A Comparative Study of the Chloroplast Genomes of Five Lepidium Medicinal Plants

Qian Zhou

Kashi University

Yun Chen

Kashi University

Mingyuan Li

Kashi University

Weijun Zeng

Xinjiang Normal University

Jilian Wang

Kashi University

Yuzhou Wang

Kashi University

Yanhong Li

Xinjiang Normal University

Huixin Zhao (✉ zhaohuixin101@sina.com)

Xinjiang Normal University

Research Article

Keywords: Chloroplast genome, Lepidium, Comparative analysis, Codon usage bias

Posted Date: September 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-916085/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Herb genomics is a rapidly developing field of medicinal plant research and development. Plant genomic studies demonstrate the unique advantage of employing plants in medicinal therapy. The genus *Lepidium* falls under the *Brassicaceae* family and it includes crucial medicinal plants. Herein, we sequenced the complete chloroplast (cp) genomes of *Lepidium apetalum* (LA) and *Lepidium perfoliatum* (LP) and assessed their genetic profiles against the reported profiles of *Lepidium sativum* (LS), *Lepidium meyenii* (LM), and *Lepidium virginicum* (LV).

Results

In particular, we examined genomic arrangement, gene number, type, and repeat sequences. Based on our annotation data, both LA and LP possessed 130 distinct genes that included 85 protein-coding, 37 transfer RNA (tRNA), and 8 ribosomal RNA (rRNA) genes. Our repeat analyses revealed that LA harbored 20 forward repeats, 16 palindrome repeats, 30 tandem repeats, and 87 simple sequence repeats, whereas LP had 15 forward repeats, 20 palindrome repeats, 4 reverse repeats, 21 tandem repeats, and 98 simple sequence repeats. Using syntenic analysis, we also revealed a high degree of sequence similarity within the coding regions of *Lepidium* cp genomes and a remarkably high degree of divergence among the intergenic spacers. Pairwise alignment and single-nucleotide polymorphism (SNP) examinations further revealed certain *Lepidium*-specific gene fragments, particularly in the intergenic regions of the *trnK-atpA*, *trnC-psbC*, *trnT-rbcL*, *ndhF-ndhH*, *ycf1-trnR*, *accD*, *ccsA*, *matK*, *ndhF*, *rpoB*, *rpoC2*, and *ycf1* genes. Moreover, following codon usage analysis, we observed that codon 14 was the most frequently used codon in the *Lepidium* CDS. In addition, correlation investigations revealed that the ENC (the effective number of codon) content was strongly associated with GC3, GC3s, and N.

Conclusion

Based on these data, LA and LP originate from very similar genetic backgrounds. Furthermore, neutrality, ENC, and PR2-plots analyses demonstrated that the CUB (the codon usage bias) of *Lepidium* cp genome was strongly influenced by mutation and natural selection. Our analysis of the cp genomic sequences of LA and LP will likely enhance breeding, species recognition, phylogenetic evolution, and cp genetic engineering of the *Lepidium* medicinal plants.

Background

Lepidium, belongs to the *Brassicaceae* family, and is widely distributed all over the world. It has around 150 species, with 15 species in China[1]. It has been used in Traditional Chinese Medicine (TCM) or ethnic medicine for decades[2]. Thus far, these medicinal plants are reported to contain large quantities of active principles like cardiac glycosides, flavonoids, terpenoids, sterols, tannins, coumarins, alkaloids, phenolic acids, volatile oils, fatty oils, amino acids, and so on[3]. As such, these plants can be beneficial for patients with cardiovascular diseases, hyperlipemia, diabetes mellitus, cancer, and so on[4]. *L. apetalum* (LA) seeds possess medicinal properties that are harnessed in TCM[5]. LA seeds are known to be aperient, diuretic, tonic, demulcent, carminative, galatagogue, and emmenagogue[6]. However, these characteristics and their subsequent pharmacological effects are common

in other *Lepidium* plants as well. *L. perfoliatum* (LP), for instance, contains multiple plants bearing anti-inflammatory, chemopreventive, insecticide, antibacterial, antiviral, and antiparasitic characteristics. Here, the hydroalcoholic extract of LP was examined against *T. gondii* (TG) *in vitro*. Our data revealed that glucoerucin protects LP from CCl₄-mediated hepatic destruction via suppression of the oxidative damage networks[7]. The water extract of *L. virginicum* (LV) contains anti-cancer and anti-inflammatory properties. Additionally, the presence of benzyl glucosinolate in this extract also exerts a bacteriostatic function[8, 9]. *L. meyenii* (LM), on the other hand, contains not only rich nutrition, but also plenty of secondary metabolites. Multiple reports indicate that LM exerts properties that resist oxidation and fatigue, as well as enhance fertility, modulate endocrine system, increase immunity, suppress tumor, relief osteoporosis, control blood sugar, and protect the nervous system[10, 11]. Lastly, *L. sativum* (LS) is a traditional medicine available in many parts of the world. Its main chemical constituents are alkaloids, flavonoids, terpenes, and glucosinolates. Its pharmacological activities include protecting against hypoglycaemia and hypertension, as well as enhancing fracture healing and drug efficacy in gastrointestinal diseases[12, 13].

Owing to maternal inheritance and evolution rate, cp gene sequences have been extensively studied in the fields of plant phylogeography and molecular evolution, as well as phylogenetics, phylogenomics, and genomic evolution[14]. Because of this, cp genome is regarded as excellent models for investigating lineage-specific molecular evolution. Recently, the complete cp genome was placed in the spotlight for its massive inherent information that can remarkably enhance resolution at lower plant taxonomic levels[15, 16]. Studies involving microstructural mutations within the cp genome, such as, indels, small inversions, and inverted repeats (IRs) have added to the existing knowledge of genomic evolution among plants[17]. Moreover, the presence of repeated sequences, particularly, simple sequence repeats (SSR), short tandem repeats (STR), homopolymeric repeats, and long repeats are speculated to originate from various pathways like gene conversion, intramolecular recombination, and slipped-strand mispairing (SSM)[18]. Genomic events like duplication, deletion, and rearrangement mostly result from repeated sequences[17, 19]. In present times, the availability of next-generation- (NGS) and third-generation sequencing (TGS) offers rapid and cost effective strategies to sequencing cp genomes[20].

Codons serve an essential function in transmitting genetic information[21]. This information is passed along from DNA to RNA to protein using codons, which are sets of three nucleotides. Subsequently, each protein unit (amino acid) is processed from 1–6 codons[22]. It is possible for several codons to encode the same amino acid. These codons are known as synonymous codons (SCs)[23]. Interestingly, the likelihood of SCs usage differs greatly during protein synthesis. In circumstances that promote natural selection or mutation preference, one or more specific SCs are used. This is called synonymous codon usage bias[24–26]. CUB is common among different organisms and it is brought on by factors like environmental factors, base mutation, gene drift, genome size, tRNA abundance, and gene expression[27, 28].

Based on our existing knowledge of the cp genome, genetic engineering using CUB can greatly enhance plant genome in a way that maximizes pharmacological benefit[29]. Till now, studies have examined cp genome CUB in various plants, including *Drynaria*[30], *Panicum* species[31], and *Oryza* species[21]. However, thus far, there have been no reports of CUB in *Lepidium*. Genomic studies, including nuclear and organelle DNA, can enhance our understanding of the origin of medicinal plants. Herein, we employed NGS to obtain the complete cp genomic sequence of LA and LP, and compared them to the reported genomic sequences of three other *Lepidium*, namely, LS, LM, and LV. We based our analysis on the genomic arrangement, presence of repeat sequences, and IR length,

and established the CUB and optimal codon of the *Lepidium* cp genome. Moreover, using cluster analysis, we generated an evolutionary tree. Our work adds to the existing knowledge of the *Lepidium* cp genome and can be instrumental in the generation of the *Lepidium* cp genomic library. This will enhance recognition of *Lepidium* and provide clues to its evolutionary origins.

Results

Genomic arrangement and gene characteristics

Raw base (approximately 3.94Gb and 4.45Gb) and clean base (approximately 4.4Gb and 3.9Gb) were obtained from LA and LP, respectively, and the resulting cp genomes MT880914 (154,846bp long) and MT880913 (154,264bp), respectively, were uploaded to the NCBI website. Both LA and LP exhibited a standard quadripartite structure that is common in most flowering plants. The structure consisted of a pair of IRs (26,450 and 26,478 bp) and 2 single-copy regions (large single-copy region (LSC): 83,391 and 83,896bp; small single-copy region (SSC): 17,973 and 17,994bp) (Fig. 1 and Table 1). Here, we annotated 130 genes in two LS cp genomes that comprised of 85 protein-coding, 37 transfer RNA (tRNA), and 4 ribosomal RNA (rRNA) genes. Among them, 15 genes (*trnK-UUU*, *trnL-UAA*, *rps16*, *trnG-UCC*, *atpF*, *trnV-UAC*, *rpoC1*, *ndhB*, *petB*, *rpl16*, *petD*, *rpl2*, *trnA-UGC*, *trnI-GAU* and *ndhA*) comprised of 1 intron and 2 genes (*clpP*, *ycf3*) had 2 introns. Moreover, one trans-splicing gene (*rps12* gene) was detected as well, as shown in Table 1.

A minor variation was identified in the base composition among the 5 *Lepidium* cp genomes (Table 2). The LP co genome was the shortest, the LA and LM were medium sized, and the LS and LV had the biggest length. In particular, the co genome lengths ranged from 154,264-154,997bp. The LSC lengths ranged from 83,391-84,035bp. The IR ranged from 26,360-26,491bp. Lastly, the SSC ranged from 17,973-18,008bp. The gene number was between 128-130 and the protein coding genes were between 83-85. All analyzed *Lepidium* cp genomes produced 37 tRNA genes and 8 ribosomal RNA genes. The GC contents were 36.40%, 36.46%, 36.31%, 36.47%, and 36.41% , respectively. All 5 *Lepidium* cp genomes exhibited more AT than GC, compared to other land plants[32-35].

A minor variation was detected in the base composition of the coding sequences (CDS) of all 5 *Lepidium* cp genomes (Table 3). The CDS GC contents were 37.03%, 37.05%, 37.09%, 37.12%, and 37.08%, respectively. The first codon GC contents were 44.81%, 44.78%, 44.90%/44.92%, and 44.87%, respectively. The second codon GC contents were 37.53%, 37.62%, 37.69%, 37.69%, and 37.61%, respectively. Finally, the third codon GC contents were 28.75%, 28.76%, 28.68%, 28.75%, and 28.76%, respectively (Table 3). All 5 *Lepidium* cp genomes exhibited a preference for AT in the third codon position, similar to what is observed in other angiosperms[36-38]. We also calculated a codon adaptation index (CAI) of 0.165, which suggests lack of bias in codon usage in all 5 species. The effective number of codons was between 49.04 and 52.93. LS displayed elevated ENC values, compared to the other 4 species.

Overall, 85 protein-coding genes carried 26,608, 26,601, 26,4512, and 26,373 codons in the LA, LP, LM and LV cp genome, respectively (Table S1). In comparison, LS displayed 26,105 codons that encoded 83 proteins. Among these codons, AAA (1149/1160/1128/1147) encoding lysine (Lys) and UGC (83/83/80/85) encoding cysteine (Cys) were the most and least used in LA, LP, LS, and LM, respectively. In addition, AUU encoding isoleucine (Ile) and UGC were the most and least used in the LV cp genome (1146 and 83).

The relative SC usage (RSCUs) among the 5 *Lepidium* cp genomes had minor variations. Several codons encoded the same amino acids, including methionine and tryptophan (Fig. 2). When RSCU > 1, 30 identical codons were identified, 29 of which were A/U-ending codons (except for UUG). When RSCU < 1, 32 identical codons were found, 29 of which were C/G-ending codons (except AUA, CUA and UGA). The largest and lowest RSCU values were for UUA and CUG, respectively. Interestingly, both codons produced leucine (Leu). In summary, we demonstrated that UUA was positively biased, whereas CUC was negatively biased.

Introns are critical for modulating gene expression and can be manipulated to enhance certain beneficial gene expression[39,40]. The 5 analyzed species contained 18 genes with introns. Among them, 15 carried one intron, whereas another 3 carried two. The *trnK-UUU* gene had the largest intron and harbored the *matK* gene. Additionally, the 5' end of the *rps12* gene resided in the LSC region and its 3' end was found in the IR region, thereby making it a trans-spliced gene. We also found that the exon lengths were conserved in all 5 species, other than the *ndhB* and *trnG-UCC* genes (Table S2).

We next compared the 18 genes of LA with the other *Lepidium* species. There were slight differences in most exon lengths. The exon lengths of the *trnG-UCC*, *rpoC1*, *ndhA*, and *ndhB* genes were unique to LA. A larger difference was seen in intron lengths, compared to the exon lengths. Of note, some intron lengths remained the same in all 5 species (*ndhB*, *rpl2*, and *rps12*).

RNA editing is a highly beneficial tool employed for enriching transcription and protein diversity[41-43]. Here, 16 genes belonging to the 5 *Lepidium* cp genomes were analyzed for possible RNA editing sites (RES). We estimated 45 RESs. Among them, 39 possessed common sites within all 5 species (Table S3). Among the 45 putative RES, 13 were located in the first position of the corresponding codon and 32 were found in the second location. We did not find any RES in the third position. In addition, we noticed that all base conversions were from C to T. These data resemble those of other land plants[44]. In terms of amino acids, the Ser to Leu conversion was the most frequent, and the Ala to Val, Arg to Trp, and Thr to Met conversions were least frequent.

SSR analysis

Owing to great advancements in NGS technology, SSR assessments can be made faster, with high precision, and in a highly cost-effective manner, even in species that were not examined previously[38]. Given that cp genome SSRs can serve as crucial phylogenetic markers and valuable resources for genomic evolution, we identified and quantified 6 types of repeat patterns in *Lepidium* (Fig. 3 and Table S4). Using the microsatellite recognition tool MISA, we screened 87, 98, 91, 87, and 96 SSRs in LA, LP, LS, LV, and LM, respectively. Among them, the most prevalent SSRs were A or T mononucleotide repeats, which were found in 67.8%, 66.32%, 69.2%, 67.8%, and 63.5% of the total SSRs, respectively. Compared to the di-, tri-, and tetranucleotides, there were fewer quantities of penta- and hexanucleotides. Moreover, SSRs were mostly identified in the LSC regions (75.6%, 75.8%, 67.0%, 70.1%, and 66.7%, respectively), followed by the SSC regions (18.2%, 16.48%, 19.7%, 16.1%, and 18.8%, respectively) and the IR regions (6.1%, 7.7%, 6.6%, 6.9%, and 7.3%, respectively).

Comparing SSRs among the 5 species revealed high level of similarities. In all 5 species, most SSRs consisted of mono and di repeat motifs. The mono and di repeats varied from 59-68 and 14-18, respectively. In contrast, >3 oligonucleotide repeats were somewhat low in quantity, but rich in types.

Furthermore, we used REPuter and Tandem Repeats Finder to compare repeat sequences among the 5 species (Fig. 4). In all, there were 66 repeats in LA, 60 in LP, 69 in LS, 67 in LV, and 64 in LM. LA consisted of 20 forward, 16 palindrome, and 30 tandem repeats. LP had 15 forward, 20 palindrome, 4 reverse, and 21 tandem repeats. LS harbored 27 forward, 15 palindrome, and 27 tandem repeats. LV possessed 20 forward, 16 palindrome, and 31 tandem repeats. Lastly, LM possessed 16 forward, 16 palindrome, 2 reverse, and 30 tandem repeats. The lengths of these repeats in all 5 species ranged from 10-30 bp, whereas the forward and palindromic repeats ranged from 30-45 bp.

Using repeat sequence analysis (REPuter), we identified 66 and 22 repeats in LA and LP (Table S5). However, part repeats were found in the *ycf2*, *psaB*, *psaA*, *ycf1*, and *trnS-GCU* coding region; part were found between the *ycf3*, *ndhA*, and *trnL-UAA* intron region, and the rest were found in the IGS regions. In LA and LP, 30 and 21 tandem repeats were detected, respectively (Table S6). Part of these repeats were located in the *ycf1*, *ycf2*, *rpl2*, *ndhA*, *trnF-GAA*, and *atpF* genes, while most of the repeats were located in the IGS regions. The longest repeat, consisting of 83 bp in length, was observed in *trnF-GAA* in the LA.

Comparing the cp genome of all 5 *Lepidium* species

Here, we performed genomic analysis of 5 complete *Lepidium* cp genomes using the mVISTA program. Based on our analysis, there was substantial homology in genomic composition and size among the species (Fig. 5). The coding regions of these species strongly resembled one another. However, the non-coding areas were variable. The overall sequence identity and detection of the divergent regions in all 5 species were done with the mVISTA and DnaSP programs (Fig. 5, Table S7 and S8). The cp genomic arrangement analysis demonstrated a high degree of synteny and gene order conservation, indicating evolutionary conservation of these genomes at the genomic level. As mentioned earlier, there was a higher degree of divergence in the non-coding, relative to coding locations. The coding areas that showed divergence among species included the *ycf1*, *matK*, and *ndhF* genes. The highest divergence in the non-coding locations was discovered in the *trnK-atpA*, *trnC-psbC*, *trnT-rbcL*, *ndhF-ndhH*, *ycf1-trnR*, *accD*, *ccsA*, *matK*, *ndhF*, *rpoB*, *rpoC2*, and *ycf1* genes. Of note, the LSC and SSC locations were more divergent, compared to IR.

IR contraction and expansion

The IR region is a highly conserved part of the cp genome and it is prevalent in plant evolutionary events involving border region contraction and expansion. Here, we assessed the IR boundaries of LA and LP against those of LS, LM, and LV (Fig.6). To do this, we aligned the LSC, IRA, SSC, and IRB borders with adjoining genes of members of the *Lepidium* species. We demonstrated that these locations hold strong similarity to all previously described cp genomes[45]. It is speculated that the expansion and contraction events between the two IR and single-copy regions give rise to genomic size variations among plant lineages[14]. Based on our data, the IR regions were more conserved than the LSC and SSC regions in the *Lepidium* cp genomes. As a result, despite expansion or contraction events within IR regions, the overall cp genomic size remained somewhat stable. Owing to expansion events, the *rps19*, *ndhF*, *trnH*, and *ycf1* genes were partly incorporated into the IR regions of the *Lepidium* species. The *rps19* and *trnH* genes, on the other hand, resided in the LSC/IRb and IRa/LSC locations, respectively. Its distance from the LSC/SSC borders was variable among the different *Lepidium* species, particularly LV. Conversely, 2204bp of the *ndhF* gene was extended into the SSC locations in all 5 analyzed species. In addition, 1036, 1036, 1039, 907, and 1039bp of the *ycf1* gene was extended into the IR regions in the

LA, LP, LS, LV, and LM species, respectively. Collectively, these evidences suggest that the expansion and contraction of the IR/SC locations in *Lepidium* do not dramatically alter their genomic patterns.

CUB analysis

We next evaluated the base composition of 423 CDS in the 5 *Lepidium* species to determine whether the base content affected CUB in these species[46]. We identified that the GC1 base content varied from 26.42-62.20% (average: 45.36%), GC2 content was between 25.64-56.83% (average: 39.37%), and GC3 content was between 12.50-37.50% (average: 27.13%). Hence, based on our analysis, GC1 frequency was the highest, whereas GC2 and GC3 were comparable. To next assess the influence, if any, of base content on CUB, we next computed the GC contents and GC3. Our data revealed that the GC content in all genes was between 28.43-44.70% (average: 37.29%) (Table S9). This suggests that the *Lepidium* genes have a higher preference for AT than GC. Next, we examined the association between GC12 and GC3, using a neutrality plot (GC12 against GC3). The GC12 content varied from 30.77-55.49%. Moreover, the regression coefficient was 0.1696 (Fig. 7), which suggested that CUB was based on natural selection, and not mutational pressure

We next evaluated the influence of RSCU and ENC on CUB. ENC in *Lepidium* was from 21.46-61.00 (average: 45.64). Using ENC values, we demonstrated that 24 genes exhibited high CUB and carried an ENC value of <35. These data indicate the presence of a common randomized codon usage in *Lepidium* genes and no CUB preference. Moreover, the RSCU values of 59 sense codons also suggested that the *Lepidium* genes have weak CUB. As depicted in Fig. S2, over 50% (29/59) of codons are underlined (RSCU>1) and they represent frequent usage. The high-frequency codons were identified as UCU, UUA, AGA, GCU, ACU, GGA, UAU, AAU, CAA, GAU, GAA, CAU, AAA, and CCU.

To elucidate the significance of GC3s in CUB, we next generated an ENC-plot. Based on our analysis, most genes had an ENC value >351. In addition, most *Lepidium* genes were located some distance away from the expected ENC-plot curve, with a few located around the curve (Fig. 8). This indicates that the mutational pressure alone does not influence CUB. In fact, CUB may be influenced by other factors like translational selection and so on.

We further assessed associations among the 6 essential parameters, namely, GC1, GC2, GC3, GC, GC3s, N, and ENC (Table 4). ENC displayed marked associations with 4 other parameters, namely, GC3, GC3s, and N ($r=0.243$, $r=0.285$, and $r=0.288$; $p < 0.001$), suggesting that the CUB of *Lepidium* genes are regulated by two primary agents: mutational pressure and translational selection[47].

We next analyzed the C/T to A/G balance within the third codon. Prior reports suggest that during mutational pressure on CUB, AU and GC are often paired at the third codon position[48]. Interestingly, our data revealed that *Lepidium* genes have vastly different frequencies of A3 (A3s: the adenine content at synonymous third codon position) and T3 (T3s: the thymine content at synonymous third codon position) or G3 (G3s: the guanine content at synonymous third codon position) and C3 (C3s: the cytosine content at synonymous third codon position). Evaluating the GC and AT pair frequencies in genes can delineate its significance in CUB. We, thus, generated a ParityRule 2 (PR2) plot to examine the presence or absence of bias in *Lepidium* genes. As depicted in Fig. 9, a majority of the distribution remained between 0-0.85. This suggests a low G3/C3 or A3/T3 bias in *Lepidium*. Additionally, we divided the plot into 4 quadrants, with 0.5 as the center of both axes. In doing this, we observed additional points in the third and fourth quadrants (the ratio of $G3/(G3+C3) < 0.5$), relative to the first and second.

The first points were located in the second quadrant. Based on these collective evidences, there was a minor, but significant, preference for C and T at the third codon position in *Lepidium* genes, thereby validating the presence of other factors like translational selection in CUB formation.

Phylogenetic analysis

To identify the location of LA and LP in the *Brassicaceae* phylogenetic tree, we analyzed 21 complete *Lepidium* cp genome sequences. Our generated phylogenetic trees, using ML, showed comparable topologies (Fig. 10). In the tree, all 5 species exhibited a strong monophyletic branch. LS and LV were clustered together in 1 terminal branch and showed more similarities to LA. LP, on the other hand, was located further away and clustered with LM.

Discussion

Great advances in NGS technology have made it easier to sequence entire cp genomes precisely and cost-effectively. In this work, we retrieved the entire sequence of the LA (154,846bp) and LP (154,264bp) cp genomes and compared them to other *Lepidium* species, namely, LS, LV, and LM. Based on our analysis, LA and LP cp genomes contained 130 unique genes that encoded 85 proteins, 8 rRNAs, and 37 tRNAs. Among these genes, the CDS of the *ycf2* gene was the largest in length. In addition, the 5' end of the *rps12* gene was discovered in the LSC location, whereas its 3' end was in the IR location, as is often seen in angiosperm cp genomes. Moreover, the LA and LP cp genomes were of a standard quadripartite nature and its gene content, genetic arrangement, and GC content were comparable to the analyzed *Lepidium* species. We also revealed that despite having a highly conserved IR (relative to LSC and SSC locations), the border expansion and contraction events did not alter cp genomic size to a great extent. In summary, the LA and LP cp genomic arrangement, GC content, gene number and type, along with codon usage, were all comparable to LS, LM, and LV.

SSRs are sequence repetitions, distributed over the entire genome, that are crucial to genomic recombination and rearrangement. Generally, cp genomic SSRs are commonly found in intergenic locations[49]. During SSR analysis, 87–98 SSR loci were identified in the intergenic region, carrying a total of 66–69 repeats. This information can be used to examine intraspecific polymorphisms and develop lineage-specific markers for future evolution- and genetic diversity-related investigations. Based on our mVISTA data, all 5 analyzed *Lepidium* species had comparable cp genomic sequences. Additionally, the IR locations exhibited more conserved sequences and content, relative to the LSC and SSC regions, likely due to IR region-based rRNA. We revealed that although the LA and LP cp genes were very similar to the other *Lepidium* species, the genetic composition was more variable in the intergenic regions. Therefore, we recommend using the intergenic regions to potentially identify between LA, LP, LM, LV, and LS.

DNA barcoding, first introduced by Hebert et al[50], is a valuable species identification tool that employs species-specific genetic recognition via genes like *ITS2*, *matK*, *psbA-trnH*, and *rbcL*. Nevertheless, species identification can sometimes become difficult due to morphological confusions within the same genus. Hence, establishing an appropriate DNA marker for species identification is crucial. Cp genomes are often examined in phylogenetic studies and species recognition since their evolutionary rate is markedly slower than nuclear genomes[51]. Here, we analyzed the pairwise cp genomic alignment and SNP of 5 *Lepidium* species cp genomes and demonstrated a large number of species-specific locations within the cp genomes. The locations included the *trnK-*atpA**, *trnC-psbC*, *trnT-rbcL*, *ndhF-ndhH*, *ycf1-trnR*, *accD*, *ccsA*, *matK*, *ndhF*, *rpoB*, *rpoC2*, and *ycf1* genes. Given their unique

specificity, these genes can be potential biomarkers for *Lepidium* identification. Despite our extensive work in this paper, we plan to further examine the *Lepidium* cp genome sequencing data to narrow our search of *Lepidium* identifiers.

Next, we examined CUB in *Lepidium*. Based on our data, the GC content order was GC1 > GC2 > GC3. In addition, most genes possessed an ENC value > 351, with a high-frequency codon number of 14. Based on our correlation analysis, the ENC content was strongly related to GC3, GC3s, and N. Using neutrality, ENC, and PR2-plots, we demonstrated that the *Lepidium* cp genome CUB was primarily influenced by mutation and natural selection. Here, we positioned LA and LP in the *Brassicaceae* phylogenetic tree via analysis of 21 complete cp genomes. Using a ML tree, we demonstrated that all 5 *Lepidium* species formed a strong monophyletic branch. We also revealed that LS and LV formed a strong monophyletic branch with LA, followed by LM, suggesting that LS and LV are closely related to LA, compared to other *Lepidium* species. LP, on the other hand, was further away and was clustered with other *Lepidium* species. This data points to a potential association between the cp genome and genetic evolution of the plant species. Our work provides essential knowledge about *Lepidium* cp genomes which can be used to manipulate gene expression and maximize beneficial effects of *Lepidium*, as well as aid in the genetic and evolutionary studies of these species.

Conclusion

We were successful in assembling, annotating, and analyzing the complete cp genomic sequence of LA and LP, a *Lepidium* species. Cp genes are generally conserved across multiple species. In addition, the LA and LP cp genomes are comparable to other *Lepidium* species. Here, we recognized 87–98 SSR loci in the *Lepidium* cp genome, which are unique to this species and can be employed as lineage-specific markers. In addition, we compared the repeated sequences like SSR among the 5 species of *Lepidium* and revealed discrete differences. We also demonstrated that the IR region border is also comparable among the 5 *Lepidium* species. Lastly, CUB is primarily influenced by mutation and natural selection. Our conclusions, in this paper, offer an extensive overview of the phylogenetic associations between the different *Lepidium* species. Systemic knowledge of these cp genomes can highly benefit future breeding and biological discovery.

Methods

Sampling and DNA sequencing

We collected fresh LA and LP leaves from the Carp mountain in Xinjiang, China (87°34'716"E, 43°50'577"N) and deposited the Voucher sample in the Xinjiang Normal University (No.SHFZDXC01 and No.SHFZDXC02, respectively). The fresh leaves were instantly dried using silica gel and then processed for DNA analysis. In short, total DNA was isolated from ~1 g of leaves using an upgraded procedure by Li et al[52]. DNA quantification was done with an Agilent 5400 spectrometer (Agilent Technologies Inc, State of California, USA). DNA quality was assessed via library preparation and then sent for sequencing at TGS-Shenzhen, China, using an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, United States) with 150 bp paired-end reads.

Genomic assembly and annotation

The DNA yield of our genomic DNA extraction was 3.94 GB. Next, we performed qualitative analysis and assembly of the paired-end reads using SPAdes 3.9.0[53], followed by elimination of contigs with low sequencing

depths. The resulting contigs contained DNA from the cp, nuclear, and mitochondrial genomes. A BLAST search, with default parameters and LS cp genome sequence as reference (Gen Bank accession number: MN176145) was used to select cp genome sequence contigs. The cp genetic annotations were done with the DOGMA tool under default parameters[54]. The five junctions between the IRs and small single copy (SSC)/large single copy (LSC) regions were verified using PCR-based product sequencing[55]. Lastly, the complete cp genome sequence was loaded into GenBank.

Characterization of repeat sequences and SSRs

Repeat sequences that includes forward, reverse, and palindromic patterns were recognized via REPuter (Hamming distance:3, minimum repeat size: 3 and 30)[56]. SSRs are repeat sequences spanning 1–6 bp that can be found throughout the genome and that harbor markedly elevated mutation rates, compared to neutral DNA sites. We estimated SSR distributions in the cp genome with the microsatellite search tool MISA[57] and the following parameters: ≥ 10 for mononucleotide repeats, ≥ 5 for dinucleotide repeats, ≥ 4 for trinucleotide repeats, and ≥ 3 for tetranucleotide repeats, pentanucleotide repeats, and hexanucleotide repeats.

Comparative Genomic analysis

To assess the sequence divergence of the cp genome among the 5 selected *Lepidium* species, entire cp genome sequences were compared using the mVISTA program in the Shuffle-LAGAN mode[58], referencing LA annotations. Differences in the cp genome length, LSC length, SSC length, GC content, encoding gene types, and gene numbers were also compared. The LSC/IR/SSC boundaries among the species were established via comparative analysis and were used to examine variations in angiosperm cp genomes.

Codon usage bias (CUB) analysis

Codon W 1.4.2 was employed for the CDS examination of relative synonymous codon usage (RSCU). The online program CUSP model (<http://imed.med.ucm.es/EMBOSS/>) was determined by the GC content of the first (GC1), second (GC2), and third (GC3) codon base, as well as total GC content (GC). The CHIPS model (<http://imed.med.ucm.es/EMBOSS/>) analyzed the ENC (effective number of codons) value. In the neutral graph, GC3 values represented abscissa; the mean GC1 and GC2 values represented GC12, which were ordinate; and each point represented an independent gene[59]. In the ENC-plot, the horizontal ordinate comprised of GC3 values, and the longitudinal coordinate comprised of the ENC values. Using these, a two-dimensional scatter plot was protracted[60]. We also assessed the nucleotide in the third position of each codon and generated a PR2-bias plot using A3/(A3 + T3) as the ordinate and G3/(G3 + C3) as the abscissa. The base composition of each gene was depicted in a plane, whereby the center point referred to the codon state under unbiased usage. In other words, A = T and C = G. Moreover, the vector distance between the rest points and the center point represented the degree and direction of bias[61].

Phylogenetic analysis

Overall, 21 complete cp genome sequences were retrieved from the NCBI Organelle Genome and Nucleotide Resource Database to determine the phylogenetic positioning of *Lepidium* in *Brassicaceae*. The cp genomes were then aligned with the MAFFT software, while MEGA6[64] was utilized for construction of the phylogenetic tree using the ML technique. A bootstrap analysis was then performed with 1000 replicates and three bisection and reconnection (TBR) branch swapping.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are openly available in GenBank of NCBI at <https://www.ncbi.nlm.nih.gov>, reference number MT880913.1 and MT880914.1. The associated BioProject, SRA, and Bio-Sample numbers are PRJNA682689, SRR13213155(LA) and SRR13213156(LP), SAMN17005489(LA) and SAMN16993569(LP) respectively.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from Scientific research program of colleges and universities in Xinjiang (No. XJEDU2021I023), Natural Science Foundation of China (No. 32160074 and No.32160408), and the Open project of Key Laboratory in Xinjiang (No. 2020D4010).

Authors' contributions

QZ conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

YC and ML analyzed the data, prepared figures and/or tables, and approved the final draft.

WZ and JW analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

YW performed the experiments, contributed reagents/materials/analysis tools.

YL conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

HZ analyzed the data, contributed reagents/materials/analysis tools.

All authors read and approved the final manuscript

Acknowledgements

The authors would like to thank Chunyan Zhou and Ming Jiang for their technical help in the lab.

References

1. Wu ZY, Raven PH, Hong DY. Flora of China. Beijing: Science Press, 2001, pp. 28-33.
2. Zhang J, Zhao X, Wang Q. Research Advances of *Lepidium* Plants. Journal of Anhui Agricultural Sciences. 2016; 44(3):159-160,194.
3. Li XX, Zhao B, Cui C, Shen W, Gao A, Gong J, et al. Overview of Pharmacological Research on *Lepidium* L. Ningxia Journal of Agriculture and Forestry Science and Technology. 2012;53(1):91-92,100.
4. Zhu J, Du Y, Sun S, Qi D, Ma J. Research Progress in Chemical Constituents and Pharmacological Activities of *Lepidium* L. Journal of Qinghai Normal University. 2015;31(02):48-53.
5. Zhou TY, Lu LL, Yang G, Al-Shehbaz IA. *Orychophragmus* Bunge. In: WuZY, RavenPH, editors. Flora of China, Vol. 8. Beijing: Science Press; St. Louis: Missouri Botanical Garden Press, 2010; pp. 29–31.
6. Bansal D, Bhasin P, Yadav OP, Punia A. 2012. Assessment of genetic diversity in *L. sativum* (Chandrasur) a medicinal herb used in folklore remedies in India using RAPD. J Genet Eng Biotechnol. 2012;10(1):39–45.
7. Mirzaee F, Mohammadi H, Azarpeik S, Amiri F, Shahani S. Attenuation of liver mitochondrial oxidative damage by the extract and desulfo glucosinolate fraction of *Lepidium perfoliatum* L. seeds. South African Journal of Botany. 2021;138:377-385
8. Chae YH, Shin DY, Park C. Induction of Apoptosis in Human Colon Carcinoma HCT116 Cells Using a Water Extract of *L. virginicum* L. Journal of the Korean Society of Food Science Nutrition. 2011;40(5):649-659.
9. Fernando C, Elizabeth B, Roberto C. Antiamoebic activity of benzyl glucosinolate from *L. virginicum*. Phytother Res. 2003;17(6):618-619.
10. Li Y, Zhou P, Zhou X, Huang L, Yang L, Chen G. Research and application progress of *L. meyenii*(maca) China. Journal of Chinese Materia Medica. 2018;43(23):4599-4607.
11. Zhang H, Qiao T, Gu J. Research Progress on Biological Function and Application of *Maca*. Food and Nutrition in China. 2016;22(7):71-73.
12. Fan Q, Guo B, Huang W. Advances in studies on chemical constituents of *L. sativum* and its pharmacological activities. Drugs and Clinic. 2013;28(1):90-9.
13. Bao X, Zhou X, Guo Q, Han H, Shu Y, Wang Y. Study on the Antitussive, Anti-inflammatory and Bacteriostatic Effects of Semen Lepidii Sativi, an Uyghur Traditional Herbal. Medicine of Xinjiang Lishizhen Medicine and Materia Medica Research. 2009;20(9):2318-2320.
14. Dong W, Xu C, Cheng T, Zhou S. Complete chloroplast genome of *Sedum sarmentosum* and chloroplast genome evolution in *Saxifragales*. Plos One 2013; 8(10):e77965.
15. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90(1):157–166.

16. Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol Ecol*. 2016;25(7):1423-1428
17. Borsch T, Quandt D. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Syst Evol*. 2009;282(3-4):169-199.
18. Ochoterena H. Homology in coding and non-coding DNA sequences: a parsimony perspective. *Plant Syst Evol*. 2009;282(3):151-168.
19. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol J*. 2014; 12(6):743-754.
20. Lima MS, Woods LC, Cartwright MW, Smith DR. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. *Mol Ecol Resources*. 2016;16(6):1279-1286
21. Chakraborty S, Yengkhom S, Uddin DA. Analysis of codon usage bias of chloroplast genes in *Oryza* species. *Planta*. 2020;4:1-20.
22. Li G, Pan Z, Gao S, He Y, Xia Q, Yan J, et al. Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis*. *Genes Genomics*. 2019; 41(10):1173–1181.
23. Duret L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*. 2003;12(6):640-649.
24. Ruth H, Dmitri A. Selection on codon bias. *Ann Rev Genet*. 2008;42:287-299.
25. Wang H, Meng T, Wei W. Analysis of synonymous codon usage bias in helicase gene from *Autographa californica* multiple nucleopolyhedrovirus. *Genes Genomics* 2018;40(7):767-780
26. Long S, Yao H, Qi W, Li G. Analysis of compositional bias and codon usage pattern of the coding sequence in *Banna virus* genome. *Virus Res*. 2018;258:68-72
27. Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucl Acids Res*. 2000;10:2084–2090.
28. Chen X, Cai X, Chen Q, Zhou H, Cai Y, Ben X. Factors Affecting Synonymous Codon Usage Bias in Chloroplast Genome of *Oncidium Gower* Ramsey. *Evol Bioinforma*. 2019;7:271–278
29. Tang D, Wei F, Cai Z, Wei Y, Wei K. Analysis of codon usage bias and evolution in the chloroplast genome of *Mesona chinensis* Benth. *Dev Genes Evol*. 2020;(4):1-9.
30. Shen Z, Lu T, Zhang Z, Cai C, Tian B. Condon preference of chloroplast genome of *Drynaria*. *Guihaia*. 2021;41(2):266-273.
31. Li G, Zhang L, Xue P. Codon usage pattern and genetic diversity in chloroplast genomes of *Panicum* species. *Gene*. 2021;25:145866.

32. Zhou T, Zhu H, Wang J, Xu Y, Xu F, Wang X. Complete chloroplast genome sequence determination of *Rheum* species and comparative chloroplast genomics for the members of *Rumiceae*. *Plant Cell Rep.* 2020;39(6):811–824.
33. Alzahrani DA, Yaradua SS, Albokhari EJ, Abba A. Complete chloroplast genome sequence of *Barleria prionitis*, comparative chloroplast genomics and phylogenetic relationships among Acanthoideae. *BMC Genomics.* 2020;21(1):393.
34. Du X, Zeng T, Feng Q, Hu L, Luo X, Weng Q, et al. The complete chloroplast genome sequence of yellow mustard (*Sinapis alba* L.) and its phylogenetic relationship to other *Brassicaceae* species. *Gene.* 2020;731:144340.
35. Yan C, Zhang L, Qi J, Zhang L. Complete Chloroplast Genome Sequence of *Hibiscus cannabinus* and Comparative Analysis of the *Malvaceae* Family. *Front Genet.* 2020;11:277.
36. Guo H, Liu J, Luo L, Wei X, Zhang J, Qi Y, et al. Complete chloroplast genome sequences of *Schisandra chinensis*: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. *Sci China Life Sci.* 2017;60 (11):1286–90.
37. Mariotti R, Cultrera NG, Díez CM, Baldoni L, Rubini A. Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. *BMC Plant Biol.* 2010;10(1):211.
38. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet.* 2011;12(7):499-510.
39. Wu JY, Xiao JF, Wang LP, Zhong J, Yin HY, Wu SX, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. *Sci China Life Sci.* 2013;56(10):968–974.
40. Xu J, Feng D, Song G, Wei X, Chen L, Wu X, et al. The first intron of rice EPSP synthase enhances expression of foreign gene. *Sci China Life Sci.* 2003;46(6):561.
41. Yan J, Zhang Q, Yin P. RNA editing machinery in plant organelles. *Sci China Life Sci.* 2018;61(2):162–169.
42. Chen C, Bundschuh R. Systematic investigation of insertional and deletional RNA-DNA differences in the human transcriptome. *BMC Genomics.* 2012;13(1):616.
43. Knoop V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci.* 2011;68(4):567–86.
44. Grennan AK. To thy proteins be true: RNA editing in plants. *Plant Physiol.* 2011;156(2):453–454.
45. Yang Y, Zhang Y, Chen Y. Complete chloroplast genome sequence of the mangrove species *Kandelia obovata* and comparative analyses with related species. *Peer J.* 2019;7(4):e7713.
46. Nie XJ, Deng PC, Feng KW, Liu PX, Du XH, You FM, et al. Comparative analysis of codon usage patterns in chloroplast genomes of the *Asteraceae* family. *Plant Mol Biol Rep.* 2014;32:828-840.

47. Huo X, Liu S, Li Y, Wei H, Gao J, Yan Y, et al. Analysis of synonymous codon usage of transcriptome database in *Rheum palmatum*. Peer J. 2021;9:e10450.
48. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87:23-29
49. Zhao Y, Yin J, Guo H, Zhang Y, Xiao W, Sun C, et al. The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. Front Plant Sci. 2015;5:696.
50. Hebert PD, Cywinska A, Ball SL, DeWaard JR. Biological identifications through DNA barcodes. Proc R Soc B. 2003;270(1512):313–321.
51. Song Y, Chen Y, Lv J, Xu J, Zhu S, Li MF, et al. Development of chloroplast genomic resources for species discrimination. Front Plant Sci. 2017;8:1854.
52. Li J, Wang S, Jing Y, Wang L, Zhou S. A modified CTAB protocol for plant DNA extraction. Chin Bull Bot. 2013;48:72-78.
53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–477.
54. Wyman SK, Jansen R, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20(17):3252–3255.
55. Dong W, Xu C, Cheng T, Lin K, Zhou S. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of *Saxifragales*. Genome Biol Evol. 2013;5:989–997.
56. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633–4642.
57. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003;106(3):411–422.
58. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:273-279.
59. Wei L, He J, Jia X, Qi Q. Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. BMC Evol Biol. 2014;14:262.
60. Wright F. The effective number of codons used in a gene. Gene. 1990;87:23–29
61. Yang GF, Su KL, Zhao YR. Analysis of codon usage in the chloroplast genome of *Medicago truncatula*. Acta Pratacul Sin. 2015;12:171–179.
64. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–2739.

Tables

Table 1. List of annotated genes in the cp genome of *L. apetalum* and *L. perfoliatum*

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	<i>rpl2*2^{bc}, rpl14, rpl16^b, rpl20, rpl22^c, rpl23*2^c, rpl32, rpl33, rpl36</i>
	Small subunit of ribosomal proteins	<i>rps2, rps3, rps4, rps7*2^c, rps8, rps11, rps12*2^c, rps14, rps15, rps16^b, rps18, rps19</i>
	DNA dependent RNA polymerase	<i>rpoA, rpoB, rpoC1^b, rpoC2</i>
	rRNA genes	<i>rrn4.5 *2^c, rrn5 *2^c, rrn16 *2^c, rrn23 *2^c</i>
	tRNA genes	<i>trnA-UGC *2^{bc}, trnC-GCA, trnD-GUC, trnE-UUC, trnF- GAA, trnfM-CAU, trnG-UCC, trnH-GUG, trnI-CAU* 2^c, trnI-GAU *2^{bc}, trnG-GCC, trnK-UUU^b, trnL-CAA *2^c, trnL-UAA^b, trnL-UAG, trnM-CAU, trnN-GUU *2^c, trnQ-UUG, trnP- UGG, trnR-ACG *2^c, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC*2^c, trnV-UAC^b, trnW-CCA, trnY-GUA</i>
Photosynthesis	Subunits of Photosystem I	<i>psaA, psaB, psaC, psal, psaJ</i>
	Subunits of Photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbl, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
	Subunits of NADH dehydrogenases	<i>ndhA^b, ndhB *2^{bc}, ndhC, ndhD, ndhE, ndhG, ndhH, ndhI, ndhJ, ndhK, ndhF</i>
	Cytochrome b6/f complex	<i>petA, petB^b, petD^b, petG, petL, petN</i>
	ATP synthase	<i>atpA, atpB, atpE, atpF^b, atpH, atpI</i>
	Rubisco	<i>rbcL</i>
Other genes	Maturase	<i>matK</i>
	Subunit Acetyl-CoA-Carboxylate	<i>accD</i>
	Envelope membrane protein	<i>cemA</i>
	Protease	<i>clpP^a</i>
	c-type cytochrome synthesis gene	<i>ccsA</i>

Unknown	Conserved Open reading frames	<i>ycf1*2^c, ycf2 *2^c, ycf3^a, ycf4</i>
---------	-------------------------------------	--

Notes: ^aGene with two introns; ^bGene with one intron; ^cGenes located in the inverted repeats.

Table 2. Summary of the complete cp genome characteristics of five species in *Lepidium*.

Species	Genome size(bp)	LSC size(bp)	SSC size(bp)	IR size(bp)	Number of genes	Protein coding genes	tRNA genes	rRNA genes	GC content (%)
<i>L. apetalum</i>	154,846	83,896	17,994	26,478	130	85	37	8	36.40%
<i>L. perfoliatum</i>	154,264	83,391	17,973	26,450	130	85	37	8	36.46%
<i>L. sativum</i>	154,997	84,007	18,008	26,491	128	83	37	8	36.31%
<i>L. virginicum</i>	154,743	84,035	17,988	26,360	130	85	37	8	36.47%
<i>L. meyenii</i>	154,839	83,943	17,978	26,459	130	85	37	8	36.41%

Table 3. Preference related parameters of analysis of five species in *Lepidium* gene condons

Species	Number of codons	ENc	GC1	GC2	GC3	GC	CAI
<i>L. apetalum</i>	26608	49.07	44.81%	37.53%	28.75%	37.03%	0.165
<i>L. perfoliatum</i>	26601	49.10	44.78%	37.62%	28.76%	37.05%	0.165
<i>L. sativum</i>	26105	52.93	44.90%	37.69%	28.68%	37.09%	0.165
<i>L. virginicum</i>	26373	49.04	44.92%	37.69%	28.75%	37.12%	0.165
<i>L. meyenii</i>	26451	52.90	44.87%	37.61%	28.76%	37.08%	0.165

Table 4. Correlative analysis of condon components for cp genes in *Lepidium*

Parameter	GC1	GC2	GC12	GC3	GC	GC3s	N	ENC
GC1	—							
GC2	0.384**	—						
GC12	0.848**	0.815**	—					
GC3	0.136**	0.110*	0.149**	—				
GC	0.789**	0.750**	0.926**	0.511**	—			
GC3s	-0.074	0.015	-0.038	0.527**	0.168**	—		
N	-0.051	0.203**	0.149**	0.244**	-0.036	0.277**	—	
ENC	0.05	-0.053	0.001	0.243**	0.093	0.285**	0.288**	—

Note: **indicate significant correlations at 0.01 levels, * indicate significant correlations at 0.05 levels; N: Represents the number of codons; ENC: Effective number of codon.

Figures

Figure 1

Gene map of the *L. apetalum* and *L. perfoliatum* cp genomes. The genes inside and outside the outer circle are transcribed in the direction of the grey arrows inside and outside at the top. Genes are classified into 17 groups according to their biological function and are shown by different colored boxes. Within the inner circle, dark grey represents GC content and light grey represents AT content.

Figure 2

Codon content for the CDS in the five *Lepidium* cp genomes. The abscissa represents 20 amino acids and terminators, while the ordinate represents the RSCU value. For each amino acid, the corresponding species from left to right are *L. apetalum*, *L. perfoliatum*, *L. sativum*, *L. virginicum*, and *L. meyenii*. The different colors of each amino acid correspond to the codon of the same color below.

Figure 3

The distribution, type and presence of simple sequence repeats (SSRs) in the five cp genome of *Lepidium*. (A) Number of different SSRs types. (B) Number of different SSRs in the LSC, SSC, and IR regions. (C) Number of identified SSR motifs in different repeat class types.

Figure 4

Repeat sequences analysis of five cp genomes. (a) Repeat types in five cp genomes; (b) tandem repeats in five cp genomes; (c) forward repeats in five cp genomes; (d) palindromic repeats in eight cp genomes. In (a), different colors show different repeat types; in (b–d), different colors show different lengths. The ordinate represents the number of repeats.

Figure 5

Visualization alignment of the cp genome sequence of five species. The identity percentages are shown in the y-axis and range from 50% to 100%, while the horizontal axis shows the position within the cp genome. Each arrow indicates the annotated genes and direction of their transcription in the reference genome. Genome regions, i.e., exons, untranslated regions (UTRs), conserved noncoding sequences (CNS) and mRNA, are color coded.

Inverted Repeats

Figure 6

Comparison of the boundaries of the LSC, SSC and IR regions of the five cp genomes. Gene names are indicated in boxes, and their lengths in the corresponding regions are displayed above the boxes.

Figure 7

Neutrality plot analysis of the GC12 and that of the third codon position (GC3) for the entire coding DNA sequence of *Lepidium*.

Figure 8

ENc-GC3 plot. The solid line represents the expected curve when codon usage bias is only affected by mutation pressure.

Figure 9

PR2-bias plot. Using the values of $A3/(A3 + U3)$ against $G3/(G3 + C3)$.

Figure 10

ML phylogenetic tree reconstruction containing the cp genomes of 21 plants.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)