Genes and Processed Paralogs Co-exist in Plant Mitochondria

Argelia Cuenca · Gitte Petersen · Ole Seberg ·
Anne Hoppe Jahren

Received: 30 September 2011 / Accepted: 23 March 2012 / Published online: 7 April 2012

Abstract RNA-mediated gene duplication has been proposed to create processed paralogs in the plant mitochondrial genome. A processed paralog may retain signatures left by the maturation process of its RNA precursor, such as intron removal and no need of RNA editing. Whereas it is well documented that an RNA intermediary is involved in the transfer of mitochondrial genes to the nucleus, no direct evidence exists for insertion of processed paralogs in the mitochondria (i.e., processed and un-processed genes have never been found simultaneously in the mitochondrial genome). In this study, we sequenced a region of the mitochondrial gene *nad*1, and identified a number of taxa were two different copies of the region co-occur in the mitochondria. The two *nad*1 paralogs differed in their (a) presence or absence of a group II intron, and (b) number of edited sites. Thus, this work provides the first evidence of co-existence of processed paralogs and their precursors within the plant mitochondrial genome. In addition, mapping the presence/absence of the paralogs provides indirect evidence of RNA-mediated gene duplication as an essential process shaping the mitochondrial genome in plants.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9496-1) contains supplementary material, which is available to authorized users.

A. Cuenca (✉) · G. Petersen · O. Seberg
Botanical Garden, Natural History Museum of Denmark,
University of Copenhagen, Sølvgade 83 Opg. S,
1307 Copenhagen K, Denmark
e-mail: argelia.cuenca@gmail.com

A. H. Jahren
Department of Geology and Geophysics, University of Hawaii,
1680 East-West Road, Honolulu, HI 96822, USA

Keywords Plant mitochondria · Processed paralogs · RNA editing · Mitochondrial intron · *nad1* · Phylogenetic analysis · Alismatales · Monocotyledons

Introduction

Gene duplication is a major driving force shaping the eukaryotic genome and is considered a primary source of evolutionary novelties. Whereas gene duplication and the fate of duplicated genes have been widely studied in the nuclear genome, comprehensive studies of duplication events in organellar genomes are rare. Duplications are frequent in the plant mitochondria (Kubo et al. 2000; Notsu et al. 2002; Handa 2003; Clifton et al. 2004; Knoop 2004; Ogihara et al. 2005; Allen et al. 2007), but almost exclusively associated with recombinant, repeated sequences (Schuster and Brennicke 1994) resulting in duplications of large contiguous areas of the genome. However, it has been proposed that the plant mitochondrial genomes are subjected to concerted evolution, which obscures and effectively prevents divergence and recognition of duplicate genes (Handa 2003; Clifton et al. 2004; Bergthorsson et al. 2004; Ogihara et al. 2005). Besides recombination-mediated gene duplication, an RNA-mediated gene duplication mechanism has been suggested (Geiss et al. 1994; Bowe and dePamphilis 1996). Under this model, an mRNA copy of the gene is reverse transcribed and reinserted (either in total or in part) into the genome. This kind of gene copy has been termed a processed paralog, indicating the presence of possible signatures left by the maturation process of the mRNA transcribed from the original gene copy.

Two post-transcriptional modifications have a direct impact on our ability to recognize processed paralogs: (i) RNA editing and (ii) splicing of group II introns. In the mitochondria of angiosperms, RNA editing involves C to U base changes at specific sites in the RNA. Because editing is more frequent in first and second codon positions, it usually causes an amino acid change in the resulting protein. Edited sites are usually conserved among taxa (Bowe and dePamphilis 1996; Petersen et al. 2006), but certain taxa have lost the requirement for editing completely or almost completely. Phylogenetic analyses of mitochondrial DNA sequences have identified clades in which C's have been replaced by T's at all or most edited sites, suggesting that all changes took place at a single event. This pattern of nucleotide substitution has been interpreted as due to incorporation of a processed paralog into the genome (Bowe and dePamphilis 1996; Petersen et al. 2006).

The second post-transcriptional modification is splicing. In contrast to its animal counterpart, the angiosperm mitochondrion possesses ca. 25 group II introns (Bonen and Vogel 2001; Knoop 2004) plus a group I intron in *cox*1 of certain taxa (Cho et al. 1998). Possibly caused by the large number of rearrangements in plant mitochondria, some exon clusters have been broken and exons belonging to a single gene have been relocated in the genome. This has led to introns not only being *cis*-spliced, but also *trans*-spliced. It has also been shown that some group II introns may need editing to be spliced (Bonen 2008), but there is limited evidence concerning the temporal order of editing and splicing. However, when lack of introns in genes that usually possess introns is coupled with the presence of T's at edited sites, this strengthens the evidence of their origin via reinsertion from a mature RNA molecule (e.g., as a processed paralog).

In addition to the presence of T's at edited sites and lack of introns, it has been proposed that processed paralogs are distinguishable by having elevated substitution rates (Bowe and dePamphilis 1996; Parkinson et al. 2005). This change in substitution rate will be especially pronounced when a processed paralog has been inserted into another genomic compartment that has a higher substitution rate than the mitochondrion (e.g., the nuclear genome in angiosperms), whereas rate differences may be less obvious if processed paralogs are reinserted into the mitochondrial genome.

It is well established that processed paralogs of mitochondrial genes may be inserted in the nuclear genome (Nugent and Palmer 1991; Adams and Palmer 2003; Liu et al. 2009), but direct evidence of the presence of processed paralogs in the plant mitochondrial genome has been lacking. Existence of processed paralogs in the mitochondrial genome has previously only been inferred by their anomalous features in certain taxa (i.e., lack of a requirement of RNA editing, increased substitution rate) compared to the apparently normal, un-processed gene sequences in other taxa. However, the two corresponding copies, viz., the original, un-processed gene, and the processed paralog, have never been recovered from the same species.

However, there is no obvious reason why a processed paralog could not be reinserted into the mitochondrial genome and potentially co-exist with the original, un-processed copy of the gene. Though a number of questions concerning the fate of the original gene and the processed paralog seem highly pertinent, e.g., (i) are the original mitochondrial genes always replaced by their processed paralogs? and (ii) is convergent evolution or homogenization (e.g., concerted evolution) actively synchronizing the two copies?

As part of a study aiming to reconstruct the phylogenetic relationships within the monocotyledon order Alismatales (Petersen et al., unpublished), we sequenced the intron nad1i477 (Dombrovska and Qiu 2004) and a short region of its flanking exons (exons B and C). In this dataset we observed not only that some species within the Alismatales lacked the intron nad1i477, but also that some species possessed two different copies of this *nad*1 region: one with an intact intron and the second one lacking it. In addition, sequence copies differed in their RNA editing pattern. Consequently the *nad*1 sequences were considered inappropriate for phylogenetic analysis of the Alismatales. Instead, we investigate the sequence characteristics of both copies of *nad*1, as well as the genome location (mitochondrial of nuclear) of the short, non-edited copy. Finally, the possibility that one of the *nad*1 copies was acquired by horizontal transfer has been explored. The results obtained may constitute the first direct evidence of co-occurrence of the original, un-processed genes and processed paralogs within the plant mitochondrial genome.

Materials and Methods

PCR and Sequencing

Sequences of the *nad*1 intron nad1i477 together with partial sequences of the two adjacent exons were obtained for 44 species of the Alismatales (Table 1) by PCR-amplification primers nad1ECr and nad1EB (Demesure et al. 1995). Internal primers placed in the intron were developed and used for sequencing and in some cases also for PCR amplifications (Table S1 in online supplementary materials). Amplification reactions were performed in 50-µl total volume using about 50 ng of template DNA, 1 U of *Taq* DNA polymerase (Ampliqon, Rødovre, Denmark), 40 pmol of each primer, 0.2 mM of each dNTP, 2.5 mM $MgCl_2$, and 10× standard buffer provided by the manufactory. The resulting PCR products were visualized on a 1 % agarose gel and purified using the Qiagen QIAquick PCR Purification Kit. If more than one amplification

Table 1 List of taxa indicating which copy of the sequence *nad*1 region they contain

Family	Species	Paralog I Intron-present	Paralog E Intron-less
Acoraceae	*Acorus calamus*	+	−
	Acorus gramineus	+	−
Araceae	*Arisaema amurense*	+	−
	Gymnostachys anceps	+	−
	Orontium aquaticum	+	−
	Symplocarpus foetidus	+	−
Alismataceae	*Alisma plantago-aquatica*	+	−
	Baldellia ranunculoides	+	−
	Caldesia oligococca	−	+
	Echinodorus cordifolius	+	+
	Echinodorus uruguayensis	+	+
	Luronium natans	+	−
	Ranalisma humile	+	−
	Sagittaria sagittifolia	+	+
Aponogetonaceae	*Aponogeton crispus*	+	+
Butomaceae	*Butomus umbellatus*	+	−
Cymodoceaceae	*Amphibolis griffithii*	+	−
	Cymodocea nodosa	+	−
	Syringodium isoetifolium	+	−
Hydrocharitaceae	*Blyxa aubertii*	+	+
	Egeria naias	−	+
	Elodea canadensis	−	+
	Elodea nutalii	−	+
	Halophila sp.	−	+
	Hydrilla verticillata	−	+
	Hydrocharis morsus-ranae	+	−
	Limnobium laevigatum	+	−
	Najas guadalupensis	−	+
	Najas sp.	−	+
	Nechamandra alternifolia	−	+
	Ottelia ovalifolia	+	+
	Stratiotes aloides	+	−
	Vallisneria sp.	−	+
Limnocharitaceae	*Hydrocleys nymphoides*	−	+
	Limnocharis flava	−	+
Juncaginaceae	*Lilaea scilloides*	+	−
	Triglochin maritima	?	+
	Triglochin palustre	+	+
Posidoniaceae	*Posidonia oceanica*	+	+
Potamogetonaceae	*Potamogeton lucens*	+	n.t.
	Potamogeton natans	+	+
	Zannichellia palustris	−	+
Ruppiaceae	*Ruppia cirrhosa*	+	+
Scheuchzeriaceae	*Scheuchzeria palustris*	+	+
Tofieldiaceae	*Tofieldia pusilla*	+	−
	Pleea tenuifolia	+	−
Zosteraceae	*Heterozostera tasmanica*	+	n.t
	Phyllospadix scouleri	+	+
	Zostera marina	+	−
	Zostera noltii	−	+

product was obtained, PCR products were cleaned from a 1 % agarose gel using the QIAquick PCR gel Purification Kit (Qiagen). Purified PCR products were sequenced using ABI PRISM BigDye Terminator Cycle Sequencing v2.0 Ready Reaction Kit (AP Biosystems). All DNA sequences generated in this study were deposited in GenBank (acc. nos. HM576827 to HM576888). Sequences supposed to lack nad1i477 will be designated Paralog E sequences and sequences containing this intron will be designated Paralog I sequences.

As we detected a lack of the nad1i477 intron in some of the sequences obtained, we decided to explore whether any other of the flanking introns (all of them *trans*-spliced) were also missing, what they might be as consequence of retro-processing of a mature RNA product. To do so, primers (Table S1 in online supplementary materials) were developed to amplify the regions exon A–exon C, and exon B–exon E (the two fragments both with a size of ca. 550 nt in the mature RNA).

Determination of Edited Sites

The presence of edited sites was verified experimentally for a subset of our taxon sampling only, using cDNA sequences generated either from fresh plants or tissue stored in RNA-later (Qiagen). Total RNA was extracted using the Total RNA extraction kit (Ambion, Austin, TX) and DNAased with 1 ul of DNAase I (Promega, Madison, WI). The one step RT-PCR kit (Qiagen) was used to generate cDNAs and to amplify *nad*1. To obtain specific cDNA sequences of Paralog I, amplification was done using primers placed in exon A and exon C, because primers placed in exon B and C tend to produce low quality cDNA sequences, probably caused by co-amplification of paralogous sequences. RT-PCR was performed in accordance to the protocol provided by the manufacturer and using 50 °C for 30 min to generate the cDNA copies and 52 °C as annealing temperature during the DNA amplification.

Genomic Location of *nad*1 Paralogs: qPCR

To test whether both paralogs I and E are located in the mitochondrial genome or whether Paralog E could possibly be nuclear-encoded, we performed a number of qPCR assays to estimate the relative copy number of mitochondrial versus nuclear genes. qPCR was performed using genomic DNA extractions, where both nuclear and mitochondrial DNA are present. The amplification curve displayed by qPCR is a function of the number of copies in which the amplified region is present in the initial sample. As the number of mitochondrial genomes is considerably higher than the number of nuclear genomes in the total DNA fraction, amplification curves are different depending on whether the target region is placed in one or the other genome compartment.

Primers used for the qPCR assays were designed to amplify Paralog E of *nad*1, two mitochondrial genes (*nad*5 and *cob*), and one nuclear gene (*phy*C) using the online PrimerQuestSM Tool (Integrated DNA Technologies, IA, USA) with default parameters. The four regions were amplified for four taxa: *Elodea canadensis*, *Zostera noltii, Potamogeton natan*s, and *Echinodorus uruguayensis* using total genomic DNA as template. This taxon sample was chosen because it includes two species where only Paralog E is present and two species where Paralog E and Paralog I co-occur. All qPCRs assays were done using the SsoFast™ EvaGreen® Supermix (Bio-rad) following the manufacturer's instructions. To insure that all primers were amplifying efficiently (>90 %), qPCR were performed in 5-steps serial dilutions of total DNA, and the r^2 between CT and the logarithm of the starting DNA was calculated. As the primers amplifying the short, intron-less copy of *nad*1 (Paralog E; ca. 150 pb) are also able to amplify the long copy (Paralog I; ca. 1150 pb), melting curves were done for each assay to insure that only one fragment was amplified. To avoid co-amplification of the long Paralog I sequences, the annealing step was kept for only 7–8 s in each cycle. In addition, all *nad*1 products were run in a 2 % agarose gel to check the size of the amplified fragments.

Sequence Comparison Between Paralog I and Paralog E

Following removal of nad1i477, exon sequences from Paralog I were aligned with Paralog E sequences in Mesquite version 2.72 (Maddison and Maddison 2009). This generated a data set of 155 pb that includes the last 39 nucleotides of exon B and the first 116 nucleotides of exon C. This dataset includes too little sequence variation to perform any meaningful character-based phylogenetic analyses; therefore, all comparisons between sequences were based on genetic distance measures, only.

Average uncorrected genetic distance comparisons were calculated in Mega4 (Tamura et al. 2007) as follows: (a) among all Paralog I sequences, (b) among all Paralog E sequences, (c) between a group including all Paralog I sequences versus a group of all Paralog E sequences, (d) between paralogs E and I when both sequences are found in the same individual, and (e) between two subgroups of Paralog E, one including all sequences that co-exist with Paralog I and the other including Paralog E sequences where only Paralog E is found. This last comparison was relevant because there is a possibility that Paralog E sequences are degenerated when another copy of the same region exists in the genome. Paralog I sequences of both species of *Acorus* (outgroup) were removed from all calculations of genetic distance, as well as the Paralog I sequences of *Pleea* and *Potamogeton natans* as the obtained exon sequences were too short or lacking. In all comparisons, missing data and gaps were removed only in pairwise comparisons. Error estimates were obtained by a bootstrap analysis with 1,000 replicates.

There is a possibility that the differences found in genetic distances are caused exclusively by differences in the editing pattern (C <–> T changes) between Paralog E and Paralog I sequences. To explore this, we performed a Pairwise Relative Ratio Test using *Tofieldia* as outgroup and constraining the synonymous substitution rate, which is less influenced by editing. *P* values were obtained by a likelihood ratio test. All the analyses were performed in HyPhy (Kosakovsky Pond et al. 2005) by the simplest codon model (Goldman and Yang 1994).

Phylogenetic Mapping of Paralogs

In order to reconstruct the history of the duplication event(s), we produced a combined gene tree for Alismatales using sequences from five mitochondrial genes (*atp*1, *cob*, *nad*5, *ccm*B, and *mtt*2) published elsewhere (Petersen et al. 2006; Cuenca et al. 2010). A few additional sequences of *nad*5, *mtt*2, and *ccm*B were generated following the protocol of Cuenca et al. (2010) (Table S2, supplementary material). Maximum likelihood (ML) searches were performed in GARLI version 0.96b8 (Zwickl 2006) using the general time-reversible (GTR) substitution model with empirical base frequencies and program estimates of the proportion of invariant sites and the shape of the rate heterogeneity distribution. Five initial runs of GARLI were performed to insure that the same topology was achieved. MacClade version 4.08 (Maddison and Maddison 2005) was used to map the absence/presence of intron (Paralog E) on the phylogenetic tree of the Alismatales.

Test of Horizontal Transfer

In order to estimate whether Paralog I sequences could possibly have originated by horizontal gene transfer, we performed a phylogenetic analysis including our own intron sequences and nad1i477 sequence data from GenBank. The analysis included a total of 71 taxa from 18 angiosperm orders (GenBank accession numbers available upon request). Given the large length variation, no unambiguous alignment of the whole intron was possible. The alignment used for phylogenetic analysis thus included the first 805 nucleotides of 5′ region of the intron, including domains I, II, III, and the first 11 to 19 nt of the domain IV, in addition to 58 nt including domains V and VI as well as the stem of domain IV [domain assignment following Michel and Ferat (1995)]. The largest part of the domain IV needed to be excluded and only an area of ca. 78 nt was readily alignable and included in the dataset. We also excluded: (1) an insertion of 25 to 50 nt present between domains ID(i) and ID(ii) for some members of Alismatales, (2) a 12 to 7-nt region just after domain ID(ii), (3) domain D3(ii) which can only be aligned with considerable doubt in the Alismatales, and (4) 89 characters were excluded in the loop of domain II where a high number of indels were found. The alignment is available upon request. In total, the dataset included 930 characters. Phylogenetic analysis was carried on by ML using the same strategy than for the five mitochondrial genes data set (see above) and an estimate of support was obtained by 100 bootstrap replications.

A similar test of the origin of Paralog E sequences was straightforward with regard to sequence alignment, but since the number of characters is considerably smaller, the result of the phylogenetic analysis is highly spurious. However, we performed a phylogenetic analysis including all Alismatales *nad*1 exon sequences (Paralog E sequences + exon regions of Paralog I sequences) and 130 additional angiosperm *nad*1 sequences (GenBank accession numbers available upon request). If present, the nad1i477 intron was removed manually. ML analyses were done by PhyML v. 3.0 (Gindon and Gascuel 2003) following the JC69+G, to avoid over-parameterization given the size of the data set.

Results

Table 1 shows the occurrence of intron-bearing (Paralog I) and intron-lacking (Paralog E) *nad*1 sequences within Alismatales.

Evidence That Paralog E is a Processed Paralog

Pattern of Post-transcriptional Editing

We have found no more than one edited site in the 40 pb of exon B and four to five in the 90 pb of exon C (Fig. 1 and

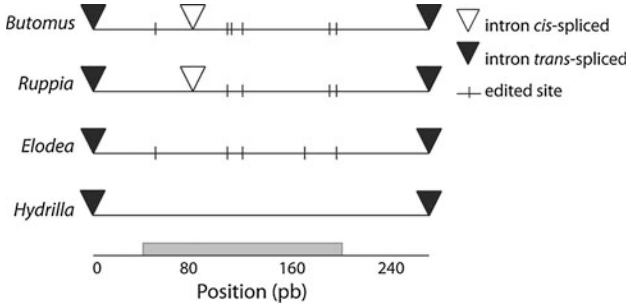


Fig. 1 Intron presence/absence and editing status *nad*1 paralogs. Graphical representation of the complete exon B and C of *nad*1, based on the sequence of *Butomus umbellatus*. *Solid* and *open triangles* represent *trans* and *cis*-spliced introns, respectively. *Vertical lines* represent edited sites found by comparison between DNA and cDNA sequences. The *gray box* indicates the area that was sequenced and included in the analyses

Table 2 Average uncorrected genetic distances of the partial exon sequences of *nad*1

	No. taxa	Average P_{DIST}	s.e[a]
Comparisons among taxa			
Among all Paralog I sequences	31[b]	0.005	0.002
Among all Paralog E sequences	26[c]	0.037	0.009
Paralog E among taxa where two paralogs co-exist	12[c]	0	0
Paralog E among taxa with only Paralog E present	14	0.063	0.057
Comparisons within taxa			
Paralog I vs. Paralog E	12[c]	0.088	0.026
Comparisons among paralogs			
Paralog I vs. Paralog E (all)	31 vs. 26	0.076	0.021
Paralog I vs. Paralog E			
When only Paralog E is present	31 vs. 14	0.063	0.017
When Paralog I and Paralog E co-exist	31 vs. 12	0.088	0.026
Paralog E single copy vs. double copy	14 vs. 12	0.051	0.013

[a] Standard error estimates were obtained by bootstraping (1,000 replicates) and missing data were removed in pairwise comparisons only

[b] The two sequences of *Acorus* (outgroup) were removed from the analysis, together with sequences of *Pleea* and *Potamogeton natans*, because exon sequence was too short or almost lacking

[c] Paralog E sequence of *Blyxa* was removed from the analysis due to alignment problems in exon C

Fig. S1 in supplementary online material). In contrast to this, most Paralog E sequences either lack editing completely or have lower editing frequency. Exceptions to this pattern are found in the sequences of *Elodea*, *Egeria*, and *Caldesia*, with five, four, or three edited sites, respectively, thus being more similar to Paralog I than to Paralog E sequences.

Extent of the Retroprocessing

The missing intron, in association with the absence or low frequency of edited sites, in the majority of the Paralog E sequences strongly supports a hypothesis of the presence of a processed paralog of *nad*1. PCR amplifications of larger *nad*1 regions including other exons (exons A, D, and E), which are all *trans*-spliced in the mitochondrial genome of *Butomus umbellatus* (data not shown), were unsuccessful indicating that if Paralog E was created by the insertion of a processed paralog, this occurred after *cis*-splicing of nad1i477, but before *trans*-splicing of the remaining *nad*1 introns.

Sequence Comparison Between Paralog I and Paralog E

Uncorrected genetic distances (P_{DIST}) between paralogs I and E in the same individual ranged from 0.078 to 0.098 substitutions per site, with an average $P_{DIST} = 0.088$ (Table 2). In comparison, the average P_{DIST} between two Paralog I sequences is only 0.005. Not a single base difference was found between the Paralog E sequences in the 12 species with two copies of *nad*1, with the exception of *Blyxa,* where the Paralog E sequence seems to be degenerated. In contrast, Paralog E sequences are considerably more variable when present as the only copy of *nad*1, with an average P_{DIST} value of 0.063 (Table 2). In addition, a relative rate test was performed for each possible pair of taxa between Paralog E and I sequences using *Tofieldia* as outgroup. In general, sequences of Paralog I show a significant different dS than sequences of Paralog E, both when they occur in isolation ($P = 0.01$–0.05) and when the copies coexist ($P = 0.001$–0.01).

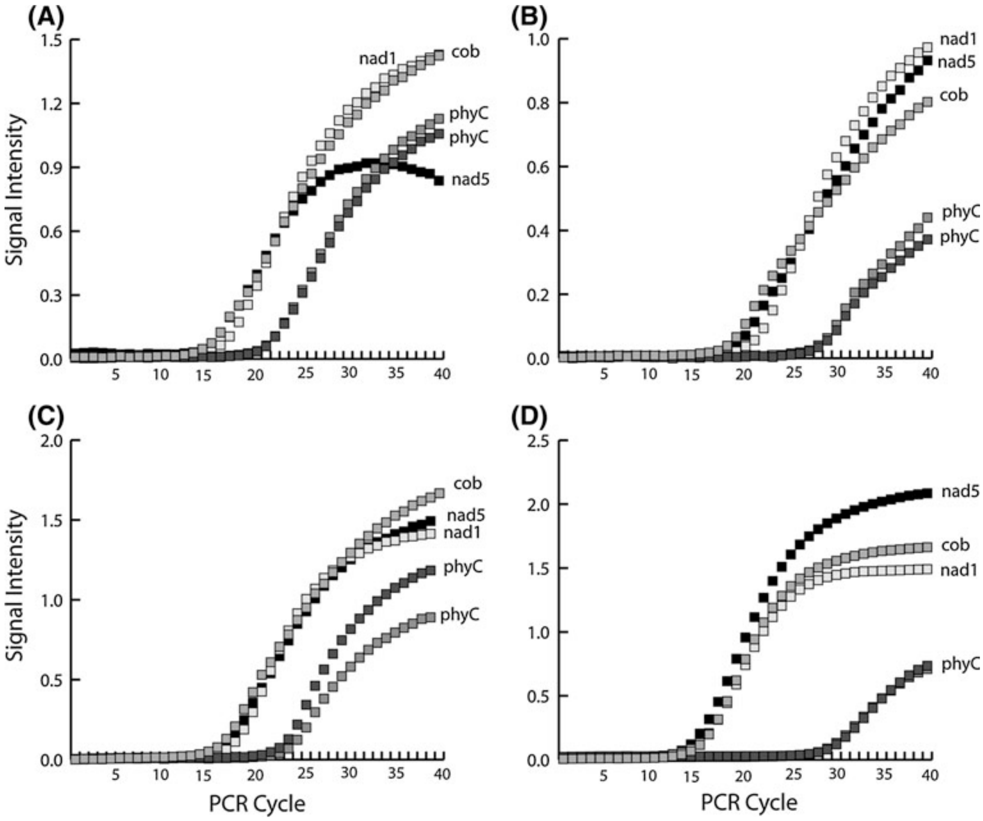
Genome Location of Paralog E

The qPCR amplification curves for mitochondrial genes are expected to differ from curves for nuclear genes due to the proportionally larger fraction of mitochondrial genomes in a sample of total DNA. Our qPCR tests of two mitochondrial genes (*nad*5 and *cob*) and one nuclear gene (*phy*C) confirm this expectation (Fig. 2). In all qPCR tests, the *nad*1 Paralog E behave as expected for a mitochondrial gene, with a clear difference in copy number compared to the nuclear gene *phy*C (Fig. 2). As nuclear sequences are largely unavailable for the Alismatales, no other taxa and/or regions could be included in the analyses. Thus, we cannot completely rule out that Paralog E sequences of other taxa than the four tested could be located in the nucleus.

Evolutionary History of Paralogs I and E, and Exploring Horizontal Transfer

Once the gene tree of the Alismatales is constructed using five other mitochondrial genes (*atp*1, *cob*, *ccm*B, *nad*5, and

Fig. 2 Genomic location of Paralog E. Quantitative PCR was used to amplify Paralog E, together with two mitochondrial loci (*cob* and *nad*5) and one nuclear locus (*phy*C). Genomic DNA was used as template and all amplifications were run in duplicate (only shown for *phy*C). The *four panels* indicate results for different taxa: **a** *Elodea canadensis*, **b** *Zostera noltii*, **c** *Echinodorus uruguayensis*, and **d** *Potamogeton natans*



*mtt*2) the occurrence of Paralog I can be optimized unambiguously, indicating the presence of three losses during the evolution of the Alismatales. In contrast to this, the presence of Paralog E could not be unambiguously optimized. ACCTRAN optimization, which force character state changes to occur as early as possible on the branches, indicates that a duplication creating Paralog E occurred once or possibly twice followed by six losses (Fig. 3). In contrast, DELTRAN optimization, which force character state changes to occur as late as possible on the branches, indicates up to five duplication events and three subsequent Paralog E losses. The number of times Paralog E is inferred to have been lost may be overestimated, as we were unable to obtain readable sequences from a few taxa.

No evidence was obtained that either Paralog I or Paralog E should have arisen through horizontal transfer from taxa outside the alismatids (Alismatales excluding Araceae and Tofieldiaceae). Although our phylogenetic analyses based on Paralog I recovered Alismatales as paraphyletic with respect to the remaining monocotyledons; not sister to them as in most recognized angiosperm phylogenies (APG-III, Bremer et al. 2009) (Fig. S2 supplementary materials). The short *nad*1 exon sequences display only little variation and the phylogenetic tree is as expected fairly unresolved, and not unexpectedly it includes some odd clades compared to the angiosperm phylogeny as generally perceived (Fig. S3 in supplementary materials). However, a clade formed by the alismatid Paralog E sequences is clearly placed together with other *nad*1 sequences from species of Alismatales. Even though this result needs to be taken with great caution (considering the low support of each clade, and the problems of reconstructing a phylogeny with so few data), it lends no support to the hypothesis that Paralog E sequences were derived through horizontal transfer.

Discussion

Evidence of the Existence of a Processed Paralog for *nad*1

Gene duplications in plant mitochondria are generally due to large segmental duplications caused by recombination either between short repetitive sequences or between subgenomic circles and/or plasmids present within the mitochondrion (Knoop 2004; Handa 2008). However, we found no evidence that duplication of exons B and C of *nad*1 is associated with recombinant regions. In addition, besides the obvious difference that one paralog has an intron and the other does not, the sequences of the two paralogs are different. In fact, sequence divergence between the exons of Paralog I and E in a single individual is greater than the divergence found among sequences of any of the paralogs when taxa from different plant families are compared. In

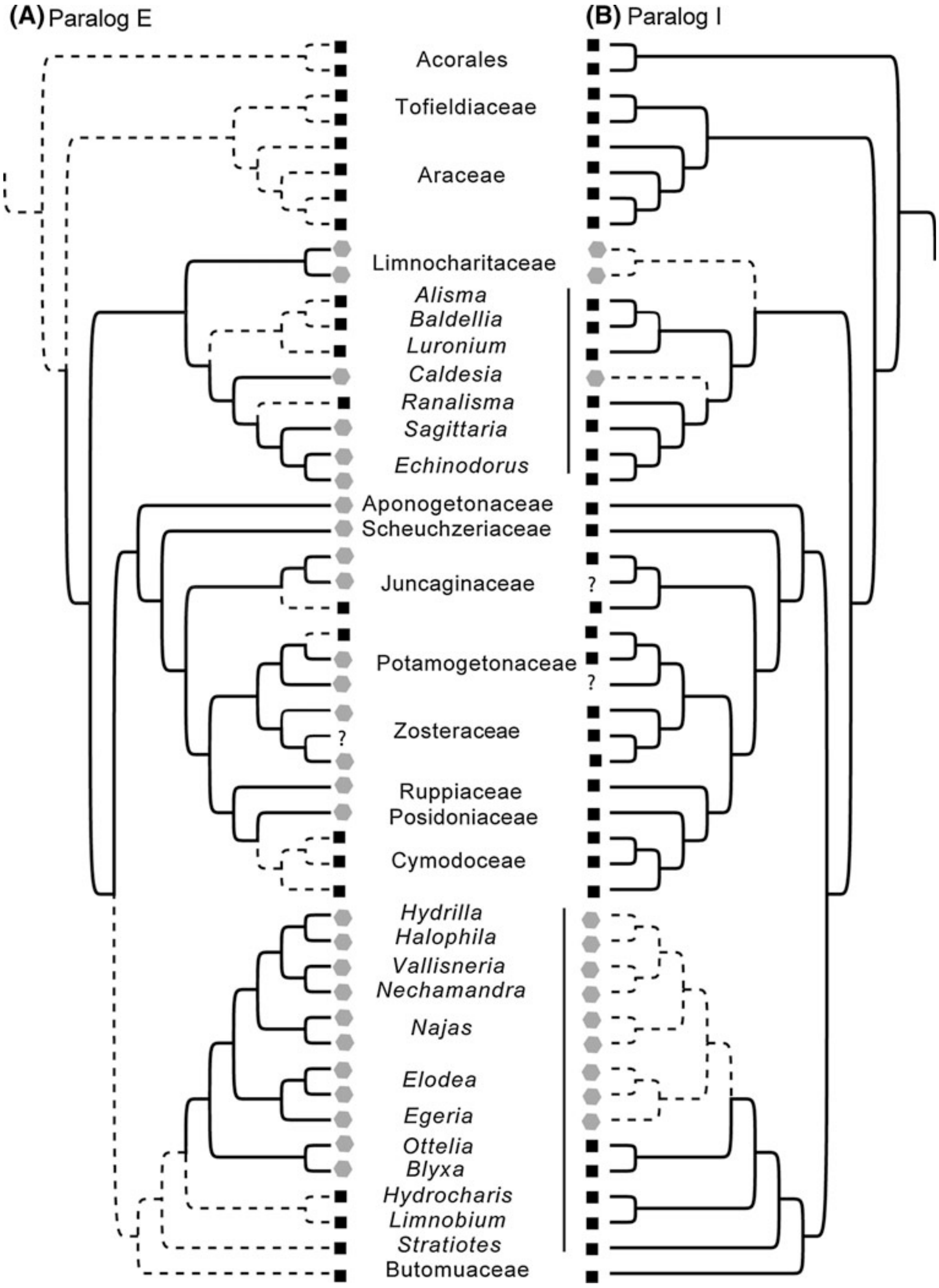


Fig. 3 Mapping of the presence/absence of paralogs E and I. ACCTRAN mapping of the presence/absence of the Paralog E (**a**) and the Paralog I (**b**). The presence is indicated by *continuous lines* and absence by *dashed lines*. *Black squares* represent the presence of Paralog I and *gray circles* the presence of Paralog E. Taxon names are given by families, except in the Alismataceae (indicated by the *vertical lines* in the upper part of the tree) and in the Hydrocharitaceae (indicated by *vertical lines* in the lower part of the tree), where generic names are used (for a complete taxon list see Table 1)

addition, neither of the copies shows evidence of being silenced (e.g., the presence of stop codons or frame shift mutations).

As previously mentioned, the two main signatures of a processed paralog are lack of introns coupled with a reduction in or a total lack of RNA editing. The presence of the nad1i477 intron is the most obvious difference between paralogs I and E, but their editing frequencies are also different. The lack of edited sites may be caused by an accelerated substitution rate for one of the *nad*1 copies, as suggested for other mitochondrial regions (Parkinson et al. 2005; Cuenca et al. 2010); however, this alternative fails to explain the associated loss of the *nad*1 intron. In cases where paralogs I and E co-exist in the mitochondrial genome, four to five edited sites are conserved in Paralog I, whereas Paralog E has lost all edited sites. The editing frequency of Paralog E is slightly different in taxa where both copies of *nad*1 are present, compared to taxa that only posses Paralog E. In the latter case, lack of or strong reduction in editing is usually found, except in *Egeria* and *Elodea*, where editing frequency is similar to that in Paralog I rather than to Paralog E. Not only their editing frequency, but also the sequences of *Egeria* and *Elodea* differ from those of the other Paralog E sequences. Thus, their uncorrected genetic distance when compared against other Paralog E sequences is $P_{\mathrm{DIST}} = 0.072 \pm 0.018$. This strongly suggests that the loss of the intron in *Egeria* and *Elodea* occurred independently of the loss that created the remaining Paralog E sequences. Indeed, the sequences of *Egeria* and *Elodea* are more similar to the Paralog I sequences ($P_{\mathrm{DIST}} = 0.02 \pm 0.009$), perhaps reflecting a secondary loss of nad1i477 in these taxa. The mechanism mediating this secondary intron loss is not clear, since nad1i477 was removed without any change in editing frequency. Thus, it remains unclear whether secondary intron loss could be due to recombination or insertion of a partially processed paralog or caused by some completely different mechanism.

The co-existence of paralogs I and E in the mitochondrial genomes is shown here in several taxa within Alismatales. Whereas the loss of intron nad1i477 has occurred independently in a number of angiosperms (Gugerli et al. 2001; Bakker et al. 2006), the co-existence of gene copies with and without this intron in the mitochondrial genomes has never been shown previously. This makes it most likely that the *nad*1 duplication occurred in the ancestor of twelve families within the Alismatales, and that one of the copies was secondarily lost repeatedly and independently (Fig. 3). However, it is possible that the number of taxa having both paralogs may be underestimated in our analyses. In cases where sequences quality was low as primers placed in the exons were amplifying more than one region, we used specific primers placed in the intron to sequence and/or amplify Paralog I. This strategy was obviously not applicable for Paralog E, thus some Paralog E copies may have gone unnoticed.

In addition to lack of introns and lost or reduced editing, changes in substitution rate have been proposed as an added characteristic of processed paralogs (Bowe and dePamphilis 1996). Changes in substitution rate are particularly obvious in processed paralogs inserted into the nuclear genome, which has an accelerated substitution rate compared to mtDNA. However, it is completely unclear what happens if a processed paralog is reinserted into the mitochondrial genome. Reverse transcriptase is an error-prone enzyme, and it has been suggested that this alone introduced mutations in the cDNA copy, which then erroneously are interpreted as evidence of an accelerated substitution rate (Parkinson et al. 2005). Even if this is true, an apparent change in substitution rate will only be observed in comparison between a processed paralog and the original copy. It will not affect comparisons involving orthologs of the processed paralog, only. Once the processed paralog with its possible cumulated mutations has been reinserted in the mitochondrial genome, the descendant copies will most probably share the original low substitution rate of most plant mitochondria. Therefore, this alone does not explain the accelerated synonymous substitution rate observed between different Paralog E sequences. The accelerated substitution rates are exclusively found in taxa where only Paralog E is present, consequently, we can rule out any hypothesis that this is caused by one of the copies being degenerated. One possible explanation is that most taxa showing an accelerated substitution rate of Paralog E also have a higher substitution rate in other mitochondrial genes, such as *mtt*2, *nad*5, *ccm*B, and *cob* (Cuenca et al. 2010); even though this accelerated substitution rate in other genes is not as pronounced as in *nad*1. Thus, the elevated substitution rates in these taxa may reflect processes affecting the entire mitochondrial genome, and are not restricted to the *nad*1 Paralog E. Still, we emphasize that the sequences of both paralogs of *nad*1 are quite short and advice caution in interpretation the results.

Further support for the hypothesis that Paralog E sequences are processed paralogs would be the presence of flanking target site duplications or direct repeats (Vanin 1985). As the present sequence data only include an internal fraction of what may assumed to be a larger retro-transcribed sequence, we do not present any such evidence here, but in future studies exploration of sequences flanking suggested processed paralogs will be recommendable.

Clearly, the hypothesis that one of the copies of *nad*1 found in this study is a processed paralog which relies heavily on an active reverse-transcriptase being present in the Alismatales mitochondria. Direct evidence of reverse-transcriptase activity in plant mitochondria has been shown in potato (Moenne et al. 1996), and a number of studies found evidence of the incorporation of reverse-transcribed RNA into other plant mitochondria (Geiss et al. 1994; Petersen et al. 2006; Cuenca et al. 2010; Sloan et al. 2010). However, the origin of the reverse-transcriptase in plant mitochondria is unknown, and no reverse-transcriptase proteins are encoded in any mitochondrial genomes sequenced so far. Thus, the protein must be imported into the mitochondrion, though the dynamics of this process are still unknown.

Genome Position of Paralog E

Judged by the number of genes absent in their mitochondria (Adams et al. 2002), gene transfer from the mitochondrion to the nucleus occurs frequently in some groups within the Alismatales. Nevertheless, this is unlikely to be the case for *nad*1 as this in one of the five respiratory genes (together with *cob*, *cox*1, *nad*4, and *nad*5) that are considered universal to all mitochondrial genomes (Adams and Palmer 2003). It has been proposed that most of the genes involved in the respiratory chain have physical characteristics impairing their transfer, e.g., high hydrophobicity

(Popot and de Vitry 1990; Adams and Palmer 2003). Proteins encoded in the nucleus and synthesized in the cytoplasm could not be re-imported into the mitochondrion, which has been shown for both *cob* (Claros et al. 1995) and *cox*2 (Daley et al. 2002). Another hypothesis is that the expression of certain genes playing key roles in electron transport and energy coupling are quickly and directly regulated by the redox state (Race et al. 1999; Allen 2003), making their export to the nucleus highly inefficient. In accordance with expectation, our qPCR results also confirm a mitochondrial location of the *nad*1 sequences. Though we were unable to perform qPCR on all taxa having Paralog E it seems highly improbable that the accelerated substitution rates found in some Paralog E sequences are caused by their transfer to the nucleus. In addition, it seems exceedingly complicated to imagine a scenario in which the sequences orthologous to exons B and C are found in the nucleus, whereas the remaining three exons are found in the mitochondrion. In the unlikely event that a partial copy of *nad*1 was transferred to the nucleus (and a complete copy remained in the mitochondrial genome), it is to be expected that the transferred copy would quickly degenerate. Paralog E sequences do not show any sign of silencing, e.g., reading frame shifts or the presence of stop codons, and most of the nucleotide changes are synonymous. However, that does not necessarily mean that Paralog E is in use when both copies are present. In plants with the presence of both paralogs, mRNA sequences were obtained for Paralog I, only. This could be a methodological artifact, given the lack of paralog-specific primers, but it could also reflect a transcriptional preference for one of the copies.

Horizontal Transfer

As several examples of horizontal transfer have been shown to occur in the plant mitochondria, one of them including *nad*1 (Won and Renner 2003), the hypothesis that one of the *nad*1 copies was obtained by horizontal transfer needed to be explored, viz., that either of both paralogs have been acquired by horizontal transfer. However, we have found no compelling evidence that this is the case. Albeit our phylogenetic analyses of Paralog I recovered Alismatales as paraphyletic, the position of the order is largely in accordance with the generally recognized history of the angiosperms (Fig. S2). Even though our analysis of the flanking exons, including Paralog E sequences, is not able to resolve even the major angiosperm groups, the sequences of Paralog E are still grouped with other monocot species, more specifically with sequences of Paralog I of Alismatales (Fig. S3). Short patch gene conversion events could obscure the phylogenetic signal; however, given the very short Paralog E sequences a meaningful analysis of gene conversion between the copies cannot be performed. Thus, with no data to its support we consider the presence of Paralog E sequences to be caused by horizontal transfer very unlikely, although we cannot completely disregard the possibility.

Phylogenetic Implications

Mixing orthologous (homologous) and paralogous sequences in a phylogenetic context has well-known pitfalls, including an obvious potential to mislead phylogenetic hypotheses (Moritz and Hillis 1990). Theoretically, any direct duplication is a synapomorphy precisely at the level where it occurs, and unless gene conversion, concerted evolution, silencing, etc., interfere, both copies should track the same evolutionary history and create no problems if all copies are extracted and analyzed simultaneously. Strictly speaking, the problem stems solely from the fact that the history and fate of the duplication are unknown and can only be inferred. The duplicated sequences cannot usually be separated a priori, and our inability to obtain a sequence from a sample is not necessarily a proof that it is not present, even though their positions in the genome most frequently reveal their history as duplications.

In contrast, processed paralogs carry distinct signatures of their own history which may be used to characterize them (e.g., no need for RNA editing and lack of introns) as duplications, and the directionality of the event is equally well-defined. However, a series of other potential problems surface, e.g., elimination of one or the other copies, transfer to the nucleus and hence accelerated substitution rates, and lack of editing, which makes the use of processed paralogous in a phylogenetic framework spurious at best (Bowe and dePamphilis 1996; Szmidt et al. 2001; Petersen et al. 2006; Duvall et al. 2008).

Acknowledgments The authors thank Charlotte Hansen and Hannah Blossom for skillful assistance in the lab, and the authors acknowledge the DNA Bank at the Royal Botanical Gardens of Kew, Jerrold I. Davis and Dennis W. Stevenson for supplying DNA samples. Financial support was provided by the Danish National Sciences Research Council Grant No. 272-06-0436.

References

Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29:380–395

Adams KL, Qiu YL, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci USA 99:9905–9912

Allen JF (2003) Why chloroplasts and mitochondria contain genomes. Comp Funct Genomics 4:31–36

Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang C, Du F, Xu D, Gibson M, Cifrese J, Clifton SW, Newton KJ (2007) Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177:1173–1192

Bakker FT, Breman F, Merckx V (2006) DNA sequence evolution in fast evolving mitochondrial DNA *nad*1 exons in Geraniaceae and Plantaginaceae. Taxon 55:887–897

Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD (2004) Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc Natl Acad Sci USA 101:17747–17752

Bonen L (2008) *Cis*- and *trans*-splicing of group II introns in plant mitochondria. Mitochondrion 8:26–34

Bonen L, Vogel J (2001) The ins and outs of group II introns. Trends Genet 17:322–331

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

Bremer B et al (2009) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121

Cho Y, Qiu YL, Kuhlman P, Palmer JD (1998) Explosive invasion of plant mitochondria by a group I intron. Proc Natl Acad Sci USA 95:14244–14249

Claros MG, Perea J, Shu Y, Samatey FA, Popot J-L, Jacq C (1995) Limitations to in vivo import of hydrophobic proteins into yeast mitochondria. Eur J Biochem 228:762–771

Clifton SW, Minx P, Fauron CM, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C, Meyer L, Wilson RK, Newton KJ (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol 136:3486–3503

Cuenca A, Petersen G, Seberg O, Davis J, Stevenson D (2010) Are substitution rates and RNA editing correlated? BMC Evol Biol 10:349. doi:10.1186/1471-2148-10-349

Daley DO, Clifton R, Whelan J (2002) Intracellular gene transfer: reduced hydrophobicity facilitates gene transfer for subunit 2 of cytochrome c oxidase. Proc Natl Acad Sci USA 99:10510–10515

Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129–134

Dombrovska O, Qiu YL (2004) Distribution of introns in the mitochondrial gene *nad*1 in land plants: phylogenetic and molecular evolutionary implications. Mol Phylogenet Evol 32: 246–263

Duvall MR, Robinson JW, Mattson JG, Moore A (2008) Phylogenetic analyses of two mitochondrial metabolic genes sampled in parallel from Angiosperms find fundamental interlocus incongruence. Am J Bot 95:871–884

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 243:97–105

Gindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Gugerli F, Sperisen C, Büchler U, Brunner I, Brodbeck S, Palmer JD, Qiu Y-L (2001) The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. Mol Phylogenet Evol 21:167–175

Handa H (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. Nucleic Acids Res 31: 5907–5916

Handa H (2008) Linear plasmids in plant mitochondria: Peaceful coexistences or malicious invasions? Mitochondrion 8:15–25

Knoop V (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet 46:123–139

Kosakovsky Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679

Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T (2000) The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNA-Cys (GCA). Nucleic Acids Res 28:2571–2576

Liu SL, Zhuang Y, Zhang P, Adams KL (2009) Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. Mol Biol Evol 26:875–891

Maddison WP, Maddison DR (2005) MacClade 4: analysis of phylogeny and character evolution. Version 4.08a

Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. Version 2.72

Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64:435–461

Moenne A, Begu D, Jordana X (1996) A reverse transcriptase activity in potato mitochondria. Plant Mol Biol 31:365–372

Moritz C, Hillis DM (1990) Molecular systematics: context and controversies. In: Hillis DM, Moritz C (eds) Molecular systematics. Sinauer Ass, Sunderland, MA, pp 1–10

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, 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 268:434–445

Nugent JM, Palmer JD (1991) RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell 66:473–481

Ogihara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, Miyashita N, Nasuda S, Nakamura C, Mori N, Takumi S, Murata M, Futo S, Tsunewaki K (2005) Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res 33:6235–6250

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

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

Popot J, de Vitry C (1990) On the microassembly of integral membrane proteins. Ann Rev Biophys Biophys Chem 19:369–403

Race H, Herrmann R, Martin W (1999) Why have organelles retained genomes? Trends Genet 15:364–370

Schuster W, Brennicke A (1994) The plant mitochondrial genome: physical structure, information content, RNA Editing, and gene migration to the nucleus. Ann Rev Plant Physiol Plant Mol Biol 45:61–78

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 185:1369–1380

Szmidt AE, Lu MZ, Wang XR (2001) Effects of RNA editing on the coxI evolution and phylogeny reconstruction. Euphytica 118: 9–18

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Ann Rev Genet 19:253–272

Won H, Renner SS (2003) Horizontal gene transfer from flowering plants to *Gnetum*. Proc Natl Acad Sci USA 100:10824–10829

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas, Austin