

Diversity of RNA Editing in Chloroplast Transcripts across Three Main Plant Clades

Aidi Zhang

The Innovative Academy of Seed Design, Chinese Academy of Sciences

Jing Fang

University of Chinese Academy of Sciences

Xiujun Zhang (✉ zhangxj@wbgcas.cn)

The Innovative Academy of Seed Design, Chinese Academy of Sciences

Research Article

Keywords: RNA editing, plant, RNA sequencing, cytosine, evolution

Posted Date: August 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1838292/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Plant Systematics and Evolution on March 28th, 2023. See the published version at <https://doi.org/10.1007/s00606-023-01849-z>.

Abstract

RNA editing is a post-transcriptional modification of an RNA nucleotide sequence. Until now, different RNA editing systems were found in the major eukaryotic clades. In the plant kingdom, RNA editing was mainly documented in the mitochondria and chloroplast genomes. However, variation among large taxonomic groups and the evolutionary trajectory in terms of intra- and inter- clades remains unclear. To gain a better understanding of RNA editing evolution, in this study, based on publicly available RNA-seq data across three clades (fern, gymnosperm, and angiosperm), we provided a detailed analysis of chloroplast RNA editing events and discussed its evolution in land plants. A total of 5,205 editing sites were determined across 21 species after rigorous screening. We found that the clustering relations of RNA editing sites across 21 species agreed with the phylogenetic tree based on protein sequences approximately, and more editing sites occurred in early-diverging lineages for all the three clades, implying they shared similar evolutionary trajectories of editing loss. We observed that the average RNA editing level varied among species as well as genes, a lowest RNA editing level (~ 0.42) was detected in *Selaginella moellendorffii*; the highest editing level (~ 0.88) was detected in *atpA* gene. The reduction of cytosine content with evolution detected in our study further suggested that the substitution of the genomic sequence was the significant driver of loss of editing for later-branching plants. Many of the identified sites in our study have not been previously reported and provided a valuable data set for the future research community. Our findings also provide valuable information for the evolution of RNA editing in plants.

Introduction

RNA editing is a post-transcription process through which the nucleotide specified in the genome template is modified to produce a different transcript, thus contributing to the restoration of functional protein and proteomic variation, and providing another mechanism for modulating gene expression (Walkley and Li 2017, Zahn 2017, Small, Schallenberg-Rudinger et al. 2020). In the plant kingdom, RNA editing was first documented over a decade ago in the mitochondria of flowering plants (Covello and Gray 1989, Gualberto, Lamattina et al. 1989), and reported in chloroplast two years later (Hoch, Maier et al. 1991). There are two types of RNA editing in plants, the most common type is Cytosine-to-Uracil (C-to-U) conversion, and the infrequent type is U-to-C conversion that is reported only in ferns, mosses, and Lycopodiaceae (Gerke, Szovenyi et al. 2020). Traditionally, RNA editing seems to occur only in organelle genome encoded transcripts, however, a recent study detected U-To-C RNA editing events for nuclear genes in *Arabidopsis thaliana* (Ruchika, Okudaira et al. 2021). RNA editing predominantly takes place at the first or second positions of codons, thereby affecting the translated regions of protein-coding transcripts. The amino acids specified by the altered codons generated by editing are generally conserved in evolution, suggesting that most RNA editing events can restore the evolutionarily conserved amino acid residues in mRNAs (Ichinose and Sugita 2017). RNA editing thereby is an important process to maintain essential functions of encoded proteins at the RNA level, for example, pigment deficiency in tobacco cybrids is caused by the editing failure of the plastid *ATP synthase alpha-subunit (atpA)* mRNA (Schmitz-

Linneweber, Kushnir et al. 2005). Dynamic response of plant RNA editing to environmental factors was detected in previous studies (Miyata and Sugita 2004, Rodrigues, da Fonseca et al. 2017, Xiong, Tao et al. 2017, Zhang, Jiang et al. 2020). Many factors are involved in plant RNA editing and considered to interact with one another to form a large protein complex, termed as editsome (Shikanai 2015). PLS subfamily members of pentatricopeptide repeat (PPR) proteins function in site recognition of the target cytosine, almost all the PPR proteins are localized in either chloroplasts or mitochondria where those proteins participate in different facets of RNA metabolism such as RNA splicing, RNA stability, and translational initiation (Yagi, Tachikawa et al. 2013, Shikanai 2015). Multiple organelle RNA editing factors (MORF) family members are also components of the RNA editosome and are required for RNA editing at multiple editing sites in plants (Yagi, Tachikawa et al. 2013, Yan, Zhang et al. 2018).

In the last two decades, RNAs were usually compared with their corresponding DNA templates to detect RNA editing sites, however, this approach is time-consuming and prone to underestimate the numbers of editing sites. In recent years, the availability of large quantities of RNA sequencing data makes it possible to identify RNA editing sites and quantify their editing level on a large scale. This strategy allows a transcriptome-wide fast detection of editing sites and has enormous potential to deepen our knowledge of transcriptional processes in the plant. Indeed, with the growth of complete plant organellar genomes and related transcriptome data in the last decade, hundreds of editing sites have been identified in more and more plants (Lo Giudice, Pesole et al. 2018, Lo Giudice, Hernandez et al. 2019, Oldenkott, Burger et al. 2020, Shtratnikova, Schelkunov et al. 2020). But this strategy is also a challenging task due to its accuracy of mapping the RNA-seq reads against genomic sequence, hence, different bioinformatic strategies have been introduced to improve the detection accuracy of RNA editing sites (Sun, Li et al. 2016, Wang, Lian et al. 2016, Zhang, Lu et al. 2017, Edera and Sanchez-Puerta 2021, Ichinose and Sugita 2021).

Evolutionary studies can help to understand the puzzling nature of RNA editing in plants. More and more recent studies demonstrated that RNA editing is a widespread phenomenon that occurred in various land plants, including the liverworts, mosses, hornworts, lycopods, ferns, and flowering plants (Edera, Gandini et al. 2018, Ishibashi, Small et al. 2019). However, no instance of RNA editing has yet been detected in algae, suggesting that RNA editing may have evolved in organelles only after the green plants established themselves on the land (Ichinose and Sugita 2017). The frequency of RNA editing sites varies from zero to hundreds across the plant kingdom, among land plants (Takenaka, Zehrmann et al. 2013, Smith 2020). The unparalleled variation in RNA editing among fern plastomes was demonstrated in several recent studies (Smith 2020, Fauskee, Sigel et al. 2021). Yet, variation in the frequency and editing level among gymnosperms and angiosperms remains unclear. Rare comparison study has been conducted in terms of intra- and inter- groups for chloroplast RNA editing except for ferns. To gain a better understanding of RNA editing in plant chloroplast, in this study, we chose diverse plant species that distributed in three main clades (fern, gymnosperm, and angiosperm), and determined thousands of editing sites based on the amount of RNA-seq data. The detailed comparison of RNA editing events provided valuable information for the evolution of plant RNA editing.

Materials And Methods

Data acquisition

We selected 21 species across three clades (fern, gymnosperm, and angiosperm) for detection of chloroplast RNA editing. For each species, the corresponding raw Illumina RNA reads were obtained from SRA database at NCBI based on two criteria: (1) paired-end reads that possess higher mapping specificity were preferred; (2) RNA reads obtained from leaves of wild type individuals were only selected. Besides, for each species, the reference file consisting of chloroplast genome sequences and corresponding gene annotation files were also downloaded from the GenBank database. Detailed information of RNA-seq data and reference files used in our study were listed in Table S1.

Read mapping and SNP calling

The identification process of RNA editing sites can be decomposed into three steps: first read alignment, second the SNP calling, and third detection of RNA editing sites. For each species, to increase sequencing depth, we merged all the replicates into one sample. The quality control of paired-end Illumina sequencing data was evaluated first by NGSQCToolkit (Patel and Jain 2012), low-quality sequence data were filtered out ($\text{cutOffQualScore} < 20$). RNA reads from each species were then mapped to reference using hisat2 software under default parameters (Kim, Langmead et al. 2015). Afterward, the alignment results were sorted, removed duplicates, indexed, and sorted by using SAMtools (Li, Handsaker et al. 2009). Finally, the resulting BAM file was then used to call DNA/RNA variants using bcftools, VCF files that describe transcriptome variation were generated (Danecek and McCarthy 2017).

Detection of RNA editing sites

For each species, based on the SNP-calling results (in “VCF” format) and genome annotation files (in “tbl” format), RNA editing sites were identified under default parameter values by using the REDO tool (Wu, Liu et al. 2018). REDO is a comprehensive application tool for identifying RNA editing events in plant organelles based on variant call format files from RNA-sequencing data. REDO works require only three input files: a file that contains the SNP-calling results (records for all sites), the genome sequence file of organelle reference (FASTA format), and its corresponding gene annotation file (feature table file, www.ncbi.nlm.nih.gov/projects/Sequin/table.html). Regarding the high false positive of editing sites, REDO uses a series of stringent criteria to filter the raw variants, as the below following: (1) quality control filter ($\text{MQ} > 255$), the low quality sites are filtered out according to the reads quality; (2) total reads depth filter ($\text{DP} > 3$); (3) alt proportion filter ($\text{alt proportion} < 0.1$); (4) multiple alt filter, only the variant with one alt allele is retained for RNA editing detection; (5) distance filter, the variant sites in short distance (< 3 bp) are filtered out due to the possible positional interference for RNA editing; (6) spliced junction filter, variants within short spliced anchor are removed; (7) indel filter, the indel variants are removed; (8) likelihood ratio (LLR) test filter, LLR test is a probabilistic test incorporating error probability of bases

(error probability is obtained using adjacent nonvariant sites in specific window) for detecting RNA editing sites (LLR <10); (9) Fisher's exact test filter (p value < 0.01), the significance for a given RNA editing site (alt reads, ref reads) by comparing its expected levels (0, alt reads + ref reads) using the Fisher exact test; (10) complicated filter model, based on the statistics results for the attributes of codon table and experiment validated RNA editing sites, a complicated filter model was built according to five features of RNA editing sites, which are RNA editing types, alt proportion, amino acids change, codon phase, and hydrophobic/hydrophilic change. Finally, all raw RNA editing sites were detected, meanwhile, their corresponding annotation information files were also generated.

Filtration of RNA-editing sites

To minimize the false positives produced by the automated approach, we manually examined all raw editing sites, only kept the C-to-U and U-to-C editing types, and excluded other mismatches, such as A-to-C, T-to-A, etc. In addition, the changes that may be derived from genomic variation, (always exhibit 99% alteration rates) with low depth were also removed. To evaluate the reliability of editing, for each species, we also used PREPACT tool to predict potential RNA editing events supplying with entire chloroplast genomes as input files, with a filter threshold of at least 80% of the references under BLASTX mode (Lenz and Knoop 2013).

Comparing RNA-editing sites

All the filtered RNA editing sites detected in 21 species were used for further statistics and feature analysis, including statistics of editing number, editing type, codon position, amino acid changes, and involved genes, and so on. To decipher the distribution of RNA editing frequency across different species, the top 30 genes with most editing sites across 21 species were selected, cluster analysis and heatmap plotting were conducted based on the matrix of RNA editing numbers. The CDS sequences of the top 30 genes across 21 species were concatenated and subjected to alignments and phylogenetic tree construction using RAxML (Stamatakis 2014). Meanwhile, the RNA editing level of the top 30 genes was also subjected to statistical analysis. The value of RNA editing level at one site was expressed as the proportion between edited transcripts and total transcripts. If one site was edited, the C/G base (wild type) should be altered to the T/A base (edited type), since one editing site could be detected hundreds of times via sequencing, the number of wild types (C/G) or edited type (T/A) of bases could then be counted at this particular site, then the editing level at one site could then be calculated by the formula: depth of edited bases (T and A)/total read depth of bases. Values of the editing level matrix were normalized by subtracting the row-wise mean from the values in each row of data and multiplying all values in each row of data by the value of standard deviation. For each clade (fern, gymnosperm, and angiosperm), a heatmap was plotted across all of its species using "pheatmap" function in R respectively, the distance matrix of different samples was calculated using "dist" function with the default Euclidean method, the hierarchical clustering was computed using "hclust" function.

Comparing cytosines content

Considering protein-coding genes varied among different species, we picked out shared edited genes for the statistics of cytosine content across 21 species. For each shared edited gene, we extracted its CDS sequence and calculated the ratio of cytosine content. For each two of the clades (fern, gymnosperm, and angiosperm), pairwise comparisons of average cytosines content were conducted. A two-tailed Wilcoxon rank-sum test was used.

Illustration of *atpA* RNA editing sites

The gene sequences of *atpA* across 21 species were collected, the intersection of all the species' RNA editing sites of *atpA* was concatenated for alignment and annotated, sequence logo of *atpA* gene was produced by WebLogo (Crooks, Hon et al. 2004), alignment was conducted using ClustalW that implemented in MEGA5 under default parameters (Kumar, Stecher et al. 2018).

Results

Identification of RNA editing sites

A series of species that represent distant evolutionary clades were selected judging by two criteria, one is enough transcriptomic data of leaf tissue in the SRA database, and another is the availability of sequenced chloroplast genome. Hence, 21 species, consisting of 6 ferns, 4 gymnosperms, and 11 angiosperms (see Table 1), and corresponding 317 SRA accessions were chosen finally. Detailed information of SRA data and chloroplast genome accessions were listed in Table S1. Mapping results (Figure S1) showed that RNA-seq data volume varied among different species, compared with angiosperms, ferns and gymnosperms have lower mapping depths, especially for *Adiantum aleuticum* and *Histiopteris incisa*, indicating that the actual number of RNA editing in ferns and gymnosperms might be underestimated. We also observed that the read density also varied widely among different genomic regions, ranging from less than 30 to more than 800 in a few species, demonstrating varied expression levels of genes or sequence biases. Based on results of RNA-seq data mapping and SNP calling, an automated bioinformatics pipeline implemented in REDO tool (Wu, Liu et al. 2018) was conducted under default thresholds. Consequently, there were a total of 6,011 raw editing sites located chloroplast genome was detected in leaf. However, despite multiple stringent filters, sequence mismatches that accord with RNA editing occasionally appeared, hence, we manually examined all mismatches to eliminate false positive, only kept C-to-U and U-to-C editing types, editing sites with editing efficiency close to 100% that may be derived from genomic variation were also eliminated. Finally, a total of 5,205 RNA editing sites, consisting of 237 U-to-C and 4,968 C-to-U edit sites, were screened out for further analysis. To evaluate the reliability of results, PREPACT webserver (Lenz, Hein et al. 2018) was also used to verify the editing sites. We found that the distribution of the predicted number of RNA editing based on PREPACT agreed with that of our prediction based on RNA-seq data basically (see Figure 1), reflecting our pipeline offered

high performance with reliable results. To describe the attributes of RNA editing sites, we illustrated by one example of samples of *Adiantum aleuticum* (see Figure S2), which depicts the reliability of RNA editing sites by REDO tools statistically.

Table 1

The summary of RNA editing sites across three clades

	Species	chloroplast genome	Number of raw editing sites	Number of filtered editing sites	Percent of edited genes	MD ¹
Ferns	<i>Pteris vittata</i>	MH500228	349	319	74%	102
	<i>Adiantum aleuticum</i>	MH173079	422	362	82%	23
	<i>Selaginella moellendorffii</i>	HM173080	1,316	1,316	96%	73
	<i>Histiopteris incisa</i>	NC_040220	588	414	85%	27
	<i>Cibotium barometz</i>	NC_037893	562	540	85%	32
	<i>Cyrtomium fortunei</i>	NC_037510	778	551	91%	40
Gymnosperms	<i>Ginkgo biloba</i>	NC_016986	301	291	81%	228
	<i>Picea abies</i>	HF937082	123	121	61%	450
	<i>Cycas revoluta</i>	NC_020319	173	132	44%	27
	<i>Pinus massoniana</i>	MF564195	96	93	59%	182
Angiosperms	<i>Liriodendron tulipifera</i>	NC_008326	320	260	75%	113
	<i>Nelumbo nucifera</i>	NC_025339	154	121	63%	295
	<i>Nicotiana tabacum</i>	Z00044	107	74	23%	134
	<i>Glycine max</i>	NC_007942	64	59	34%	195
	<i>Populus tremula</i>	KP861984	93	89	43%	244
	<i>Arabidopsis thaliana</i>	KX551970	50	45	24%	389
	<i>Gossypium hirsutum</i>	NC_007944	127	97	37%	171
	<i>Helianthus annuus</i>	NC_007977	59	57	29%	148
	<i>Phoenix dactylifera</i>	NC_013991	121	118	42%	781
	<i>Zea mays</i>	NC_001666	105	79	37%	814
	<i>Oryza sativa</i>	NC_001320	103	65	26%	550

Note: ¹ mapping depth.

Distribution of RNA editing across three clades

After manual inspection and elimination of mismatches, filtered RNA editing sites across 21 species were screened out, the summary was listed in Table 1, Table S2, detailed corresponding annotation

information was produced simultaneously, as listed in Table S3. All the filtered editing sites were located in 1,038 genes across all species, the average percent of edited genes for the three clades (ferns, gymnosperms, and angiosperms) are 0.85, 0.61, and 0.39 respectively. Compared to the latter two clades, more sites and genes were edited in chloroplast transcripts from ferns remarkably (Figure 2a). In ferns, *Selaginella moellendorffii* has the largest number of editing sites, up to 1,316, which is nearly 100-fold more abundant than that of flowering plants, nearly all the chloroplast genes (77 genes, ~96%) suffered effective editing. Differ from *Selaginella moellendorffii* that exclusively belongs to Lycopsidea, the other five fern plants are members of Leptosporangiopsida, have relatively smaller numbers of editing sites, represented by *Cyrtomium fortune*, which owned the second-largest number of editing sites, with 551 editing sites and 79 edited genes (~91%). Whereas for gymnosperms, the average number of editing sites and percent of edited genes were all less than that of ferns. Compared with the other three gymnosperms, *Ginkgo biloba* has the most editing sites, with 291 editing sites and 68 edited genes (~81%). On the opposite end, angiosperms have the lowest average numbers of editing sites, only a part of genes were effectively edited, lower than 50%. It was noticeable that *Liriodendron tulipifera* and *Nelumbo nucifera* distinguished them from other angiosperms with more editing events. A total of 260 editing sites were detected in *Liriodendron tulipifera*, which is nearly 3-fold more abundant than that of other angiosperms and gymnosperms except for *Ginkgo biloba*, the percent of edited genes was up to 75%, well above the average. The numbers of editing sites among other 9 angiosperms showed no significant differences. The above results illustrated the diversity of RNA editing distribution across the three clades, early-diverging clades/species showed higher numbers of editing sites compared with that of later-diverging. The statistics of RNA editing sites showed that 21.97%, 67.16%, and 10.58% of sites were edited at first, second, and third codon positions respectively (Figure 2b). In terms of editing types, C-to-U was the dominant editing type (nearly ~95.1%), the next is U-to-C type, and other mismatch types were rare (Figure 2c). The statistics of editing types showed that the majority (~95%) of the editing events resulted in non-synonymous codon changes, and the changed amino acids tend to be hydrophobic, the frequency of changes from hydrophilic to hydrophobic was the highest, followed by changes from hydrophobic to hydrophobic (Figure 2d). The most common amino acid change types were Ser-to-Leu and Pro-to-Leu, serine is hydrophilic, whereas leucine and proline are both hydrophobic. The above results demonstrated that the RNA editing exhibited a selective advantage in the overall increase in hydrophobicity of the resulting proteins, which was also in good agreement with our previous studies (Zhang, Jiang et al. 2020).

Variability of RNA editing among species and genes

To further explore the evolutionary trajectory of RNA editing for the three clades, for each gene, we summed the number of its editing sites across 21 species (see Table S4), and picked out the top 30 genes with most editing sites for cluster analysis. Base on the matrix of numbers of editing sites across 21 species, a hierarchically clustered heatmap was plotted in Figure 3. The clustering relations showed that the 21 species were divided into three subgroups. The remarkable abundant of editing events make two exceptional species (*Selaginella moellendorffii* and *Liriodendron tulipifera*) cluster far away from their

own clades. We found that not every gene were edited in all the species, by grouping the genes based on their function, genes encoding membrane subunits of the *chloroplast NDH complex (ndh)* exhibited larger average numbers of editing sites, this is consistent with previous studies that RNA editing occurred preferentially in genes encoding membrane-bound proteins under strong selection (Mower and Palmer 2006). Whereas *ribosomal subunit genes* showed lower numbers of editing sites. Due to the well-studied background and abundant editing sites in the plant, *ndhB* gene is assumed to be a good case for the study of RNA editing evolution, in our study, *ndhB* was confirmed to possess the most editing sites, with 333 editing sites spread across 17 species. Despite this, there is a biased distribution of RNA editing sites in *ndhB* among three clades. In ferns, 50 sites were detected in *Selaginella moellendorffii*, only 4~17 sites in the other five ferns; in angiosperms, there were about 20 editing sites for each of the 11 angiosperm species; in gymnosperms, RNA editing in *ndhB* was only detected in *Ginkgo biloba*, for *Picea abies* and *Pinus massoniana*, no *ndhB* gene annotated in their chloroplast genome, whereas for *Cycas revolute*, no RNA editing events were detected in its *ndhB* gene, which may result from loss of editing or too low depth around genomic regions of its *ndhB* gene. Based on alignments of merged protein sequence for top 30 genes across 21 species, phylogenetic tree by Maximum Likelihood method was constructed, as shown in Figure 3b, which showed that the clustering relations agreed with that of the matrix of numbers of editing sites roughly. The above observation implied the conservation of RNA editing evolution across different clades. Hierarchically clustered heatmaps of numbers of RNA editing sites from all/each clade were shown in Figure S3-6 respectively.

Variability of RNA editing level

RNA editing level was used to measure to what extent the edited transcripts among all transcriptomes for one gene. In this study, we explored the distribution of editing levels among three aspects: codon positions, species, and edited genes. We observed that the distribution of RNA editing level for three codons did not comply with the normal distribution, featuring a peak around ~0.2 and fat tails, as shown in Figure 4. The comparison between codon positions showed that the average editing level in the second codon position (~0.72) is higher than that of the first (~0.63) and third codon positions (~0.61) significantly, suggesting non-synonymous substitution occurred in second codon position tend to be effectively edited, it was higher editing level that dominated the landscape of RNA editing. In addition, we also found that the average editing level also varied widely across 21 species, ranging from 0.42 to 0.85 (Figure 5a), the trend of plotting of filtered sites was consistent with that of raw sites roughly. *Selaginella moellendorffii* has the lowest editing level (~0.42), far below that of other species. In gymnosperms and angiosperms, *Ginkgo biloba* and *Nelumbo nucifera* had the lowest editing efficiencies (~0.6) respectively. It seemed that abundant editing sites detected in those early-branching plants might have a negative impact on their editing level. However, as an ancient plant in angiosperms, *Liriodendron tulipifera* was one exception, with an editing level up to 0.81. The editing level was also analyzed in each gene individually, we averaged the RNA editing level among the top 30 genes across 21 species, the result also demonstrated the diversity of distribution, as shown in Figure 5b, *Maturase K gene (matK)* has the lowest editing level (~0.5), oppositely, editing level of *atpA* gene was the highest, up to 0.88.

Reduced cytosines content with the evolution of plants

Considering the large differences in the scale of RNA editing events along with evolution, we analyzed the nucleic acid base composition of edited genes shared by the 21 species. There were a total of 51 genes shared in all the chloroplast genomes of 21 species. The percent of cytosines for each gene was calculated by the formula: the number of cytosines (C)/total number of bases (A/T/C/G), as listed in Table S5. Afterward, the cytosines content of genes was averaged for each clade, remarkable significances ($p\text{-value} < 0.05$) were detected for pairwise comparisons between every two clades except for gymnosperms-angiosperms, as shown in Figure 6a. The percent of cytosines in ferns was far below that of angiosperms, followed by gymnosperms. The highest average of cytosines content was found in *Selaginella moellendorffii* (~0.26), the lowest average was found in *Glycine max* (~0.17). *Ginkgo biloba* and *Liriodendron tulipifera* have the highest averages (~0.19) in their own clade, showing a positive correlation with their high numbers of RNA editing sites. We further compared the cytosines content for each gene across the 21 species, as shown in Figure 6b, which demonstrated that the percent of cytosine dramatically declined roughly along with evolution with a few exceptions, such as *50S ribosomal protein L23 (rpl23)* gene. One striking example was *Photosystem II reaction center protein I (psbI)* gene, which has the highest percent of cytosines in *Selaginella moellendorffii* (~0.37), and dropped to about 0.18 in other species. The above results indicated that the declination of cytosine content with evolution was the significant driver of loss of editing sites for later-branching plants.

Illustration of atpA RNA editing

We illustrated the RNA editing of *atpA* gene to help understand the evolution trajectory vividly. All editing sites identified in *atpA* gene were marked by a yellow color, as shown in Figure 7, it demonstrated that numbers of editing sites, as well as cytosine content, declined from ferns to angiosperms. We found that editing sites at the third codon position only occurred in certain species, were poorly conserved, for example, RNA editing in locus-a (Figure 6) only occurred in *Adiantum aleuticum* despite the existence of cytosine in other fern members, indicating that synonymous substitutions at third codon position were not conserved. In contrast, RNA editing at the second codon position was relatively conserved, RNA editing in locus-b and -c occurred in all three clades. Whereas in locus-d, RNA editing occurred in two members of gymnosperms, and the base types of other 19 species were all corrected to thymine in the genome, furthermore, we found many sites occurred in certain species for the three clades, implying the independent origins of RNA editing across different clades. In extreme cases, when all the cytosines with potential editing capacity of *atpA* gene were completely edited in fern plants, the overall uracil levels in edited transcripts will increase to the comparable levels of angiosperms and gymnosperms on the whole. However, in ferns' transcripts, most of the sites were partly edited, hence, their uracil level is still lower than that of angiosperms and gymnosperms.

Discussion

As a post-transcription process, RNA editing can modify the genome template to produce a different transcript (Ichinose and Sugita 2017). Numerous studies have proved that RNA editing occurred in nearly all plants in the kingdom, RNA editing exhibited dynamics response to different stress factors and development stages, such as flower development and male sterile (Miyata and Sugita 2004, Brenner, Mader et al. 2019, Lo Giudice, Hernandez et al. 2019, Zhang, Jiang et al. 2020, Fang, Jiang et al. 2021). RNA editing may also cause secondary structure transformation of transcripts (Farre, Akinin et al. 2012). In this study, to gain a better understanding of RNA editing in plant chloroplast, we collected a large amount of RNA-seq data and performed a series of procedures to investigate RNA editing in 21 diverse plants distributed in three clades. We determined a total of 5,205 editing sites located in chloroplast genes across 21 species and quantified their editing level, demonstrating the powerfulness of the bioinformatics approach in studying RNA editing, many of the identified sites have not been previously reported, thus providing a valuable data resource for future research. We found that the clustering relation of numbers of RNA editing sites agreed with the phylogenetic tree based on gene sequences approximately, verifying that RNA editing across plant kingdom are comparatively conservative and accord with laws of evolution roughly. A great deal of variability of RNA editing numbers, as well as the RNA editing level, were detected among species, genes, codon positions.

In total, numbers of editing sites declined with the evolution of plants, editing events occur more often in the early-diverging plant than later-branching ones for each clade. As one of the most ancient flowering trees, *Liriodendron tulipifera* possessed the highest number of editing sites in angiosperms. In gymnosperms and ferns, one gymnosperm plant, *Ginkgo biloba*, and one fern plant, *Selaginella moellendorffii* both exhibited ancient features, such as higher numbers of sites. The above observation was consistent with a previous study, which found that the profile of chloroplast RNA editing of *Amborella* represented an ancestral RNA editing pattern in angiosperms (Ishibashi, Small et al. 2019). These observations implied that RNA editing may break out in early-branching plants from different clades simultaneously and suffer a lot of loss during evolution. We also found that a reasonable percentage of editing sites occurred in certain clades, and were lost in other clades whose cytosine already corrected to thymine in the genome. The declination of cytosine content with evolution detected in our study further indicated that the substitution of genomic sequence (C-to-T) was the significant driver of loss of editing for later-branching plants. For certain species, its lack of editing at a few genes may be explained by two reasons, one is the absence of the genes that are annotated in the chloroplast genome, another is no RNA editing occurred in the genes. We found that new editing sites also occurred in certain higher plants occasionally that lack in early-diverging species, indicating the diversity of RNA editing evolution. Even though the substitution of C-to-T in the genome occurred in all gymnosperms and certain angiosperms, there were still some angiosperms that needed the RNA editing strategy to make functional protein, such as the locus-c of *atpA* gene.

RNA editing level also showed variability in three aspects: codon positions, species, and edited genes. It seemed that abundant editing sites detected in early-branching plants have a negative impact on their editing levels, such as *Selaginella moellendorffii*, *Ginkgo biloba*, and *Nelumbo nucifera*, these early-diverging species had a relatively lower editing level. It is speculated that the efficiency of RNA editing

might be at the mercy of the limited expression of RNA editing factors. RNA editing activities also showed a varied degree of conservation among different codon positions, the higher conservation and editing level in the second codon position suggested non-synonymous substitution tended to be effectively edited, it was a higher editing level that dominated the landscape of RNA editing. To a certain degree, the discrepancy of editing level among genes reflected their importance of function.

Until now, there are two viewpoints about the nature of RNA editing, one is a contribution to variations in proteomic sequence and modulating gene expression; another point thinks that RNA editing in plants might be a repair mechanism to correct genomic point mutations at the post-transcription level, thus increase the substitution rate that is extremely low in the organellar genome (Takenaka, Zehrmann et al. 2014, Tang and Luo 2018). Previous studies revealed that organelle genomes have a slower evolutionary rate than the nuclear genome, thus accumulating many T-to-C mutations that constitute a prerequisite for the generation of plant organellar RNA editing (Barbrook, Howe et al. 2010), which could correct those T-to-C mutations to restore the evolutionarily conserved amino acid residues in mRNAs. During the evolution of land plants, most mutations would have been finally corrected to thymine in the genome especially for higher plants, eliminating the need for editing at certain sites, thus the number of editing sites showed a remarkable reduction. But some sites still need to be edited to thymine or remain to be cytosine for coding for different amino acids. Hence, the genome mutations are the driving force behind the evolution of editing sites in plants, the increasing modification of C-to-T at the genome level might be more accurate to describe the evolution trajectory instead of RNA editing.

Conclusions

In this study, based on a bioinformatics pipeline, we provided a detailed analysis of RNA editing events in chloroplast genomes across three main plant clades, and discussed the evolution of RNA editing in land plants. Our study represents a valuable data set for the research community, and thus helps understand the puzzling nature of RNA editing in plants.

Abbreviations

C-to-U: Cytosine-to-uracil; *PPR*: *Pentatriconic peptide repeat*; *ndhB*: *subunit of the NADH dehydrogenase-like complex*; *MORF*: *Multiple organelle RNA editing factors*; *atpA*: *ATP synthase alpha-subunit*.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the NCBI repository (<https://www.ncbi.nlm.nih.gov/>) with the accessions shown below:

RNA-seq data	Chloroplast genome
PRJNA398727	AY522330.1
PRJEB21674	HF937082.1
PRJEB21674	HM173080.1
PRJEB29012	KP861984.1
PRJEB29012	KX551970.1
PRJEB29530	MF564195.1
PRJNA196471	MH173079.1
PRJNA253977	MH500228.1
PRJNA273274	NC_001319
PRJNA281136	NC_001666
PRJNA281136	NC_007942.1
PRJNA301536	NC_007944.1
PRJNA343084	NC_007977.1
PRJNA349800	NC_008326.1
PRJNA350792	NC_013991.2
PRJNA356942	NC_016986.1
PRJNA384992	NC_025339.1
PRJNA396655	NC_037510.1
PRJNA472694	NC_037893.1
PRJNA735245	NC_040220.1
PRJNA802195	Z00044.2

Competing interests

The authors declare no conflict of interest.

Funding

This research was funded by the National Natural Science Foundation of China, grant numbers 32070682, the National Science & Technology Innovation Zone Project, grant numbers 1716315XJ00200303 and 1816315XJ00100216, and CAS Pioneer Hundred Talents Program.

Authors' contributions

ADZ and XJZ designed the experiments. ADZ and JF performed statistical analyses of gene expression data. ADZ performed statistical analyses. ADZ drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We would like to thank the members of Bioinformatics Group of Wuhan Botanical Garden, Chinese Academy of Sciences, China for the discussion and suggestion to improve the manuscript.

References

1. Barbrook, A. C., C. J. Howe, D. P. Kurniawan and S. J. Tarr (2010). "Organization and expression of organellar genomes." *Philos Trans R Soc Lond B Biol Sci* 365(1541): 785-797.
2. Brenner, W. G., M. Mader, N. A. Muller, H. Hoenicka, H. Schroeder, I. Zorn, M. Fladung and B. Kersten (2019). "High Level of Conservation of Mitochondrial RNA Editing Sites Among Four *Populus* Species." *G3-Genes Genomes Genetics* 9(3): 709-717.
3. Covello, P. S. and M. W. Gray (1989). "Rna Editing in Plant-Mitochondria." *Nature* 341(6243): 662-666.
4. Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). "WebLogo: a sequence logo generator." *Genome Res* 14(6): 1188-1190.
5. Danecek, P. and S. A. McCarthy (2017). "BCFtools/csq: haplotype-aware variant consequences." *Bioinformatics* 33(13): 2037-2039.
6. Edera, A. A., C. L. Gandini and M. V. Sanchez-Puerta (2018). "Towards a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria." *Plant Mol Biol*.
7. Edera, A. A. and M. V. Sanchez-Puerta (2021). "Computational Detection of Plant RNA Editing Events." *Methods Mol Biol* 2181: 13-34.
8. Fang, J., X. H. Jiang, T. F. Wang, X. J. Zhang and A. D. Zhang (2021). "Tissue-specificity of RNA editing in plant: analysis of transcripts from three tobacco (*Nicotiana tabacum*) varieties." *Plant Biotechnology Reports* 15(4): 471-482.
9. Farre, J. C., C. Aknin, A. Araya and B. Castandet (2012). "RNA Editing in Mitochondrial Trans-Introns Is Required for Splicing." *Plos One* 7(12).
10. Fauskee, B. D., E. M. Sigel, K. M. Pryer and A. L. Grusz (2021). "Variation in frequency of plastid RNA editing within *Adiantum* implies rapid evolution in fern plastomes." *Am J Bot* 108(5): 820-827.
11. Gerke, P., P. Szovenyi, A. Neubauer, H. Lenz, B. Gutmann, R. McDowell, I. Small, M. Schallenberg-Rudinger and V. Knoop (2020). "Towards a plant model for enigmatic U-to-C RNA editing: the

- organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*." *New Phytol* 225(5): 1974-1992.
12. Gualberto, J. M., L. Lamattina, G. Bonnard, J. H. Weil and J. M. Grienemberger (1989). "RNA editing in wheat mitochondria results in the conservation of protein sequences." *Nature* 341(6243): 660-662.
 13. Hoch, B., R. M. Maier, K. Appel, G. L. Igloi and H. Kossel (1991). "Editing of a chloroplast mRNA by creation of an initiation codon." *Nature* 353(6340): 178-180.
 14. Ichinose, M. and M. Sugita (2017). "RNA Editing and Its Molecular Mechanism in Plant Organelles." *Genes* 8(1).
 15. Ichinose, M. and M. Sugita (2021). "Substitutional RNA Editing in Plant Organelles." *Methods Mol Biol* 2181: 1-12.
 16. Ishibashi, K., I. Small and T. Shikanai (2019). "Evolutionary Model of Plastidial RNA Editing in Angiosperms Presumed from Genome-Wide Analysis of *Amborella trichopoda*." *Plant Cell Physiol* 60(10): 2141-2151.
 17. Kim, D., B. Langmead and S. L. Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements." *Nat Methods* 12(4): 357-360.
 18. Kumar, S., G. Stecher, M. Li, C. Knyaz and K. Tamura (2018). "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." *Mol Biol Evol* 35(6): 1547-1549.
 19. Lenz, H., A. Hein and V. Knoop (2018). "Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0." *BMC Bioinformatics* 19(1): 255.
 20. Lenz, H. and V. Knoop (2013). "PREPACT 2.0: Predicting C-to-U and U-to-C RNA Editing in Organelle Genome Sequences with Multiple References and Curated RNA Editing Annotation." *Bioinform Biol Insights* 7: 1-19.
 21. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25(16): 2078-2079.
 22. Lo Giudice, C., I. Hernandez, L. R. Ceci, G. Pesole and E. Picardi (2019). "RNA editing in plants: A comprehensive survey of bioinformatics tools and databases." *Plant Physiol Biochem* 137: 53-61.
 23. Lo Giudice, C., G. Pesole and E. Picardi (2018). "REDIdb 3.0: A Comprehensive Collection of RNA Editing Events in Plant Organellar Genomes." *Front Plant Sci* 9: 482.
 24. Miyata, Y. and M. Sugita (2004). "Tissue- and stage-specific RNA editing of rps 14 transcripts in moss (*Physcomitrella patens*) chloroplasts." *J Plant Physiol* 161(1): 113-115.
 25. Mower, J. P. and J. D. Palmer (2006). "Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*." *Mol Genet Genomics* 276(3): 285-293.
 26. Oldenkott, B., M. Burger, A. C. Hein, A. Jorg, J. Senkler, H. P. Braun, V. Knoop, M. Takenaka and M. Schallenberg-Rudinger (2020). "One C-to-U RNA Editing Site and Two Independently Evolved Editing Factors: Testing Reciprocal Complementation with DYW-Type PPR Proteins from the Moss

- Physcomitrium (Physcomitrella) patens and the Flowering Plants Macadamia integrifolia and Arabidopsis." Plant Cell 32(9): 2997-3018.
27. Patel, R. K. and M. Jain (2012). "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data." PLoS One 7(2): e30619.
 28. Rodrigues, N. F., G. C. da Fonseca, F. R. Kulcheski and R. Margis (2017). "Salt stress affects mRNA editing in soybean chloroplasts." Genetics and Molecular Biology 40(1): 200-208.
 29. Ruchika, C. Okudaira, M. Sakari and T. Tsukahara (2021). "Genome-Wide Identification of U-To-C RNA Editing Events for Nuclear Genes in Arabidopsis thaliana." Cells 10(3).
 30. Schmitz-Linneweber, C., S. Kushnir, E. Babiyshuk, P. Poltnigg, R. G. Herrmann and R. M. Maier (2005). "Pigment deficiency in nightshade/tobacco cybrids is caused by the failure to edit the plastid ATPase alpha-subunit mRNA." Plant Cell 17(6): 1815-1828.
 31. Shikanai, T. (2015). "RNA editing in plants: Machinery and flexibility of site recognition." Biochimica Et Biophysica Acta-Bioenergetics 1847(9): 779-785.
 32. Shtratnikova, V. Y., M. I. Schelkunov, A. A. Penin and M. D. Logacheva (2020). "Mitochondrial genome of the nonphotosynthetic mycoheterotrophic plant Hypopitys monotropa, its structure, gene expression and RNA editing." PeerJ 8: e9309.
 33. Small, I. D., M. Schallenberg-Rudinger, M. Takenaka, H. Mireau and O. Ostersetzer-Biran (2020). "Plant organellar RNA editing: what 30 years of research has revealed." Plant J 101(5): 1040-1056.
 34. Smith, D. R. (2020). "Unparalleled Variation in RNA Editing among Selaginella Plastomes." Plant Physiology 182(1): 12-14.
 35. Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics 30(9): 1312-1313.
 36. Sun, Y., X. Li, D. Wu, Q. Pan, Y. Ji, H. Ren and K. Ding (2016). "RED: A Java-MySQL Software for Identifying and Visualizing RNA Editing Sites Using Rule-Based and Statistical Filters." PLoS One 11(3): e0150465.
 37. Takenaka, M., A. Zehrmann, D. Verbitskiy, F. Glass, N. Brehme and A. Brennicke (2014). "The World of Rna Editing in Mitochondria and Chloroplasts in Plants." Biocell 38: 58-59.
 38. Takenaka, M., A. Zehrmann, D. Verbitskiy, B. Hartel and A. Brennicke (2013). "RNA Editing in Plants and Its Evolution." Annual Review of Genetics, Vol 47 47: 335-352.
 39. Tang, W. and C. Luo (2018). "Molecular and Functional Diversity of RNA Editing in Plant Mitochondria." Mol Biotechnol 60(12): 935-945.
 40. Walkley, C. R. and J. B. Li (2017). "Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs." Genome Biol 18(1): 205.
 41. Wang, Z., J. Lian, Q. Li, P. Zhang, Y. Zhou, X. Zhan and G. Zhang (2016). "RES-Scanner: a software package for genome-wide identification of RNA-editing sites." Gigascience 5(1): 37.
 42. Wu, S., W. Liu, H. A. Aljohi, S. A. Alromaih, I. O. Alanazi, Q. Lin, J. Yu and S. Hu (2018). "REDO: RNA Editing Detection in Plant Organelles Based on Variant Calling Results." J Comput Biol 25(5): 509-

516.

43. Xiong, J., T. Tao, Z. Luo, S. Yan, Y. Liu, X. Yu, G. Liu, H. Xia and L. Luo (2017). "RNA Editing Responses to Oxidative Stress between a Wild Abortive Type Male-Sterile Line and Its Maintainer Line." *Front Plant Sci* 8: 2023.
44. Yagi, Y., M. Tachikawa, H. Noguchi, S. Satoh, J. Obokata and T. Nakamura (2013). "Pentatricopeptide repeat proteins involved in plant organellar RNA editing." *RNA Biol* 10(9): 1419-1425.
45. Yan, J., Q. Zhang and P. Yin (2018). "RNA editing machinery in plant organelles." *Sci China Life Sci* 61(2): 162-169.
46. Zahn, L. M. (2017). "The evolution of edited RNA transcripts." *Science* 355(6331): 1278-1279.
47. Zhang, A., X. Jiang, F. Zhang, T. Wang and X. Zhang (2020). "Dynamic response of RNA editing to temperature in grape by RNA deep sequencing." *Funct Integr Genomics* 20(3): 421-432.
48. Zhang, F., Y. Lu, S. Yan, Q. Xing and W. Tian (2017). "SPRINT: an SNP-free toolkit for identifying RNA editing sites." *Bioinformatics*.

Additional File 6

Supplementary Figure S1 is not available with this version

Figures

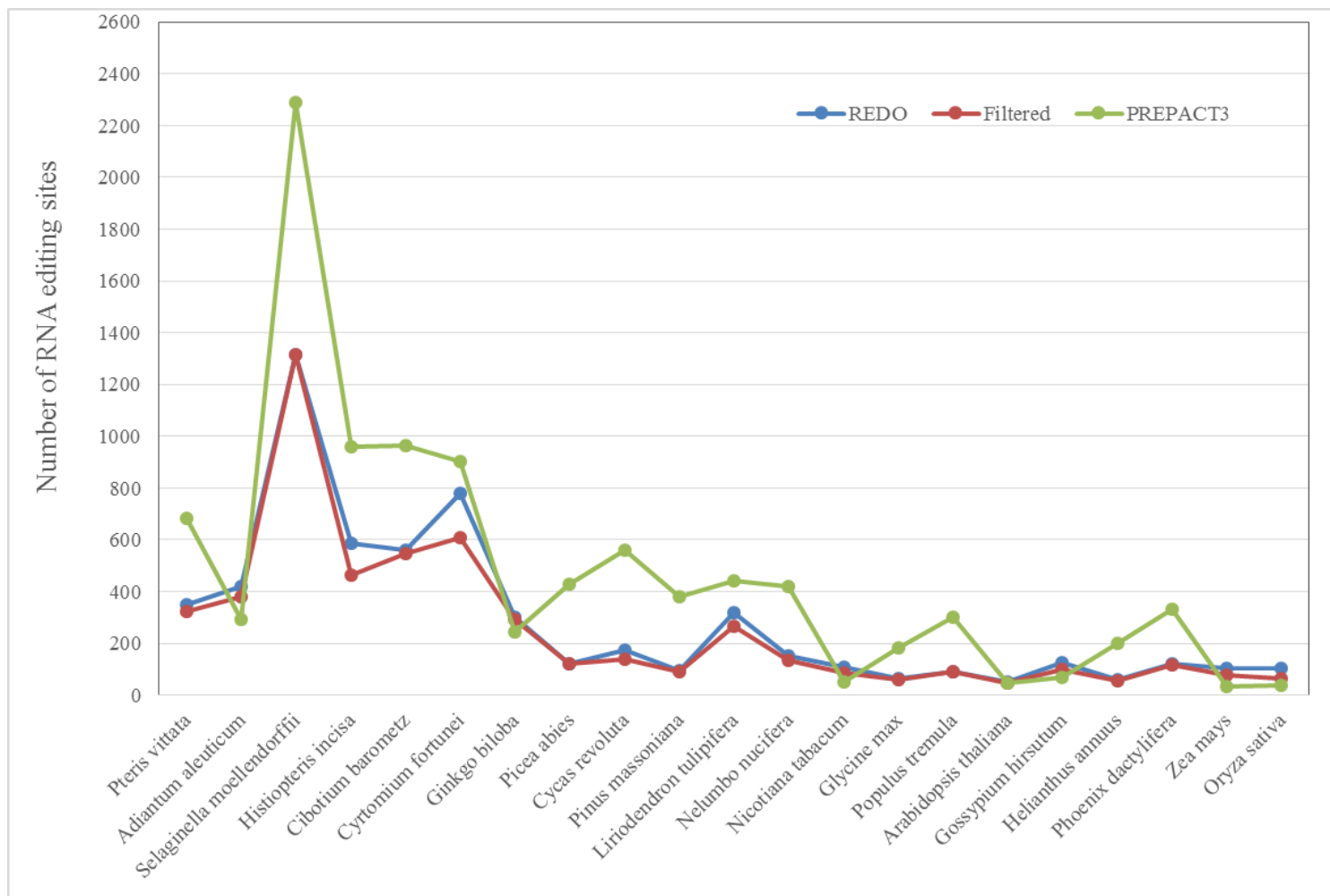


Figure 1

Comparison of predicted numbers of RNA editing sites by RNA-seq data with PREPACT website server.

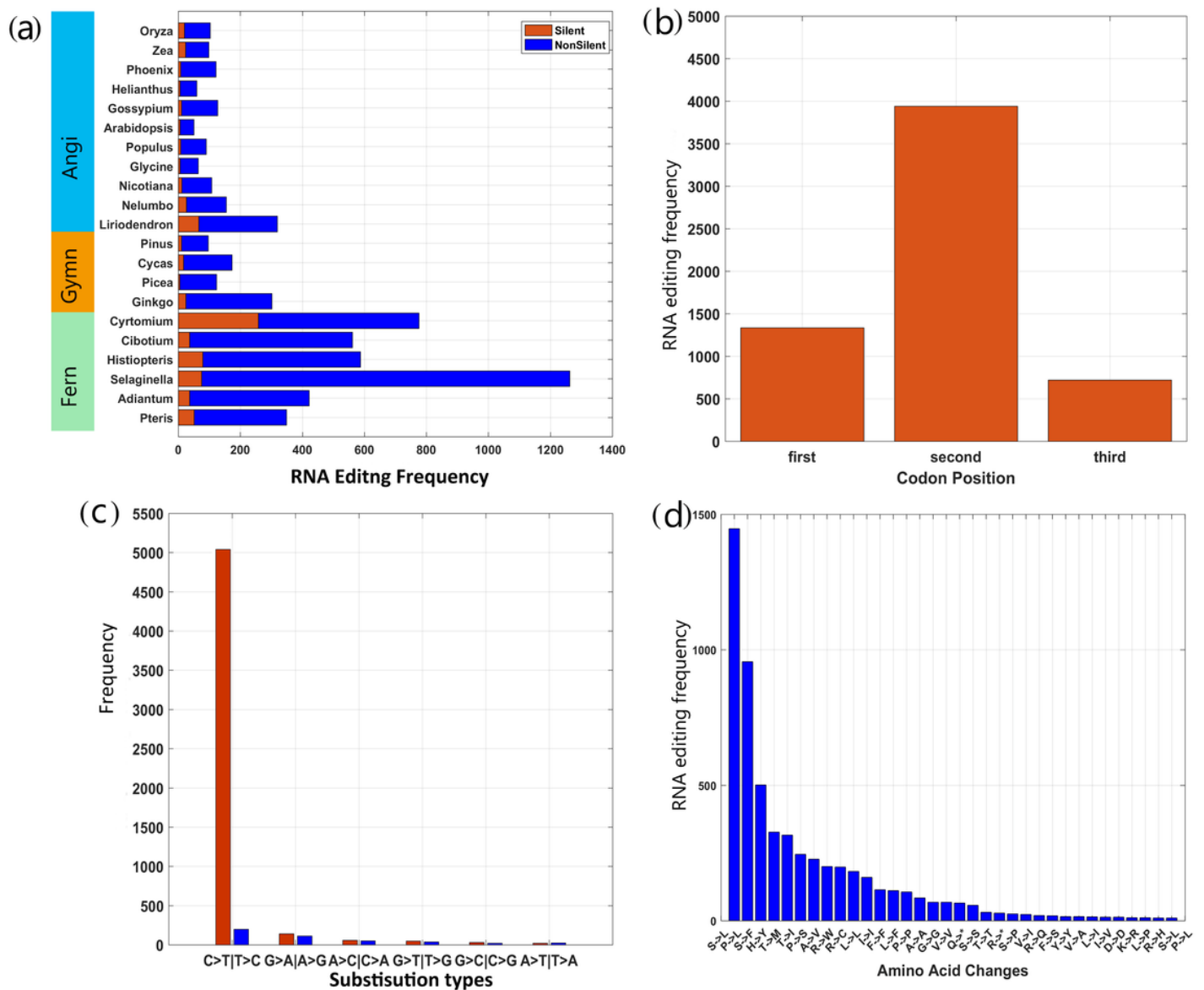


Figure 2

The statistics of identified RNA editing sites across 21 species. **(a)** The total number of editing sites for each species. Stacked bars depict numbers of nonSilent editing sites (blue), and silent editing sites (red) respectively. The symbol of each species is represented by its first word of the genus name, such as *Oryza* – *Oryza sativa*. **(b)** Codon position statistics of RNA editing sites. **(c)** Statistics of 12 nucleotide substitution types, each pair is classified by two-color bars (blue and red). **(d)** Statistics of amino acid change types.

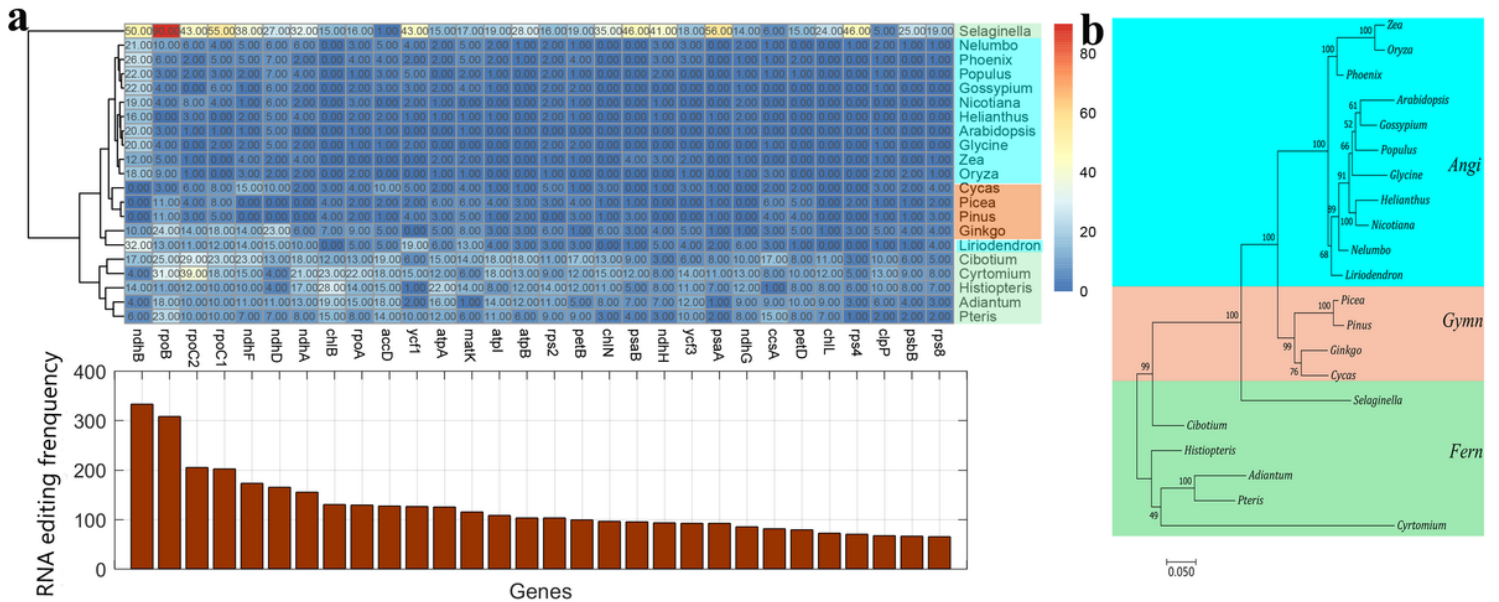


Figure 3

Distribution of numbers of RNA editing sites in top 30 genes with most editing sites across three clades. (a) A hierarchical cluster of numbers of RNA editing sites is shown above, the x-axis represents genes, the y-axis represents species, and species from different clades are highlighted by different colors. The total number of editing sites of each gene is shown below correspondingly; (b) Phylogenetic tree by Maximum Likelihood method based on alignments of merged protein sequence for top 30 genes across three clades, species from different clades are highlighted by different colors.

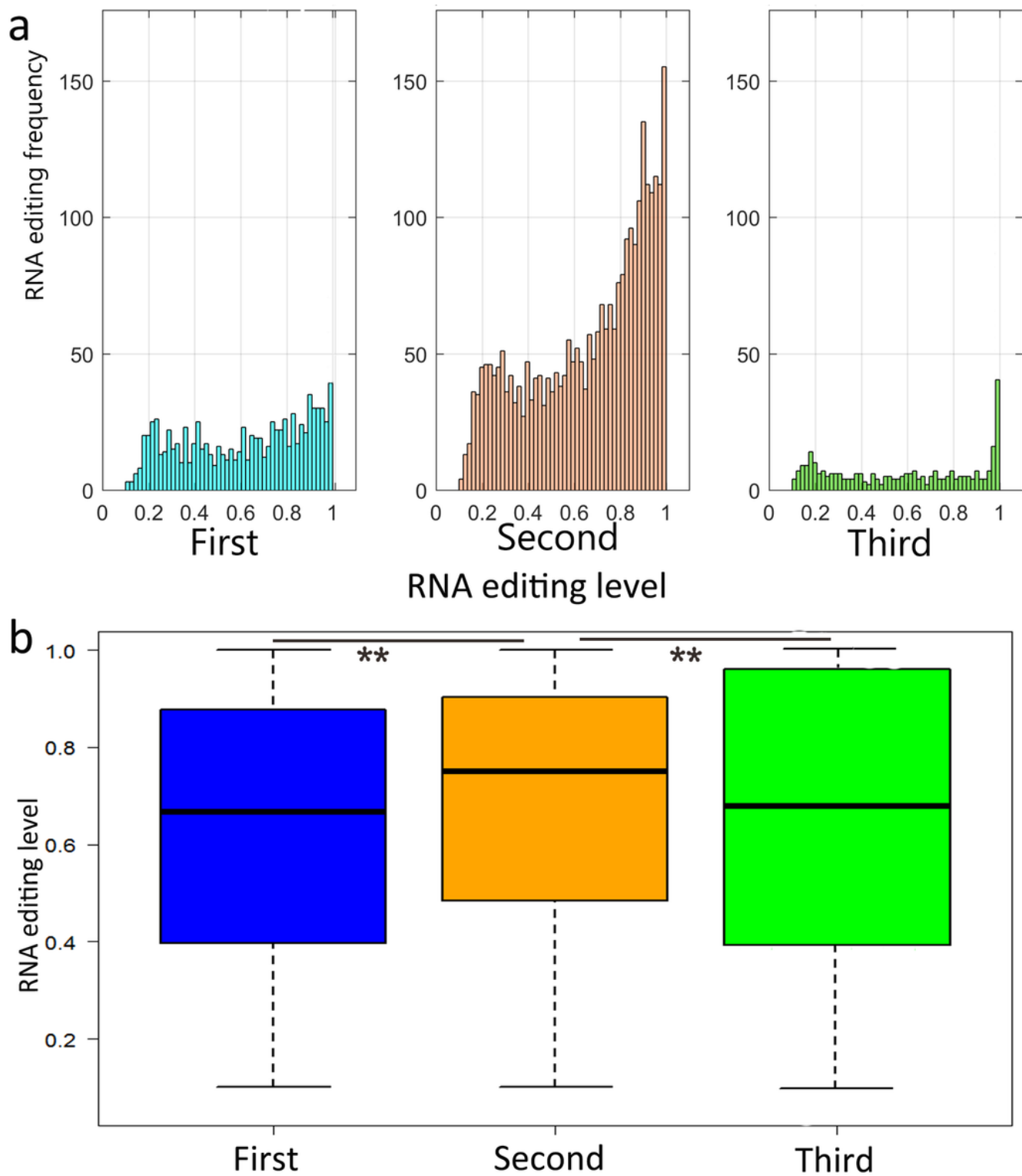


Figure 4

Distribution of RNA editing level per codon position. **(a)** Frequency distribution of RNA editing level per codon position, the x-axis represents RNA editing level, and the y-axis represents frequency; **(b)** Boxplots of RNA editing level per codon position, the x-axis represents codon position, and the y-axis represents RNA editing level. Asterisks denote significant differences: * $p\text{-value} < 0.05$; ** $p\text{-value} < 0.01$.

used. (b) Line plots of cytosine content for each shared edited gene across 21 species. The average cytosine content of all edited genes for each species is indicated by bold red lines. Three genes (*rpl23*, *psbI*, and *psbN*) are indicated by black arrows, species from different clades are highlighted by different colors.

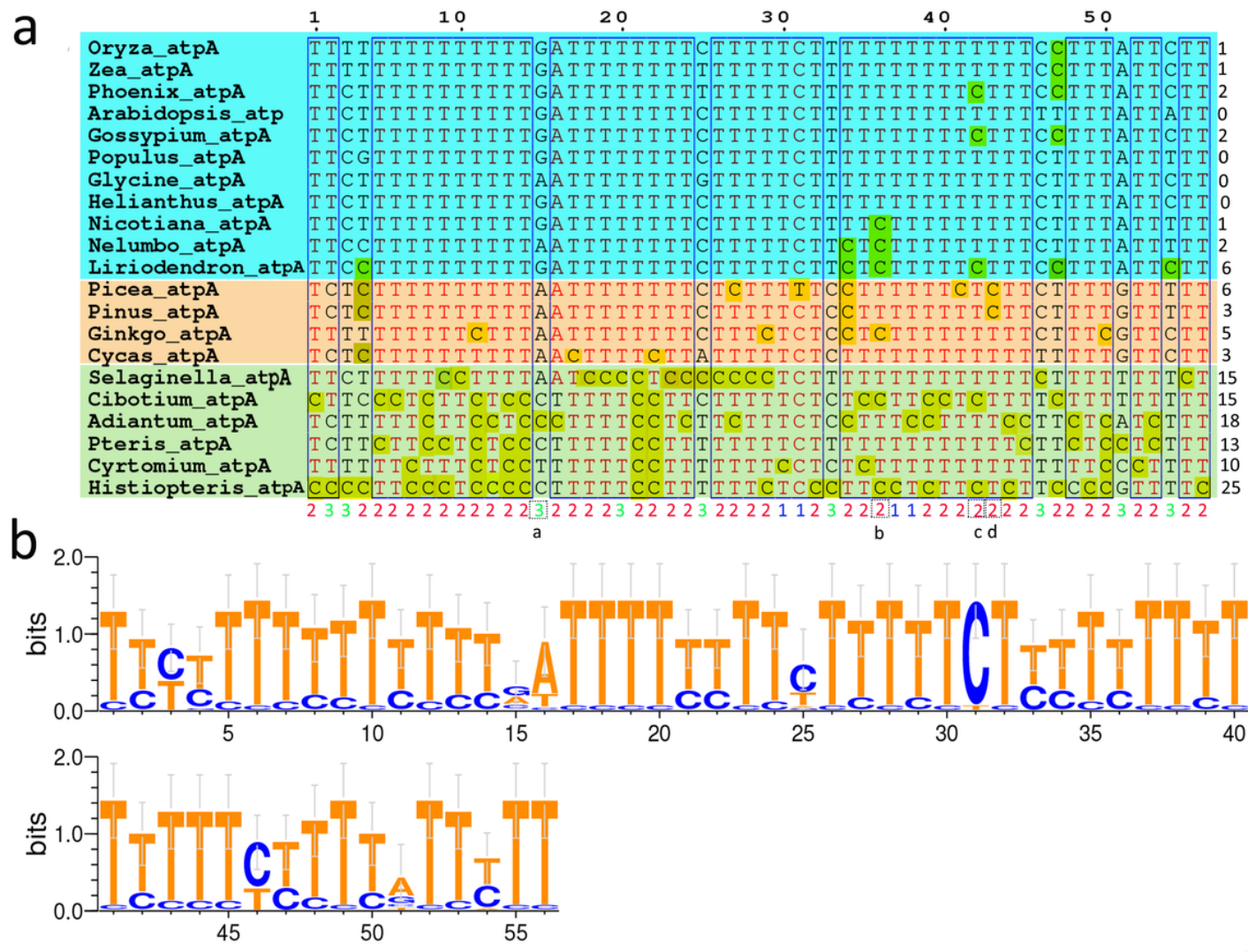


Figure 7

Illustration of RNA editing sites from *atpA* gene across 21 species. (a) Alignments of RNA editing sites of *atpA* gene, edited cytosines are marked by yellow. Codon positions (1, 2, and 3) are labeled below each column of the site by blue, red, and green colors. Locus-a, b, c, d were indicated at the bottom of codon positions. (b) Sequence logo of RNA editing sites from *atpA* gene.

Discussion

Supplementary Files

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

- [SupplementaryTableS1.xlsx](#)
- [SupplementaryTableS2.xlsx](#)
- [SupplementaryTableS3.xlsx](#)
- [SupplementaryTableS4.xlsx](#)
- [SupplementaryTableS5.xlsx](#)