

Assembly and comparative analysis of the complete mitochondrial genome of three *Macadamia* species (*M. integrifolia*, *M. ternifolia* and *M. tetraphylla*)

Yingfeng Niu^{1*}, Yongjie Lu^{2*}, Weicai Song², Xiyong He¹, Ziyang Liu¹, Cheng Zheng¹, Shuo Wang², Chao Shi^{2#}, Jin Liu^{1#}

¹Yunnan Institute of Tropical Crops, Xishuangbanna, China.

²Qingdao University of Science & Technology, Qingdao, China.

Correspondence: Jin Liu (liujin2416@163.com); Chao Shi (chsh1111@aliyun.com)

* These authors contributed equally: Yingfeng Niu, Yongjie Lu

Abstract

Background: *Macadamia* is a true dicotyledonous plant that thrives in a mild, humid, low wind environment. It is cultivated and traded internationally due to its high-quality nuts thus, has significant development prospects and scientific research value. However, information on the genetic resources of *Macadamia* spp. remains scanty.

Results: The mitochondria (mt) genomes of three economically important *Macadamia* species, *Macadamia integrifolia*, *M. ternifolia* and *M. tetraphylla*, were assembled through the Illumina sequencing platform. The results showed that each species has 71 genes, including 42 protein-coding genes, 26 tRNAs, and 3 rRNAs. Repeated sequence analysis, RNA editing site prediction, and analysis of genes migrating from chloroplast (cp) to mt were performed in the mt genomes of the three *Macadamia* species. Phylogenetic analysis based on the mt genome of the three *Macadamia* species and 35 other species was conducted to reveal the evolution and taxonomic status of *Macadamia*. Furthermore, the characteristics of the plant mt genome, including genome size and GC content, were studied through comparison with 36 other plant species. The final non-synonymous (Ka) and synonymous (Ks) substitution analysis showed that most of the protein-coding genes in the mt genome underwent negative selections, indicating their importance in the mt genome.

Conclusion: The findings of this study provide a better understanding of the *Macadamia* genome and will inform future research on the genus.

Keywords: *Macadamia*; Mitochondrial genome; Genome assembly; Repeats; RNA editing; Phylogenetic analysis

1. Introduction

Macadamia spp belongs in the family Proteaceae, class *Magnoliopsida*, and order *Proteales*. The Proteaceae family has five subfamilies, 80 genera, and over 1600 species [1]. Most of them are distributed in Oceania and South Africa, while a few are produced in East Asia and South America. Notably, more than 100 species in the Proteaceae family produce flowers that are traded internationally. Besides, the species grown in the northeastern part of Oceania are also rich in nuts. The genus *Macadamia* comprises four species: *Macadamia integrifolia*, *M. jansonii*, *M. ternifolia*, and *M. tetraphylla*. These species are naturally distributed in the subtropical rain forests from southeastern Queensland in Australia to northeastern New South Wales [2,3]. Among them, *M. integrifolia* and *M. tetraphylla* produce edible nuts; thus, most commercial cultivars are either these two species or their hybrids [4]. The other two species, *M. Jansonii* and *M. ternifolia* produce non-edible nuts containing high levels of bitter cyanide glycosides, thus has not been used to guide the breeding [5-7]. *Macadamia* seeds are sweet with high nutritional and medicinal value. Therefore, they have enjoyed the reputation of "King of Thousand Fruits". They are also used in international transactions due to their high economic value [8].

Mitochondria (mt) are organelles that primarily convert biomass energy in living cells into chemical energy to fuel biological activities. Besides, they participate in other biological processes, including cell differentiation, cell apoptosis, cell growth, and cell division [9-12]. Therefore, mt are the core hub of life activities within individual cells and the entire living body [13]. Both plastids and mt harbor genetic information and are thought to have evolved through endosymbiosis of freely living bacteria [14-16]. In most seed plants, nuclear genetic information is inherited from both parents, while cp and mt are derived from maternal genes [17]. Thus, we can temporarily ignore the influence of paternal genes, thereby reducing the difficulty of genetic research and promoting the research of genetic mechanisms [18].

Studies have shown that the size of the mt genome varies significantly between different species. For example, plants have a larger mt genome than animals [19].

Furthermore, mt genome size in seed plants can vary by at least one order of magnitude ranging from ~ 222 Kbp in *Brassica napus* [20] and ~ 316 Kbp in *Allium cepa* [21] to ~ 3.9 Mbp in *Amborella trichopoda* [22] and a striking ~ 11.3 Mbp in *Silene conica* [23]. This phenomenon may be caused by the abundance of non-coding regions and repeated elements in the plant mt genome [24]. DNA recombination between homologous sequences produces small circular sub-genomic DNA. The circular genomic DNA coexists with the complete "master" genome in the cell. These genomes typically have several kb repeats, leading to multiple heterogeneous forms of the genome [25-31]. The mutation rate of plant mt genomes is very low; however, their rearrangement rate is so high that there is almost no conservation of synteny [27; 32-34].

The development of cost-effective and more efficient DNA sequencing methods like high-throughput sequencing has accelerated mt genome sequencing. So far (until June 2021), the mt genomes of 618 green plant species have been released in the NCBI database. Long-term mutually beneficial symbiosis caused the mt to lose some of the original DNA, possibly by transfer, leaving only the DNA encoding it [35, 36]. Mt DNA integrates DNA from various sources by intracellular and horizontal transfer [37]. Therefore, regardless of the length, gene sequence and content, mt genome varies remarkably among different plant species [38]. The mt genome length of the smallest terrestrial plant is about 66 Kb, and that of the largest terrestrial plant is 11.3 Mb [39,40]; the number of genes is usually between 32 and 67 [41]. In this study, the mt genomes of three *Macadamia* species were sequenced, assembled, and annotated. Also, their genomic and structural features were analyzed and compared with other angiosperms (and gymnosperms). This study improves our understanding of *Macadamia* genetics and provides crucial data to inform future research on the evolution of mt genomes of land plants.

2. RESULTS AND DISCUSSION

2.1 Genomic features of the mt genomes of the three *Macadamia* species

The mt genomes of *M. integrifolia*, *M. ternifolia* and *M. tetraphylla* have a typical

terrestrial plant genome ring structure (Fig 1). A total of 71 unique genes were identified in the mt genomes of the three *Macadamia* species, including 42 protein-coding, 26 tRNA, and 3 rRNA genes (Table 1). In addition, two copies of *rRNA26*, *ccmB*, *rps19*, *trnN-GTT*, and *trnH-GTG*, and seven copies of *trnM-CAT* were identified. It has been established that the mt genomes of land plants contain a variable number of introns [42]. In the present study, the three mt genomes had ten genes with introns, length ranging from 13bp (*rps3*) to 31841bp (*cox2*) where *ccmFC*, *rpl2*, *rps3*, and *rps10* had two introns, *cox2* had three, *nad1*, *nad4*, and *nad5* had four and *nad2* and *nad7* had five introns. Besides, in all protein-coding genes, except *atp6*, *cox1*, *nad1*, *nad4L*, *rps4*, and *rps10*, which had ACG as the start codon, all the others had ATG as their start codon. In addition, the stop codons in all the protein-coding genes were: TAA 45.2%, TGA 28.6%, TAG 14.3%, CAA 9.5%, and CGA 2.4%.

The size and GC content of mt genome are the primary characteristics. Here, we compared the size and GC content of mt genomes between three *Macadamia* species and 36 other green plants, including four phycophyta, three bryophytes, two gymnosperms, four monocots, and 23 dicots (Table S1). The size of the mt genomes ranged from 22,897 bp (*Chlamydomonas moewusii*) to 2,709,526 bp (*Cucumis melo*) (Fig 2). Compared to phycophyta and bryophytes, the mt genomes of the three *Macadamia* species are bigger. The GC content in the mt genomes was also highly variable, ranging from 32.24% in *Sphagnum palustre* to 50.36% in *Ginkgo biloba*. Overall, the GC content of angiosperm mt genome (including monocots and dicots) is higher than that in bryophytes but less than in gymnosperms, implying that the GC contents fluctuated following the angiosperms divergence from bryophytes and gymnosperms. Interestingly, the GC content significantly fluctuated in algae and was mostly conserved in angiosperms, although their genome sizes vary significantly.

2.2 Repeat sequences analysis

Microsatellites or simple sequence repetitions (SSRs) are DNA fragments composed of short sequence repeating units of 1-6 base pairs [43]. Their unique value is created by their polymorphism, relative abundance, codominant inheritance, large-scale

genome coverage, and PCR detection simplicity [44]. Based on the SSRs analysis, we identified 87 SSRs with SSRs monomers and dimers accounting for 70.11% of the total SSRs. Adenine (A) was the most repeated monomer with 19 (38%) out of the 50 identified monomer SSRs. The AT repeat was the most common dimer SSR, accounting for 66.67% of all the identified dimers. However, one hexamer [ATTAGG(X3)] was present in the mt genomes of three *Macadamia* species.

Among the reference mt genome, only *Nelumbo nucifera* has been published in the NCBI database. *N. nucifera* belongs to the family Nelumbonaceae and the same order (Proteales) with *Macadamia*. Therefore, the mt genome of *N. nucifera* was used as a reference for comparative analysis in the present study. The monomers in *N. nucifera* were lower than in the three *Macadamia* species, while pentamers and hexamers in *N. nucifera* were significantly higher than in the three *Macadamia* species (Fig 3a). Moreover, the SSRs in mt genomes of *M. integrifolia*, *M. ternifolia*, *M. tetraphylla*, and *N. nucifera* were mainly single-nucleotide A/T motifs, and dimer AT/TA motifs. Within the *Macadamia* genus, the mt SSRs among the different species are highly similar (Fig 3b). However, compared with *N. nucifera*, there were both differences and similarities. For example, the single nucleotide A/T in the three *Macadamia* species has 23 unit repeats, while *N. nucifera* has only nine. Nevertheless, their single-nucleotide C/G numbers were the same (two-unit repeats) (Fig 3b). In addition, the AG/CT and AT/AT motifs unit repetitions are the same, although *N. nucifera* also has an AC/GT motif, lacking in the three *Macadamia* species. Interestingly, the pentanucleotide AATGT/ACATT, ACTAG/AGTCT, and ACATT/AGTAT also had the same number of repetitions in the three *Macadamia* species and *N. nucifera*. Overall, the greater the nucleotide motif, the greater the difference between the three *Macadamia* species and *N. nucifera*.

Core repeating units ranging from 1 to 200 bases (tandem repeats) are widely present in eukaryotes and some prokaryotes genomes [45]. In the present study, 25, 21, and 20 tandem repeats (10-33 bp) were identified in the *M. integrifolia*, *M. ternifolia*, and *M. tetraphylla* with a match greater than 95% (Table S2, S3, and S4). The tandem repeats (11-20 bp and 21-30 bp) significantly varied among the three

Macadamia species (Fig 3c), where *M. ternifolia* had the least number of repetitions, while *M. integrifolia* and *M. tetraphylla* had a very similar number of repetitions. However, *N. nucifera* had the least (11-20bp and 21-30bp) and had the highest (0-10bp, 31-40bp, 41-50bp) tandem repeated compared to the three *Macadamia* species. Besides, no repetitions ranged from 51-60 bp among the four genomes, while the number of repetitions was the same for 60-70 bp and above.

2.3 The prediction of RNA editing

RNA editing is a post-transcriptional process entailing the addition, deletion, or conversion of bases in the coding region of a transcribed RNA. The conversion of cytosine to uridine is common in cp and mt genomes of plants [46-50], which improves protein preservation in plants. The accurate detection of ribonucleic acid editing is inseparable from the proteomics data. However, Mower's software PREP computationally predicts the RNA edit site [51].

In the present study, we predicted 42 protein-coding genes (including two multi-copy genes: *ccmB* and *rps19*) in the mt genomes of the three *Macadamia* species using the PREP-mt program. The findings revealed that the RNA editing sites were 688, 689, and 688 (Fig 4). Among the protein-coding genes, *nad4* had the most RNA editing sites (59 sites), while *atp8*, *rpl2*, *rpl10*, *rps1*, *rps2*, *rps7*, *rps10*, *rps11*, *rps13*, *rps14*, *rps19*, *sdh3*, and *sdh4* had less than 10 RNA editing sites. Up to 34.3% (236 sites) RNA editing sites occurred in the first base position of the codon, 68.6% (472 sites) appeared in the second base position, and there was no RNA editing in the third base position. *M. ternifolia* had more than one RNA editing site, unlike the other two *Macadamia* species.

The RNA editing increases the diversity at the start and stop codons in protein-coding genes. However, even with RNA editing, 30.2% (208 positions) of amino acid hydrophobicity and 12.5% (86 positions) of amino acid hydrophilicity remained unchanged in the *M. integrifolia* and *M. tetraphylla* mt genomes. However, 6.7% (46 positions) of amino acids were converted from hydrophobic to hydrophilic, and 47.9% (330 positions) from hydrophilic to hydrophobic. In addition, five amino

acids were converted from glutamine to stop codons and two from arginine to stop codons (Table 2). The findings in this study revealed that most amino acids were converted from serine to leucine (23.3%, 160 sites), proline to leucine (22.4%), and serine to phenylalanine (15.3%). The remaining 269 RNA editing sites included other RNA editing types, such as Ala-Val, His-Tyr, Leu-Phe, Pro-Phe, Pro-Ser, Arg-Cys, Arg-Trp, Thr-Ile, Thr- Met, Gln-X, and Arg-X (X=stop codon). Compared to *M. integrifolia* and *M. tetraphylla*, *M. ternifolia* only had one more RNA-edited site (Leu-Phe).

2.4 DNA migration from cp to mt

The cp-like sequences in the mt genome were detected by comparing against the complete cp genome sequence of *M. integrifolia* obtained from the NCBI database (Fig 5). We detected 28 fragments in the mt genome of *M. integrifolia*, ranging in size from 32 bp to 5210 bp. The cp-like sequence had 36,902 bp, accounting for 5.4% of the mt genome. Five complete annotated tRNA genes were detected, namely *trnH-GTG*, *trnM-CAT*, *trnW-CCA*, *trnD-GTC*, and *trnN-GTT*, with some fragments of *rrn18* genes. The findings also revealed that 28 insertion regions accounted for 23.2% of the cp genome, including seven complete protein-coding genes (*petL*, *petG*, *ndhE*, *rps15*, *rpl23(X2)*, *rpl2*) and eight complete tRNA genes (*trnH-GUG*, *trnD-GUC*, *trnM-CAU*, *trnW-CCA*, *trnP-UGG*, *trnP-GGG*, *trnI-CAU*, *trnN-GUU*). Besides, several protein-coding genes were also identified, including *psbA*, *rpoB*, *psbD*, *psbC*, *ndhC*, *rpl2*, *ycf2(X2)*, *ndhB*, *rps7(X2)*, *ndhD*, *ndhB* and *ycf1*, and some tRNA genes (*trnI-GAU*, *trnA-UGC*, *trnN-GUU*), which migrated from the cp genome into the mt genome. But, most of these genes lost their integrity during the evolution process, and only their partial sequences were found in the mt genome. Furthermore, most cp-like sequences were located in the spacer region of the mt genome. These findings are consistent with previous research, where during evolution, tRNA genes were more conserved than the protein-coding genes and rRNA genes since they play an important role in mt genome [52].

2.5 Phylogenetic analysis within higher plant mt genomes

Phylogenetic analysis was performed to understand the evolution of the three *Macadamia* species compared to 29 dicots, four monocots, and two gymnosperms (out-groups). The phylogenetic tree was constructed based on the comparisons in the data matrix of 23 conserved protein-coding genes (Fig 6). The findings revealed that the phylogenetic tree strongly supports the separation of Proteales from rosids and asterids, the separation of eudicots from monocots and angiosperms from gymnosperms. The evolutionary relationships among all the taxa separated into 20 families (Leguminosae, Cucurbitaceae, Apiaceae, Apocynaceae, Solanaceae, Rosaceae, Caricaceae, Brassicaceae, Salicaceae, Bataceae, Malvaceae, Vitaceae, Lamiaceae, Nelumbonaceae, Proteaceae, Butomaceae, Arecaceae, Poaceae, Cycadaceae, and Ginkgoaceae) were efficiently deduced in the phylogenetic tree (Fig 6).

2.6 The substitution rates of protein-coding genes

In genetics, non-synonymous (K_a) and synonymous (K_s) substitution rates help understand the evolutionary dynamics of protein-coding genes among similar species since the K_a to K_s ratio indicates gene selection [53,54]. In the present study, *N. nucifera* was used as a reference species to calculate the K_a/K_s ratio of 40 protein-coding genes present in the mt genome of three *Macadamia* species. The K_s of *atp9* and *rps14*, and the K_a of *rps12* was 0. Besides, in most protein-coding genes, the K_a/K_s ratio was significantly less than 1 (Fig 7). However, the K_a/K_s ratio of *nad4*, *rpl2*, *rps3*, *rps4*, and *rps10* was greater than 1, with the *rps3* ratio being 2.34, implying that these genes might have undergone surgery positive selection following *Macadamia* and *N. nucifera* differentiation from their last common ancestor [55]. Besides, the ATP synthase, Cytochrome C biogenesis, Ubichinol Cytochrome C reductase, and Maturases of K_a/K_s ratios were below 1, implying that the negative selection acted on these genes (Table 2). Therefore, these genes may be highly conserved during the evolution of higher plants.

3. CONCLUSIONS

The complete mt genomes of *M. integrifolia*, *M. ternifolia* and *M. tetraphylla* share many common features with angiosperm mt genomes. In this study, we found that the mt genomes of the three *Macadamia* species were circular like most mt genomes. Compared them with the GC content of the mt genome of 36 other green plants, the results supported the conclusion that the GC content in the *Macadamia* species and angiosperms are highly conserved. In addition, we conducted studies on SSRs and longer tandem repeats in the three sets of data. Besides, 688 RNA editing sites were identified in 42 protein-coding genes, providing important clues for predicting gene function with new codons. By detecting gene migration, we observed 28 fragments (with five complete tRNA genes) were transferred from the cp genome to mt genome. The subsequent phylogenetic analysis results also showed their accuracy in plant classification. Moreover, based on the Ka/Ks substitution of protein-coding genes, most coding genes have undergone negative selection, indicating that the protein-coding genes in the mt genome are conserved in *Macadamia* species. The findings of this study provide information on the mt genome of *Macadamia* species, which is key in understanding the evolutionary history of the family Proteaceae.

4. MATERIALS AND METHODS

4.1 Genome sequencing

The three *Macadamia* species examined in this study were collected from Xishuangbanna. Agarose electrophoresis was used to assess the quality of the extracted DNA samples. Meanwhile, a Nanodrop spectrophotometer was used to detect the concentration and purity of the DNA samples. The qualified DNA samples were used for Illumina DNA library construction, according to the standard procedure. Subsequently, a paired-end sequencing library with an insert size of 350 bp was constructed. The Illumina Hiseq 4000 high-throughput sequencing platform was used for sequencing. The sequencing strategy involved PE150 (Pair-End 150) and the sequencing data volume of not less than 1 Gb. Illumina high-throughput sequencing

results initially existing as original image data files were converted into Raw Reads. CASAVA software was used for Base Calling.

4.2 Genome assembly and annotation

SPAdes v.3.5.0 software was used to splice and assemble mt genome sequences [56]. DOGMA [57] and ORF Finder were used to annotate the mt genome. The Blastn and Blastp method was used to compare mt gene-encoding protein and rRNAs among related species. The accuracy of the results was verified, and corrections were made accordingly. tRNA scan-SE2.0 [58] and ARWEN [59] were used to annotate tRNA. The tRNAs with unreasonable length and incomplete structure were eliminated. Subsequently, a tRNA secondary structure diagram was generated. The final mt genomes of *M. integrifolia*, *M. ternifolia*, and *M. tetraphylla* have been deposited in the GenBank (Accession number: MW566570/MW566571/MW566572).

4.3 Analysis of repeat structure and sequence

Microsatellites within the mt genomes of the three *Macadamia* species were identified using MISA [60]. The minimum number of repeats for the motif length of 1, 2, 3, 4, 5, and 6 were 10, 6, 5, 4, 3, and 3, respectively, were identified in this analysis. The tandem repeats were detected using Tandem Repeats Finder v4.09 software [61] with default parameters.

4.4 DNA transformation from cp to mt and RNA editing analyses

The cp genome of *M. integrifolia* (NC_025288) was downloaded from the NCBI database. Chloroplast-like sequences were identified and the genome was mapped using TBtools [62]. The online program Predictive RNA Editor for Plants (PREP) suite [63] was adopted to identify the possible RNA editing sites in the protein-coding genes of the three *Macadamia* species. The cutoff value was set as 0.2 to ensure accurate prediction. The protein-coding genes from other plant mt genomes were used as references to reveal the RNA editing sites in the mt genomes of the three *Macadamia* species.

4.5 Phylogenetic tree construction and Ka/Ks analysis

The genome sequences of the three *Macadamia* species were compared with those of 35 other plant species to further verify their phylogenetic position. Notably, the complete mt genome sequences of these species were available in the NCBI database. Phylogenetic analyses were performed on 23 conserved protein-coding genes (*atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *ccmB*, *ccmC*, *ccmFc*, *ccmFn*, *cob*, *cox1*, *cox2*, *cox3*, *matR*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7* and *nad9*) that were extracted from the mt genomes of the 35 plant species using TBtools [62]. These conserved genes were then aligned using Muscle [64] implemented in MEGA X [65]; the alignment was modified manually to eliminate gaps and missing data. The GTR + G + I model was determined to be the best model based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) calculated by ModelFinder [66]. The Maximum Likelihood (ML) algorithm in MEGA X [65] was used to construct a phylogenetic tree. The bootstrap consensus tree was inferred from 1000 replications. *Cycas taitungensis* and *Ginkgo biloba* were designated as the outgroup in this analysis.

The Ka and Ks replacement rates of protein-coding genes in mitochondrial genomes of the three *Macadamia* species and other higher plants were analyzed. blastn in TBtools [62] was used to extract the sequences of corresponding protein-coding genes in *Macadamia* and *N. nucifera* genomes. The Ka and Ks replacement rates of each protein-coding gene were estimated using *N. nucifera* genome as a reference.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 31760215 and No. 31801022) and the Technology Innovation Talents Project of Yunnan Province (2018HB086).

Conflict of interest

The authors declare that they have no conflict of interest.

Author Contributions

Chao Shi and Jin Liu conceived the study and managed the project. Yingfeng Niu and Yongjie Lu designed the scientific objectives. Xiyong He collected the samples and extracted the genomic DNA. Ziyan Liu and Cheng Zheng estimated the genome size, assembled the genome and carried out the gene annotation. Yongjie Lu wrote the manuscript, and all authors contributed to writing and editing the final manuscript.

Data availability

The assembled mitochondrial genome sequence and annotation have been deposited at GenBank under the accession MW566570, MW566571 and MW566572.

References

- [1] Stevens, P. F. (2001 onwards). Angiosperm Phylogeny Website. Version 14, July 2017 [and more or less continuously updated since]. <http://www.mobot.org/MOBOT/research/APweb/>.
- [2] Hardner, C.M., Peace, C., Lowe, A.J., Neal, J., Pisanu, P., Powell, M., Schmidt, A., Spain, C. and Williams, K. (2009). Genetic Resources and Domestication of Macadamia. In Horticultural Reviews, J. Janick (Ed.). <https://doi.org/10.1002/9780470593776.ch1>
- [3] Topp B.L., Nock C.J., Hardner C.M., Alam M., O' Connor K.M. (2019) Macadamia (*Macadamia* spp.) Breeding. In: Al-Khayri J., Jain S., Johnson D. (eds) Advances in Plant Breeding Strategies: Nut and Beverage Crops. Springer, Cham. https://doi.org/10.1007/978-3-030-23112-5_7
- [4] Hardner, C.; Topp, B. *Macadamia integrifolia* and *M. tetraphylla*. In *Trees for Life in Oceania: Conservation and Utilisation of Genetic Diversity*; Lee, T., John, D., Clarke, B., Eds.; Australian Centre for International Agricultural Research: Canberra, Australia, 2018; pp. 147 – 151.
- [5] Dahler JM , Mcconchie C Turnbull CGN (1995) Quantification of Cyanogenic Glycosides in Seedlings of Three *Macadamia* (Proteaceae) Species . Australian Journal of Botany 43, 619-628. <https://doi.org/10.1071/BT9950619>
- [6] Vivian-Smith, A. Cyanogenesis and the Fruit Growth of Three *Macadamia* Species. Bachelor's Thesis, University of Queensland, Brisbane, Australia, 1995.
- [7] Gross, C.L.; Weston, P .H. *Macadamia jansenii* (Proteaceae), a new species from central Queensland. Aust. Syst.Bot. 1992, 5, 725 – 728. [CrossRef]
- [8] *Macadamia ternifolia* F. Muell. Trans. Philos. Inst. Vict. 2: 72. 1858; Sleum. in Fl. Malesiana 5 (1) : 195, f. 21. 1955; 广州植物志 173. 1956.
- [9] Bonora M, De Marchi E, Patergnani S, Suski J, Celsi F, Bononi A, Giorgi C, Marchi S, Rimessi A, Duszyński J. Tumor necrosis factor- α impairs oligodendroglial differentiation through a mitochondria-dependent process. Cell Death Differentiation. 2014;21(8):1198 – 208.

- [10] van Loo G, Saelens X, Van Gurp M, MacFarlane M, Martin S, Vandenabeele P. The role of mitochondrial factors in apoptosis: a Russian roulette with more than one bullet. *Cell Death Differentiation*. 2002;9(10):1031 – 42.
- [11] Kroemer G, Reed JC. Mitochondrial control of cell death. *Nat Med*. 2000;6(5):513 – 9.
- [12] Rehman J, Zhang HJ, Toth PT, Zhang Y, Marsboom G, Hong Z, Salgia R, Husain AN, Wietholt C, Archer SL. Inhibition of mitochondrial fission prevents cell cycle progression in lung cancer. *FASEB J*. 2012;26(5):2175 – 86.
- [13] Ogiwara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, Miyashita N, Nasuda S, Nakamura C, Mori N. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. *Nucleic Acids Res*. 2005;33(19):6235 – 50.
- [14] Greiner S, Bock R. 2013. Tuning a ménage à trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. *Drug & Alcohol Dependence* 35:354 – 365. DOI 10.1002/bies.201200137.
- [15] Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society of London* 365:729 – 748. DOI 10.1098/rstb.2009.0103.
- [16] Sagan L. 1967. On the origin of mitosing cells. *Journal of Theoretical Biology* 14:255 – 274.
- [17] Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences of the United States of America* 92:11331 – 11338 DOI 10.1073/pnas.92.25.11331.
- [18] Wallace DC, Singh G, Lott mitochondrial, Hodge JA, Schurr TG, Lezza A, Elsas LJ, Nikoskelainen EK. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science*. 1988;242(4884):1427 – 30.
- [19] Ward, B. L., R. S. Anderson and A. J. Bendich, 1981 The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). *Cell* 25: 793 - 803.
- [20] Handa H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the

mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Res.* 2003;31(20):5907 – 16.

[21] Kim B, Kim K, Yang TJ, Kim S. Completion of the mitochondrial genome sequence of onion (*Allium cepa* L.) containing the CMS-S male-sterile cytoplasm and identification of an independent event of the *ccmFN* gene split. *Curr Genet.* 2016;62(4):873 – 85.

[22] Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science.* 2013;342(6165):1468 – 73.

[23] Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, DE MC, Palmer JD, et al. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 2012; 10(1):e100124.

[24] Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci U S A.* 2015;112(33):10177 – 84.

[25] Folkerts, O., and M. R. Hanson, 1989 Three copies of a single recombination repeat occur on the 443 kb master circle of the *Petunia hybrida* 3704 mitochondrial genome. *Nucleic Acids Res* 17: 7345 - 7357.

[26] Klein, M., U. Eckert - Ossenkopp, I. Schmiedeberg, P. Brandt, M. Unseld et al., 1994 Physical mapping of the mitochondrial genome of *Arabidopsis thaliana* by cosmid and YAC clones. *Plant J.* 6: 447 - 455.

[27] Palmer, J. D., and L. A. Herbon, 1988 Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J Mol Evol* 28: 87 - 97.

[28] Palmer, J. D., and C. R. Shields, 1984 Tripartite structure of the *Brassica campestris* mitochondrial genome. *Nature* 307: 437.

[29] Siculella, L., F. Damiano, M. R. Cortese, E. Dassisti, G. Rainaldi et al., 2001 Gene content and organization of the oat mitochondrial genome. *Theoretical and Applied Genetics* 103: 359-365.

[30] Sloan, D. B., A. J. Alvers, H. Storchova, J. D. Palmer and D. R. Taylor, 2010

Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. *BMC Evol Biol* 10: 274.

[31] Stern, D. B., and J. D. Palmer, 1986 Tripartite mitochondrial genome of spinach: physical structure, mitochondrial gene mapping, and locations of transposed chloroplast DNA sequences. *Nucleic Acids Res* 14: 5651 - 5666.

[32] Drouin, G., H. Daoud and J. Xia, 2008 Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol Phylogenet Evol* 49: 827 - 831.

[33] Richardson, A. O., D. W. Rice, G. J. Young, A. J. Alverson and J. D. Palmer, 2013 The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol* 11: 29.

[34] Wolfe, K., W. Li and P. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. *Proc. Natl. Acad. Sci. U S A* 84: 9054 - 9058.

[35] Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann Entomol Soc Am.* 1994;87(6):651 - 701.

[36] Knoop V. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr Genet.* 2004;46(3):123 - 39.

[37] Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. *Proc Natl Acad Sci.* 2005;101(51):17747 - 52.

[38] Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD. The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol.* 2013;11(1):1 - 17.

[39] Skippington E, Barkman TJ, Rice DW, Palmer JD. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has

lost all nad genes. *Proc Natl Acad Sci.* 2015;112(27):E3515 – 24.

[40] Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 2012;10(1):e1001241.

[41] Hsu CL, Mullin BC. Physical characterization of mitochondrial DNA from cotton. *Plant Mol Biol.* 1989;13(4):467 – 8.

[42] Liao X, Zhao Y, Kong X, Khan A, Zhou B. Complete sequence of kenaf (*Hibiscus cannabinus*) mitochondrial genome and comparative analysis with the mitochondrial genomes of other plants. *Sci Rep.* 2018;8(1):1 – 13.

[43] Y-c L, Liu S, Liu D-C, Wei Y-X, Liu C, Yang Y-M, Tao C-G, Liu W-S. Exploiting EST databases for the development and characterization of EST-SSR markers in blueberry (*Vaccinium*) and their cross-species transferability in *Vaccinium* spp. *Sci Hortic.* 2014;176:319 – 29.

[44] Powell W, Machray GC, Provan J. Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* 1996;1(7):215 – 22.

[45] GAO H, KONG J. Distribution characteristics and biological function of tandem repeat sequences in the genomes of different organisms. *Zool Res.* 2005;26(5):555 – 64.

[46] Bock R, Khan MS. 2004. Taming plastids for a green future. *Trends in Biotechnology* 22:311 – 318 DOI 10.1016/j.tibtech.2004.03.005.

[47] Chen H, Deng L, Jiang Y, Lu P, Yu J. 2011. RNA editing sites exist in protein-coding genes in the chloroplast genome of *Cycas taitungensis*. *Journal of Integrative Plant Biology* 53:961 – 970 DOI 10.1111/j.1744-7909.2011.01082.x.

[48] Raman G, Park S. 2015. Analysis of the complete chloroplast genome of a medicinal plant, *dianthus superbus* var. *longicalyncinus*, from a comparative genomics perspective. *PLOS ONE* 10:e0141329 DOI 10.1371/journal.pone.0141329.

[49] Wakasugi T, Hirose T, Horihata M, Tsudzuki T, Kossel H, Sugiura M. 1996. Creation of a novel protein-coding region at the RNA level in black pine cps: the pattern of RNA editing in the gymnosperm chloroplast is different from that in

angiosperms. Proceedings of the National Academy of Sciences of the United States of America 93:8766 – 8770 DOI 10.1073/pnas.93.16.8766.

[50] Zanduetta-Criado A, Bock R. 2004. Surprising features of plastid *ndhD* transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. Nucleic Acids Research 32:542 – 550 DOI 10.1093/nar/gkh217.

[51] Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 2009;37(suppl_2):W253 – 9.

[52] Cheng Yan, He Xiaoxue, Priyadarshani S V G N, Wang Yu, Ye Li, Shi Chao, Ye Kangzhuo, Zhou Qiao, Luo Ziqiang, Deng Fang, Cao Ling, Zheng Ping, Aslam Mohammad, Qin Yuan. Assembly and comparative analysis of the complete mitochondrial genome of *Suaeda glauca*. [J]. BMC genomics, 2021, 22(1):

[53] Fay JC, Wu C-I. Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genomics Hum Genet. 2003;4(1):213 – 35.

[54] Zhang Z, Li J, Zhao XQ, Wang J, Wong KS, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4(4):259 – 63.

[55] Bi C, Paterson AH, Wang X, Xu Y, Wu D, Qu Y, Jiang A, Ye Q, Ye N. 2016. Analysis of the complete mitochondrial genome sequence of the diploid cotton *Gossypium* Ye et al. (2017), PeerJ, DOI 10.7717/peerj.3148 22/27.

[56] Lapidus A., Antipov D., Bankevich A., Gurevich A., Korobeynikov A., Nurk S., Prjibelski A., Safonova Y., Vasilinets I., Pevzner P. A. New Frontiers of Genome Assembly with SPAdes 3.0. (poster), 2014.

[57] Automatic annotation of organellar genomes with DOGMA. Wyman SK, Jansen RK, Boore JL, Bioinformatics 2004 20 (17) : 3252-3255.

[58] Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucl. Acids Res. 44: W54-57.

[59] Laslett, D. & Canbäck B. (2008) ARWEN, a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24;172-175.

- [60] Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. 2017;33(16):2583 – 5.
- [61] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573 – 80.
- [62] Chen C, Xia R, Chen H, He Y. 2018. TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. *bioRxiv* 289660 DOI 10.1101/289660.
- [63] Mower JP. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Research* 37:W253 – 259 DOI 10.1093/nar/gkp337.
- [64] Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792 – 1797 DOI 10.1093/nar/gkh340.
- [65] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33(7):1870 – 4.
- [66] Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14:587 – 589 DOI 10.1038/nmeth.4285.

Figure legends:

Fig. 1. The circular map of three *Macadamia species* mitochondrial genome. Gene map showing 71 annotated genes of different functional groups

Fig. 2. The sizes and GC contents of 39 plant mitochondrial genomes. The blue dots represent the genome size and the orange trend line shows the variation of GC content across the different taxa.

Fig. 3. The comparison of microsatellites and oligonucleotide repeats in three *Macadamia* species and *N. nucifera* mitochondrial genomes.

Fig. 4. The distribution of RNA-editing sites in the mt protein-coding genes of three species of *Macadamia*. The bars of different colors represent the number of RNA-editing sites of each gene.

Fig 5. Schematic representation of mitochondrial genome, chloroplast genome

and chloroplast-like sequence of *M. integrifolia*. Dots and heat maps inside the two chromosomes show where genes are located. The green lines in the circle show the regions of chloroplast-like sequences inserted from the chloroplast genome into the mt genome.

Fig 6. The phylogenetic relationships of three species of *Macadamia* with other 35 plant species. The Maximum Likelihood tree was constructed based on the sequences of 23 conserved protein-coding genes. Colors indicate the families that the specific species belongs.

Fig 7. The Ka/Ks values of 40 protein-coding genes of three *Macadamia* species.

Tables

Table 1. Gene profile and organization of three *Macadamia* species(*M. integrifolia*, *M. ternifolia* and *M. tetraphylla*).

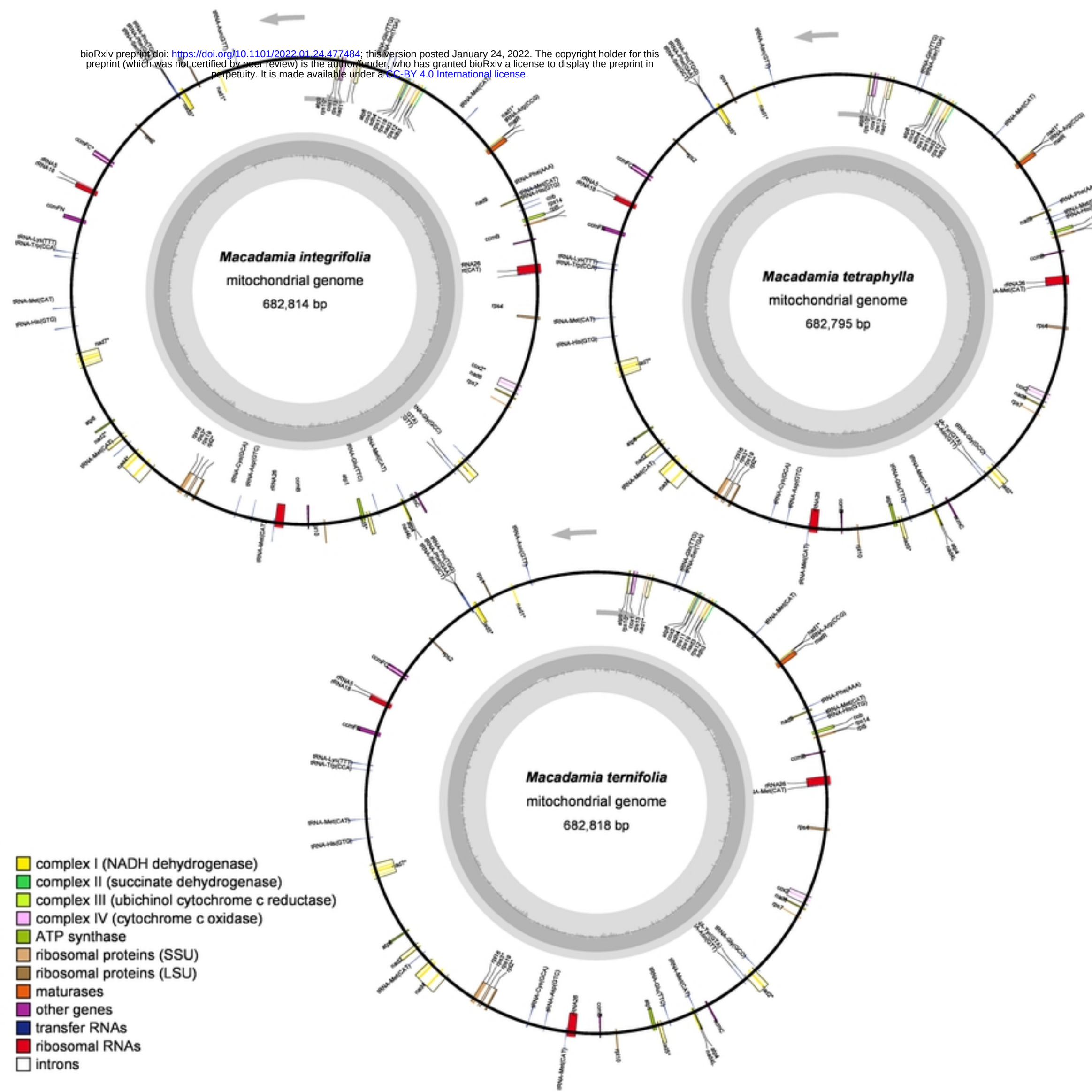
Group of genes	Gene/ element	Size(bp)	GC_ Percent	AminoAcids (aa)	InferredInitiation Codon	Inferred TerminationCodon
ATP synthase	<i>atp1</i>	1530	45.29%	509	ATG	TGA
	<i>atp4</i>	597	43.05%	198	ATG	TAG
	<i>atp6</i>	783	39.08%	260	ACG	TAA
	<i>atp8</i>	480	40.63%	159	ATG	TAA
	<i>atp9</i>	225	46.67%	74	ATG	CAA
Cytochrome c biogenesis	<i>ccmB(2)</i>	621,621	42.83%	206	ATG	TGA
	<i>ccmC</i>	771	44.23%	256	ATG	TAA
	<i>ccmFC^a</i>	1356	46.53%	451	ATG	TAA
	<i>ccmFN</i>	1734	47.58%	577	ATG	TGA
Ubichinol cytochrome c reductase	<i>cob</i>	1182	42.39%	393	ATG	TGA
Cytochrome c oxidase	<i>cox1</i>	1584	44.26%	527	ACG	TAA
	<i>cox2^a</i>	822	42.34%	273	ATG	TAG
	<i>cox3</i>	798	45.11%	265	ATG	TGA
Maturases	<i>matR</i>	1968	52.64%	655	ATG	TAG
NADH dehydrogenase	<i>nad1^a</i>	978	44.99%	325	ACG	TAA
	<i>nad2^a</i>	1467	40.90%	488	ATG	TAA
	<i>nad3</i>	357	41.74%	118	ATG	TAA
	<i>nad4^a</i>	1488	42.67%	495	ATG	TGA
	<i>nad4L</i>	303	37.29%	100	ACG	TAA
	<i>nad5^a</i>	1989	41.78%	662	ATG	TAA
	<i>nad6</i>	630	40.95%	209	ATG	TGA
	<i>nad7^a</i>	1185	45.23%	394	ATG	TAG
	<i>nad9</i>	573	42.93%	190	ATG	TAA
Ribosomal proteins (LSU)	<i>rpl2^a</i>	999	52.15%	332	ATG	CAA
	<i>rpl5</i>	561	44.74%	186	ATG	TAA
	<i>rpl10</i>	516	46.32%	171	ATG	TAA
	<i>rpl16</i>	492	43.09%	163	ATG	TAA
Ribosomal proteins (SSU)	<i>rps1</i>	606	43.56%	201	ATG	TAA
	<i>rps2</i>	648	39.20%	215	ATG	CAA
	<i>rps3^a</i>	1692	43.91%	563	ATG	TAG
	<i>rps4</i>	1059	40.51%	352	ACG	TAA
	<i>rps7</i>	447	43.18%	148	ATG	TAA
	<i>rps10^a</i>	333	39.04%	110	ACG	CGA
	<i>rps11</i>	444	45.27%	147	ATG	CAA
	<i>rps12</i>	378	45.50%	125	ATG	TGA
	<i>rps13</i>	351	39.60%	116	ATG	TGA
	<i>rps14</i>	303	40.92%	100	ATG	TAG
Transport membrane protein	<i>sdh3</i>	336	37.20%	111	ATG	TGA
	<i>sdh4</i>	450	41.33%	149	ATG	TGA
Ribosomal RNAs	<i>rrn5</i>	119	52.94%			
	<i>rrn18</i>	2061	55.12%			
	<i>rrn26(2)</i>	3989, 3989	53.02%			
Transfer RNAs	<i>tmR-CCG</i>	75	57.33%			
	<i>tmN-GTT^b(2)</i>	75,72	49.33%			
	<i>tmD-GTC^b</i>	74	63.51%			
	<i>tmC-GCA</i>	76	52.63%			
	<i>tmQ-TTG</i>	72	47.22%			
	<i>tmE-TTC</i>	72	50.00%			
	<i>tmG-GCC</i>	74	54.05%			
	<i>tmH-GTG^b(2)</i>	75,75	54.67%			
	<i>tmK-TTT</i>	75	46.67%			
	<i>tmM-CAT^b(7)</i>	72,75,73,72, 77,72,72	59.72%,46.67%, 43.84%,59.72%, 44.16%,59.72%, 59.72%			
	<i>tmF-AAA</i>	75	49.33%			
	<i>tmF-GAA</i>	74	47.30%			
	<i>tmP-TGG</i>	75	54.67%			
	<i>tmS-TGA</i>	88	51.14%			
	<i>tmS-GCT</i>	91	46.15%			
	<i>tmW-CCA^b</i>	74	51.35%			
	<i>tmY-GTA</i>	84	51.19%			

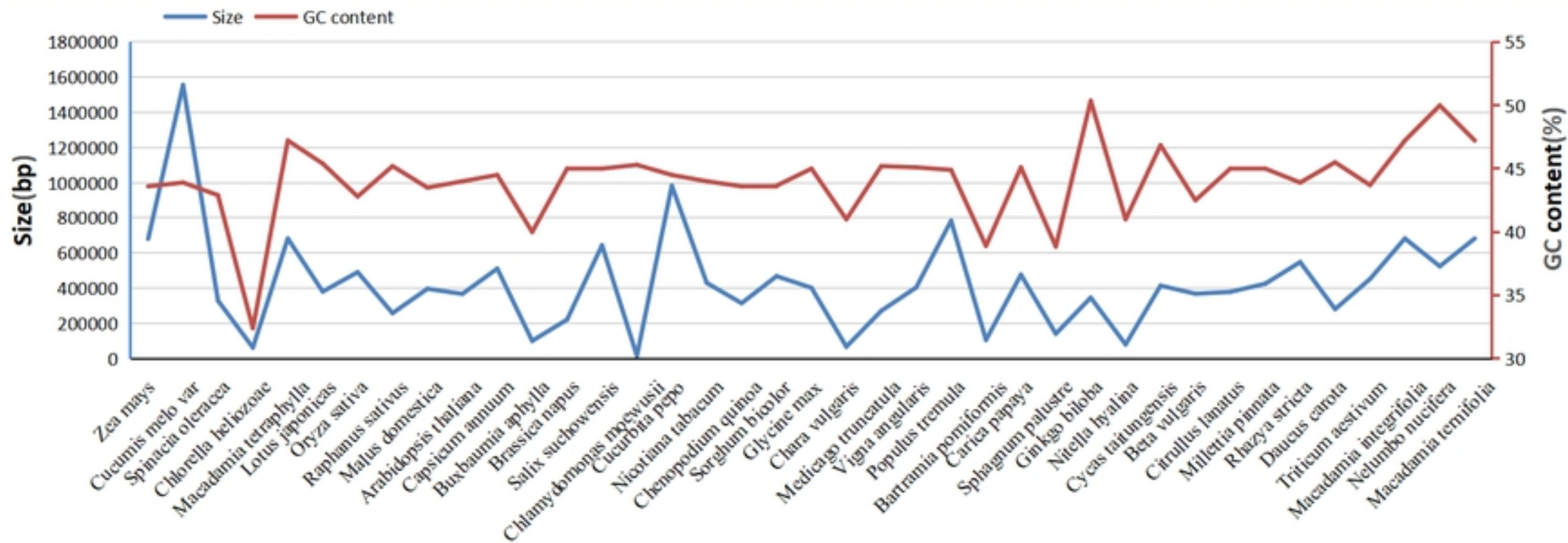
Notes: The numbers after the gene names indicate the duplication number. Lowercase a indicates the genes containing introns, and lowercase b indicates the chloroplast-derived genes.

Table 2. Prediction of RNA editing sites

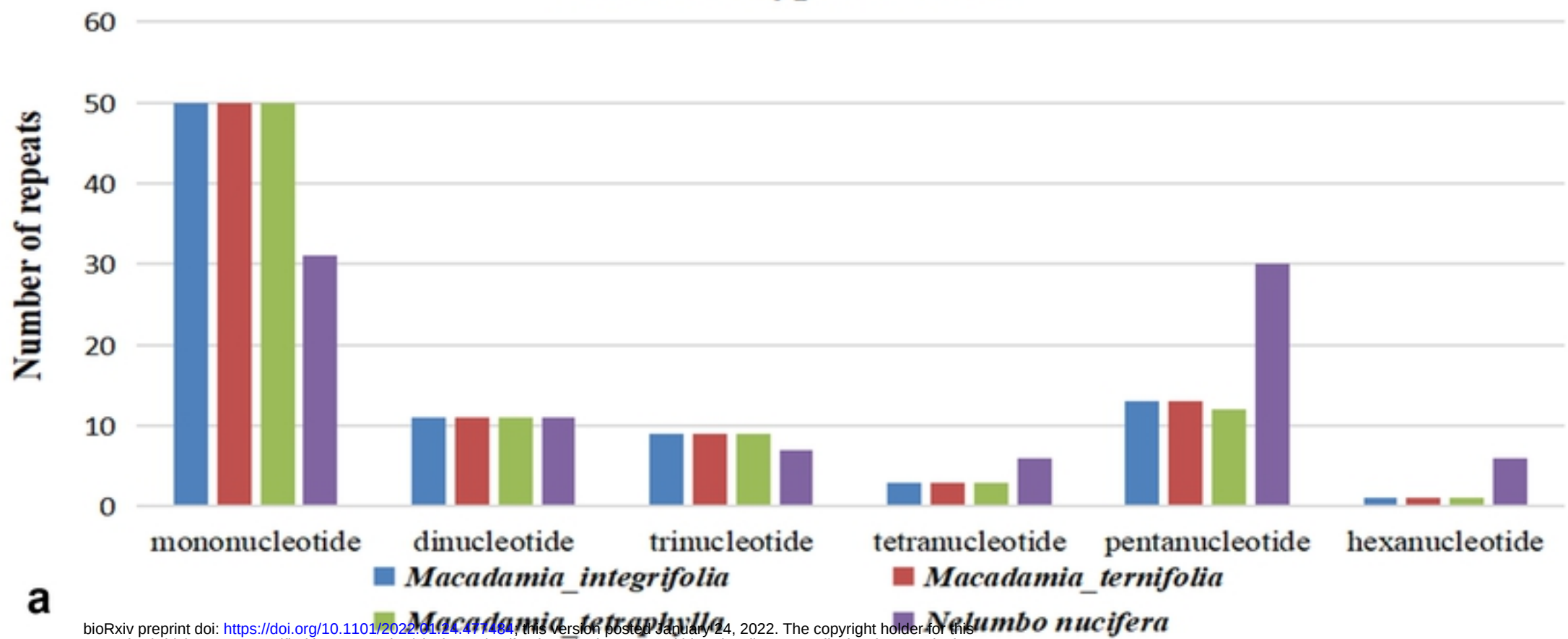
Type	RNA-editing	Number	Percentage
hydrophobic	GCA (A) => GTA (V)	1	30. 23%
	GCG (A) => GTG (V)	6	
	GCT (A) => GTT (V)	4	
	CTC (L) => TTC (F)	7	
	CTT (L) => TTT (F)	16	
	CCC (P) => TTC (F)	6	
	CCT (P) => TTT (F)	14	
	CCA (P) => CTA (L)	61	
	CCC (P) => CTC (L)	14	
	CCG (P) => CTG (L)	44	
	CCT (P) => CTT (L)	35	
hydrophilic	CAT (H) => TAT (Y)	24	12. 50%
	CAC (H) => TAC (Y)	11	
	CGC (R) => TGC (C)	15	
	CGT (R) => TGT (C)	36	
hydrophobic-hydrophilic	CCA (P) => TCA (S)	16	8. 28%
	CCC (P) => TCC (S)	13	
	CCG (P) => TCG (S)	6	
	CCT (P) => TCT (S)	22	
hydrophilic-hydrophobic	CGG (R) => TGG (W)	43	47. 97%
	TCC (S) => TTC (F)	47	
	TCT (S) => TTT (F)	58	
	TCA (S) => TTA (L)	101	
	TCG (S) => TTG (L)	59	
	ACA (T) => ATA (I)	7	
	ACC (T) => ATC (I)	1	
	ACG (T) => ATG (M)	8	
	ACT (T) => ATT (I)	6	
hydrophilic-stop	CGA (R) => TGA (X)	2	1. 02%
	CAG (Q) => TAG (X)	1	
	CAA (Q) => TAA (X)	4	

Notes: Compared with the other two species of *Macadamia*, *M. ternifolia* had only one more RNA-editing site(CTT (L) =>TTT (F)).

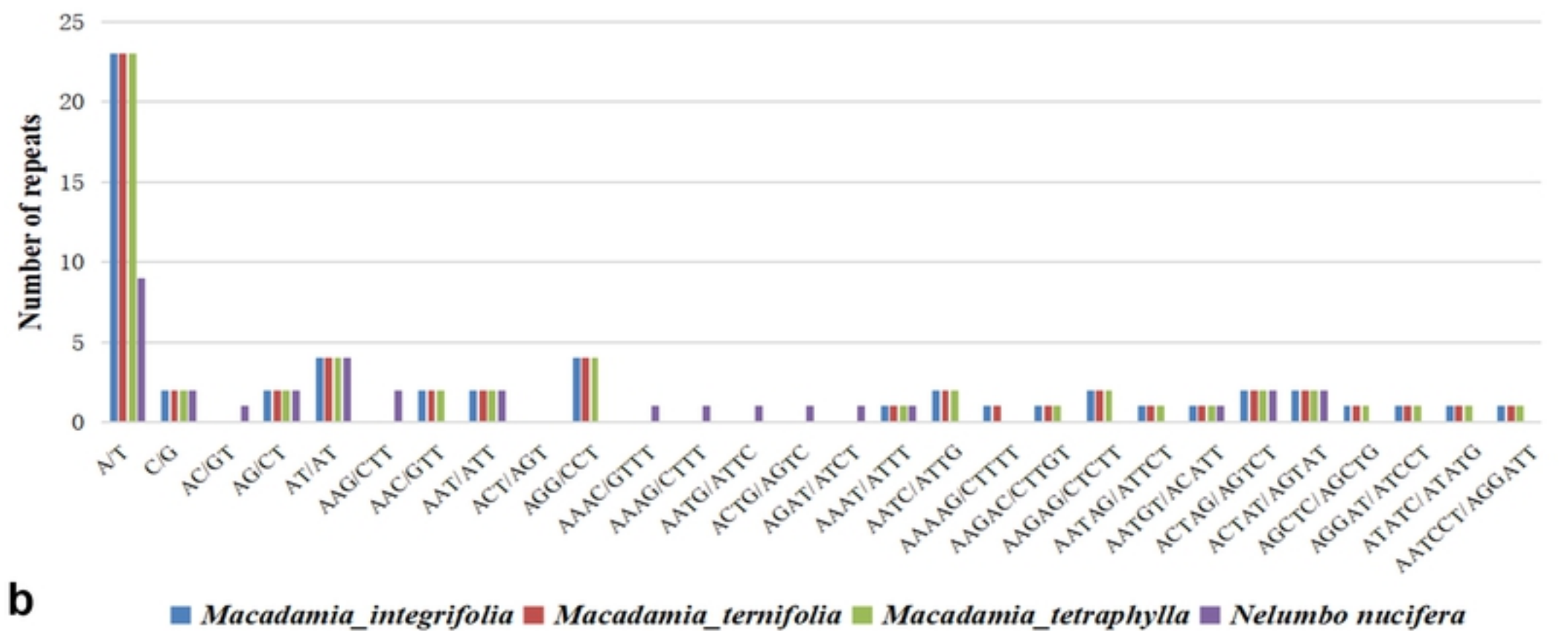




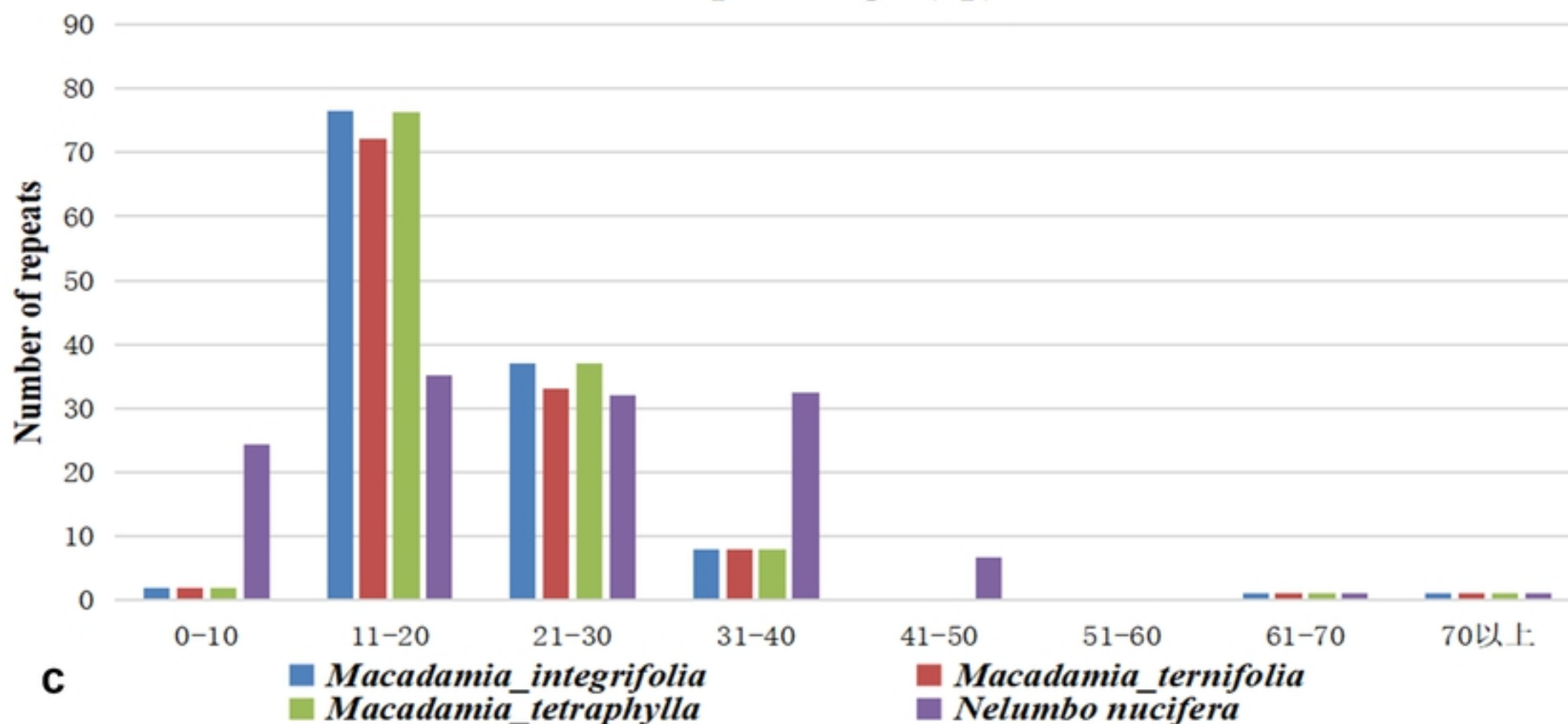
Different types of SSRs



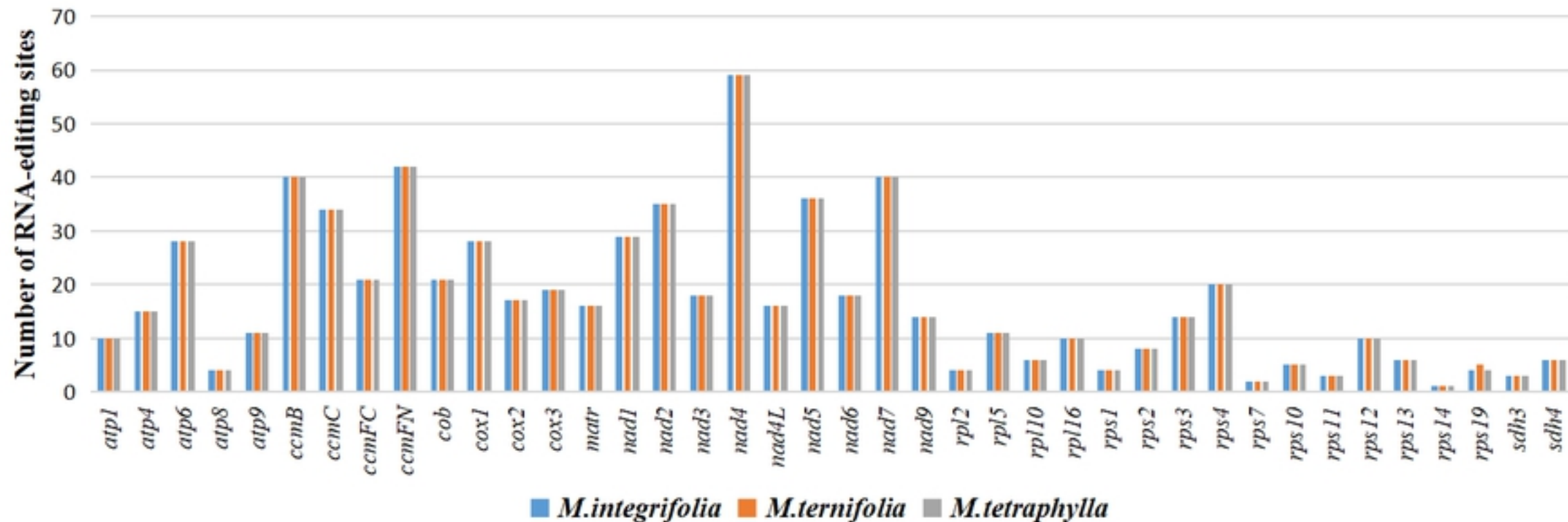
Different types of nucleotide motifs

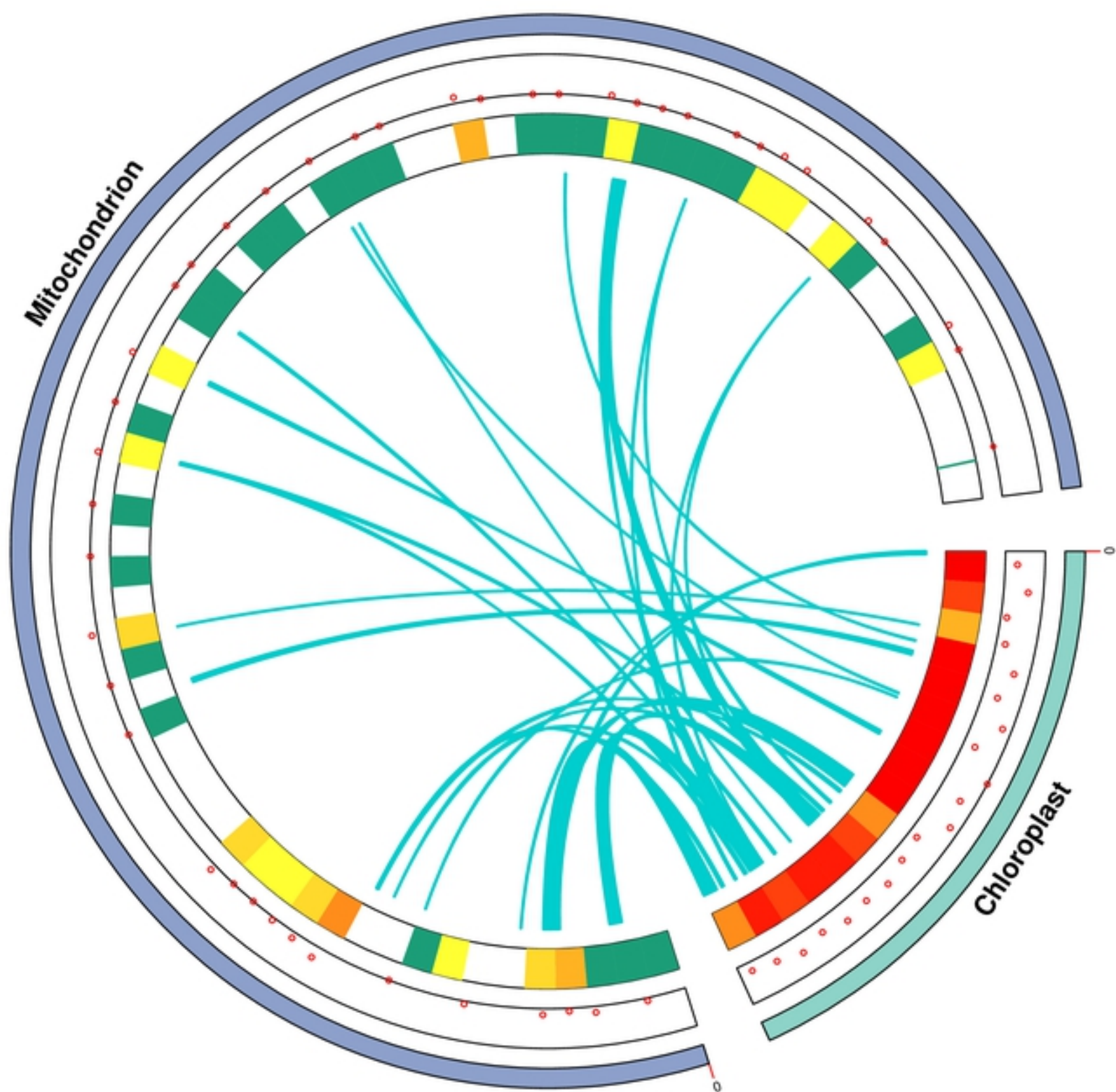


Repeat length(bp)



Number of RNA-editing sites

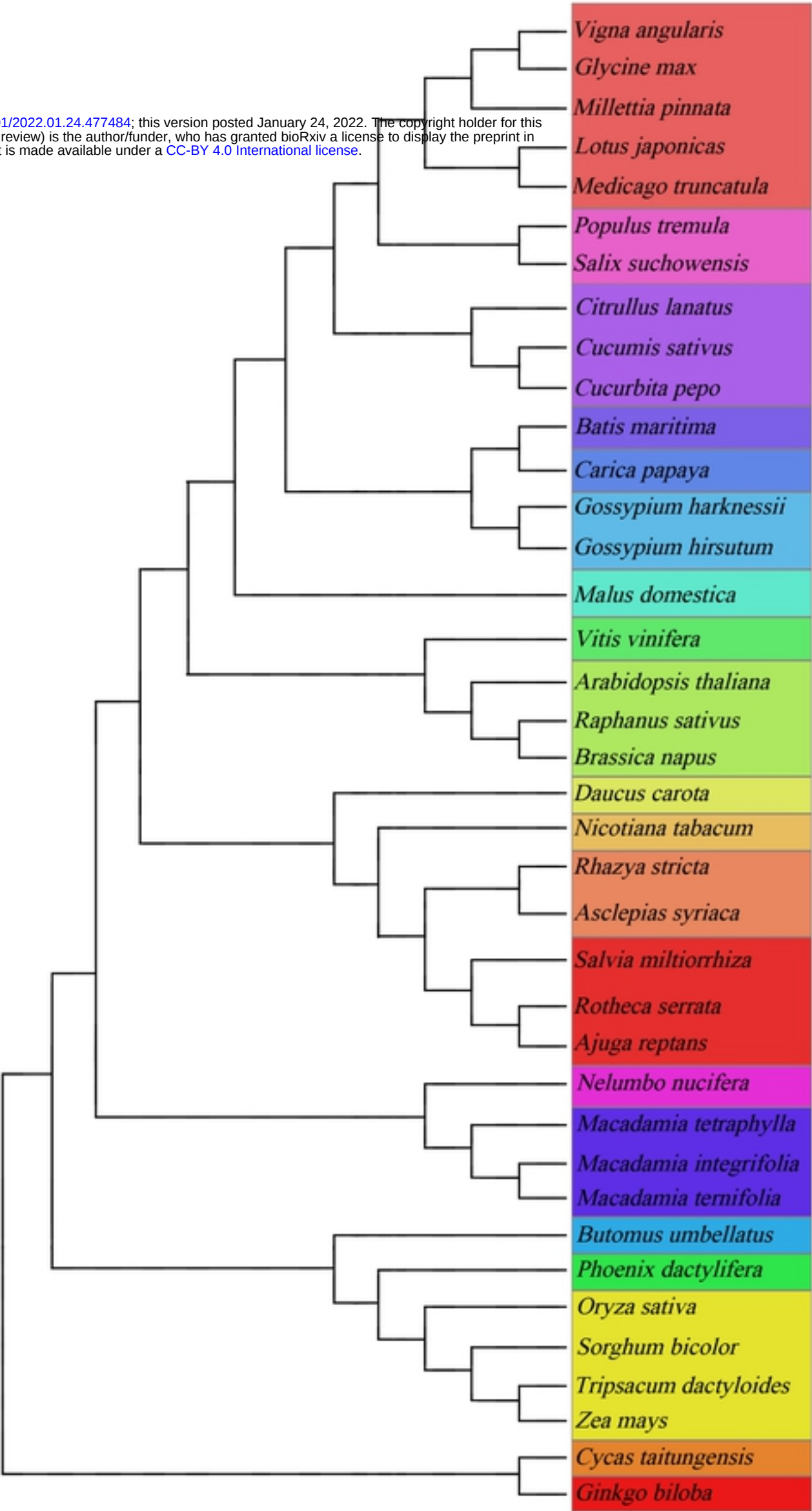




Family

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.24.477484>; this version posted January 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#).

- Leguminosae*
- Salicaceae*
- Cucurbitaceae*
- Bataceae*
- Caricaceae*
- Malvaceae*
- Rosaceae*
- Vitaceae*
- Brassicaceae*
- Apiaceae*
- Solanaceae*
- Apocynaceae*
- Lamiaceae*
- Nelumbonaceae*
- Proteaceae*
- Butomaceae*
- Arecaceae*
- Poaceae*
- Cycadaceae*
- Ginkgoaceae*



Eudicots

Monocots

Gymnospernae

M.integrifolia *M.ternifolia* *M.tetraphylla*

