A comprehensive study on chloroplast RNA editing by performing a broad-spectrum RNA-seq analysis

aidi zhang
Chinese Academy of Sciences

jing fang
Wuhan Botanical Garden

xiaohan jiang
Wuhan Botanical Garden

tengfei wang
Wuhan Botanical Garden

xiujun zhang (✉ zhangxj@wbgcas.cn)
Wuhan Botanical Garden

Research article

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

Posted Date: July 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-38570/v1

Abstract

Background

RNA editing is a post-transcriptional modification that complement variation at the DNA level. Until now, different RNA editing systems were found in the major eukaryotic lineages. However, the evolution trajectory in plant chloroplast remains unclear. To gain a better understanding of RNA editing in plant chloroplast, in this study, based on publicly available RNA-seq data across three clades (fern, gymnosperm, and angiosperm), we provided a detailed analysis of RNA editing events in plant chloroplasts and discussed the evolution of RNA editing in land plants.

Results

There were a total of 5,389 editing sites located in leaf chloroplast identified across 21 plants after rigorous screening. We found that the cluster of RNA editing sites across 21 plants complied with the phylogenetic tree based on linked protein sequences approximately, there is a common phenomenon that more editing sites occurred in ancient plants for all the three clades. Statistics results revealed that majority (~ 95%) of the editing events resulted in non-synonymous codon changes, RNA editing occurred in second codon position was mainly the largest, and RNA editing caused an overall increase in hydrophobicity of the resulting proteins. The analyses also revealed that there was an uneven distribution of editing sites among species, genes, and codon positions, the average RNA editing extent varied among different plants as well as genes, a lowest RNA editing extent (0.43) was detected in *Selaginella moellendorffii*. Finally, we found that the loss of editing sites along angiosperm evolution is mainly occurring by reduce of cytosines content, where fern plants has the highest cytosine content.

Conclusions

Many of the RNA sites identified in our study have not been previously reported and provide a valuable data set for future research community. Our findings also provide valuable information for evolution of RNA editing in plants.

Background

RNA editing is a post-transcription process through which the nucleotide specified in the genome template is modified to produce a different transcript, thus contribute to proteomic sequence variation and provide another mechanism for modulating gene expression [1, 2]. In the plant kingdom, RNA editing was first documented over a decade ago in the mitochondria of flowering plants [3, 4], and reported in angiosperm chloroplasts two years later [5], no editing seems to occur in nuclear genome-encoded transcripts. There are two types of RNA editing in plants, the most common type is C-to-U conversion, and the infrequent type is U-to-C conversion that reported only in ferns, mosses, and Lycopodiaceae. RNA editing predominantly take place at the first or second positions of codons, thereby affects the translated regions of protein coding transcripts. Occasionally, it can produce a new protein with a different function from the genome-encoded transcript. 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 [6]. RNA editing thereby is an essential process to maintain essential functions of encoded proteins at the RNA level. For example, pigment deficiency in nightshade/tobacco cybrids is caused by the failure of editing the plastid *ATPase alpha-subunit mRNA* [7].

Numerous studies have demonstrated that RNA editing plays an important role in various plant fundamental processes, such as organelle biogenesis, adaptation to environmental changes, and signal transduction [8–10]. A number of factors are involved in plant RNA editing, and considered to interact with one another to form a large protein complex, termed as editsome, PLS subfamily members of pentatrico peptide repeat (PPR) proteins function in site recognition of the target C, multiple organelle RNA editing factors (MORF) family members are components of the RNA editosome and also required for RNA editing at multiple editing sites in plants [11, 12]. 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 [13]. Despite progress in increasing editing sites identified and the mechanism underlying editing target recognition, the biological function of RNA editing in plants remains the fundamental question.

More and more recent studies demonstrated that RNA editing is a widespread phenomenon that occurred in the various land plants, including the liverworts, mosses, hornworts, lycopods, ferns and flowering plants. However, no instance of RNA editing has yet been detected in *Marchantiidae* and *algae*, suggesting that RNA editing may have evolved in organelles only after the green plants established themselves on the land [6]. Many excellent studies in different plants have recently appeared and described their mechanistic and functional aspects [14–16]. The frequency of organellar RNA editing varies from zero to thousands of sites across plant kingdom, among land plants. The early-diverging lineages show the highest numbers of editing sites compared with higher plants that undergoing an extensive loss of editing sites through the substitution of genomic editable cytidines to thymidines [17]. The evidence gathered to date suggests that RNA editing in plant organelles evolved independently from RNA processes in distant evolutionary lineages of animals, fungi, and protozoans when the first plants moved from the aquatic environment onto the land. Evolutionary studies can help to understand the puzzling nature of RNA editing in plants.

The straightforward way to detect RNA editing sites is to compare RNAs with their corresponding DNA templates. Sanger sequencing of cDNAs was widely used in the last two decades, though it is time-consuming and prone to underestimate editing sites. In recent years, the advent of next-generation sequencing technologies and the availability of large quantities of RNA sequencing data makes it possible to identify RNA editing sites and quantify their editing extent 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 plant. Indeed, the number of complete plant organellar genomes and related transcriptome data have considerably grown in the last decade, and tens of thousands of editing sites have been identified in more and more plants [18, 19]. However, this strategy is also a challenging task due to its accuracy of mapping the RNA-seq reads against genomic sequence, hence, so different bioinformatic strategies have been introduced to improve the detection of RNA editing sites [20–22].

To gain a better understanding of RNA editing in plant chloroplast, we investigated RNA editing located in chloroplast genome across diverse plants that distributed in three clades: fern, gymnosperm, and angiosperm. Through systematically exploiting enormous RNA-seq data derived from public database, thousands of RNA editing sites were detected, with a lot of these sites have not been previously reported, the data set provides valuable information for the research community. Additionally, we provided a detailed analysis of data in land plant chloroplasts and propose a model for the evolution of RNA editing in land plants.

Results

Identification of RNA editing sites

We selected a series of plants to represent distant evolutionary lineages judging by two criterions, one is enough transcriptomic data in the SRA database at NCBI, another is availability of sequenced chloroplast genome. Hence, 21 plants, consisting of 6 ferns, 4 gymnosperms and 11 angiosperms, and corresponding 317 SRA accessions were chosen finally. Detailed information of SRA data and chloroplast genome accessions were listed in Additional file 1: Table S1. Based on results of RNA-seq data mapping and SNP calling, an automated bioinformatics pipeline implemented in REDO tools was conducted under default thresholds. Consequently, there were a total of 6,011 raw editing sites located in leaf chloroplast genome were detected. However, in spite of multiple stringent filters, sequence mismatches that accord with RNA editing occasionally appear. Hence, to eliminate these false positive mismatches, we manually examined all mismatches, only C-to-U and U-to-C editing types were kept, therefore, a total of 5,389 RNA editing sites, consisting of 264 U-to-C and 5,125 C-to-U edit sites, were finally screened out for further analysis.

For the reason that RNA-seq data volume varies among different species, for several species, such as *Adiantum aleuticum*, *Histiopteris incisa*, lacked enough RNA-seq data, caused the read depth of these two species was lower than that of other species', consequently result in less number of editing sites identified in these two species. Furthermore, the read density also varied widely between different genomic regions, ranges from less than 30 to more than 800 in a few species, demonstrating various expression level of genes, the average depth distribution across 21 species was shown in Additional file 6: Figure S1. Compared with angiosperms, ferns and gymnosperms have more lower depth, indicating the actual number of RNA editing in ferns and gymnosperms might be underestimated. To evaluate the reliability of editing sites number, PREPACT webserver was also used to predict editing sites, we found that the distribution of predicted number of RNA editing based on PREPACT accorded with that of our prediction based on RNA-seq data basically (see Additional file 6: Figure S2), reflecting our pipeline offered a high performance with reliable results. To describe the attributes of RNA editing sites, we illustrated by one example of samples of *Adiantum aleuticum* (see Additional file 6: Figure S3), which depicts the reliability of RNA editing sites in each species by REDO tools statistically.

Characteristics of the statistics for RNA editing sites

The statistics of raw results showed that C-to-U was the dominate editing type (nearly ~ 95.1%), the next is U-to-C type, other mismatches types were rare (see Fig. 1c). After manual inspection of raw results and elimination of mismatches, filtered RNA editing sites across 21 plants were screened out, detailed corresponding annotation information files were produced simultaneously, as listed in Additional file 2: Table S2, the statistical result was also listed in Additional file 3: Table S3, Table 1 and Fig. 1. In terms of codon position, 21.97%, 67.16%, and 10.58% of sites were edited at first, second, and third codon positions respectively (Fig. 1b). Furthermore, the statistics of editing types showed that the majority (~ 95%) of the editing events resulted in non-synonymous codon changes, and the amino acids changes tend to be hydrophobic, the change from hydrophilic to hydrophobic was the highest, followed by the change from hydrophobic to hydrophobic. The most common amino acid change types were Ser-to-Leu and Pro-to-Leu, as shown in Fig. 1d, Serine is hydrophilic, whereas Leucine and Proline are both hydrophobic. The above results demonstrated that the RNA editing exhibited a selective advantage in overall increase in hydrophobicity of the resulting proteins, which was also in good agreement with previous studies [10].

Table 1
The statistical result of RNA editing sites across 21 plants

	Species	Genome	Raw	Filtered	Percent of edited Genes	Depth
Fern	*Pteris vittata*	MH500228	349	324	0.74	102
	Adiantum aleuticum	MH173079	422	380	0.82	23
	Selaginella moellendorffii	HM173080	1316	1316	0.96	73
	Histiopteris incisa	NC_040220	588	466	0.85	27
	Cibotium barometz	NC_037893	562	549	0.85	32
	Cyrtomium fortunei	NC_037510	778	610	0.91	40
Gymn	*Ginkgo biloba*	NC_016986	301	291	0.81	228
	Picea abies	HF937082	123	121	0.61	450
	Cycas revoluta	NC_020319	173	140	0.44	27
	Pinus massoniana	MF564195	96	93	0.59	182
Angi	*Liriodendron tulipifera*	NC_008326	320	265	0.75	113
	Nelumbo nucifera	NC_025339	154	134	0.63	295
	Nicotiana tabacum	Z00044	107	88	0.23	134
	Glycine max	NC_007942	64	59	0.34	195
	Populus tremula	KP861984	93	89	0.43	244
	Arabidopsis thaliana	KX551970	50	45	0.24	389
	Gossypium hirsutum	NC_007944	127	100	0.37	171
	Helianthus annuus	NC_007977	59	57	0.29	148
	Phoenix dactylifera	NC_013991	121	118	0.42	781
	Zea mays	NC_001666	105	79	0.37	814
	Oryza sativa	NC_001320	103	65	0.26	550

All the filtered editing sites were located in 1,038 genes across all species, the average percent of edited genes in the three lineages (ferns, gymnosperms and angiosperms) are 0.85, 0.61, and 0.39 respectively (Table 1). Compared to the latter two lineages, more sites and genes were edited in chloroplast transcripts from ferns remarkably, see Fig. 1a. In ferns, the largest number of editing sites was found in *Selaginella* *moellendorffii*, with 1,316 editing sites, exclusively of the C-to-U type, which is nearly 100-fold more abundant than that of flowering plants, a total of 77 genes (~ 96%) were edited, and the overwhelming majority (94.3%) of the 1,241 silent editing events altered codon directly. Differ from *Selaginella moellendorffii* that exclusively belonging to taxa of *lycopsida*, the other five fern plants are members of *Leptosporangiopsida*, has relatively smaller numbers of editing sites, represented by *Cyrtomium fortune*, owning second-largest number of editing sites among all specifies, with 610 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 other three species of the same taxa, *Ginkgo biloba* has the most editing sites, with 291 editing sites and 68 edited genes (~ 81%) in chloroplast, reflecting its ancient nature of "living fossil". On the opposite end, angiosperms have the lowest average numbers of editing sites and only a part of genes were effectively edited. It was noticeable that *Liriodendron tulipifera* and *Nelumbo nucifera* distinguished them from other angiosperms with more editing events. *Liriodendron tulipifera* is one of the most ancient flowering trees as the genus dates back about 65 million years, a total of 265 editing sites were detected, which is nearly 3-fold more abundant than that of other angiosperms and gymnosperms except *Ginkgo biloba*, the percent of edited genes was up to 75%, well above the average, reflecting early angiosperms possess much more diversified editing sites. The average number of editing sites among other 9 angiosperms showed no significant differences among them, but they were significantly lower than those of *Liriodendron tulipifera* and *Nelumbo nucifera.* The above results illustrated the differential distribution of RNA editing among varied plants, the cases of *Selaginella moellendorffii* and *Liriodendron tulipifera* also implies independent origins and subsequent evolutionary trajectories of editing processes.

Variability of RNA editing among plants and genes

To further explore the evolutionary trajectory of RNA editing among different species, for each gene, we summed the number of the editing sites across 21 plants (see Additional file 4: Table S4), and picked out the top 30 genes with most editing sites across 21 plants for cluster analysis, as shown in Fig. 2. We found that not every gene were edited in all the species, for certain species, the lack of editing at a few genes may be explained by two reasons, one is absence of the genes that annotated in chloroplast genome, another is no RNA editing occurred in the genes actually. By grouping the genes based on their function, genes encoding membrane subunits of the chloroplast NDH complex and RNA polymerase exhibited the largest average numbers of editing sites, while ribosomal subunits showed the lowest numbers, this is consistent with previous studies that RNA editing occurred preferentially in genes encoding membrane-bound proteins under strong selection [24]. Due to the well-studied background and abundant editing sites in plant, *ndhB* gene is assumed to be a good case for the study of RNA editing evolution, in our study, *ndhB* was also confirmed to possess the most editing sites based transcriptome data, with 333 editing sites spread across 17 species. In spite of this, there is a biased distribution of RNA editing sites in *ndhB* among three clades, in fern group, 50 sites were detected in *Selaginella moellendorffii*, and about 10 sites in the other ferns, in angiosperms, there were about 20 editing sites in each of the 11 angiosperm plants. 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.

To compare the number of editing sites among the top 30 genes, a matrix of numbers of editing sites across 21 species was produced, a hierarchically-clustered heatmap was plotted in Fig. 2, which showed that the clustering relationships agreed with their phylogenetic tree based on sequence alignments roughly. 21 plants were divided into three clustering groups with two exceptions, *Selaginella moellendorffii* and *Liriodendron tulipifera* were clustered far away from their own clades respectively, further implied independent origins and subsequent evolutionary trajectories of editing processes. The above results suggest that RNA editing in chloroplast may break out in early-branching plants from different clades simultaneously and suffer a lot of loss during evolution. Hierarchically clustered heatmaps of numbers of all RNA editing genes in chloroplast across 21 plants and each clade were shown in Additional file 6: Figure S4-7 respectively.

Uneven distribution of RNA editing extent

RNA editing extent was used to measure to what extent the edited transcripts among all transcriptome for one gene, if one site was edited, the C/G base (wild type) should be changed to the T/A base (edited type), thus its editing extent can be calculated by the formula: depth of edited bases (T and A)/total read depth of bases. In this study, we explored the distribution of editing extent among codon positions, species and edited genes. First, the comparison between codon positions showed that the distribution of RNA editing extent among them was uneven, and did not comply with the normal distribution, featuring an peak around ~ 0.2 and fat tails, as shown in Fig. 3. The average editing extent in second codon position (~ 0.78) is higher than that of first (~ 0.69) and third codon positions (~ 0.66), suggesting non-synonymous substitution occurred in second codon position tend to be effectively edited, it was higher editing extent that dominated the landscape of RNA editing. Second, The average editing extent also varied widely across 21 plants, ranging from 0.43 to 0.87 (see Fig. 4a), *Selaginella moellendorffii* has the lowest editing extent (~ 0.43), 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 extent. However, as an ancient plant in angiosperms, *Liriodendron tulipifera was* one exception, with editing extent up to 0.81. The editing extent was also analyzed in each gene individually, we averaged the RNA editing extent among the top 30 genes across 21 plants and the result also demonstrated an uneven distribution, as shown in Fig. 4b, *matK* gene has the lowest editing extent (~ 0.5), oppositely, editing extent 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 the evolution of plants, we analyzed the nucleic acid base content of involving genes shared by the 21 plants. There were a total of 51 genes annotated in all the chloroplast genome of 21 plants, for each plant, its corresponding gene sequences were extracted, the percent of cytosines for each gene was calculated by the formula: number of cytosines (C)/total number of bases (A/T/C/G), as listed in Additional file 5: Table S5. Afterwards, comparison between each two clades were conducted, the cytosines content of each gene were averaged across all the members of each clade, and two-tailed Wilcoxon rank-sum test was used to perform pairwise comparisons. The statistical result showed that a remarkable significance ($p < 0.05$) was detected in the comparison of cytosines content between each two clades except for gymnosperms-angiosperms, the percent of cytosines of ferns was far below than that of angiosperms, followed by gymnosperms, as shown in Fig. 5a. The percent of cytosines of each gene across the 21 plants were further plotted in Fig. 5b, which also demonstrated that the percent of cytosines dramatically declined roughly along with the evolution, with a few exceptions (such as *rpl23* gene). One striking example was *psbI* gene, which has the highest percent of cytosines in *Selaginella moellendorffii* (~ 0.37), and dropped to about 0.18 in other species. While the highest average of 51 shared genes was found in *Selaginella moellendorffii (~ 0.26)*, the smallest average was found in *Glycine max* (~ 0.17). Furthermore, *Ginkgo biloba* and *Liriodendron tulipifera* have higher average in angiosperms, correspond to ~ 0.192 and ~ 0.188 respectively, showed a positive correlation with their high numbers of RNA editing sites. The above results indicated the number of editing sites declined dramatically with the evolution of plants, which maybe due to loss of cytosine content in chloroplast gene for later-branch plants.

We illustrated one RNA editing example of *atpA* gene that may help to understand the evolution trajectory across plants vividly. The gene sequences of *atpA* across 21 species were collected, intersection of all the species' RNA editing sites of *atpA* was concatenated for alignment and annotated in Fig. 6. We marked all editing sites identified in *atpA* gene by yellow color, the distribution of its editing sites demonstrated that numbers of editing sites as well as cytosine content declined from ferns to angiosperms. We found that editing sites at third codon positions were poorly conserved, for example, in Fig. 6a, RNA editing in site a was only occurred in *Adiantum aleuticum* in spite of existence of cytosine in other fern members, indicating that synonymous substitution at third codon positions were not active in re-establishing functional proteins. However, sites b and c are both located at second codon position, in contrast to site b that occurred in all three clades, RNA editing in site c occurred in several members of ferns and angiosperms, was absent in gymnosperms where the cytosine have already corrected to thymine in the genome. Whereas in site 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. The above results demonstrated the diversity of RNA editing evolution, further validated that RNA editing in plant might evolve independently in distant evolutionary lineages, and in certain higher plants, new editing sites may occur occasionally.

Discussion

As a post-transcription process, RNA editing can modify the genome template to produce a different transcript [6]. For plants, significant progress in RNA editing has been made in recent years, numerous studies have proved that RNA editing occurred in nearly all plants in the kingdom, and demonstrated that RNA editing played roles not only in abiotic stress tolerance but also likely in the plant development, such as flower development and male sterile [9, 10, 14, 19], RNA editing may also cause secondary structure transformation of transcripts [25]. Until now, there are two viewpoints about the nature of RNA editing, one is contribution to variations in proteomic sequence, and play roles in modulating gene expression; another point thinks that RNA editing in plants might be a repair mechanism to correct genomic point mutations at post-transcription level, thus increase the substitutional rate that is extremely low in organellar genome [26, 27].

With prior knowledge, to detect editing sites, RT-PCR primers should be designed to flank a region that contains the sites of interest, the editing sites can be determined by comparing PCR products with the genomic DNA sequence. However, this traditional approach is a time-consuming procedure requiring larger experiments to perform, and prone to underestimate number of editing sites and overestimate editing extent, because for site with editing extent less than 10%, to find one edited transcript, more than 10 clones have to be sequenced when comparing its cDNA with genomic sequences. With the advent of sequencing technology and bioinformatics, more and more RNA-seq data was generated, combined with the sequencing data and bioinformatics tools, much progress has been made in identifying all the potential RNA editing sites and quantifying their editing extent especially for the sites with very low extent. Hence, RNA editing sites were identified in more and more organisms based on RNA deep sequencing [14, 19, 28].

In this study, to gain a better understanding of RNA editing in plant chloroplast, we collected a large mount of RNA-seq data and performed a series of bioinformatics procedures to investigate RNA editing in 21 diverse plants that distributed in three clades. A total of 5,389 editing sites located in leaf chloroplast genes across 21 plants were identified and quantified their editing extent, demonstrating the powerfulness of bioinformatics approach, many of the identified sites have not been previously reported, thus provided a valuable data resource for future research. The statistics results revealed RNA editing sites that occurred in second codon position was mainly the largest, and majority (~ 95%) of the editing events resulted in non-synonymous codon changes, additionally, editing significantly increased the hydrophobic amino acid with a selective advantage. We found that the cluster relationship of numbers of RNA editing sites complied with the phylogenetic tree based on gene sequences approximately, further verified that the RNA editing across plant kingdom are comparatively conservative and accord with laws of evolution roughly. An uneven distribution of editing sites among species, genes, codon positions were found for numbers of editing sites as well as the average RNA editing extent. In total, numbers of editing sites declined with the evolution of plants, editing events occur more often in ancient plant than higher plant, such as *Ginkgo biloba*, *Liriodendron tulipifera*, which both owned highest number of editing sites in each clade. Compared with other species, fern plant *Selaginella moellendorffii* was identified to own the highest number of sites and the lowest editing extent (~ 0.43). We also found that a reasonable percentage of editing sites occurred in certain clades, and lost in other clades whose cytosine already corrected to thymine in the site of genome, RNA-editing activities affecting third-codon positions showed a higher evolutionary variability. The decrease of cytosine content in chloroplast gene for later-branch plants might explain the reason for declined number of editing sites with the evolution of plants.

Previous studies revealed that organelle genomes have a more slower evolutionary rate than nuclear genome, thus accumulated a number of T-to-C mutation that constitutes a prerequisite for generation of plant organellar RNA editing [29], 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, but some sites still need to be edited to thymine or remained to be cytosine for coding for different amino acids. During the evolution of higher plants, genome mutations of C-to-T eliminate the need for editing at certain sites, thus the number of editing sites showed a remarkable reduction, since substitution of genome are more constantly than that of mRNA level. Hence, the genome mutations is actually 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 accurately to describe the evolution trajectory instead of loss of RNA editing sites, The uneven distribution of editing sites among species, genes, codon positions implied their independent origins across three clades, suggesting that in early-branching plants of different clades, chloroplast RNA editing may break out simultaneously and suffer a lot of loss during evolution.

Conclusion

To illuminate the evolution mechanism on a genome-wide scale, we performed a systematic characterization and comparison of RNA editing events across three major clades of plants. Based on a large amount of RNA-seq data, a relaxed automated approach combined with manual inspection that eliminated false positives was used, and finally screened out thousands of RNA editing sites, demonstrating the advantages of bioinformatics approach in detection of RNA editing sites. Many of the identified sites have not been previously reported so far and provide a valuable data set for future research. The distribution of editing sites showed an heterogeneity characteristics among species, genes, and codon positions. The average RNA editing extent also varied among different plants as well as genes. The genome-wide distributions of chloroplast RNA editing across three clades suggest that plants have undergone drastic changes in both the numbers and patterns of editing. Ancient plants have more editing sites and lower editing extent, while constant genome mutations occurring by replacing cytosines with thymidines might be the underlying force of loss of editing sites in higher plants. Our comparative study provided valuable information for evolution of RNA editing in plants. However, further mechanistic studies are still needed to characterize the RNA-binding proteins that involved in site recognition and editing across different clades.

Methods

Data collection

All data sets used in this study are publicly available. We selected 21 plants across three clades (fern, gymnosperm, and angiosperm) for analysis of chloroplast RNA editing events. For each plant, the corresponding transcriptome data was downloaded from SRA database at NCBI based on two criteria: first, to increase reliability of editing sites, SRA accessions with paired-end reads that possess higher mapping specificity were preferred, second, for the reason that more chloroplast mRNA in leaf were extracted and sequenced compared with other tissues, RNA-seq data obtained from leaves of wild type individuals were only selected. Besides, for each plant, the reference file consisting of chloroplast genome sequences and corresponding gene annotation file in ‘tbl’ format were also downloaded from the GenBank database. Detailed information of SRA data and reference files used in our study were listed in Additional file 1: Table S1.

Read mapping and SNP calling

In total, 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 plant, in order to increase sequencing depth, we merged all the SRA accessions from the same species into one sample. The quality control of paired-end Illumina sequencing data were evaluated first by NGSQCToolkit, low quality sequence data were filtered out (cutOffQualScore < 20) [30], then transcriptome data from each plant was aligned against its chloroplast reference by hisat2 software under default parameters [31]. Afterwards, the alignment results were sorted, removed duplicates, indexed, and sorted by using SAMtools [32]. Finally, the bcftools tool was used to identify SNPs, and VCF files that describe transcriptome variation were generated [33].

Detection of RNA editing sites

The principles of RNA editing detection is similar to that of transcriptome variation in SAMtools. Thus, we used the preliminary output from variant calling software to identify RNA editing sites. For each plant, based on the SNP-calling results (in “VCF” format) and genome annotation files (in “tbl” format), the RNA editing sites were identified under default parameter values by using the REDO tool [21]. 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 only works requiring 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 in “tbl” format, www.ncbi.nlm.nih.gov/projects/Sequin/table.html). Regarding the high false positive of editing sites, REDO uses a series of stringent criterias to filter the raw variants, as the below following: (1) quality control filter (MQ > 255), the low quality sites are filtered out according to the reads quality, (2) total reads depth filter (DP > 1), (3) alt proportion filter (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 (< 2) 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.

To minimize the number of false negatives with the automated approach, for the produced raw RNA editing sites, we manually examined all mismatches to eliminate false positives, and only kept the C-to-U and U-to-C editing types and excluded mismatches with other editing types, such as A-to-C, T-to-A, etc. In addition, to evaluate the reliability of editing sites number, for each plant, we also used PREPACT tools to predict potential RNA editing events supplying with entire chloroplast genomes as input files, with filter threshold at least 80% of the references under BLASTX mode, PREPACT originally relied on BLASTX hits in manually assembled collections of reference protein sequences [34].

Characteristic statistics

All the filtered RNA editing sites detected in 21 plants were used for further statistics and feature analysis, including statistics of editing number, editing type, codon position, amino acid changes, involved genes and so on. In order to decipher the distribution of RNA editing frequency across different species, the top 30 genes with most editing sites across 21 plants were selected, cluster analysis and heatmap plotting were also provided based on the matrix of RNA editing numbers of top 30 genes across 21 plant. The CDS sequences of top 30 genes across 21 plants were concatenated and subjected to alignments and phylogenetic tree construction using MEGA [35]. Meanwhile, the RNA editing extent of top 30 genes were also subjected to statistical analysis. In terms of RNA editing extent, its value 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 type (C/G) or edited type (T/A) of bases could then be counted at this particular site, then the editing extent at one site could then be calculated by the formula: depth of edited bases (T and A)/total read depth of bases. Values of editing extent 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 value of standard deviation. For each lineage (fern, gymnosperm, and angiosperm), a heatmap was plotted across all of its species respectively using “pheatmap” function in R, the distance matrix of different samples was calculated using “dist” function with the default Euclidean method, and the hierarchical clustering was computed using “hclust” function.

Considering protein coding genes varied among different plant, we picked out shared edited genes for the statistics of cytosine content across 21 plant. For each shared edited gene, we extracted its CDS sequence, and calculated the ratio of cytosine content, and further performed pairwise comparisons between any two of lineages (fern, gymn, and angi). Two-tailed Wilcoxon rank-sum test was used. We illustrated *atpA* gene as an example, sequence logo of *atpA* gene was produced by WebLogo [36], alignment was constructed by using MEGA under default parameters [35].

Abbreviations

C-to-U
Cytosine-to-uracil; PPR:Pentatrico peptide repeat; ndhB:NADH dehydrogenase subunit 2; MORF:multiple organelle RNA editing factors

Declarations

Acknowledgments

We thank Yuepeng Han (Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture) for providing invaluable assistance in design of this study.

Funding

This work was supported by grants from the CAS Pioneer Hundred Talents Program, the National Natural Sciences Foundation of China (No. 61402457, No. 31702322), and the National Science and Technology Innovation Zone Project (No. 1716315XJ00200303, 1816315XJ00100216).

Availability of supporting data

Supporting data are included as Additional files.

Authors’ contributions

ADZ, JF,XHJ, FPZ and XJZ conceived and designed the experiments, ADZ, JF performed data analysis, ADZ, and JF wrote the manuscript. XHJ, TFW and XJZ provided many critical suggestions. All authors reviewed the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Zahn LM. The evolution of edited RNA transcripts. Science. 2017;355(6331):1278–9.
2. Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. Genome Biol. 2017;18(1):205.
3. Covello PS, Gray MW. Rna Editing in Plant-Mitochondria. Nature. 1989;341(6243):662–6.
4. Gualberto JM, et al. RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature. 1989;341(6243):660–2.
5. Hoch B, et al. Editing of a chloroplast mRNA by creation of an initiation codon. Nature. 1991;353(6340):178–80.
6. Ichinose M, Sugita M. RNA Editing and Its Molecular Mechanism in Plant Organelles. Genes, 2017. 8(1).
7. Schmitz-Linneweber C, et al. Pigment deficiency in nightshade/tobacco cybrids is caused by the failure to edit the plastid ATPase alpha-subunit mRNA. Plant Cell. 2005;17(6):1815–28.
8. Xiong J, et al., RNA Editing Responses to Oxidative Stress between a Wild Abortive Type Male-Sterile Line and Its Maintainer Line. Front Plant Sci, 2017. 8: p. 2023.
9. Miyata Y, Sugita M. Tissue- and stage-specific RNA editing of rps 14 transcripts in moss (Physcomitrella patens) chloroplasts. J Plant Physiol. 2004;161(1):113–5.
10. Zhang A, et al. Dynamic response of RNA editing to temperature in grape by RNA deep sequencing. Funct Integr Genomics. 2020;20(3):421–32.
11. Yan J, Zhang Q, Yin P. RNA editing machinery in plant organelles. Sci China Life Sci. 2018;61(2):162–9.
12. Yagi Y, et al. Pentatricopeptide repeat proteins involved in plant organellar RNA editing. RNA Biol. 2013;10(9):1419–25.
13. Shikanai T. RNA editing in plants: Machinery and flexibility of site recognition. Biochimica Et Biophysica Acta-Bioenergetics. 2015;1847(9):779–85.
14. Brenner WG, et al., High Level of Conservation of Mitochondrial RNA Editing Sites Among Four Populus Species. G3-Genes Genomes Genetics, 2019. 9(3): p. 709–717.
15. Wang WQ, et al., RNA Editing in Chloroplasts of Spirodela polyrhiza, an Aquatic Monocotelydonous Species. Plos One, 2015. 10(10).
16. Hein A, Polsakiewicz M, Knoop V. Frequent chloroplast RNA editing in early-branching flowering plants: pilot studies on angiosperm-wide coexistence of editing sites and their nuclear specificity factors. Bmc Evolutionary Biology, 2016. 16.
17. Takenaka M, et al. RNA Editing in Plants and Its Evolution. Annu Rev Genet. 2013;47:335–52. 47.
18. Lo Giudice C, Pesole G, Picardi E. REDIdb 3.0: A Comprehensive Collection of RNA Editing Events in Plant Organellar Genomes. Front Plant Sci. 2018;9:482.

19. Lo Giudice C, et al. RNA editing in plants: A comprehensive survey of bioinformatics tools and databases. Plant Physiol Biochem. 2019;137:53–61.

20. Sun Y, et al. RED: A Java-MySQL Software for Identifying and Visualizing RNA Editing Sites Using Rule-Based and Statistical Filters. PLoS One. 2016;11(3):e0150465.

21. Zhang F, et al., SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics, 2017.

22. Wang Z, et al. RES-Scanner: a software package for genome-wide identification of RNA-editing sites. Gigascience. 2016;5(1):37.

23. Oldenkott B, et al. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte Selaginella uncinata. Rna-a Publication of the Rna Society. 2014;20(10):1499–506.

24. Mower JP, Palmer JD. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Mol Genet Genomics. 2006;276(3):285–93.

25. Farre JC, et al., RNA Editing in Mitochondrial Trans-Introns Is Required for Splicing. Plos One, 2012. 7(12).

26. Tang W, Luo C. Molecular and Functional Diversity of RNA Editing in Plant Mitochondria. Mol Biotechnol. 2018;60(12):935–45.

27. Takenaka M, et al., The World of Rna Editing in Mitochondria and Chloroplasts in Plants. Biocell, 2014. 38: p. 58–59.

28. The chloroplast and. mitochondrial C-to-U RNA editing in Arabidopsis thaliana shows signals of adaptation. 2019.

29. Barbrook AC, et al. Organization and expression of organellar genomes. Philos Trans R Soc Lond B Biol Sci. 2010;365(1541):785–97.

30. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7(2):e30619.

31. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.

32. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

33. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33(13):2037–9.

34. Lenz H, Knoop V. 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. 2013;7:1–19.

35. Kumar S, et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547–9.

36. Crooks GE, et al. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.

Figures

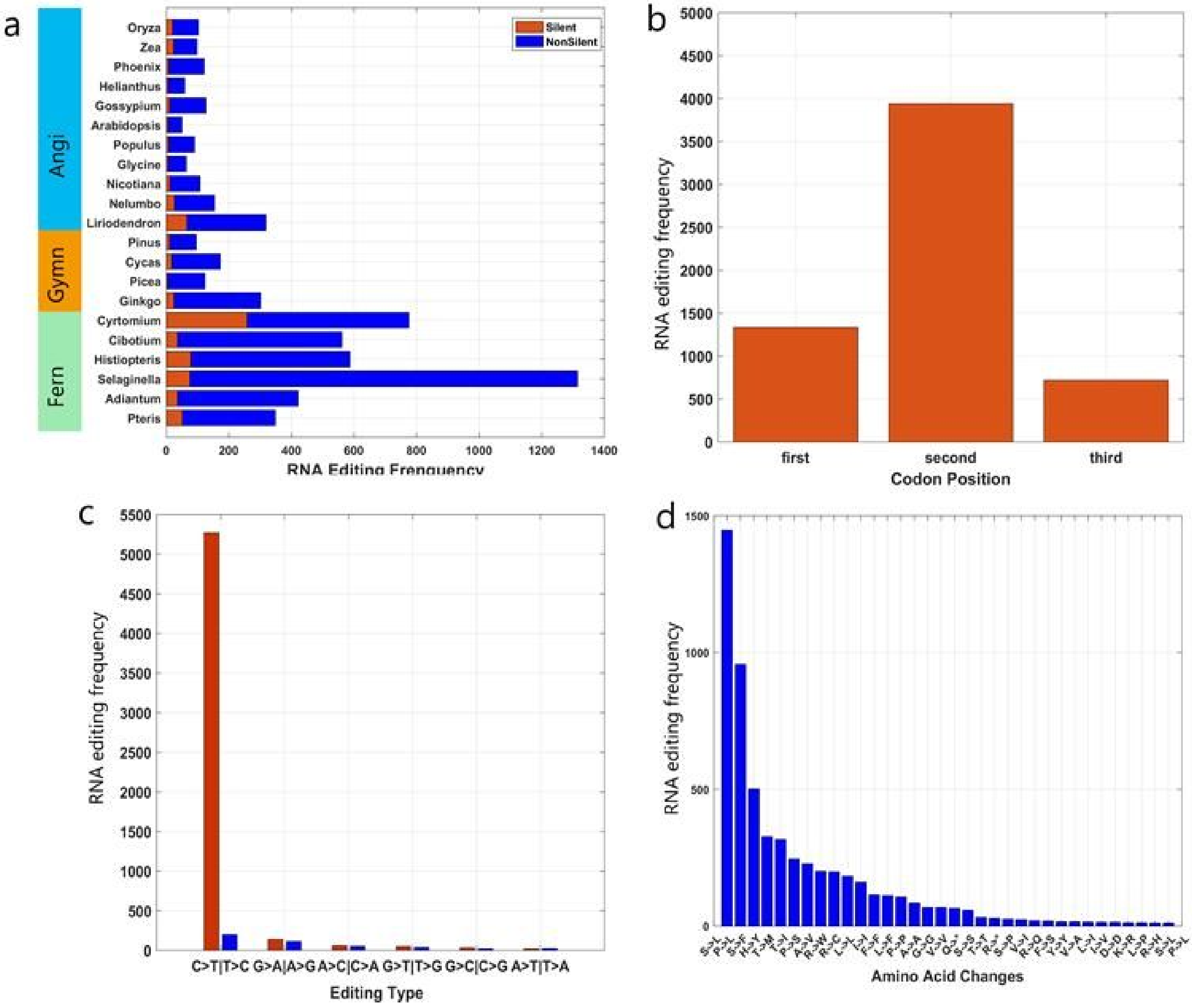

Figure 1

The statistics of identified RNA editing sites in chloroplast across 21 plants. (a) Total number of editing sites in all protein-coding genes across 17 angiosperms. Stacked bars depict numbers of nonSilent editing sites (blue), and silent editing sites (red) respectively. To simplify our presentation, the symbol of each species was represented by its first word of scientific name, such as Oryza – Oryza sativa. (b) Codon position statistics of RNA editing sites. (c) Statistics of 12 RNA editing types, each pair was classified by two color bars (blue and red). (d) Statistics of amino acid change types.

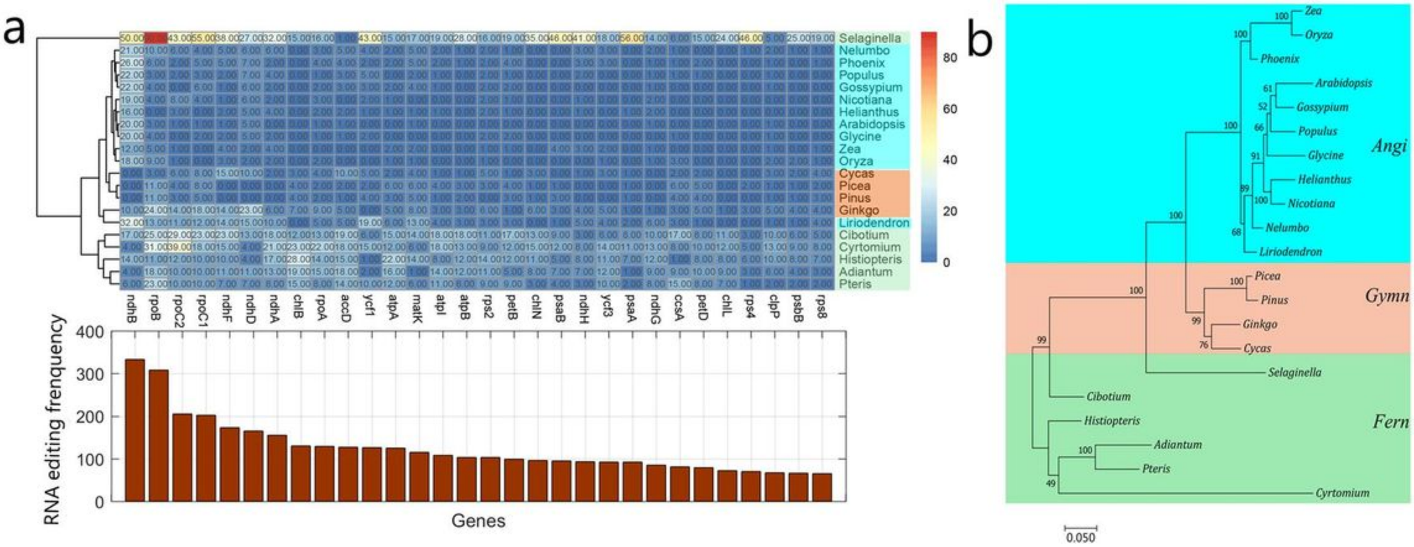

Figure 2

Analysis of numbers of RNA editing sites in top 30 genes with most editing sites across 21 plants. (a) Hierarchical cluster of numbers of RNA editing sites in top 30 genes was shown above, the x axis represents different genes, and the y axis represents plant, the total number of editing sites in each gene was shown below correspondly. (b) Phylogenetic tree by Maximum Likelihood method based on alignments of merged protein sequence for top 30 genes across 21 plants.

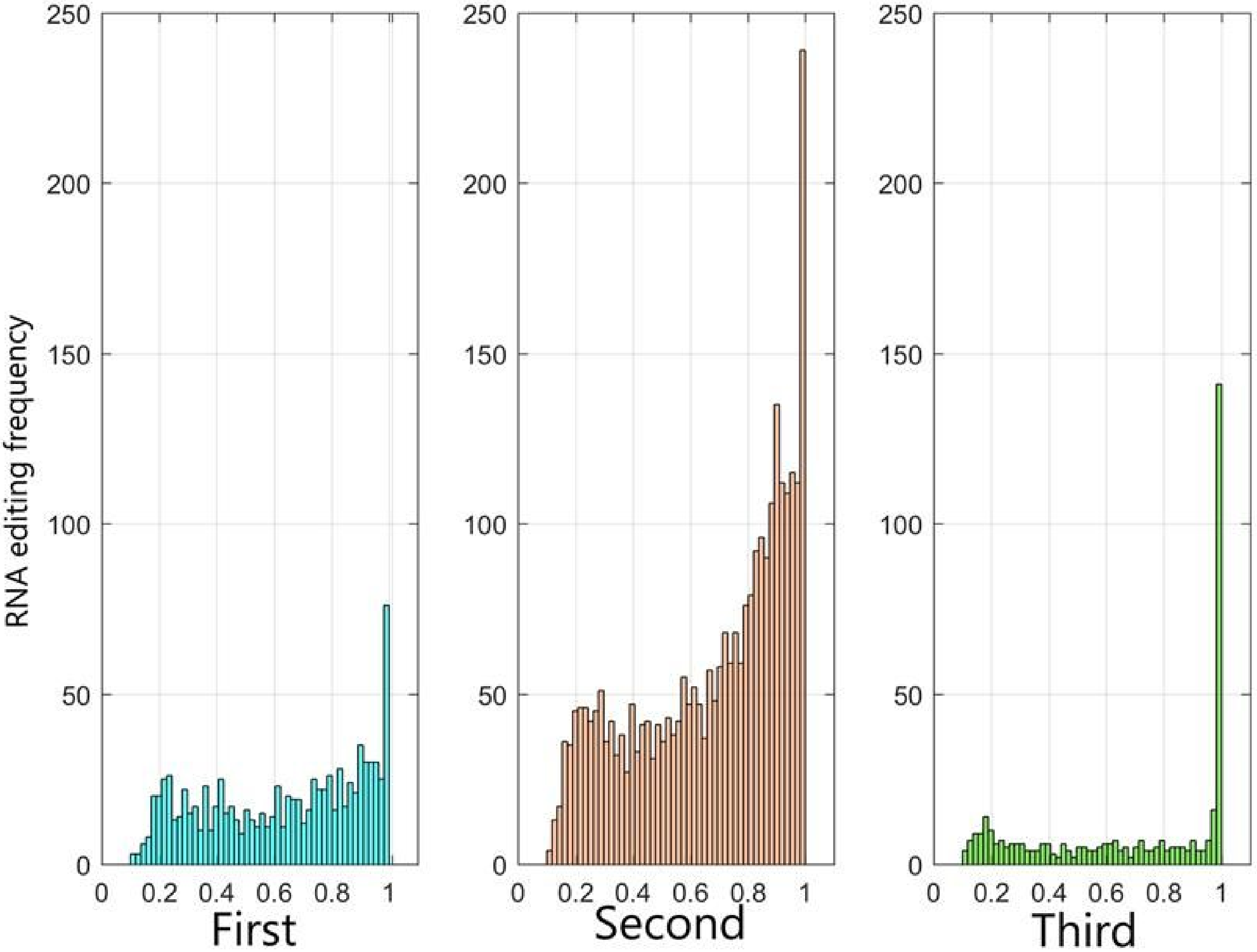

Figure 3

Editing extent of each identified editing site per codon position. The x axis represents editing extent, and the y axis represents frequency.

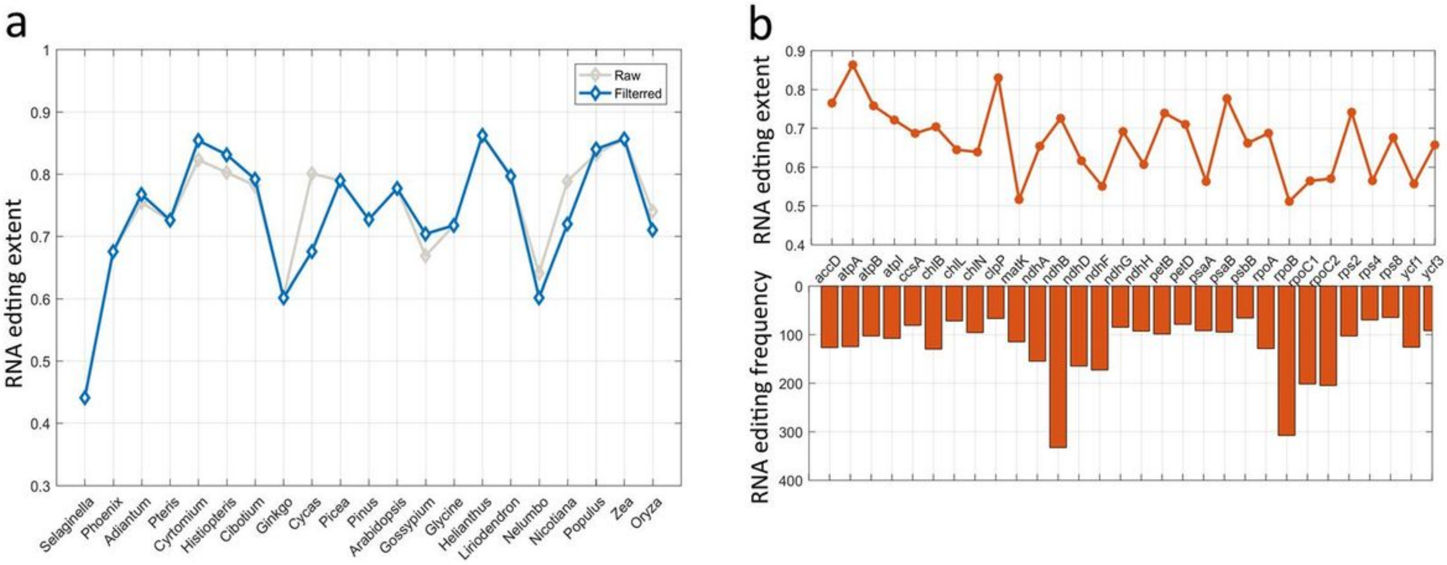

Figure 4

The statistics of identified RNA editing sites in chloroplast across 21 plants. (a) Distribution of average editing extent of all identified RNA editing sites in each plant specie. Grey and dark blue plots depict average editing extent for raw, filtered sites, respectively. (b) Distribution of average editing extent of top 30 genes. The above shows average editing extent of top 30 genes, the below shows numbers of RNA editing sites in top 30 genes correspondly.

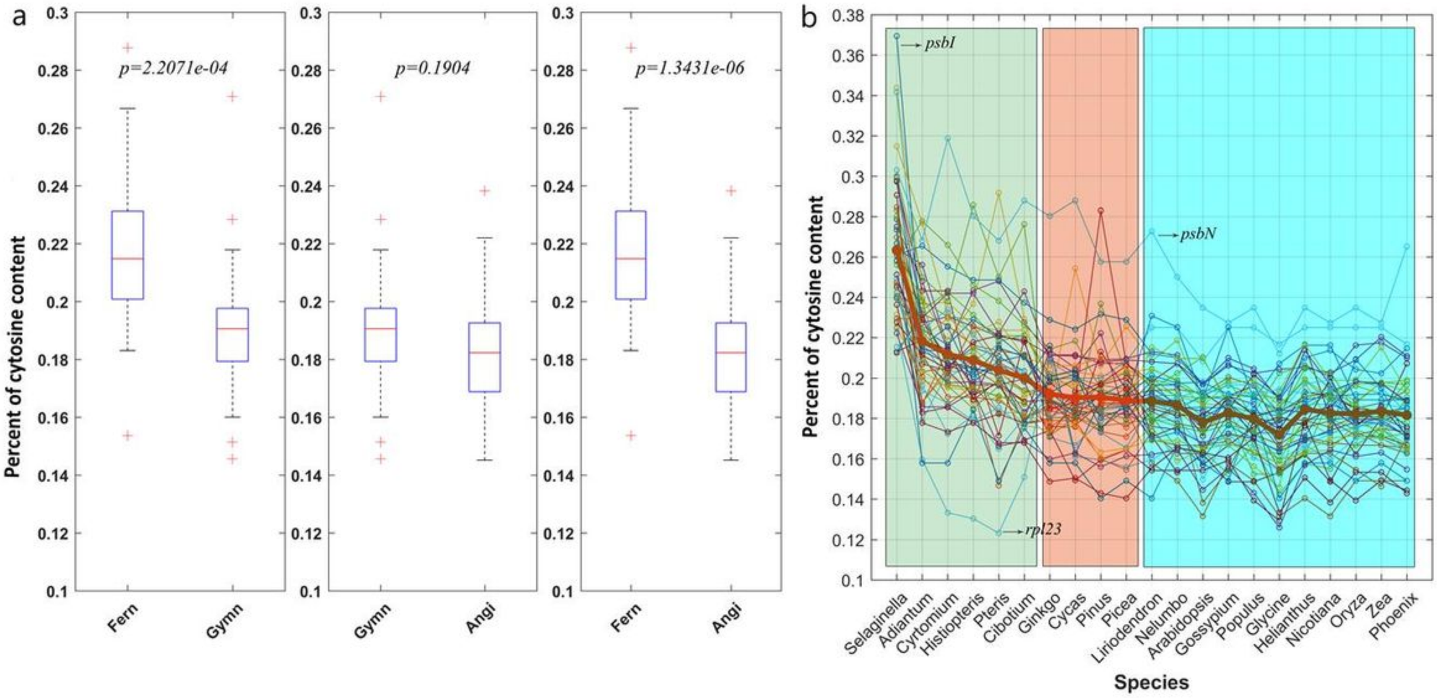

Figure 5

The statistics of cytosine content in shared RNA edited genes across 21 plants. (a) Bar plots of pairwise comparisons between each two of clades (fern, gymn, and angi). Two-tailed Wilcoxon rank-sum test was used to perform the pairwise comparison. (b) Line plots of cytosine content for each shared RNA edited genes across 21 plants. Average cytosine content of RNA edited genes for each specie is indicated by bold red lines. Two genes (psbI and psbN) are indicated by black arrows.

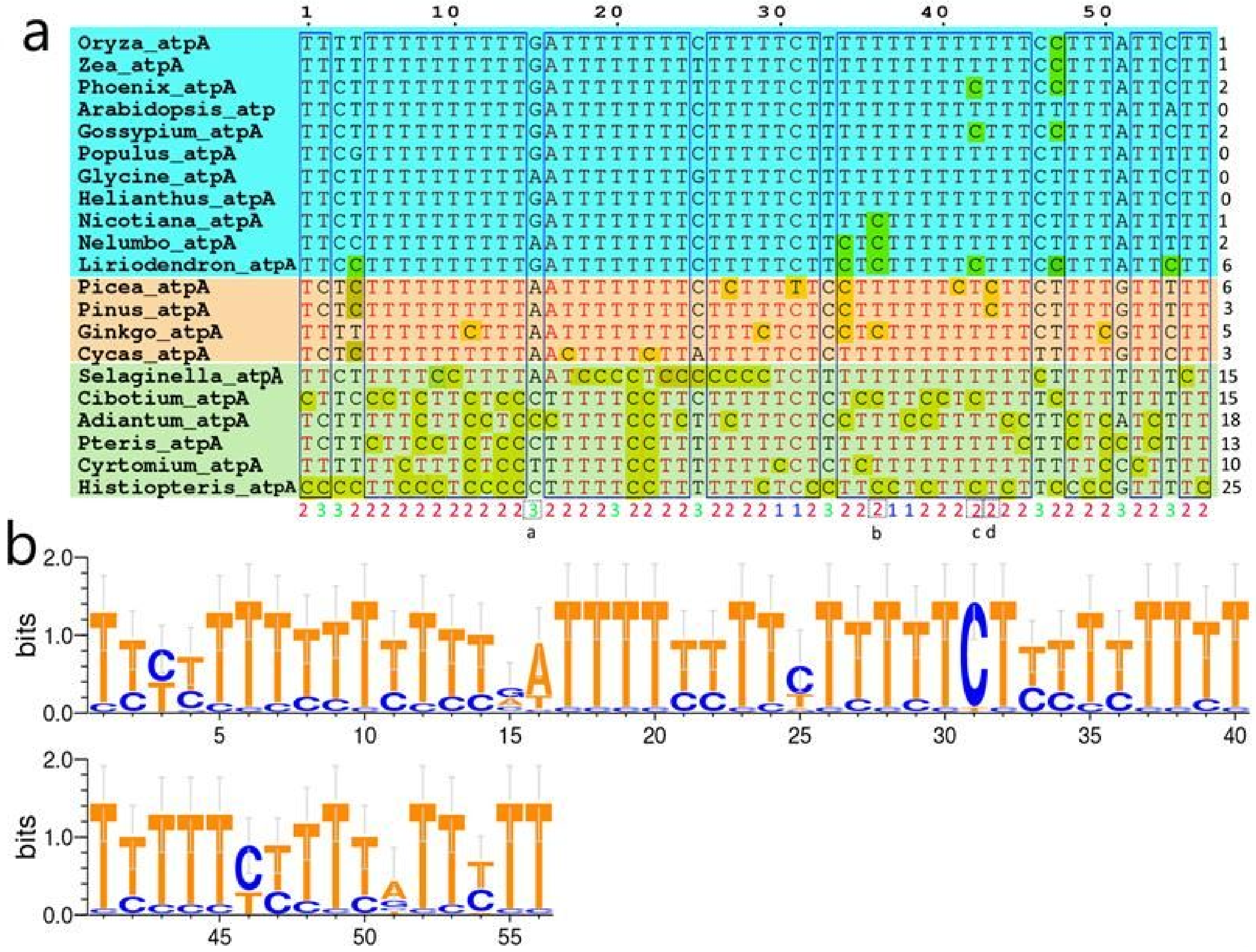

Figure 6

The statistics of cytosine content illustrated by RNA editing gene atpA. (a) Alignments of RNA editing sites extracted from initial RNA sequences in each specie, edited cytosines are marked by yellow. Codon positions are labeled by blue, red and green number under each column of site. (b) Sequence logo for RNA editing sites of atpA gene.

Supplementary Files

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

- supplementary.docx

- SupplementaryTableS5.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS1.xlsx