the *Hyssopus* and *Dracocephalum* genera together in an ML tree. These results support the division of *Hyssopus* and *Dracocephalum* genera, according to the IR expansion and contraction and ancestral range estimation analysis. Our current research also supported this point. The phylogenetic tree had high bootstrap support at all nodes, which showed the accuracy and reliability of phylogenetic relationships. After studying the complete cpDNA of *H. cuspidatus*, we found the unique regions of the gene sequence of this species and put forward a view on the classification of the genus *Hyssopus*. Our research on the complete cpDNA of the genus *Hyssopus* will be helpful in clarifying the taxonomic relationship among genera in Labiatae.

Our analysis of the *H. cuspidatus* intact CP genes identified a unique region in the gene sequence that distinguishes this species from other *Hyssopus* species. This finding has important implications for the classification of the *Hyssopus* because it provides evidence that *H. cuspidatus* is different from other *Hyssopus* species. Our study suggested that the taxonomy of the *Hyssopus* genus should be modified to reflect the genetic diversity within this population. Specifically, we propose to place *H. cuspidatus* as a unique species within the genus. Overall, our findings highlight the importance of using molecular data in taxonomic studies to classify and understand the relationships between species and genera accurately. By identifying the unique genetic traits of different species, we can improve our understanding of evolutionary history and the diversity of life.

Conclusion

Herein, we assembled the complete CP genome of *H. cuspidatus* through Illumina sequencing. Since the genome of the *Hyssopus* genus is not completely assembled, this article can support future research in this genus. Compared with the complete CP genome of *H. officinalis*, we found that the sequence length of the two species was the same. However, the gene sequence of *H. cuspidatus* is quite different at 106,797–127,350 bp, and the genes located in this sequence are highly specific. Therefore, *H. cuspidatus* is unique, whether compared with a species from the same genus (*H. officinalis*) or with species from *Dracocephalum*. The ML obtained from systematic analysis found that the *Dracocephalum* and *Hyssopus* genera were closely related, which was consistent with the conclusions obtained by previous researchers.

Our research shows the role of protein-coding genes in the phylogenetic process of *H. cuspidatus*. Through processing and analyzing the complete CP genome, we found that the regions with high variation can be used as potential molecular markers, which is conducive to a more comprehensive understanding of *Hyssopus* species. Through phylogenetic analysis, we found a close relationship between the *Hyssopus* and *Dracocephalum* genera and proposed a new taxonomy for *Hyssopus*. The results of this study enrich the genetic research of *Hyssopus*, provide more reference data for further research, identification, and classification of *Hyssopus*, and advance a new view on the classification of the *Hyssopus* genus from the aspect of phylogenetic analysis. Thus, further study of the characteristics of the gene sequences of *Hyssopus* genus would be beneficial.

Funding acknowledgement

This work was supported by the CAMS Innovation Fund for Medical Sciences (CIFMS, 2022-I2M-1-017), National Natural Science Foundation of China (U1812403-1, 82073960, 82211540726 and 82274045), the Open Fund of State Key Laboratory of Southwestern Chinese Medicine Resources (SKLTCM2022015), National Science & Technology Fundamental Resources Investigation Program of China (2018FY100701), which are gratefully acknowledged.

Conflict of interest

No potential conflict of interest was reported by the authors.

Reference

- AHMADI, H., BABALAR, M., SARCHESHMEH, M.A.A., MORSHEDLOO, M.R., SHOKRPOUR, M., 2020: Effects of exogenous application of citrulline on prolonged water stress damages in hyssop (*Hyssopus officinalis* L.): Antioxidant activity, biochemical indices, and essential oils profile. *Food Chem.* 333, 127433. DOI: 10.1016/j.foodchem.2020.127433
- ARELLA, D., DILUCCA, M., GIANANTI, A., 2021: Codon usage bias and environmental adaptation in microbial organisms. *Mol. Genet. Genomics* 296, 751–762. DOI: 10.1007/s00438-021-01771-4
- ARSENEAU, J.R., STEEVES, R., LAFLAMME, M., 2017: Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Mol. Ecol. Resour.* 17, 686–693. DOI: 10.1111/1755-0998.12616
- DANIELL, H., LIN, C.S., YU, M., CHANG, W.J., 2016: Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 134. DOI: 10.1186/s13059-016-1004-2
- HU, H., DONG, B., FAN, X., WANG, M., WANG, T., LIU, Q., 2023: Mutational bias and natural selection driving the synonymous codon usage of single-exon genes in rice (*Oryza sativa* L.). *Rice (N Y)* 16, 11. DOI: 10.1186/s12284-023-00627-2
- JAKOBSSON, M., SÄLL, T., LIND-HALLDÉN, C., HALLDÉN, C., 2007: Evolution of chloroplast mononucleotide microsatellites in *Arabidopsis thaliana*. *Theor. Appl. Genet.* 114, 223–235. DOI: 10.1007/s00122-006-0425-9
- JIANG, Q., WANG, Z., HU, G., YAO, X., 2022: Genome-wide identification and characterization of AP2/ERF gene superfamily during flower development in *Actinidia eriantha*. *BMC Genomics* 23, 650. DOI: 10.1186/s12864-022-08871-4
- KHAN, H.A., ARIF, I.A., BAHKALI, A.H., AL FARHAN, A.H., AL HOMAIDAN, A.A., 2008: Bayesian, maximum parsimony and UPGMA models for inferring the phylogenies of antelopes using mitochondrial markers. *Evol. Bioinform. Online* 4, 263–270. DOI: 10.4137/ebo.s934
- MAQSOOD, H., MUNIR, F., AMIR, R., GUL, A., 2022: Genome-wide identification, comprehensive characterization of transcription factors, cis-regulatory elements, protein homology, and protein interaction network of DREB gene family in *Solanum lycopersicum*. *Front Plant Sci.* 13, 1031679. DOI: 10.3389/fpls.2022.1031679
- MASTERS, B.C., FAN, V., ROSS, H.A., 2011: Species delimitation – a geneious plugin for the exploration of species boundaries. *Mol. Ecol. Resour.* 11, 154–157. DOI: 10.1111/j.1755-0998.2010.02896.x
- NGUYEN, L.T., SCHMIDT, H.A., VON HAESLER, A., MINH, B.Q., 2015: IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. DOI: 10.1093/molbev/msu300
- PROVAN, J., 2000: Novel chloroplast microsatellites reveal cytoplasmic variation in *Arabidopsis thaliana*. *Mol. Ecol.* 9, 2183–2185.
- RAVEN, J.A., ALLEN, J.F., 2003: Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biol.* 4, 209. DOI: 10.1186/gb-2003-4-3-209
- ROZAS, J., FERRER-MATA, A., SÁNCHEZ-DELBARRIO, J.C., GUIRAO-RICO, S., LIBRADO, P., RAMOS-ONSINS, S.E., SÁNCHEZ-GRACIA, A., 2017: DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34, 3299–3302. DOI: 10.1093/molbev/msx248
- ROZEWICKI, J., LI, S., AMADA, K.M., STANDLEY, D.M., KATOH, K., 2019: MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* 47, W5–W10. DOI: 10.1093/nar/gkz342
- SHARIFI-RAD, J., QUISPE, C., KUMAR, M., AKRAM, M., AMIN, M., IQBAL, M., KOIRALA, N., SYTAR, O., KREGIEL, D., NICOLA, S., ERTANI, A., VICTORIANO, M., KHOSRAVI-DEHAGHI, N., MARTORELL, M., ALSHEHRI, M.M., BUTNARIU, M., PENTEA, M., ROTARIU, L.S., CALINA, D., CRUZ-MARTINS, N., CHO, W.C., 2022: *Hyssopus* essential oil: an update of its phytochemistry, biological activities, and safety profile. *Oxid. Med. Cell Longev.* 2022, 8442734. DOI: 10.1155/2022/8442734
- SHARP, P.M., LI, W.H., 1987: The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. DOI: 10.1093/nar/15.3.1281
- SHOMIRZOEVA, O., LI, J., NUMONOV, S., ATOLIKSHOEVA, S., AISA, H.A., 2020: Chemical components of *Hyssopus cuspidatus* Boriss.: isolation and identification, characterization by HPLC-DAD-ESI-HRMS/MS, antioxidant activity and antimicrobial activity. *Nat. Prod. Res.* 34, 534–540. DOI: 10.1080/14786419.2018.1488710
- SHOMIRZOEVA, O., LI, J., NUMONOV, S., MAMAT, N., SHATAER, D., LU, X., AISA, H.A., 2019: Chemical components of *Hyssopus seravshanicus*: antioxidant activity, activations of melanogenesis and tyrosinase, and

- quantitative determination by UPLC-DAD. *Nat. Prod. Res.* 33, 866-870.
DOI: [10.1080/14786419.2017.1408105](https://doi.org/10.1080/14786419.2017.1408105)
- WANG, J., XU, W., LIU, Y., BAI, Y., LIU, H., 2023: Comparative mitochondrial genomics and phylogenetics for species of the snakehead genus *Channa* Scopoli, 1777 (Perciformes: Channidae). *Gene* 857, 147186.
DOI: [10.1016/j.gene.2023.147186](https://doi.org/10.1016/j.gene.2023.147186)
- YUAN, F., LIU, R., HU, M., RONG, X., BAI, L., XU, L., MAO, Y., HASIMU, H., SUN, Y., HE, J., 2019: JAX2, an ethanol extract of *Hyssopus cuspidatus* Boriss, can prevent bronchial asthma by inhibiting MAPK/NF- κ B inflammatory signaling. *Phytomedicine* 57, 305-314.
DOI: [10.1016/j.phymed.2018.12.043](https://doi.org/10.1016/j.phymed.2018.12.043)
- ZHAO, L., JI, Z., LI, K., WANG, B., ZENG, Y., TIAN, S., 2020: HPLC-DAD analysis of *Hyssopus cuspidatus* Boriss extract and mensuration of its antioxygenation property. *BMC Complement Med. Ther.* 20, 228.
DOI: [10.1186/s12906-020-03016-0](https://doi.org/10.1186/s12906-020-03016-0)
- ZHOU, X., HAI-YAN, G., TUN-HAI, X., TIAN, S., 2010: Physicochemical evaluation and essential oil composition analysis of *Hyssopus cuspidatus* Boriss from Xinjiang, China. *Pharmacogn. Mag.* 6, 278-281.
DOI: [10.4103/0973-1296.71790](https://doi.org/10.4103/0973-1296.71790)

Address of the corresponding authors:

Key Laboratory of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China

E-mail: Linfang Huang: lfhuang@implad.ac.cn

E-mail: Zhaocui Sun: zcsun@implad.ac.cn

© The Author(s) 2024.


 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/deed.en>).

Table S1: The direct and palindromic repeat sequences

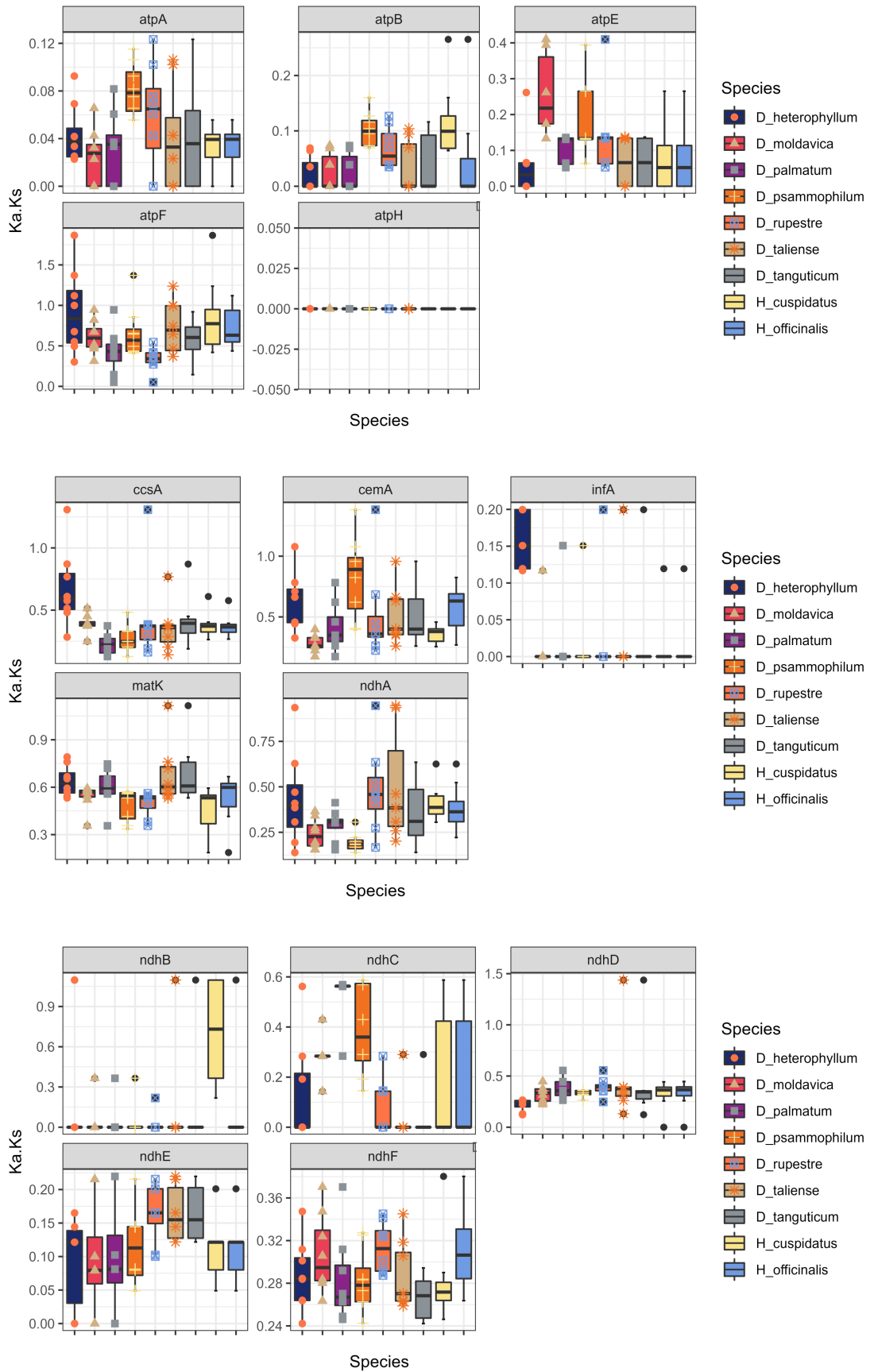
Repeat sequence length	Repeating unit	Repeating type	Repeat sequence two length	Repeat the starting point of unit 2	Repeating unit interval	e-value
53	64,919	P	53	64,919	-3	4.91E-17
42	28,244	P	42	28,244	0	3.26E-16
43	88,709	D	43	88,727	-1	1.05E-14
43	88,709	P	43	142,478	-1	1.05E-14
43	88,727	P	43	142,496	-1	1.05E-14
43	142,478	D	43	142,496	-1	1.05E-14
41	95,596	P	41	114,841	-2	9.62E-12
41	114,841	D	41	135,611	-2	9.62E-12
42	42,134	P	42	114,841	-3	1.01E-10
39	42,137	D	39	95,598	-2	1.39E-10
39	42,137	P	39	135,611	-2	1.39E-10
31	42,145	D	31	95,606	-1	1.27E-07
31	42,145	P	31	135,611	-1	1.27E-07
31	95,606	P	31	114,841	-1	1.27E-07
30	7,895	P	30	43,833	-1	4.92E-07
33	50,017	P	33	50,022	-3	1.26E-05
30	9,407	D	30	34,882	-2	2.14E-05
30	104,361	D	30	104,392	-2	2.14E-05
30	104,361	P	30	126,826	-2	2.14E-05
30	104,392	P	30	126,857	-2	2.14E-05
30	126,826	D	30	126,857	-2	2.14E-05
31	44,931	P	31	44,938	-3	1.66E-04
30	7,895	D	30	34,034	-3	5.99E-04
30	53,310	P	30	63,439	-3	5.99E-04
30	53,427	D	30	68,111	-3	5.99E-04
30	86,321	D	30	86,363	-3	5.99E-04
30	86,321	P	30	144,855	-3	5.99E-04
30	86,363	P	30	144,897	-3	5.99E-04
30	144,855	D	30	144,897	-3	5.99E-04

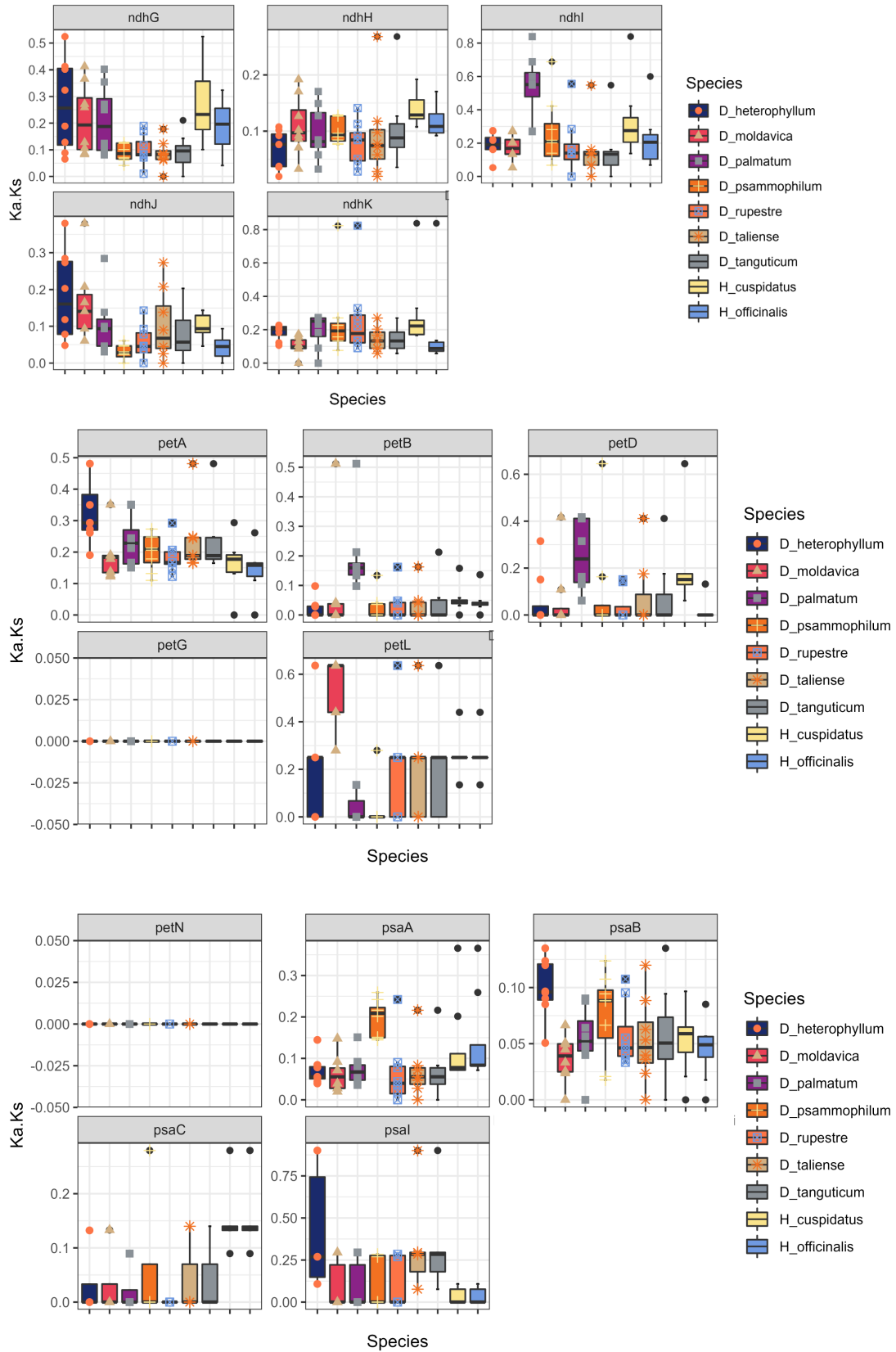
Table S2: Frequency of identified SSR motifs

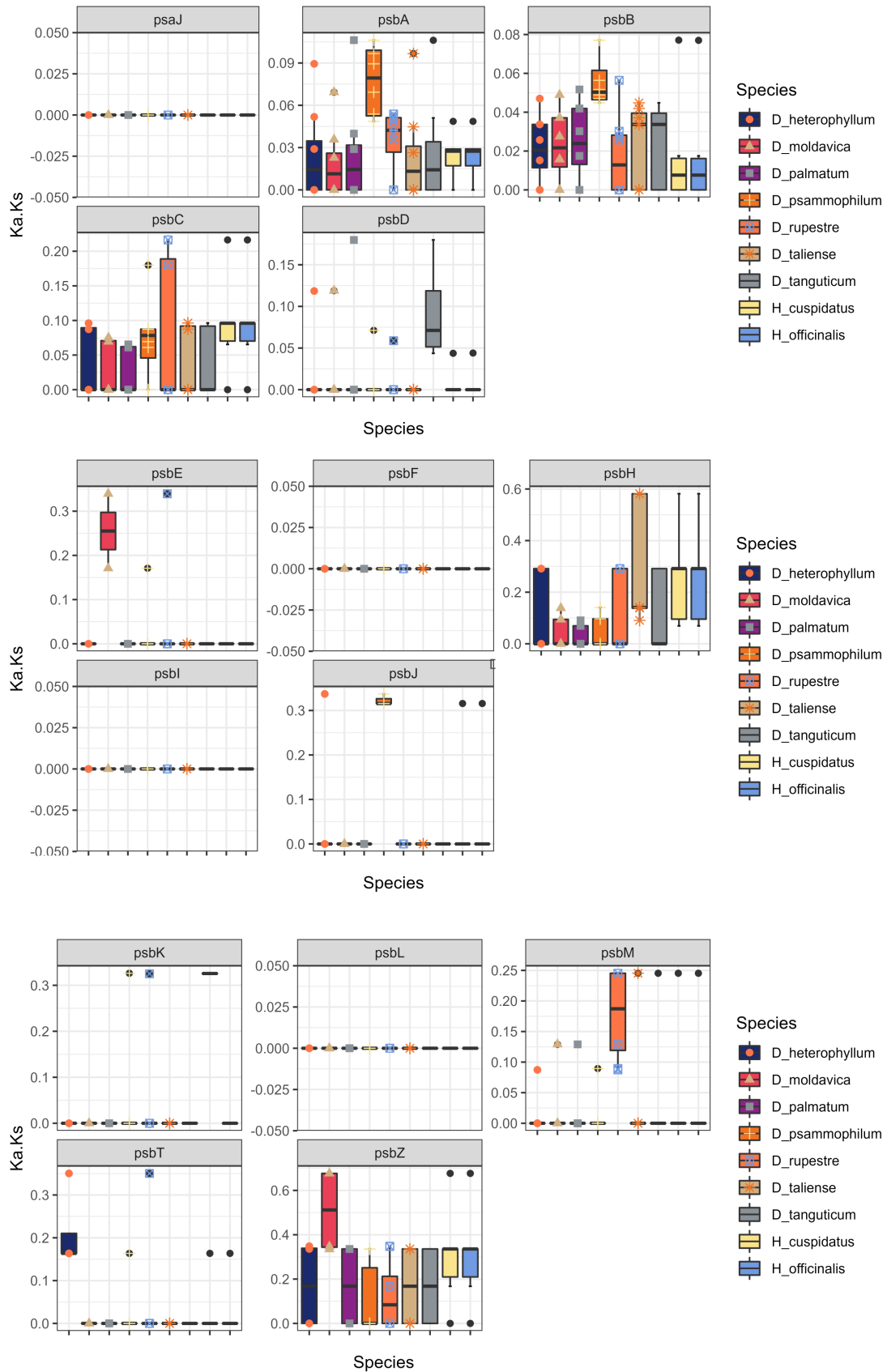
Repeats	5	6	7	8	9	10	11	12	13	14	15	16	17	total
A	-	-	-	-	-	9	7	2	4					22
G	-	-	-	-	-		1							1
T	-	-	-	-	-	11	7		1		1	1	2	23

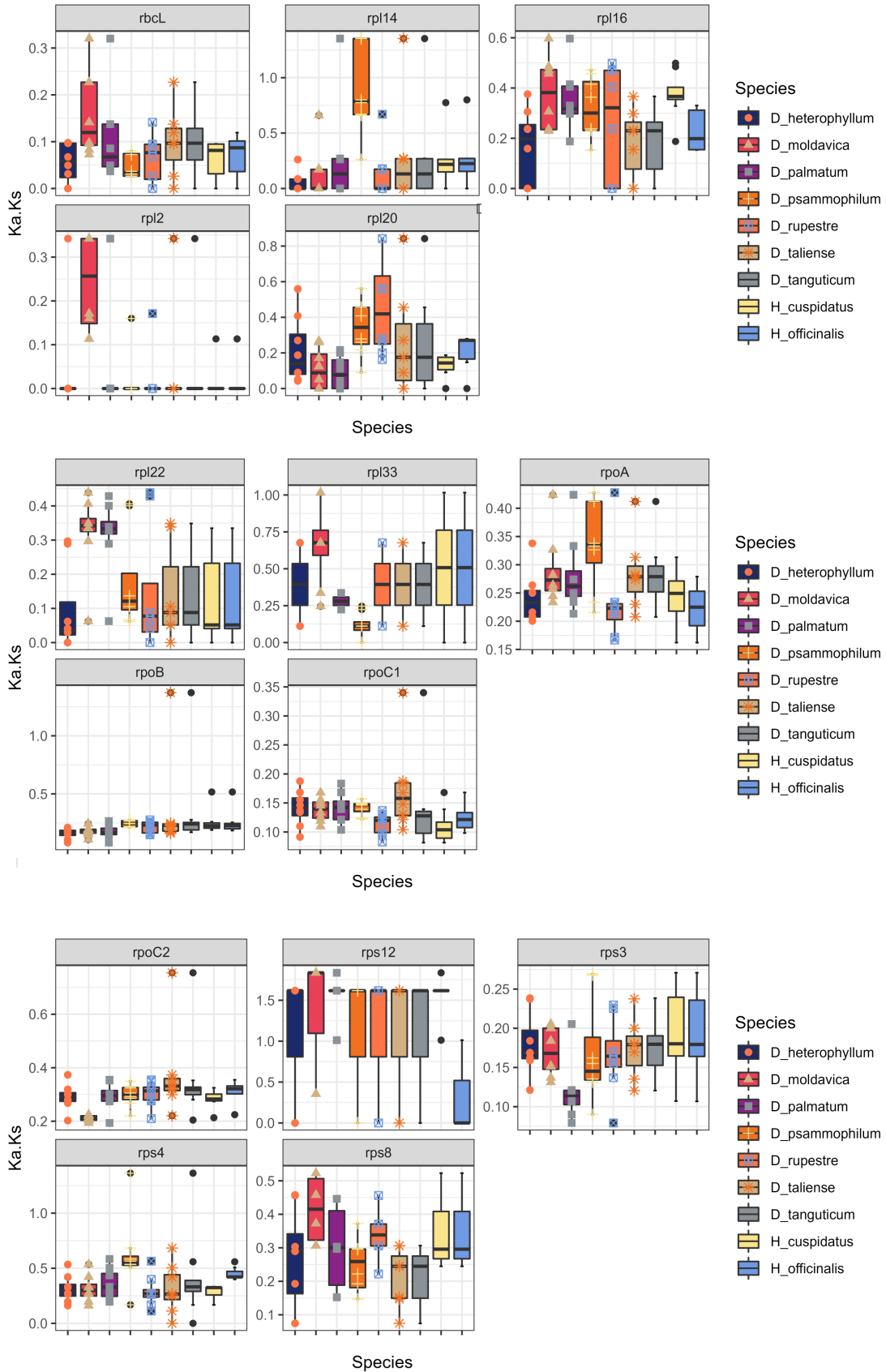
Table S2: Codon Usage in *Hyssopus cuspidatus* chloroplast genome

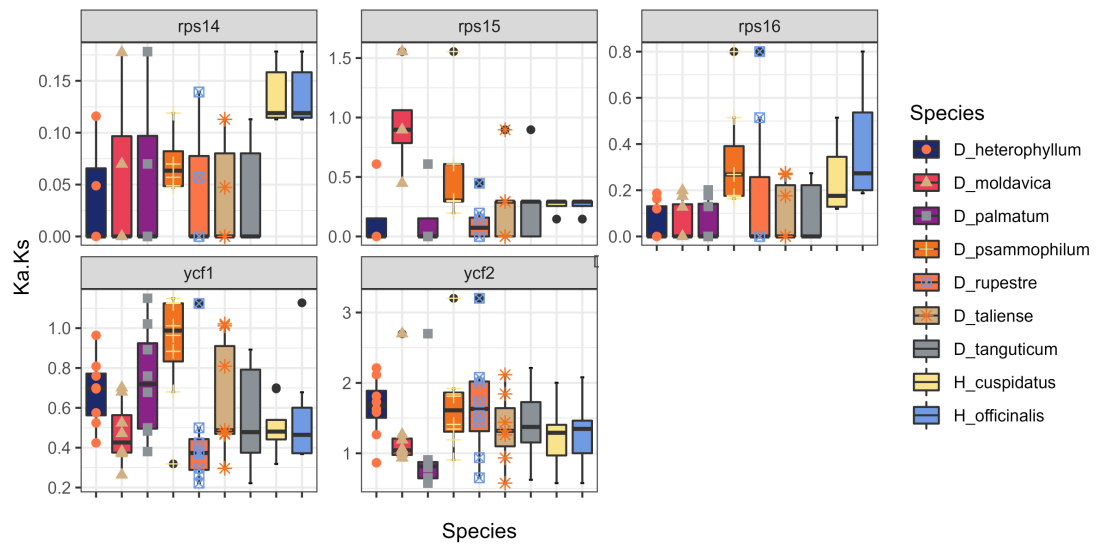
Codon	Amino acid	Frequency	Number
GCA	A	13.959	683
GCC	A	8.768	429
GCG	A	6.376	312
GCT	A	21.561	1,055
TGC	C	3.454	169
TGT	C	8.87	434
GAC	D	7.684	376
GAT	D	31.392	1,536
GAA	E	39.035	1,910
GAG	E	12.303	602
TTC	F	19.64	961
TTT	F	39.996	1,957
GGA	G	24.137	1,181
GGC	G	8.277	405
GGG	G	13.591	665
GGT	G	18.68	914
CAC	H	6.458	316
CAT	H	17.597	861
ATA	I	25.485	1,247
ATC	I	15.716	769
ATT	I	39.996	1,957
AAA	K	40.18	1,966
AAG	K	15.819	774
CTA	L	14.347	702
CTC	L	7.133	349
CTG	L	7.296	357
CTT	L	21.745	1,064
TTA	L	32.168	1,574
TTG	L	23.094	1,130
ATG	M	23.299	1,140
AAC	N	11.833	579
AAT	N	35.684	1,746
CCA	P	11.629	569
CCC	P	9.728	476
CCG	P	7.337	359
CCT	P	13.836	677
CAA	Q	27.325	1,337
CAG	Q	8.89	435
AGA	R	17.903	876
AGG	R	8.236	403
CGA	R	12.344	604
CGC	R	4.251	208
CGG	R	6.049	296
CGT	R	11.649	570
AGC	S	5.539	271
AGT	S	15.839	775
TCA	S	15.042	736
TCC	S	12.978	635
TCG	S	8.134	398
TCT	S	21.439	1,049
ACA	T	14.184	694
ACC	T	9.299	455
ACG	T	5.212	255
ACT	T	19.824	970
GTA	V	19.415	950
GTC	V	7.112	348
GTG	V	7.684	376
GTT	V	19.579	958
TGG	W	18.578	909
TAC	Y	7.112	348
TAT	Y	29.409	1,439
TAA	*	3.638	178
TAG	*	3.045	149
TGA	*	2.187	107











Supplementary figure 1: The Ka/Ks ratios of 65 genes