Plant Molecular Biology (2018) 97:215–231
https://doi.org/10.1007/s11103-018-0734-9

Towards a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria

Alejandro A. Edera[1] · Carolina L. Gandini[1] · M. Virginia Sanchez-Puerta[1,2]

Received: 16 November 2017 / Accepted: 2 May 2018 / Published online: 14 May 2018
© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract

Key message **Our understanding of the dynamic and evolution of RNA editing in angiosperms is in part limited by the few editing sites identified to date. This study identified 10,217 editing sites from 17 diverse angiosperms. Our analyses confirmed the universality of certain features of RNA editing, and offer new evidence behind the loss of editing sites in angiosperms.**

Abstract RNA editing is a post-transcriptional process that substitutes cytidines (C) for uridines (U) in organellar transcripts of angiosperms. These substitutions mostly take place in mitochondrial messenger RNAs at specific positions called editing sites. By means of publicly available RNA-seq data, this study identified 10,217 editing sites in mitochondrial protein-coding genes of 17 diverse angiosperms. Even though other types of mismatches were also identified, we did not find evidence of non-canonical editing processes. The results showed an uneven distribution of editing sites among species, genes, and codon positions. The analyses revealed that editing sites were conserved across angiosperms but there were some species-specific sites. Non-synonymous editing sites were particularly highly conserved (~ 80%) across the plant species and were efficiently edited (80% editing extent). In contrast, editing sites at third codon positions were poorly conserved (~ 30%) and only partially edited (~ 40% editing extent). We found that the loss of editing sites along angiosperm evolution is mainly occurring by replacing editing sites with thymidines, instead of a degradation of the editing recognition motif around editing sites. Consecutive and highly conserved editing sites had been replaced by thymidines as result of retroprocessing, by which edited transcripts are reverse transcribed to cDNA and then integrated into the genome by homologous recombination. This phenomenon was more pronounced in eudicots, and in the gene *cox1*. These results suggest that retroprocessing is a widespread driving force underlying the loss of editing sites in angiosperm mitochondria.

Keywords RNA editing · Mitochondrial genome · Angiosperm · Retroprocessing

Introduction

RNA editing is a post-transcriptional event that alters the content of plant organellar RNAs. The specific RNA positions affected by RNA editing, along with their corresponding DNA positions, are usually known as editing sites. In angiosperms, this post-transcriptonal event substitutes cytidines (C) for uridines (U) mainly in mitochondrial transcripts, in which ca. 400 editing sites have been reported in *Arabidopsis* (Bentolila et al. 2013; Giegé and Brennicke 1999), in comparison to the ca. 30 editing sites found in chloroplasts (Chateigner-Boutin and Small 2007; Lutz and Maliga 2001; Fujii and Small 2011; Tillich et al. 2006). Most editing sites are located in protein-coding genes but also in non-coding regions, tRNAs, introns, or UTRs (Castandet et al. 2010; Grimes et al. 2014).

Editing sites are surrounded by a sequence motif involved in the binding of the editosome (Cummings and Myers 2004; Farré et al. 2001; Giegé and Brennicke 1999; Mulligan et al. 2007; Neuwirt et al. 2005). The editosome is composed of distinct nuclear-encoded proteins with variable specificity

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11103-018-0734-9) contains supplementary material, which is available to authorized users.

✉ Alejandro A. Edera
aedera@mendoza-conicet.gob.ar

[1] IBAM, Facultad de Ciencias Agrarias, CONICET, Universidad Nacional de Cuyo, M5528AHB Chacras de Coria, Argentina

[2] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo, 5500 Mendoza, Argentina

in editing site recognition (Bentolila et al. 2013; Takenaka et al. 2012). The PLS-class proteins (Lurin et al. 2004) belong to the family of the pentatricopeptide repeat (PPR) proteins (Aubourg et al. 2000; Small and Peeters 2000), and are involved in the recognition of specific editing sites (Barkan et al. 2012; Takenaka et al. 2013; Yagi et al. 2013b). In protein-coding genes, editing sites are predominantly found at non-synonymous positions: more frequently at second followed by first codon positions (Cuenca et al. 2010; Giegé and Brennicke 1999; Mulligan et al. 2007; Sloan et al. 2010; Picardi et al. 2010; Yura and Go 2008). RNA editing can change the genomically encoded amino acids (Covello and Gray 1989; Gualberto et al. 1989; Hiesel et al. 1989; Mower 2009). Such changes tend to increase overall amino acid conservation, modify physicochemical properties (Giegé and Brennicke 1999; Jobson and Qiu 2008; Tillich et al. 2006; Kugita et al. 2003) and even alter protein folding (Yura and Go 2008; Yura et al. 2009), which means that editing sites are essential for the proper function of proteins.

The frequency of editing sites is uneven among land plants. With no editing sites reported in green algae, more than 2000 sites have been identified in lycophytes (Hecht et al. 2011), about 500 sites in gymnosperms (Salmans et al. 2010), and between 200 and 700 sites in angiosperms (Mower 2009; Mower and Palmer 2006; Sloan et al. 2010; Richardson et al. 2013). Among angiosperms, early-diverging lineages show the highest numbers of editing sites (Rice et al. 2013; Richardson et al. 2013). In contrast to other land plants, angiosperms are undergoing an extensive loss of editing sites through the substitution of genomic editable cytidines to thymidines (Cuenca et al. 2010; Mower 2008; Shields and Wolfe 1997; Sloan et al. 2010). Retroprocessing, i.e. gene conversion with a cDNA generated by reverse transcription of an edited RNA seems to be one of the molecular mechanisms responsible for editing site loss (Bowe and dePamphilis 1996; Cuenca et al. 2010, 2016; Hecht et al. 2011; Mulligan et al. 2007; Parkinson et al. 2005; Sloan et al. 2010). However, a key role of point mutations has also been proposed for the loss of editing sites favored by natural selection (Mower 2008).

To identify editing sites, C-to-U mismatches are detected by comparing RNAs with their corresponding DNA templates. For this purpose, Sanger sequencing of cDNAs has been widely used in the last two decades (Giegé and Brennicke 1999; Gualberto et al. 1989; Takenaka and Brennicke 2003, 2007), though it is time-consuming and prone to underestimate editing sites (Mower and Palmer 2006). In recent years, next-generation sequencing of transcriptomes (RNA-seq) has been used as an alternative approach. One of the main advantages of RNA-seq based identification is that it allows a fast and large-scale recognition of editing sites over whole genomes (Cuenca et al. 2016; Fang et al. 2012; Grewe et al. 2014; Grimes et al. 2014; Islam et al. 2013; Kazakoff et al. 2012; Park et al. 2014; Picardi et al. 2010; Sahraeian et al. 2017; Shearman et al. 2014; Wu et al. 2015). In this approach, the mapping of RNA-seq reads to genomic sequences is crucial to identify editing sites and their editing levels accurately (Bahn et al. 2012; Diroma et al. 2017). This is a challenging task because stringent mapping settings may lead to false negatives (Guo et al. 2015), while more relaxed settings may increase the number of false positives. As a result, different strategies have been discussed to improve the detection of RNA editing sites (Stone and Štorchová 2015; Guo et al. 2015; Diroma et al. 2017). Nevertheless, the detection of RNA editing events through RNA-seq data has enormous potential to deepen our knowledge of the transcriptional process in plant mitochondria.

To gain a better understanding of RNA editing in angiosperm mitochondria, this study identified the editing sites of diverse angiosperm mitochondria by systematically exploiting publicly available RNA-seq data. Through an in-house computational pipeline, 10,217 C-to-U editing events were identified in all protein-coding genes of 17 mitochondrial genomes belonging to five angiosperm lineages: magnoliids, monocots, basal eudicots, rosids, and asterids. Many of these sites have not been previously reported and represent a valuable data set for the research community. We aim to: (i) analyze the distribution of editing sites across the main angiosperm lineages, and a wide range of mitochondrial genes; (ii) study the amino acid changes produced by RNA editing; (iii) examine the conservation of editing sites across angiosperm lineages; (iv) compare the editing extent at synonymous and non-synonymous editing sites; (v) evaluate the role of retroprocessing in the loss of editing sites; and (vi) study the dynamics of editing site loss and gain along the evolution of angiosperm mitochondria.

Results

Identification of C-to-U editing sites

To identify editing sites, a sample of diverse angiosperm mitochondria was analyzed. The selection of angiosperms was based on the complete mitochondrial genomes publicly available in the GenBank organellar genome database[1] and on the availability of transcriptomic data in the SRA (Sequence Read Archive) database[2] at NCBI. From those lineages with genomic and transcriptomic data available, we selected 17 diverse angiosperm species [3] (Fig. 1) so that: (i) there was sufficient transcriptomic data; (ii) they did not

[1] https://www.ncbi.nlm.nih.gov/genome/organelle.

[2] https://www.ncbi.nlm.nih.gov/sra.

[3] Data harvested January 2018.

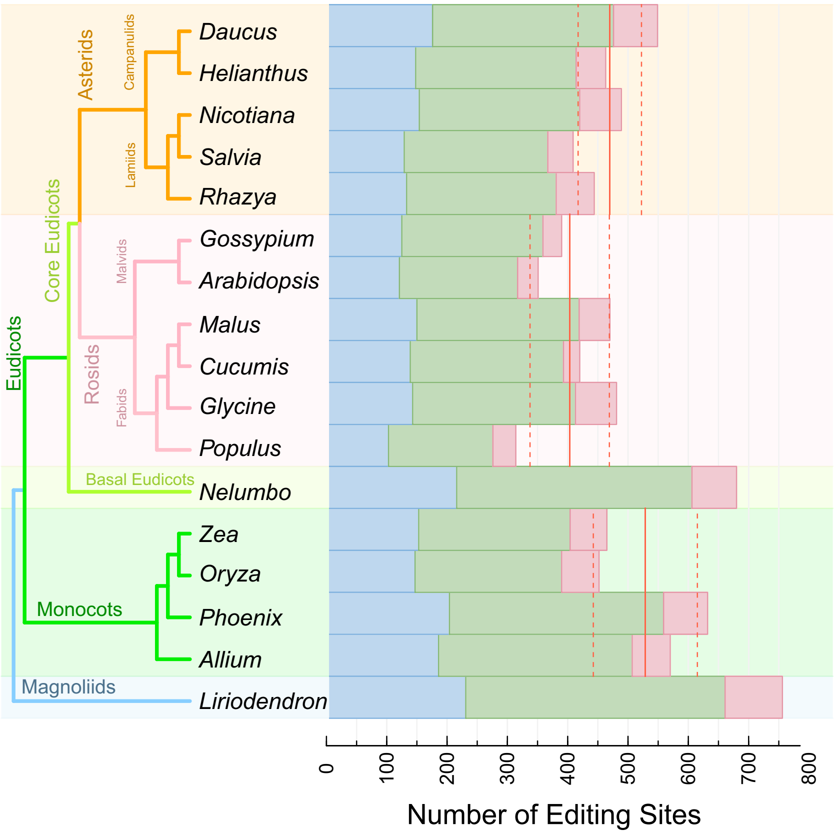


Fig. 1 Total number of editing sites in 24 protein-coding genes across 17 angiosperms. Stacked bars depict numbers of editing sites at first (blue), second (green), and third (red) codon positions, respectively. Solid and dashed red vertical lines represent average and standard deviation values, respectively, calculated for the angiosperm lineages asterids, rosids, and monocots

have odd mutation rates, such as *Silene* spp. (Sloan et al. 2010); and (iii) they were evenly distributed across the main angiosperm lineages. By using an in-house computational pipeline ("Identification of RNA editing sites from transcriptomic data" section), 10,217 C-to-U mismatches, or editing sites, were identified by comparing transcriptomic data with mitochondrial protein-coding sequences (ESM 1). For the comparative analyses described below, we used the 8318 editing sites found in the 24 protein-coding genes conserved across the 17 angiosperm mitochondria (ESM 2). These editing sites were annotated as uppercase C nucleotides in the DNA sequence alignments of each protein-coding gene (ESM 3).

Given that for some of the studied angiosperms RNA editing sites had been previously identified by Sanger sequencing of RT-PCR products, we were able to make comparisons with our identification method based on RNA-seq data (ESM4). The editing sites we detected in *Arabidopsis, Liriodendron, Nicotiana*, and *Oryza* were contrasted to those obtained previously by Sanger sequencing of cDNAs (Giegé and Brennicke 1999; Richardson et al. 2013; Sugiyama et al. 2005; Notsu et al. 2002). On average, there was an agreement of about 82, 82, 83, and 80% in *Arabidopsis, Liriodendron, Nicotiana*, and *Oryza*, respectively (ESM 4). The majority of the disagreements represented novel editing sites (i.e., sites identified in this study), a common result observed in related studies (Grimes et al. 2014; Picardi et al. 2010). On average, about 45% of the disagreements were at third codon positions in the four species, where editing sites showed poor conservation across plant species (Bentolila et al. 2013; Guo et al. 2017; Mower and Palmer 2006; Picardi et al. 2010; Stone et al. 2017; Wu et al. 2015). The disagreements caused by the sites identified by RT-PCR but not by RNA-seq showed an average editing extent between 0 and 7.5%, indicating that they were overlooked by our identification method owing to their low editing extents, i.e. < 10%, or low read-depth, i.e. < 10 reads ("Identification of RNA editing sites from transcriptomic data" section). The 50–70% of the novel editing sites identified by RNA-seq were supported by conservation levels about 50% on average, across the 17 angiosperm mitochondrial genomes, or predictions made by PREP-Mt (Mower 2009) under a cutoff of 0.2 (ESM 4).

Identification of other types of mismatches

Following the same procedure to identify the 8318 C-to-U mismatches, a total of 599 unconventional mismatches were also recognized in the 17 × 24 protein-coding genes. Overall, C-to-U mismatches were the most frequent, representing 93% of all the mismatches found, followed by G-to-A (1.2%) and G-to-U (0.8%) (ESM5A). Almost 50 unconventional mismatches were located in exon–intron boundaries in the genomic sequences of several intron-bearing genes, and were supported by > 85% of the reads and by at least two SRA experiments. After detailed analyses, we concluded that they were likely the result of errors in the annotation of splicing sites. Other mismatches supported by 100% of the reads were likely due to polymorphisms between the individuals used for the genomic and transcriptomic analyses (ESM5B). For example, 98 of the 153 (64%) unconventional mismatches found after mapping RNA-seq data of *Liriodendron chinense* on the mitochondrial genome of *L. tulipifera* were supported by 100% of reads in at least one SRA experiment (ESM5B). The remaining mismatches could be the result of: (i) dual transcription of nuclear-encoded mitochondrial genes (or pseudogenes) and mitochondrial-encoded copies (Adams et al. 1999; Daley et al. 2002; Picardi et al. 2010); (ii) RNA sequencing errors (in particular, those mismatches found only in reads from one SRA experiment and in sites containing more than one type of mismatch) (Wu et al. 2015); or (iii) non-canonical editing processes. Indeed, a large number of mismatches (50 in total) were found in the protein-coding region of *Glycine cox2* due to the documented simultaneous expression of the nuclear and mitochondrial copies of the gene *cox2* in *Glycine* (Adams et al. 1999; Daley et al. 2002). It is possible that other nuclear-encoded copies are also being transcribed simultaneously with the mitochondrial-encoded homologs. Recently, other examples of dual transcription of homologs in the nuclear and mitochondrial genomes have been reported: *rps13* in Silene vulgaris (Sloan et al. 2012), *sdh4* in Salicaceae and *Lupinus* (Choi et al. 2006; Havird et al. 2016), *rpl5* and *rps19* in Poaceae (Atluri et al. 2015; Wu et al. 2017). We identified a few candidates for dual transcription (*Nicotiana atp1* and *matR, Daucus nad3*) based on the fact that those genes showed mismatches at several positions and were supported by more than two SRA experiments, as observed for *Glycine cox2*. Sequence similarity searches of DNA-seq data revealed regions showing high similarity to genes of mitochondrial origin in the nuclear genome of tobacco and carrot. Interestingly, these sequences contained the mismatches identified by RNA-seq in the genes mentioned above. Further experimental data is needed to test the function of the nuclear transcripts of the mitochondrial genes. Overall, we did not find solid evidence of non-canonical editing processes.

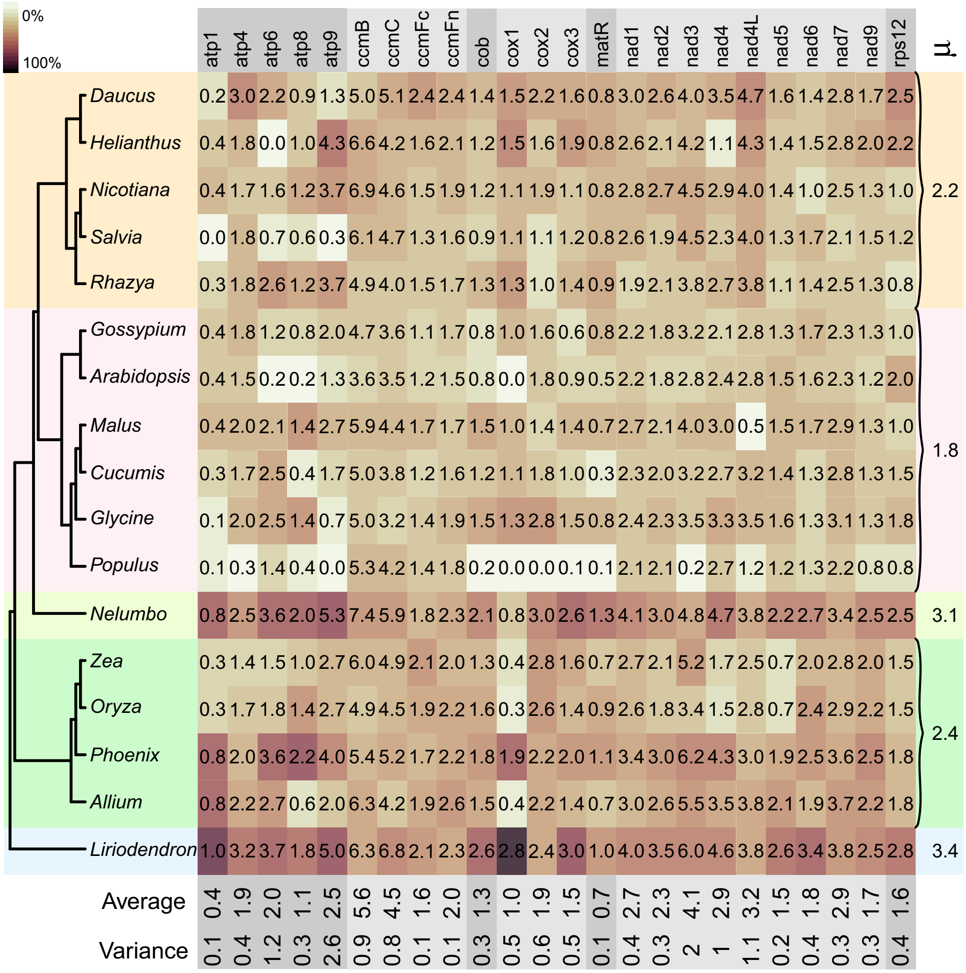
Variability of RNA editing among plants and genes

The total number of editing sites in each plant species was calculated by adding the editing sites of the 24 protein-coding genes together. The resulting totals varied two fold across the angiosperm phylogeny (Fig. 1 and ESM 2). This variability has been reported in other studies (Mower 2008; Picardi et al. 2010). The largest numbers of editing sites were found in *Liriodendron* and *Nelumbo*, with 755 and 679 editing sites, respectively. On the opposite end, two members of the rosids, *Arabidopsis* and *Populus*, had the lowest values: 349 and 313 editing sites, respectively. The average number of editing sites among monocots, rosids, and asterids showed no significant differences among them, but they were significantly lower than those of *Liriodendron* and *Nelumbo* (Fig. 1).

The number of editing sites was compared among the 24 protein-coding genes (ESM 6). The average numbers of editing sites varied greatly among genes, as observed in previous works (Cuenca et al. 2010; Giegé and Brennicke 1999; Picardi et al. 2010). While the largest average, ~ 43 editing sites, was found in the gene *nad4*, the smallest average, ~ 5 editing sites, was found in *atp8*. By grouping the genes based on their function, cytochrome C biogenesis and respiratory complex I genes exhibited the largest average numbers of editing sites, while ribosomal subunits and respiratory complex V genes showed the lowest numbers. This is consistent with other studies in which RNA editing occurred preferentially in genes encoding membrane-bound proteins and genes under strong selection (Jobson and Qiu 2008; Mower and Palmer 2006).

The average density of editing sites, expressed as the number of editing sites per 100-bp region, was calculated individually for each of the 17 × 24 protein-coding genes (Fig. 2). In total, six genes from a number of eudicots were devoid of editing sites (Fig. 2 and ESM 2). These genes were *atp9, cox1*, and *cox2* from *Populus; cox1* from *Arabidopsis; atp1* from *Salvia*; and *atp6* from *Helianthus*. The lack of editing sites in those genes was also confirmed by predictions made by PREP-Mt (Mower 2009). Previous studies have also recognized the lack of editing sites in the aforementioned *Arabidopsis* and *Salvia* genes (Bentolila et al. 2008; Giegé and Brennicke 1999; Wu et al. 2017). Editing sites in *Populus* have not been analyzed before; however, no editing sites were identified in the gene *atp9* of *Salix suchowensis*, which is closely related to *Populus* (Ye et al. 2017).

The analysis of the editing site densities of some genes showed remarkable variations across the 17 angiosperms (Fig. 2). For example, *Liriodendron* had 3.7 editing sites per 100-bp in *atp6*, while other angiosperms had almost none. Another stark disparity was observed between *Liriodendron* and *Populus nad3*, with 5.8 and 0.2 editing site densities, respectively. This variability was systematically analyzed by calculating the average and variance values of the editing site densities in each gene (bottom row in Fig. 2). The highest average densities were observed in cytochrome C biogenesis and Complex I genes. In particular, the gene *ccmB* had the highest average density value, while the gene *atp1* had the smallest. These results agree with the total numbers of editing sites described above.



	atp1	atp4	atp6	atp8	atp9	ccmB	ccmC	ccmFc	ccmFn	cob	cox1	cox2	cox3	matR	nad1	nad2	nad3	nad4	nad4L	nad5	nad6	nad7	nad9	rps12	μ
Daucus	0.2	3.0	2.2	0.9	1.3	5.0	5.1	2.4	2.4	1.4	1.5	2.2	1.6	0.8	3.0	2.6	4.0	3.5	4.7	1.6	1.4	2.8	1.7	2.5	2.2
Helianthus	0.4	1.8	0.0	1.0	4.3	6.6	4.2	1.6	2.1	1.2	1.5	1.6	1.9	0.8	2.6	2.1	4.2	1.1	4.3	1.4	1.5	2.8	2.0	2.2	
Nicotiana	0.4	1.7	1.6	1.2	3.7	6.9	4.6	1.5	1.9	1.2	1.1	1.9	1.1	0.8	2.8	2.7	4.5	2.9	4.0	1.4	1.0	2.5	1.3	1.0	
Salvia	0.0	1.8	0.7	0.6	0.3	6.1	4.7	1.3	1.6	0.9	1.1	1.1	1.2	0.8	2.6	1.9	4.5	2.3	4.0	1.3	1.7	2.1	1.5	1.2	
Rhazya	0.3	1.8	2.6	1.2	3.7	4.9	4.0	1.5	1.7	1.3	1.3	1.0	1.4	0.9	1.9	2.1	3.8	2.7	3.8	1.1	1.4	2.5	1.3	0.8	
Gossypium	0.4	1.8	1.2	0.8	2.0	4.7	3.6	1.1	1.7	0.8	1.0	1.6	0.6	0.8	2.2	1.8	3.2	2.1	2.8	1.3	1.7	2.3	1.3	1.0	1.8
Arabidopsis	0.4	1.5	0.2	0.2	1.3	3.6	3.5	1.2	1.5	0.8	0.0	1.8	0.9	0.5	2.2	1.8	2.8	2.4	2.8	1.5	1.6	2.3	1.2	2.0	
Malus	0.4	2.0	2.1	1.4	2.7	5.9	4.4	1.7	1.7	1.5	1.0	1.4	1.4	0.7	2.7	2.1	4.0	3.0	0.5	1.5	1.7	2.9	1.3	1.0	
Cucumis	0.3	1.7	2.5	0.4	1.7	5.0	3.8	1.2	1.6	1.2	1.1	1.8	1.0	0.3	2.3	2.0	3.2	2.7	3.2	1.4	1.3	2.8	1.3	1.5	
Glycine	0.1	2.0	2.5	1.4	0.7	5.0	3.2	1.4	1.9	1.5	1.3	2.8	1.5	0.8	2.4	2.3	3.5	3.3	3.5	1.6	1.3	3.1	1.3	1.8	
Populus	0.1	0.3	1.4	0.4	0.0	5.3	4.2	1.4	1.8	0.2	0.0	0.0	0.1	0.1	2.1	2.1	0.2	2.7	1.2	1.2	1.3	2.2	0.8	0.8	
Nelumbo	0.8	2.5	3.6	2.0	5.3	7.4	5.9	1.8	2.3	2.1	0.8	3.0	2.6	1.3	4.1	3.0	4.8	4.7	3.8	2.2	2.7	3.4	2.5	2.5	3.1
Zea	0.3	1.4	1.5	1.0	2.7	6.0	4.9	2.1	2.0	1.3	0.4	2.8	1.6	0.7	2.7	2.1	5.2	1.7	2.5	0.7	2.0	2.8	2.0	1.5	2.4
Oryza	0.3	1.7	1.8	1.4	2.7	4.9	4.5	1.9	2.2	1.6	0.3	2.6	1.4	0.9	2.6	1.8	3.4	1.5	2.8	0.7	2.4	2.9	2.2	1.5	
Phoenix	0.8	2.0	3.6	2.2	4.0	5.4	5.2	1.7	2.2	1.8	1.9	2.2	2.0	1.1	3.4	3.0	6.2	4.3	3.0	1.9	2.5	3.6	2.5	1.8	
Allium	0.8	2.2	2.7	0.6	2.0	6.3	4.2	1.9	2.6	1.5	0.4	2.2	1.4	0.7	3.0	2.6	5.5	3.5	3.8	2.1	1.9	3.7	2.2	1.8	
Liriodendron	1.0	3.2	3.7	1.8	5.0	6.3	6.8	2.1	2.3	2.6	2.8	2.4	3.0	1.0	4.0	3.5	6.0	4.6	3.8	2.6	3.4	3.8	2.5	2.8	3.4
Average	0.4	1.9	2.0	1.1	2.5	5.6	4.5	1.6	2.0	1.3	1.0	1.9	1.5	0.7	2.7	2.3	4.1	2.9	3.2	1.5	1.8	2.9	1.7	1.6	
Variance	0.1	0.4	1.2	0.3	2.6	0.9	0.8	0.1	0.1	0.3	0.5	0.6	0.5	0.1	0.4	0.3	2	1	1.1	0.2	0.4	0.3	0.3	0.4	

Fig. 2 Editing site densities in 24 protein-coding genes of 17 angiosperms. The numbers of editing sites per 100-bp (editing site densities) are associated to a color gradient defined from the maximum and minimum editing site densities in each column. Higher densities are depicted by darker colors, and lower ones with lighter colors. Average editing site densities and their variances are shown in the bottom rows, respectively. The rightmost column depicts the average value μ calculated from all the editing site densities within each angiosperm lineage

RNA editing at each codon position

Editing sites also showed an uneven distribution among codon positions (ESM2). Globally, 32, 56, and 12% of the 8318 identified editing sites were found at first, second, and third codon positions, respectively, in agreement with previous studies (Cuenca et al. 2010; Giegé and Brennicke 1999; Mulligan et al. 2007; Sloan et al. 2010; Picardi et al. 2010). This uneven distribution among codon positions was also observed in each angiosperm mitochondria (stacked bars in Fig. 1), and in each gene across the phylogeny with a few exceptions (e.g., the genes *atp8* and *cob* in ESM 6). Nevertheless, as another study has observed (Cuenca et al. 2010), unexpected codon position distributions were found when the analysis was carried out in each gene of each plant species individually (ESM 7). In some cases, all editing sites were found at a single codon position (dashes in ESM 7).

Because editing sites represent a sample of the cytidines in a gene, the proportion of editing sites per codon position may be simply the result of how cytidines are distributed among codon positions. We compared the distributions of editing sites and cytidines among first and second codon positions for each gene of each plant (chi-squared test, two degrees of freedom, and significance level of 0.05). Only 70 out of 17 × 24 genes, i.e. ~ 15% of genes, were distributed significantly different with respect to cytidines (dots in ESM 7) suggesting a selective advantage given by a greater increase in hydrophobicity with editing at second codon position (Cuenca et al. 2010; Jobson and Qiu 2008). In contrast, in about 85% of the studied genes both distributions were statistically similar and the larger proportion of editing sites at second than first codon positions could be the result of random sampling from a biased distribution of cytidines between codon positions.

To study the physicochemical modifications generated by RNA editing at the protein level, amino acid changes produced by the identified editing sites were analyzed ("Amino acid changes produced by RNA editing" section). Our results are consistent with previous studies (Giegé and Brennicke 1999; Jobson and Qiu 2008; Tillich et al. 2006; Kugita et al. 2003). The majority (~ 85%) of the editing events resulted in non-synonymous codon changes (ESM 8) and mainly involved three amino acid changes: Pro → Leu, Ser → Leu, and Ser → Phe. In addition, based on

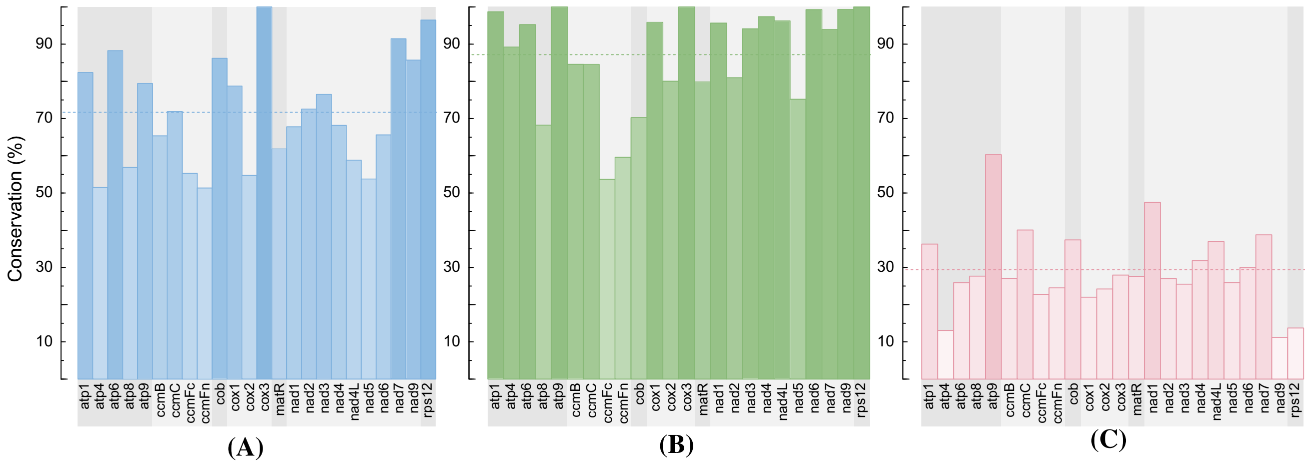


Fig. 3 Average conservation of editing sites in each mitochondrial gene at first (**a**), second (**b**), and third (**c**) codon positions, respectively. Dashed horizontal lines represent the global average of conservation of editing sites at each codon position

the physicochemical properties of amino acids (Kyte and Doolittle 1982), codon alterations were classified according to the hydrophobicity of the resulting amino acids (ESM8). About ~ 55% of the amino acid changes were hydrophilic → hydrophobic produced by editing sites mainly at second codon positions. As shown in previous studies, RNA editing causes an overall increase in hydrophobicity of the resulting proteins (Giegé and Brennicke 1999; Mower 2008; Picardi et al. 2010; Yura and Go 2008; Yura et al. 2009).

In a few cases, RNA editing at first codon positions led to the creation of stop codons, affecting 19 CAA, 25 CGA, and three CAG codons (ESM1). About 74% of such edits were required to form proteins of a conserved length across angiosperms and were edited efficiently (i.e. most reads supported the editing event). For example, in the gene *atp6 of Malus*, a premature stop codon is generated through an editing site located 72 nucleotides upstream of the genomic stop codon. This editing site showed 97% of editing extent and 100% of conservation. In 10 cases, premature stop codons were created, probably as a result of erroneous editing (Schuster and Brennicke 1991). Those sites were only found edited in a single SRA experiment in most cases, and when edited, less than 20% of the reads showed the editing event. For example, a premature stop codon in the gene *nad2* of *Nicotiana* is produced about 560 bases upstream of the genomic stop codon due to an editing site poorly conserved (i.e., only present in one out of 17 angiosperms) and with 10% of editing extent. The generation of premature stop codons may be the result of meaningless editing (Bentolila et al. 2013; Hammani and Giegé 2014; Mower and Palmer 2006; Picardi et al. 2010; Sun et al. 2016), or may have a functional role in plant mitochondria (Schuster and Brennicke 1991).

Conservation of editing sites

We set out to analyze the phylogenetic distribution of RNA editing sites in 24 mitochondrial genes of 17 diverse angiosperms. We investigated to which extent individual mitochondrial RNA editing sites are conserved among plant lineages. The conservation of editing sites across the studied angiosperms was evaluated as the proportion of edited cytidines or genomic thymidines at each position in the gene alignments (ESM 1, ESM 9). For each gene, the average conservation of editing sites was measured separately for each codon position (Fig. 3). The conservation levels varied among codon positions, with average values across all genes of ~ 70, ~ 85, and ~ 30% for first, second, and third codon positions, respectively (dashed lines in Fig. 3). Editing sites at third codon positions were poorly conserved, as expected for synonymous sites with no obvious impact in the resulting protein. Interestingly, cytochrome C biogenesis genes had conservation levels below the average values for first and second codon positions.

Even though editing sites at first and second codon positions showed a high level of conservation across the studied plant species and for most genes, we expected even greater conservation given the apparent effect of editing at these sites. Looking in detail at the gene alignments (ESM 9), we noticed that the level of conservation is diminished because of species-specific editing sites. For example, a particular plant species was edited at a position in which all other species had a genomic C (i.e. an unedited cytidine), resulting in little editing site conservation at that site and a reduction in the amino acid conservation in that particular codon. For example, *Arabidopsis, Gossypium*, and *Populus* had a first-codon-position editing site at positions

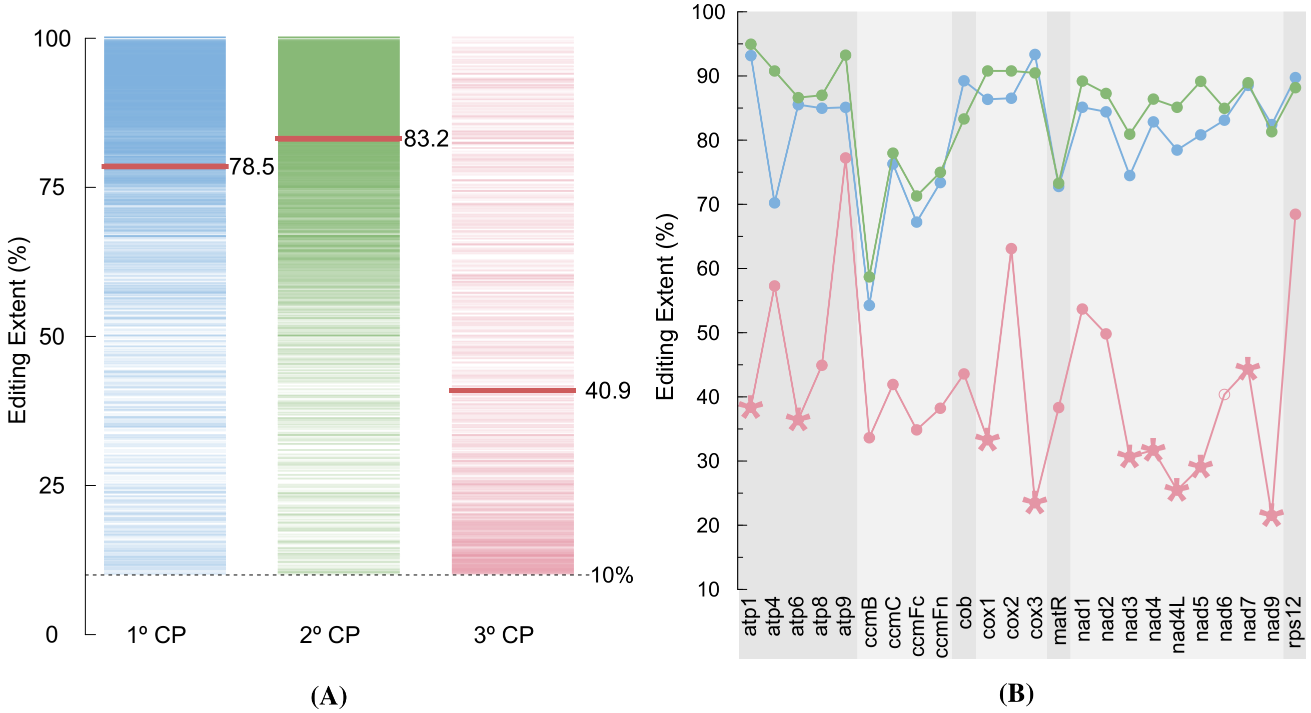


Fig. 4 Editing extent of each identified editing site per codon position. **a** Editing extent at each editing site according to codon positions (CP). Thick red lines show average values of editing extent at each codon position. **b** Average editing extent for each gene according to the codon position of the editing sites. Blue, green, and red points depict values for first, second, and third codon position, respectively. Asterisks indicate significant differences among the average editing extents at each codon position within each gene. Dashed lines indicate average values from the editing extent across genes for each codon position

382, 379, and 382 of the gene *atp4*, respectively, while this site had a genomic C in the remaining 14 angiosperms.

Low editing extent at third codon positions

The extent of RNA editing was measured for each editing site in the 24 mitochondrial genes of 17 angiosperms (ESM 10). When grouping the editing sites according to their codon positions, different levels of editing extent were observed (Fig. 4a). Note that the editing extent ranges between 10 and 100% because only editing sites with a minimum editing extent of 10% were considered. The average editing extent was around 80% for first and second codon positions, while it was around 40% for third codon positions (thick red lines in Fig. 4a). These differences across codon positions were also observed in an in-depth analysis of the transcriptome of *Arabidopsis* (Bentolila et al. 2013) as well as in other studies (Guo et al. 2017; Mower and Palmer 2006; Picardi et al. 2010; Stone et al. 2017; Wu et al. 2015). The average editing extent of sites at third codon positions was significantly different from those at first or second codon positions (one-sample Student's t-test, *p*-values equal to 8.6 × 10e−219 and 4.1 × 10e−251 for first and second codon positions, respectively). The lower editing extent at third codon position agrees with a low conservation of editing sites at this codon position across angiosperms (previous section, Fig. 3). This suggests that editing at synonymous positions is not required for the translation of those transcripts, although this partial editing could also be relevant for the regulation of gene expression (Castandet et al. 2010; Stone et al. 2017).

The editing extent was also analyzed in each gene individually. Average editing extents were discriminated by codon position (Fig. 4b). The levels observed at first and second codon positions barely differed across all genes, while average values for third codon positions were always smaller and more variable. Interestingly, the editing extent in cytochrome C biogenesis genes at first and second position was much lower than the rest of the genes—as observed previously (Giegé and Brennicke 1999; Zehrmann et al. 2008), in agreement with a lower conservation of editing sites in these genes along the angiosperm phylogeny (Fig. 3). Also *matR* showed low editing efficiencies for both codon positions as reported for several other angiosperms (Mower and Palmer 2006).

Based on the 8318 editing sites in the 24 protein-coding genes conserved across the 17 mitochondrial genomes, the correlation between the editing extent and conservation level was assessed by means of Pearson's correlation. As a result, both variables showed a strong positive correlation (rho

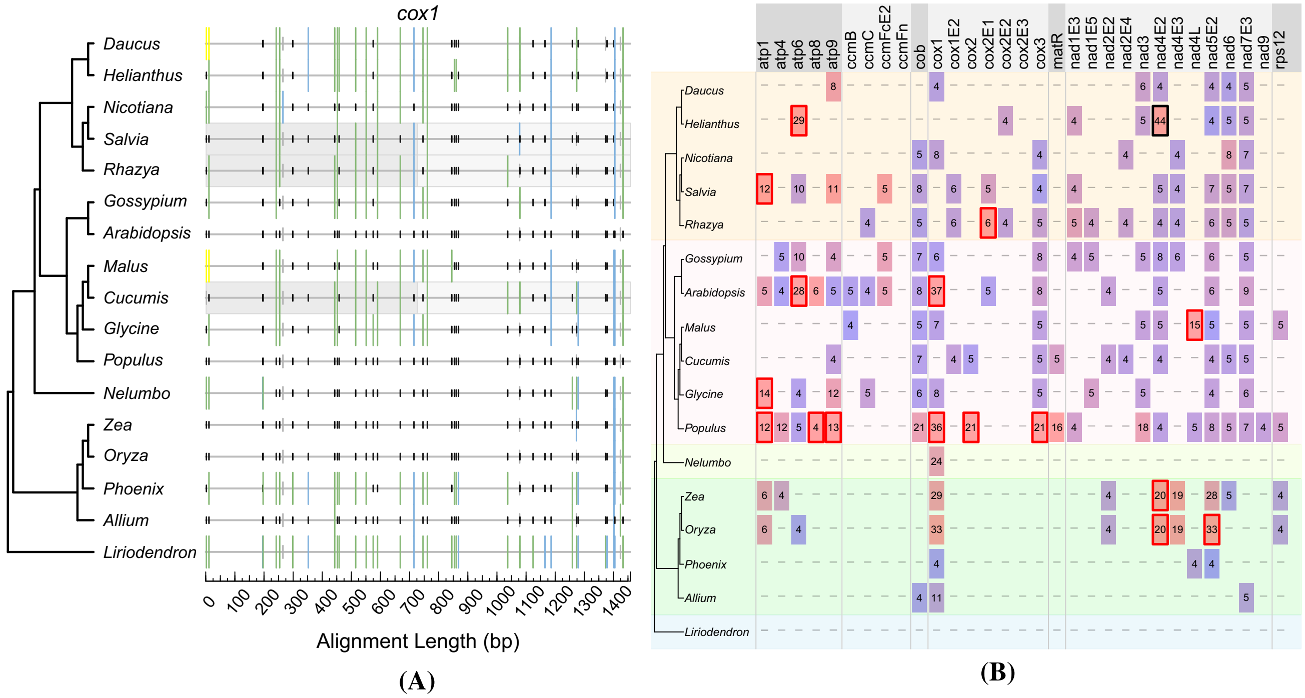


Fig. 5 **a** Localization of editing sites in the cox1 alignment. Editing sites at first (blue) or second (green) codon positions are depicted as vertical bars that cross horizontal lines (gene sequences). For positions in which at least one angiosperm contains an editing site, the nucleotides of other angiosperms at this site are marked with blue or green vertical bars for editing sites at first and second codon positions, respectively; with black, gray or yellow vertical bars when a T, a C or a gap is present, respectively; and no marks for A or G bases. Exons are represented by gray boxes of different intensity. **b** Presence of thymidine footprints in each gene. Thymidine footprints are depicted by colored squares. For each square, the number of thymidines replacing highly conserved editing sites is shown. The color of a square indicates the ratio of the footprint length and the gene/exon length. A color gradient from blue (ratio = 0) to red (ratio = 1) is used. Squares with red borders indicate footprints that span the whole gene/exon (ratio = 1). Exons without thymidine footprints in any plant species are not shown. Squares with black borders indicate intron loss by retroprocessing

equal to 0.65 and *p*-value equal to 0). When this analysis was repeated considering only editing sites in each species or each gene, the correlation estimates fluctuated along a wide range. For example, while *Nicotiana* showed a very strong correlation (rho equal to 0.82 and p-value 0), *Cucumis* showed a moderate-weak correlation (rho equal to 0.40 and p-value 0). Similarly, the correlation resulted very strong for *cox1* (rho equal to 0.85 and p-value 0), while a weak correlation was detected in *ccmB* (rho equal to 0.33 and p-value 0).

Thymidine footprints

Editing sites are being mainly lost via silent replacements to thymidines in genomes (Cuenca et al. 2010; Mower 2008; Shields and Wolfe 1997; Sloan et al. 2010). We observed that several gene sequences showed regions containing numerous consecutive thymidines at positions of highly conserved editing sites ("Identification of thymidine footprints" section). These regions were named *thymidine footprints* and are likely the result of loss of editing sites through retroprocessing (Bowe and dePamphilis 1996; Cuenca et al. 2016, 2010; Hecht et al. 2011; Mulligan et al. 2007; Parkinson et al. 2005; Sloan et al. 2010). For example, the gene *cox1* had long thymidine footprints in several angiosperms (Fig. 5a). In *Arabidopsis* and *Populus* the thymidine footprints spanned the whole gene as the *cox1* gene is devoid of editing sites (ESM 9). *Nelumbo, Zea*, and *Oryza* exhibited long footprints (~ 1 kb long) with 24, 28, and 32 consecutive thymidines at positions of highly conserved editing sites, respectively.

A systematic analysis identified more than 150 thymidine footprints on gene alignments encompassing more than a thousand of thymidines at positions of highly conserved editing sites (Fig. 5b). We observed that all protein-coding genes except for *ccmFn* had at least one angiosperm showing thymidine footprints (Fig. 5b and ESM 9). Strikingly, 16 genes or exons were fully covered by thymidine footprints (cells with red borders in Fig. 5b). It is also noticeable that most footprints were found in eudicots and none in *Liriodendron* (Fig. 5b). *Populus* showed the largest number of thymidine footprints, distributed in 19 out of 24 genes. Three of them (*atp9, cox1*, and *cox2*) lack editing sites even at third codon positions, while the other 3 genes had 1 or 2 editing sites at third codon positions, which were only poorly conserved across the phylogeny. When analyzing each gene across the phylogeny, we observed that the gene *cox1* exhibited thymidine footprints most frequently, followed by *nad4, nad5*, and *nad7*.

We also looked at the intron status to assess any possible correlation between the loss of editing sites (thymidine footprints) and nearby introns. We found that the gene *nad4* of *Helianthus* lack editing sites and introns in a ~ 1 kb region that contains introns 2 and 3 in the other asterids (ESM 9). In contrast, *Zea* and *Oryza* showed no editing sites in exon 2 but retained all introns in the gene *nad4*.

Retroprocessing behind the loss of editing sites

Based on the identified RNA editing sites, different evolutionary hypotheses about the gain and loss of these sites were tested. This was carried on by means of non-reversible Markov models based on five nucleotide states ("Estimation of Markov models" section), in which editing sites were the fifth nucleotide state E (Mower 2008). The gain of editing sites was defined as the rate ratio ρ_{TE} at which thymidines are substituted by editable cytidines, whereas a loss of editing was defined as the rate ratio ρ_{ET} at which editable cytidines are substituted by thymidines. The rationale behind these definitions is that the E-to-T and T-to-E substitutions are synonymous at the protein level, and thereby theoretically free of selection (Mower 2008; Shields and Wolfe 1997). If editing sites are being lost during angiosperm evolution, the rate ratio ρ_{ET} is expected to be greater than the rate ratio ρ_{TE} (Cuenca et al. 2010; Mower 2008; Shields and Wolfe 1997; Sloan et al. 2010).

To test this hypothesis, we first evaluated the best way of modeling the nucleotide alignments containing editing sites by using a single Markov model for the three codon positions (M_0 in ESM 11A) or three different Markov models for each codon position ($M_{1,}$ $M_{2,}$ and, M_3 in ESM 11A for first, second, and third codon positions, respectively). The selection between the two models (M_0 vs $M_1 + M_2 + M_3$) was carried out by comparing their likelihoods with a likelihood-ratio test (LRT) with degree of freedom (df) equal to the difference in their numbers of rate ratios. The LRT indicated that the use of a Markov model for each codon position ($M_{1,}$ $M_{2,}$ and M_3 in ESM 11A) significantly increased the fit to the data (df = 40; *p*-value equals 0). Therefore, we used these models to test the hypothesis that editing sites are lost more rapidly than gained in angiosperm evolution. In the three Markov models ($M_{1,}$ $M_{2,}$ and M_3 in ESM 11A), the rate ratio ρ_{TE} was significantly greater than the rate ratio ρ_{ET} (df = 1; *p*-values of 2 × e−145, 1 × e−246, and 4 × e-−32 for first, second, and third codon position, respectively), indicating that editing sites are being lost faster than gained in the three codon positions. The significance of the comparison between the two rate ratios was assessed by estimating an alternative model for each codon position, in which the rate ratio ρ_{TE} was constrained to be equal to the rate ratio ρ_{ET} (M_4, M_5, M_6 in ESM 11A for first, second, and third codon positions, respectively). The results were significant for all three codon positions.

However, these models assume that editing sites are subjected to the same mutation rate as the other sites and do not consider the effect of retroprocessing, which might be affecting the rate ratio ρ_{ET} in particular (Mulligan et al. 2007; Sloan et al. 2010). To quantify the effect of retroprocessing in each of the 20 rate ratios, 20 new alternative Markov models were estimated for each codon position (M_{1a-t}, M_{2a-t}, and M_{3a-t} in ESM 11B) based on modified alignments, in which all the thymidine footprints along with their homologous sites were removed from the initial alignments. For each codon position, each alternative model was estimated with one of the 20 rate ratios fixed with the value of the same rate ratio estimated in the initial model (M_1, M_2, or M_3). Comparisons of these alternative models with the initial ones using LRTs showed that the rate ratio ρ_{ET} for the first and second codon positions was significantly affected by the presence of thymidine footprints (df = 1; *p*-values of 3 × 10e−13 and 4 × 10e−23, respectively).

To avoid the confounding effect of retroprocessing, the difference between the editing site gain and loss rates was tested again using the alignment without thymidine footprints. For each codon position, we compared the (null) models (M_7, M_8, M_9 in ESM 11A) with alternative ones, in which ρ_{TE} was equal to ρ_{ET} (M_{10}, M_{11}, and M_{12} in ESM 11A). Their comparisons revealed that, even with thymidine footprints removed, the rate ratio ρ_{ET} was greater than ρ_{TE} and that the two rate ratios were significantly different to each other for the three codon positions (df = 1; *p*-values of 4 × 10e−37, 5 × 10e−55, 5 × 10e−10, respectively). Nevertheless, the new rate ratio ρ_{ET} showed a drastic reduction in comparison to that in the first testing that included thymidine footprints. For example, the new rate ratio ρ_{ET} has almost reduced by half in comparison to the values obtained originally (from 32.5 to 18.5). Even though thymidine footprints were removed, it is possible that the rate ratio ρ_{ET} may still be affected by retroprocessing owing to the large number of more lax thymidine footprints (i.e., those with less than 4 consecutive thymidines) observed in the alignments. The number of thymidine footprints containing 3 thymidines, for example, doubled in number those containing four or more thymidines. We were unable to compare the rate ratios ρ_{TE} and ρ_{ET} for alignments without footprints containing 3 consecutive thymidines because they were almost devoid of editing sites, preventing us from estimating further Markov models.

We also studied the evolution of editing sites by analyzing the rate ratios ρ_{CE} and ρ_{EC} (Mower 2008). These rate ratios indicate the gain and loss of editing site recognition

Table 1 Features of the different types of editing site gains inferred through ancestral state reconstruction

Fraction of gains	Total	(%)	1st codon position (%)	2nd codon position (%)	3rd codon position (%)	Synonymous editing sites (%)	Non-synonymous editing sites (%)	Average editing extent (%)	Average conservation (%)
T ➔ E	34	5	21	56	24	24	76	75	97
C ➔ E	651	92	25	13	62	72	28	37	24
A ➔ E	18	3	28	17	56	61	39	50	34
G ➔ E	5	1	60	20	20	40	60	45	21
Total	708								

motifs, respectively (Cummings and Myers 2004; Farré et al. 2001; Giegé and Brennicke 1999; Mower 2008; Mulligan et al. 2007; Neuwirt et al. 2005). Using the Markov models M_1, M_2, and M_3 as the null models, alternative models (M_{13}, M_{14}, M_{15} in ESM 11A) were estimated from the initial alignments for each codon position by constraining the rate ratios ρ_{CE} and ρ_{EC} to be equal. The results showed that the rate ratio ρ_{CE} was not significantly different to the rate ratio ρ_{EC} for first and second codon positions (ESM 11A). These findings indicate that editing recognition motifs are lost and gained at the same rate for editing sites at first and second codon positions.

Comparing the rate ratios estimated by the Markov models, the rates of editing site loss (ρ_{ET}) were greater than the loss rates of editing site recognition motif (ρ_{EC}) for first and second codon positions based on the alignments with or without thymidines (ESM 11A). In addition, the rates of gain of editing site recognition motif (ρ_{CE}) were much larger than the rates of editing site gain (ρ_{TE}) for all codon positions based on the alignments with or without thymidines (ESM 11A).

Understanding editing site gains

To learn more about the gain of editing sites during the evolution of plant mitochondria, we performed ancestral state reconstruction at each node of the angiosperm phylogeny. A total of 708 transitions to “Edited” were estimated along the evolution of the 24 mitochondrial protein-coding genes in the 17 selected angiosperms (ESM 12, Table 1). In agreement with the elevated rate ratio for C to E described above, most of the gained editing sites (92%) arose from an ancestral C. In this case, ancestral cytidines would be, later in evolution, recognized by the editosome and edited to thymidines. About 5% of the new editing sites arose from an ancestral T, in which case the edited transcript ended up unchanged through the acquisition of the edited C. Editing sites gained from ancestral A or G were the least represented (~ 3 and ~ 1% respectively). These rare gains involved either synonymous or non-synonymous editing at sites with poor amino acid conservation among angiosperms.

The gains of editing sites were further analyzed respect to their level of conservation and the effect of editing in the translated protein. The majority of the C-to-E gains (72%) were at positions in which editing was synonymous, mainly found at third codon positions. Accordingly, the gained editing sites from an ancestral C showed low editing extents and low levels of conservation (Table 1). For example, *Phoenix nad3* gained a third-codon position editing site (CDS position 225 in ESM1) with an editing extent and conservation of 12 and 6%, respectively. In contrast, most T-to-E gains (76%) represented non-synonymous editing that resulted in an evolutionarily conserved amino acid. These gains were in general edited very efficiently and those sites were extremely conserved across the angiosperm phylogeny (Table 1). Finally, we examined the nucleotides flanking the gained editing sites for the presence of conserved motifs involved in editing site recognition. Previous studies observed that editing sites in plant mitochondria are often preceded by pyrimidines and in some cases followed by purines (Cummings and Myers 2004; Giegé and Brennicke 1999; Mulligan et al. 2007). We found that 86% of the gained editing sites had pyrimidines preceding the editing site, while 42% showed the full YCR motif. These values are slightly lower than those for all the identified editing sites, in which, 93 and 55% exhibited the YCN and YCR motif, respectively. Interestingly, gained editing sites at second codon positions showed the lowest frequency of the YCN motif (67%), while 86 and 90% of those at first and third codon positions had the YCN motif.

Discussion

To achieve a comprehensive picture of mitochondrial C-to-U RNA editing sites, 17 mitochondrial genomes from diverse angiosperms were analyzed. By comparing the mitochondrial protein-coding genes with transcriptomes available as RNA-seq data in public databases, 10,217 editing sites, or C-to-U changes, were identified. Although other types of nucleotide mismatches were also found, these unconventional mismatches represented less than 7% in comparison to the number of C-to-U mismatches. Many of the unconventional mismatches were the result of errors in the annotation of splicing sites, or due to the dual transcription of mitochondrial and nuclear copies of mitochondrial genes. No strong evidence was found to support non-canonical RNA editing in the studied angiosperm mitochondria.

For the comparative and evolutionary analyses, we focused on 8318 editing sites located in the 24 protein-coding genes that are present in all 17 angiosperm mitochondrial genomes. C-to-U RNA editing sites were unevenly distributed across the 17 plant species. Mitochondrial genes in the magnoliid *Liriodendron* and the basal eudicot *Nelumbo* were the most heavily edited in stark contrast to those in rosids. Also, mitochondrial protein-coding genes showed different frequency of C-to-U mismatches. Genes involved in cytochrome C biogenesis and in the respiratory complex I exhibited the highest editing densities, while ribosomal subunits and respiratory complex V genes showed the lowest editing frequencies. Although some protein-coding genes had consistent numbers of editing sites across the phylogeny, others showed differences (e.g., *atp9, nad3*, and *nad4).* These results contrast with previous observations based on a limited number of taxa, in which the frequency of editing sites across species appeared to be fairly consistent and homogeneous (Mower and Palmer 2006). In agreement with other studies (Cuenca et al. 2010; Giegé and Brennicke 1999; Mulligan et al. 2007; Sloan et al. 2010; Picardi et al. 2010; Yura and Go 2008), the identified C-to-U RNA editing sites were unevenly distributed among the three codon positions, being preferably located at second followed by first codon positions.

RNA editing at first and second codon positions is mostly non-synonymous and tends to increase amino acid conservation. Editing sites at second codon positions, and to a lesser extent at first codon positions, were highly conserved across the studied plant species and were efficiently edited. Moreover, non-synonymous editing sites led to amino acid changes that increased the level of hydrophobicity in protein products, which is consistent with previous observations (Cuenca et al. 2010; Covello and Gray 1993; Giegé and Brennicke 1999; Jobson and Qiu 2008). The high conservation and editing efficiency we observed at editing sites at first and second codon positions highlight the essential role of non-synonymous editing sites for the proper function of mitochondrial-encoded proteins. All together, these results suggest that natural selection seems to be maintaining non-synonymous editing in mitochondrial genomes (Cuenca et al. 2010; Jobson and Qiu 2008; Mower 2008).

In contrast to first and second codon positions, editing sites at third codon positions were relatively few (12%) and led to synonymous codon changes. They were poorly conserved and were only partially edited exhibiting the lowest editing extents (40% in average). These observations suggest that they are not directly maintained by natural selection, as observed previously (Bentolila et al. 2008, 2013; Guo et al. 2017; Mower and Palmer 2006; Picardi et al. 2010; Stone et al. 2017; Wu et al. 2015; Zehrmann et al. 2008; Sloan et al. 2010). Furthermore, it has been shown that the editing extent at synonymous positions varies more often among individuals, tissue types or developmental stages than at non-synonymous sites (Bentolila et al. 2008; Zehrmann et al. 2008; Sloan et al. 2010). So, why are third codon editing sites edited? They are likely edited accidentally by the RNA editing machinery or editosome (Bentolila et al. 2013; Hammani and Giegé 2014; Mower and Palmer 2006; Picardi et al. 2010; Sun et al. 2016). To perform editing, the editosome specifically binds messenger RNAs through molecular signals flanking editing sites (Cummings and Myers 2004; Farré et al. 2001; Giegé and Brennicke 1999; Mulligan et al. 2007; Neuwirt et al. 2005). Third codon position editing sites could be surrounded by degenerated signals which are partially recognized by the editosome (Bentolila et al. 2013; Picardi et al. 2010; Takenaka et al. 2013). Alternatively, synonymous edits could be playing different roles in the expression of mitochondrial genes (Sloan and Taylor 2010). Previous research has shown that, in some cases, such editing sites affect the formation and stability of secondary structures in transcripts (Kindgren et al. 2015; Stone et al. 2017), which is crucial to ensure splicing in some genes (Castandet et al. 2010). Our data agree with these observations, showing 23 editing sites at third codon positions with conservation and editing extent greater than 90%. For example, the editing site at position 399 in *Daucus ccmC* showed an editing extent of 92% and a conservation of 94%. A detailed examination revealed the presence of other editing sites within a 3 nucleotide window surrounding 7 of the 23 editing sites, suggesting that their successful editing could be required for the RNA editing of the nearby, non-synonymous, editing sites (Kindgren et al. 2015). .

Our results showed that early-diverging lineages had a significant tendency towards having more editing sites in their mitochondrial genes than late-diverging ones, as observed in other studies (Cuenca et al. 2010; Mower 2008; Sloan et al. 2010). We found that the loss of editing sites along angiosperm evolution is mainly occurring by replacing editing sites with thymidines, instead of a degradation of the editing recognition motif around editing sites. The loss rates of editing sites (E to T) was greater than the loss of editing site recognition (E to C) in our evolutionary analyses. This was expected because the replacement of editing sites by genomic thymidines does not alter the encoded amino acid. This replacement could take place by C-to-T point-mutations or by retroprocessing (Mower 2008; Shields and Wolfe 1997; Sloan et al. 2010; Grewe et al. 2010; Parkinson et al. 2005; Ran et al. 2010; Grewe et al. 2014). Under the model of retroprocessing, the reverse transcription of edited mRNAs produces cDNA molecules, in which editing sites are replaced by thymidines. When these cDNAs undergo homologous recombination with the mitochondrial genome, consecutive adjacent editing sites are lost simultaneously. Although in vivo evidence is yet lacking in angiosperm mitochondria, abundant indirect evidence of the acting of retroprocessing has accumulated (Bégu et al. 1998; Grewe et al. 2010; Farré and Araya 1999; Odom and Herrin 2013; Ran et al. 2010; Wahleithner et al. 1990). The role of retroprocessing is apparent when long sequence stretches lost consecutive editing sites (thymidine footprints) and even introns (Cuenca et al. 2010, 2016; Grewe et al. 2010; Ran et al. 2010).

A systematic analysis revealed the presence of a number of thymidine footprints in mitochondrial genes with a minimum of four consecutive thymidines in positions of highly conserved editing sites. If retroprocessing was indeed responsible for the thymidine footprints, our results indicate that retroprocessing is a widespread phenomenon in rosids, moderate in monocots, and almost absent in the magnoliid *Liriodendron* and the basal eudicot *Nelumbo*. The spatial localization of the identified thymidine footprints on genes revealed that, even though in a few cases all editing sites were lost from a gene (e.g. *cox1*), retroprocessing seemed to act locally, eliminating editing sites from different internal regions of a gene (Cuenca et al. 2016; Geiss et al. 1994; Grewe et al. 2009; Itchoda et al. 2002; Lopez et al. 2007; Ran et al. 2010; Sloan et al. 2010). Different models have been proposed to explain this localized effect of retroprocessing, each entailing different predictions regarding the frequency and location of intron and editing site losses along genes (Cuenca et al. 2016; Zhang et al. 2010). One is the reverse transcription of variable-length cDNA molecules from completely or partially edited transcripts. Another is the presence of full-length cDNA molecules, which recombine only partially. Recent evidence suggests that microconversion could be responsible for the localized loss of edited sites (Cuenca et al. 2010; Hao and Palmer 2009). We agree with previous studies that retroprocessing is an important mechanism behind editing and intron loss in mitochondrial genes of plants (Cuenca et al. 2016; Mower 2008; Sloan et al. 2010). Here, we identified the loss of introns in the gene *nad4* of *Helianthus* and did not detect a bias in the location of thymidine footprints on genes. Our observations agree with a retroprocessing model, in which reverse-transcribed transcripts recombine with the genome but gene conversion takes place most frequently in interior regions of the genes rather than at the ends (Zhang et al. 2010). Previous reports showed that the frequency of recombination varies along genomic regions and that some genes or gene regions may recombine preferentially (Lawson et al. 2010; Paigen and Petkov 2010). Similarly, we observed a preferential loss of editing sites in the gene *cox1* across the angiosperm phylogeny.

To further evaluate the retroprocessing role in editing site losses, statistical tests compared evolutionary hypotheses based on data sets including and excluding the thymidine footprints. These tests showed a significant decrease in the rate ratio E-to-T for first and second codon positions when thymidine footprints were excluded, while all other nucleotide changes remained the same. These results provide evidence that retroprocessing is, at least partly, responsible for the E to T changes. If retroprocessing indeed played a major role in the loss of editing sites, then the role of natural selection in purging RNA editing sites through independent point mutations during angiosperm mitochondrial evolution (Mower 2008) should be reconsidered. Nevertheless, further work will be necessary to gather additional evidence to test the hypothesis of loss of editing sites by retroprocessing.

The gain and loss rates of editing site recognition (C to E and E to C, respectively) were also estimated and compared. Our analyses showed that the rates of acquisition and degradation of editing site recognition motifs were relatively elevated and it was 5–15 times higher in the third than in first and second codon positions. These observations agree with a previous study (Mower 2008) and may relate to the dynamic nature of the editing process and relaxed specificity of the editing machinery (Bentolila et al. 2013; Takenaka et al. 2012). Many of those sites that are variably recognized by the editosome along angiosperm evolution may not represent real acquisitions or degradation of editing motifs. Instead, they may be the result of the inconsistency of the editing process within species and tissues, particularly at synonymous sites. This hypothesis is supported by the fact that the rate ratios C-to-E and E-to-C are greater in editing sites at third codon positions, which are synonymous and show overall a low editing extent and are poorly conserved.

By ancestral state reconstruction, we examined the gain of editing sites and found that the majority of the gains arose from ancestral Cs and in a much lesser extent from ancestral Ts. This result was unexpected because a C to E gain results in a different nucleotide in the transcript, which may lead to a change in protein sequence. In contrast a T to E gain results in an identical transcript and protein sequence. A more detailed analysis revealed that the gains from ancestral Cs were mostly synonymous (72%). Also, these C to E gains showed low average conservation (24%) and low average editing extent (37%). These observations indicate that cytidines that are not under strong natural selection are being edited erratically, as previously observed for editing at third codon positions. It is also possible that putative gains of editing recognition in C to E gains may be an artifact resulting from the natural variability of editing efficiency at those sites with predominant synonymous editing. Then, if cytidines are poorly edited (< 10% editing extent; i.e. not considered editing sites) in several species while a C at the same alignment position shows an editing extent bearly greater than 10% in a single species, it will appear as a C-to-E gain. For example, one site of the gene *cox3* (alignment position 603) was identified as an editing site in *Glycine* and *Populus* with 12–18% of editing extent, while it was considered a non-edited C in 11 angiosperms that showed 1–6% of editing extent. The opposite is true for T to E gains, in which most of the editing is non-synonymous (76%) and these editing sites are efficiently edited (75%) and highly conserved (97%).

Materials and methods

Identification of RNA editing sites from transcriptomic data

To identify editing sites, a custom computational pipeline was coded to take two inputs for each plant species: a reference file with mitochondrial protein-coding genes and a file with transcriptomic data. The reference file consisted in a multi-FASTA file containing the CDS of protein-coding genes. These sequences were extracted from complete mitochondrial genomes available in the GenBank organellar genome database (Accession Numbers in ESM 13). To increase the read coverage at the end of the CDS, their 5′ and 3′ ends were extended with 100-bp flanking regions, which were discarded after read mapping. The transcriptomic data consisted in a SAM file, which contained reads aligned on the reference gene. These reads were obtained from RNA-seq data in the SRA database at NCBI (SRA accessions in ESM 13). Based on these two inputs, editing sites were identified by counting the number of A, C, G, and U nucleotides in reads aligned at each position of the reference gene. The *editing extent,* #U/(#A+#C+#G+#U), was calculated at each C-containing reference position. If a C-containing reference position showed an editing extent equal to or greater than 10%, it was defined as an editing site. The same procedure was used to identify unconventional mismatches between mitochondrial genes and transcriptomes. The threshold of 10% was chosen because it led to a high number of editing sites conserved across all the studied plant species (ESM 14).

The SRA accessions were chosen to match the same plant species of the reference gene; however, different individuals or even different ecotypes had to be used. For the genus *Liriodendron*, of particular interest given its evolutionary position, different species were used because a single mitochondrial genome (from *L. tulipifera*) and a single transcriptomic study (from *L. chinense*) were available. We expected a low number of disagreements between these species due to the extremely low mutation rate in the mitochondrial DNA of *Liriodendron* (Richardson et al. 2013). Also, it was expected that C to T polymorphisms would show 100% Ts in the mapped reads. Only 6 out of 991 editing sites identified in *Liriodendron* showed editing extent equal to 100% (ESM1), and those were highly conserved editing sites across the angiosperm phylogeny. In addition, the identification of about 90% of the editing sites at first and second codon positions in *Liriodendron* was supported by predictions made by PREP-Mt (Mower 2009) under a cutoff of 0.2. Thus, the putative misidentification of editing sites in *Liriodendron* due to the use of two different species is negligible.

To reliably identify editing sites, only SRA accessions with paired-end reads were selected, which increased the specificity of the read mapping, reducing the number of false positives. Whenever it was possible, RNA-seq data obtained from leaves or roots of wild type individuals were preferred, avoiding those from mutant or virus-infected organisms. About 70% of the 99 SRA accessions were from root or leaf tissues (ESM 13). Because the SRA database included a variable amount of SRA accessions for each of the 17 selected angiosperm species, a number of SRA accessions was chosen so that 80% of the C-containing positions in each mitochondrial gene had at least 10 aligned reads. To minimize the number of false negatives in editing site identification, any C-containing position with a read depth < 10× was annotated as missing data (a gap) in the gene sequences. If a single SRA accession met the listed criteria, an additional SRA accession was selected from a different individual to reduce any potential bias in the transcriptome. On average, ~ 6 SRA accessions per species were chosen, resulting in a total of 99 SRA accessions (ESM13). Considering all the selected SRA accessions, the average read coverage was calculated for all the C-containing positions in a gene in each species (ESM 15). The minimum average read coverage was ~ 12 reads for *Glycine nad6*, whereas the maximum read coverage was of ~ 900 reads for *Gossypium rps12*.

The SAM file of each gene was generated using Magic-BLAST v1.1.0 (https://ncbi.github.io/magicblast). Magic-BLAST was used instead of other read aligners because it runs remotely on NCBI servers. Thus, SAM files were created without downloading RNA-seq files into local computers. To guarantee mapping reads from mature mRNAs, avoiding partially edited transcripts, penalty of − 4, gap open of 50, and gap extend of 50 were set. To discard unreliable mapped reads, duplicated and unpaired reads were filtered out from each SAM file using SAMtools v0.1[4] When more than one SAM file was associated to a reference file, they were merged together into a single SAM file with Picard tools v2.9.0[5]. Then, each SAM file was converted into a

[4] https://github.com/genome/bam-readcount.

[5] http://broadinstitute.github.io/picard.

BAM file and sorted with SAMtools. With a sorted BAM file, bam-readcount v0.8.0 was employed to count the number of A, C, G, and U nucleotides at each position.

Amino acid changes produced by RNA editing

To analyze amino acid changes produced by editing sites, codons containing editing sites were extracted from each protein-coding gene sequence of each studied plant species. Each codon with more than one editing sites was copied as many times as its number of editing sites. Editing sites in codon copies were substituted by unedited cytidines so that each copy had a single editing site from the original codon. Codons before and after editing were translated into amino acids according to the standard genetic code and the resulting amino acids of both groups were compared to identify physicochemical changes.

Measuring conservation of editing sites

Mitochondrial protein-coding genes were automatically aligned with Mafft v7.305 (Katoh and Standley 2013), manually adjusted with AliView v1.18 (Larsson 2014), and their unalignable 5′ and 3′ ends were trimmed (ESM3). Given a gene alignment, a position (column) was eliminated unless one plant species contained an editing site at that position. The conservation of each editing site was measured by calculating a ratio between the number of editing sites or genomic thymidines (because in both cases the translated amino acid would be the same), and the total number of nucleotides in that column.

Identification of thymidine footprints

Thymidine footprints were identified from each gene alignment considering only the editing sites at first or second codon positions. Given a gene alignment, candidate regions were screened from each gene sequence. A candidate region was any region between two consecutive editing sites. For each candidate region, the number of thymidines at positions in which other species had an editing site was counted. Those regions containing > 3 thymidines at positions of highly conserved editing sites were defined as thymidine footprints.

Estimation of markov models

All nucleotide alignments were concatenated into a single alignment and the editing sites were annotated as a fifth nucleotide state E (Mower 2008). From the concatenated alignment, three additional alignments were obtained by partitioning positions based on their codon positions. The gain and loss of editing sites can be modeled by non-reversible Markov models (Mower 2008). Given five nucleotide states (A, C, G, T, and E), a Markov model represents the probability to go from any nucleotide state i to any other nucleotide state j over an infinitesimal period of time t. Formally, it defines a probability distribution $\mathrm{P}(t) = \exp(\mathrm{Q}t)$, which is entirely determined by the 5×5 matrix Q called the *instantaneous rate matrix*. Each cell, or *instantaneous rate*, is the product of two parameters: $\rho_{ij}\pi_j$, the *rate ratio* ρ_{ij}, the rate at which state i is substituted by state j, and π_j, the frequency at which state j occurs. The alignments were used to estimate instantaneous rate matrices with BayesTraits v3.0[6]. An MLTries equal to 100 and ScaleTrees equal to 1 were used. For all the Markov models that were estimated, the angiosperm phylogeny (Fig. 1) was used as tree-based hypothesis, whose branch lengths were optimized by IQ-TREE (Nguyen et al. 2014) with an UNREST evolutionary model using the concatenated nucleotide alignment with editing sites annotated as cytidines.

The ancestral states of RNA editing at each node of the angiosperm phylogeny were estimated with the method “Most Recent Common Ancestor” implemented in Bayestraits v3.0.1 using a Bayesian Markov-chain Monte Carlo (MCMC) approach. A Bayes Multistate model with five possible states (A, C, G, T, and E) was conducted for 150,000 iterations, with the first 50,000 discarded as burn-in, and sampled every 500 cycles over the topology of our phylogeny. A hyperprior with a gamma prior (seeded from a uniform distribution on the interval 0–100) was used to reduce uncertainty and arbitrariness of choosing priors in the MCMC analysis. The mean value of the posterior probabilities for each site was used to reconstruct the ancestral states. The estimated gains of editing sites along the evolution were obtained and further analyzed (ESM 12).

Acknowledgements This work was funded by the National Scientific and Technical Research Council (CONICET) in Argentina.

Author contributions MVSP and AAE conceived and designed the experiments. AAE and CLG designed the editing site annotation pipeline. AAE implemented computational scripts, and performed the experiments. CLG performed experiments of ancestral state reconstruction. MVSP and AAE analyzed the data. AAE prepared figures and supplementary materials. AAE, MVSP, and CLG wrote the manuscript. All authors read and approved the final manuscript.

References

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD (1999). Intracellular gene transfer in action: dual transcription and multiple silencings of nuclear and

[6] http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html.

mitochondrial *cox2* genes in legumes. Proc Natl Acad Sci USA *96*(24):13863–13868

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27(6):1436–1448

Atluri S, Rampersad SN, Bonen L (2015) Retention of functional genes for S19 ribosomal protein in both the mitochondrion and nucleus for over 60 million years. Mol Genet Genomics 290(6):2325–2333

Aubourg S, Boudet N, Kreis M, Lecharny A (2000) In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. Plant Mol Biol 42(4):603–613

Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res 22(1):142–150

Barkan A, Small I (2014) Pentatricopeptide repeat proteins in plants. Annu Rev Plant Biol 65:415–442

Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012). A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. PLoS Genet 8(8):e1002910

Bégu D, Mercado A, Farré J-C, Moenne A, Holuigue L, Araya A, Jordana X (1998) Editing status of *mat-r* transcripts in mitochondria from two plant species: C-to-U changes occur in putative functional RT and maturase domains. Curr Genet 33(6):420–428

Bentolila S, Elliott LE, Hanson MR (2008) Genetic architecture of mitochondrial editing in *Arabidopsis thaliana*. Genetics 178(3):1693–1708

Bentolila S, Oh J, Hanson MR, Bukowski R (2013). Comprehensive high-resolution analysis of the role of an *Arabidopsis* gene family in RNA editing. PLoS Genet, 9(6):e1003584

Bowe LM, dePamphilis CW (1996) Effects of RNA editing and gene processing on phylogenetic reconstruction. Mol Biol Evol 13(9):1159–1166

Castandet B, Araya A (2012) The nucleocytoplasmic conflict, a driving force for the emergence of plant organellar RNA editing. IUBMB Life 64(2):120–125

Castandet B, Choury D, Bégu D, Jordana X, Araya A (2010) Intron RNA editing is essential for splicing in plant mitochondria. Nucleic Acids Res 38(20):7112–7121

Chateigner-Boutin AL, Small I (2007). A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. Nucleic Acids Res 35(17):e114

Choi C, Liu Z, Adams KL (2006) Evolutionary transfers of mitochondrial genes to the nucleus in the *Populus* lineage and coexpression of nuclear and mitochondrial *Sdh4* genes. New Phytol 172(3):429–439

Covello PS, Gray MW (1989) RNA editing in plant mitochondria. Nature 341(6243):662

Covello PS, Gray MW (1993) On the evolution of RNA editing. Trends Genet 9(8):265–268

Cuenca A, Petersen G, Seberg O, Davis JI, Stevenson DW (2010) Are substitution rates and RNA editing correlated? BMC Evol Biol 10(1):1

Cuenca A, Ross TG, Graham SW, Barrett CF, Davis JI, Seberg O, Petersen G (2016) Localized retroprocessing as a model of intron loss in the plant mitochondrial genome. Genome Biol Evol 8(7):2176–2189

Cummings MP, Myers DS (2004) Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. BMC Bioinform 5(1):1

Daley DO, Adams KL, Clifton R, Qualmann S, Millar AH, Palmer JD, Elke P, Whelan J (2002) Gene transfer from mitochondrion to nucleus: novel mechanisms for gene activation from *Cox2*. Plant J 30(1):11–21

Diroma MA, Ciaccia L, Pesole G, Picardi E (2017). Elucidating the editome: bioinformatics approaches for RNA editing detection. Brief Bioinform. https://doi.org/10.1093/bib/bbx129.

Fang Y, Wu H, Zhang T, Yang M, Yin Y, Pan L, Yu X, Zhang X, Hu S, Al-Mssallem IS et al. (2012). A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. PLoS ONE, 7(5):e37164

Farré J-C, Araya A (1999) The *mat-r* open reading frame is transcribed from a non-canonical promoter and contains an internal promoter to co-transcribe exons *nad1e* and *nad5III* in wheat mitochondria. Plant Mol Biol 40(6):959–967

Farré J-C, Leon G, Jordana X, Araya A (2001) *Cis* recognition elements in plant mitochondrion RNA editing. Mol Cell Biol 21(20):6731–6737

Fujii S, Small I (2011) The evolution of RNA editing and pentatricopeptide repeat genes. New Phytol 191(1):37–47

Geiss KT, Abbas GM, Makaroff CA (1994) Intron loss from the NADH dehydrogenase subunit 4 gene of lettuce mitochondrial DNA: evidence for homologous recombination of a cDNA intermediate. Mol Gen Genet MGG 243(1):97–105

Giegé P, Brennicke A (1999). RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci USA 96(26):15324–15329

Grewe F, Viehoever P, Weisshaar B, Knoop V (2009) A *trans*-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoetes engelmannii*. Nucleic Acids Res 37(15):5093–5104

Grewe F, Herres S, Viehöver P, Polsakiewicz M, Weisshaar B, Knoop V (2010) A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte Isoetes engelmannii. Nucleic Acids Res 39(7):2890–2902

Grewe F, Edger PP, Keren I, Sultan L, Pires JC, Ostersetzer-Biran O, Mower JP (2014) Comparative analysis of 11 Brassicales mitochondrial genomes and the mitochondrial transcriptome of *Brassica oleracea*. Mitochondrion 19:135–143

Grimes BT, Sisay AK, Carroll HD, Cahoon AB (2014) Deep sequencing of the tobacco mitochondrial transcriptome reveals expressed ORFs and numerous editing sites outside coding regions. BMC Genomics 15(1):1

Gualberto JM, Lamattina L, Bonnard G, Weil JH, Grienenberger JM (1989) RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature 341(6243):660–662

Guo W, Grewe F, Mower JP (2015). Variable frequency of plastid RNA editing among ferns and repeated loss of uridine-to-cytidine editing from vascular plants. PLoS ONE, 10(1):e0117075

Guo W, Zhu A, Fan W, Mower JP (2017) Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. New Phytol 213(1):391–403

Hammani K, Giegé P (2014) RNA metabolism in plant mitochondria. Trends Plant Sci 19(6):380–389

Hao W, Palmer JD (2009). Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc Natl Acad Sci USA 106(39):16728–16733

Havird JC, Sloan DB (2016) The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial versus nuclear genomes. Mol Biol Evol 33(12):3042–3053

Hecht J, Grewe F, Knoop V (2011) Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genom Biol Evol 3:344–358

Hiesel R, Wissinger B, Schuster W, Brennicke A (1989) RNA editing in plant mitochondria. Science 246(4937):1632–1634

Islam MS, Studer B, Byrne SL, Farrell JD, Panitz F, Bendixen C, Bendixen C, Møller IM, Asp T (2013) The genome and transcriptome of perennial ryegrass mitochondria. BMC Genomics 14(1):202

Itchoda N, Nishizawa S, Nagano H, Kubo T, Mikami T (2002) The sugar beet mitochondrial *nad4* gene: an intron loss and its phylogenetic implication in the Caryophyllales. Theor Appl Genet 104(2–3):209–213

Jobson RW, Qiu Y-L (2008) Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? Biol Direct 3(1):1

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772–780

Kazakoff SH, Imelfort M, Edwards D, Koehorst J, Biswas B, Batley J, Scott PT, andGresshoff PM (2012). Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. PLoS ONE, 7(12):e51687

Kim B, Kim K, Yang TJ, Kim S (2016) Completion of the mitochondrial genome sequence of onion (*Allium cepa* L.) containing the CMS-S male-sterile cytoplasm and identification of an independent event of the ccmFN gene split. Curr Genet 62(4):873–885

Kindgren P, Yap A, Bond CS, Small I (2015) Predictable alteration of sequence recognition by RNA editing factors from *Arabidopsis*. Plant Cell 27(2):403–416

Knie N, Grewe F, Fischer S, Knoop V (2016) Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns–a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. BMC Evol Biol 16(1):134

Knoop V (2011) When you can't trust the DNA: RNA editing changes transcript sequences. Cell Mol Life Sci 68(4):567–586

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31(9):2417–2423

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132

Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30(22):3276–3278

Lawson MJ, Jiao J, Fan W, Zhang L (2010). A pattern analysis of gene conversion literature. Comp Funct Genomics 2009:761512

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Lopez L, Picardi E, Quagliariello C (2007) RNA editing has been lost in the mitochondrial *cox3* and *rps13* mRNAs in Asparagales. Biochimie 89(1):159–167

Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Lecharny A (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16(8):2089–2103

Lutz KA, Maliga P (2001) Lack of conservation of editing sites in mRNAs that encode subunits of the NAD (P) H dehydrogenase complex in plastids and mitochondria of *Arabidopsis thaliana*. Curr Genet 40(3):214–219

Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. Science 311(5768):1727–1730

Mower JP (2008) Modeling sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria. Mol Biol Evol 25(1):52–61

Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37(suppl 2):W253–W259

Mower JP, Palmer JD (2006) Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*. Mol Genet Genomics 276(3):285–293

Mulligan RM, Chang KLC, Chou CC (2007) Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. Mol Biol Evol 24(9):1971–1981

Neuwirt J, Takenaka M, der Merwe JA, Brennicke A (2005) An in vitro RNA editing system from cauliflower mitochondria: editing site recognition parameters can vary in different plant species. RNA 11(10):1563–1570

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Kadowaki K (2002) The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Genet Genomics 268(4):434–445

Odom OW, Herrin DL (2013) Reverse transcription of spliced *psbA* mRNA in Chlamydomonas spp. and its possible role in evolutionary intron loss. Mol Biol Evol 30(12):2666–2675

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet 11(3):221–233

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol 28(1):87–97

Park S, Ruhlman TA, Sabir JSM, Mutwakil MHZ, Baeshen MN, Sabir MJ, Jansen RK (2014) Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. BMC Genomics 15(1):1

Parkinson CL, Mower JP, Qiu Y-L, Shirk AJ, Song K, Young ND, Palmer JD et al (2005) Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5(1):73

Perrotta G, Regina TM, Quagliariello C, Ceci LR (1996) Conservation of the organization of the mitochondrialnad3 andrps12 genes in evolutionarily distant angiosperms. Mol Gen Genet MGG 251(3):326–337

Petersen G, Seberg O, Davis JI, Stevenson DW (2006) RNA editing and phylogenetic reconstruction in two monocot mitochondrial genes. Taxon 55(4):871–886

Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G (2010) Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. Nucleic Acids Res 38(14):4755–4767

Ran J-H, Gao H, Wang X-Q (2010) Fast evolution of the retroprocessed mitochondrial *rps3* gene in Conifer II and further evidence for the phylogeny of gymnosperms. Mol Phylogenet Evol 54(1):136–149

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Palmer JD (2013) Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science 342(6165):1468–1473

Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD (2013) The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11(1):29

Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, et al. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun 8:59

Salmans ML, Chaw SM, Lin CP, Shih ACC, Wu YW, Mulligan RM (2010) Editing site analysis in a gymnosperm mitochondrial genome reveals similarities with angiosperm mitochondrial genomes. Curr Genet 56(5):439–446

Schuster W, Brennicke A (1991) RNA editing makes mistakes in plant mitochondria: editing loses sense in transcripts of a *rps19* pseudogene and in creating stop codons in *coxl* and *rps3* mRNAs of *Oenothera*. Nucleic Acids Res 19(24):6923–6928

Shearman JR, Sangsrakru D, Ruang-Areerate P, Sonthirod C, Uthaipaisanwong P, Yoocha T, Poopear S, Theerawattanasuk K, Tragoonrung S, Tangphatsornruang S (2014) Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. BMC Plant Biol 14(1):1

Shields DC, Wolfe KH (1997) Accelerated evolution of sites undergoing mRNA editing in plant mitochondria and chloroplasts. Mol Biol Evol 14(3):344–349

Shukla P, Singh NK, Gautam R, Ahmed I, Yadav D, Sharma A, Kirti PB (2017). Molecular approaches for manipulating male sterility and strategies for fertility restoration in plants. Mol Biotechnol 59:445

Sloan DB (2017). Nuclear and mitochondrial RNA editing systems have opposite effects on protein diversity. Biol Lett 13:20170314

Sloan DB, Taylor DR (2010) Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing. J Mol Evol 70(5):479–491

Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR (2010) Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: selection vs. retroprocessing as the driving force. Genetics 185(4):1369–1380

Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, Taylor DR (2012) De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (Caryophyllaceae). Mol Ecol Resour 12(2):333–343

Small ID, Peeters N (2000) The PPR motif–a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25(2):45–47

Stone JD, Storchova H (2015) The application of RNA-seq to the comprehensive analysis of plant mitochondrial transcriptomes. Mol Genet Genomics 290(1):1–9

Stone JD, Koloušková P, Sloan DB, Štorchová H (2017). Non-coding RNA may be associated with cytoplasmic male sterility in *Silene vulgaris*. J Exp Bot 68(7):1599–1612

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272(6):603–615

Sun F, Wang X, Bonnard G, Shen Y, Xiu Z, Li X, Gao D, Zhang Z, Tan B-C (2015) *Empty pericarp7* encodes a mitochondrial E–subgroup pentatricopeptide repeat protein that is required for *ccmFN* editing, mitochondrial function and seed development in maize. Plant J 84(2):283–295

Sun T, Bentolila S, Hanson MR (2016) The unexpected diversity of plant organelle RNA editosomes. Trends Plant Sci 21(11):962–973

Takenaka M, Brennicke A (2003) In vitro RNA editing in pea mitochondria requires NTP or dNTP, suggesting involvement of an RNA helicase. J Biol Chem 278(48):47526–47533

Takenaka M, Brennicke A (2007) RNA editing in plant mitochondria: assays and biochemical approaches. In: Methods in enzymology, vol 424. Elsevier, pp 439–458

Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Härtel B, Brennicke A (2012). Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. Proc Natl Acad Sci USA 109(13):5104–5109

Takenaka M, Zehrmann A, Brennicke A, Graichen K (2013a). Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. PLoS ONE, 8(6):e65343

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013b) RNA editing in plants and its evolution. Annu Rev Genet 47:335–352

Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Mol Biol Evol 23(10):1912–1921

Tseng CC, Lee CJ, Chung YT, Sung TY, Hsieh MH (2013) Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant Mol Biol 82(4–5):375–392

Wahleithner JA, MacFarlane JL, Wolstenholme DR (1990). A sequence encoding a maturase-related protein in a group II intron of a plant mitochondrial *nad1* gene. Proc Natl Acad Sci USA 87(2):548–552

Wu Z, Stone JD, Štorchová H, Sloan DB (2015) High transcript abundance, RNA editing, and small RNAs in intergenic regions within the massive mitochondrial genome of the angiosperm *Silene noctiflora*. BMC Genomics 16(1):1

Wu Z, Sloan DB, Brown CW, Rosenblueth M, Palmer JD, Ong HC (2017) Mitochondrial retroprocessing promoted functional transfers of *rpl5* to the nucleus in grasses. Mol Biol Evol 34:2340

Yagi Y, Tachikawa M, Noguchi H, Satoh S, Obokata J, Nakamura T (2013a) Pentatricopeptide repeat proteins involved in plant organellar RNA editing. RNA Biol 10(9):1419–1425

Yagi Y, Hayashi S, Kobayashi K, Hirayama T, Nakamura T (2013b). Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. PLoS ONE, 8(3):e57286

Ye N, Wang X, Li J, Bi C, Xu Y, Wu D, Ye Q (2017) Assembly and comparative analysis of complete mitochondrial genome sequence of an economic plant *Salix suchowensis*. PeerJ 5:e3148

Yura K, Go M (2008) Correlation between amino acid residues converted by RNA editing and functional residues in protein three-dimensional structures in plant organelles. BMC Plant Biol 8(1):79

Yura K, Sulaiman S, Hatta Y, Shionyu M, Go M (2009) RESOPS: a database for analyzing the correspondence of RNA editing sites to protein three-dimensional structures. Plant Cell Physiol 50(11):1865–1873

Zehrmann A, van der Merwe JA, Verbitskiy D, Brennicke A, Takenaka M (2008) Seven large variations in the extent of RNA editing in plant mitochondria between three ecotypes of *Arabidopsis thaliana*. Mitochondrion 8(4):319–327

Zhang L-Y, Yang Y-F, Niu D-K (2010) Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus fungi*. J Mol Evol 71(5–6):364–373