Effects of RNA Editing and Gene Processing on Phylogenetic Reconstruction

L. Michelle Bowe and Claude W. dePamphilis
Vanderbilt University

RNA editing is a ubiquitous phenomenon affecting most mitochondrial and chloroplast, and some nuclear genomes, where mutations in genomic DNA are "corrected" in the mRNA during transcriptional processing. Most editing in plants and animals corrects T-to-C substitutions at nonsynonymous first or second base positions, and the overall effect is an mRNA and protein sequence that differs from that predicted by the DNA. It has been suggested that genomic sequences that undergo editing should not be used in phylogenetics. We contend that editing will have little or no effect on DNA-based phylogenetic reconstruction because it is an intrinsic transcriptional process that does not affect the historical information in the DNA sequence. The only effect of editing on protein-coding DNA should be an increase in the rate of T-to-C transitions. Here we test the effects of RNA editing on phylogenetic reconstruction, using two data sets with high levels of editing, plant *coxII* and *coxIII*. Even with high levels of editing, phylogenies based on DNA and edited mRNA are virtually identical. The two types of sequences should not be used in the same analysis, however, because the particular forms of the gene will tend to group together. We also examine the effects of processed paralogs—a term proposed for mRNA sequences that are reverse transcribed and reinserted into the genome as intact gene sequences. Processed paralogs result in a distinct and underappreciated source of conflict among gene trees because of RNA editing. Analyses with unidentified processed paralogs may yield incorrect phylogenies, and the sequences may evolve at different rates if the gene has been transferred from one genetic compartment (nuclear, mitochondrial, chloroplast) to another. Although RNA editing itself is not a problem in phylogenetic reconstruction, analyses should not combine mRNAs with DNAs, and processed paralogs should be either excluded or analyzed with caution.

Introduction

Geneticists and phylogeneticists have generally relied on the assumptions that genomic DNA is the source of stored historical information, that cDNA and genomic DNA have the same sequences, and that given a specific gene at a specific locus the DNA sequences will be homologous. However, the recent discovery of RNA editing has radically changed our impression of how information is stored in DNA. RNA editing in plants is a natural transcriptional process where (usually) certain C's in the DNA are changed to U's in the mRNA, making the translated sequence different from the original DNA sequence (Araya, Begu, and Litvak 1994; Yokobori and Pääbo 1995).

Editing has been well documented in chloroplast genes (see reviews by Gray and Covello 1993; Araya, Begu, and Litvak 1994; Schuster and Brennicke 1994), trypanosome, invertebrate, mammalian (Blum, Bakalara, and Simpson 1990; Arts et al. 1993; Hajduk, Harris, and Pollard 1993; Yokobori and Pääbo 1995), and plant mitochondrial protein-coding and tRNA genes (Araya, Begu, and Litvak 1994). RNA editing occurs in all known vascular plant groups (Hiesel, Combettes, and Brennicke 1994; Hiesel, von Haeseler, and Brennicke 1994), and is hypothesized to be a primitive process in parasitic protists (Landweber and Gilbert 1994). Although phylogenetic patterns of editing have not yet been fully determined in plants, highly edited mitochondrial sequences are being used in phylogenetic analysis (Hiesel, von Haeseler, and Brennicke 1994; Wilson et al. 1994).

Orthologous sequences are useful for tracing organismal phylogeny because they diverged as species diverged and can be traced to a common ancestor; paralogous sequences, however, should be avoided because they can be traced to duplication events, and do not share the same evolutionary history (Fitch 1970). RNA editing and processed paralogs alter our traditional perception of homology because cDNAs and DNAs are not identical if RNA editing occurs in any of the sequences used and because duplicated and reinserted sequences are not orthologous to the other sequences.

The question remains as to how the two types of sequences will behave in phylogenetic analysis. Hiesel, von Haeseler, and Brennicke (1994) assert that cDNAs, not genomic DNAs, should be used in analysis of plant mitochondrial gene sequences because protein sequences are predicted by the cDNA. Hiesel, von Haeseler, and Brennicke (1994) also assert that the pattern of RNA editing may not be consistent among different lineages, and that genomic DNA does not evolve "reliably enough" to use for phylogenetic analysis. However, Hiesel, von Haeseler, and Brennicke (1994) only showed cDNA trees, and the actual effects of editing on phylogenetic analysis have not yet been examined.

To address this question, we will (1) briefly discuss the mechanism of RNA editing and methods of detecting edited sites, (2) discuss processed paralogy, (3) test potential problems with two data sets that have been influenced by RNA editing (*coxIII* and *coxII*), and (4) discuss treatment of edited sequences in phylogenetic reconstruction. With each data set, we compare tree topology, strength of the phylogenetic hypothesis, and long branch problems of mRNA and DNA parsimony

Key words: RNA editing, RNA processing, phylogeny, paralogy, *coxII, coxIII,* mitochondrial DNA, gene transfer.

Address for correspondence and reprints: L. Michelle Bowe, Vanderbilt University, Station B Box 1812, Nashville, Tennessee 37235. E-mail: bowelm@ctrvax.vanderbilt.edu.

Mol. Biol. Evol. 13(9):1159–1166. 1996

Table 1
Number of Known Edited Sites in *coxIII* and *coxII* for Each Taxon, Based on Direct Comparison of DNA and cDNA Sequences

	No. of Edited Sites	% Edited Sites
coxIII (381 bp)	52	13.6
Seed plants	37	9.7
Angiosperms	9	2.4
Monocots	6	1.6
Triticum sativum	6	1.6
Zea mays	5	1.3
Dicots (*Oenothera berteriana*)	8	2.1
Gymnosperms	35	9.2
Conifer (*Picea abies*)	16	4.2
Ginkgo (*Ginkgo biloba*)	20	5.2
Cycad (*Cycas revoluta*)	19	5.0
Seedless vascular plants	32	8.4
Ferns	17	4.5
Asplenium nidus	8	2.1
Osmunda claytoniana	13	3.4
Psilotophyte (*Psilotum nudum*)	10	2.6
Spenophyte (*Equisetum arvense*)	10	2.6
Lycophytes	1	0.26
Lycopodium squarrosum	1	0.26
Selaginella elegans	0	0.0
coxII (792 bp)	35	4.4
Angiosperms	35	4.4
Monocots	13	1.6
Triticum sativum	9	1.1
Zea mays	12	1.5
Dicots	18	2.3
Oenothera biennis	16	2.0
Pisum sativum	8	1.0

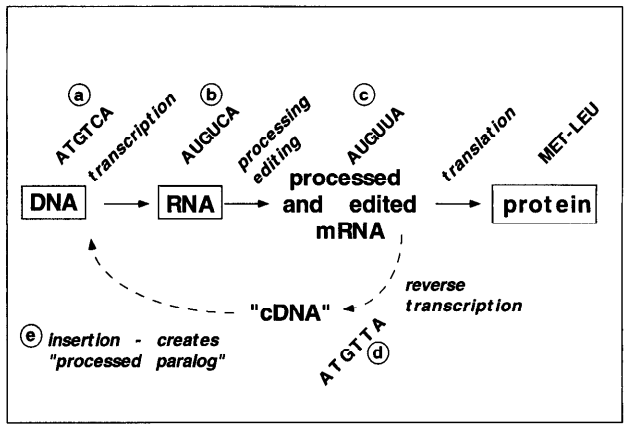

FIG. 1.—Cartoon showing the steps of transcription including RNA editing and processed paralogy. *a,* A genomic DNA sequence that would code for Met-Ser. *b,* The unprocessed mRNA sequence. *c,* The sequence after normal transcriptional processing and RNA editing: introns are removed and the C of the serine codon is edited to U. The translated sequence is now Met-Leu. *d,* Occasionally, reverse transcription occurs, creating a cDNA sequence with the edited codons. *e,* If the cDNA is inserted back into the genome, it becomes a processed paralog.

trees. Our experiments were designed to address the following questions in the context of RNA editing. (1) Is there a difference between cDNA and genomic DNA phylogenies? (2) If only one taxon is edited, will it change the resulting parsimony tree, and will its branch length be noticeably longer? (3) Will two edited taxa cluster together in a parsimony tree, and does it matter which two are edited? (4) What is the effect of combining cDNAs and genomic DNAs arbitrarily? (5) What is the effect of processed paralogs on phylogenetic analysis?

RNA Editing and Processed Paralogy

In RNA-edited genomes, editing is usually required to preserve highly conserved amino acids or loop structures; consequently, in protein-coding sequences, mostly first and second bases in a codon are affected (Covello and Gray 1990). We observed that in plant *coxI, coxII* and *coxIII* sequences the most common amino acids that are preserved by editing are phenylalanine, leucine, serine, and tryptophan; these and a few others were also named by Araya, Begu, and Litvak (1994) as the most commonly edited considering all plant mitochondrial genes. Araya, Begu, and Litvak (1994) also noted that most of these amino acids are hydrophobic and, with the exception of tryptophan (UGG), they can all be derived from C to U editing of proline (CCX) codons.

Although *coxII* and *coxIII* are edited in most plants, the largest number of edits appears to occur in the gymnosperms (Hiesel, Combettes, and Brennicke 1994). We found that 9.2% of the 391-bp sequence of *coxIII* studied by Hiesel, Combettes, and Brennicke (1994) is edited in gymnosperms, while only 2.4% is edited in the angiosperms (table 1). A total of 8.4% of the sequence is edited in the seedless plants, but only between 0.26% and 4.5% is edited in any given taxon. In *coxII,* a total of just 4.4% of the sites are edited, but all the available information is from angiosperms. The mitochondrial *coxI* sequence from *Thuja* also shows a high level of gymnosperm editing (Glaubitz and Carlson 1992).

In transcription, the antisense strand of genomic DNA is used as a template to produce an mRNA sequence identical to the DNA sense strand. During transcriptional processing, in addition to the usual events, RNA editing machinery changes some of the erroneous C's to U's (usually) in the mRNA (Sutton et al. 1991) (see fig. 1). Although the exact mechanism in plants is not known, guide sequences of RNA (gRNA) have been found in trypanosomes (Blum, Bakalara, and Simpson 1990; Arts et al. 1993; Hajduk, Harris, and Pollard 1993) that direct the editing. Consensus sequences for proposed gRNAs have been observed in both mitochondrial and chloroplast genes of a wide variety of plant groups (Gualberto, Weil, and Grienenberger 1990; Maier et al. 1992). During the course of this study, we observed the consensus sequence, ATATGGTTC, in the mitochondrial *Ginkgo, Oenothera, Triticum,* and *Zea coxIII* sequences published by Hiesel, Combettes, and Brennicke (1994) and in the chloroplast *Epifagus rps12* sequence (Ems et al. 1995). The terminal C in the consensus sequence is edited, and the consensus sequence is only found in the *coxIII* sequences that are actually edited; all of the other taxa have a T at the edited po-

sition and some of them do not have the consensus sequence that precedes the terminal T.

Obviously, the only way to be certain of RNA-edited sites in a sequence is to sequence the mRNA as well as the genomic DNA, but one can infer edited sites in several ways. Scanning the sequence and finding positions where all but one or two taxa have C's and T's will signal potential editing sites, especially if the sites are at first or second base positions. A comparison across many taxa can indicate which amino acids are the most highly conserved and may "need" editing if a base is changed (Covello and Gray 1990). Another option is to search for consensus sequences such as those identified by Gualberto, Weil, and Grienenberger (1990) and Maier et al. (1992) to find probable editing sites, but not all editing sites have been shown to contain these sequences. All of these methods are conservative and do not take into consideration recent lineages which have evolved other suitable amino acids for the position in question. The best method of inferring edits without knowing both the mRNA and DNA sequences is to compare the DNA sequence to related sequences for which the editing sites have already been determined.

If most editing sites in a given DNA sequence already have a T in the edited position, one might suspect a processed paralog. We use this term to refer to a sequence that was reverse transcribed and reinserted into the genome after RNA editing (Bowe and dePamphilis 1995) (fig. 1). The term "processed paralog" is similar to the terms "processed gene" and "processed pseudogene" (Li and Graur 1991, pp. 182–188), but refers to a functional gene in a phylogenetic framework: a processed paralog is not orthologous to most other copies of the same gene. Examples of plant processed paralogs can be found in soybean (Covello and Gray 1992), mungbean and cowpea (Nugent and Palmer 1991) *coxII, Selaginella coxIII,* and part of lettuce *nad4* (Geiss, Abbas, and Makaroff 1994) from the mitochondrial genome. Some other examples of gene transfer that may have occured via an mRNA intermediate are the *tufA* and *rbcS* chloroplast (now nuclear) genes (Baldauf and Palmer 1990) and primate genes that were transferred from mitochondrial to nuclear DNA (Collura and Stewart 1995; Zischler et al. 1995). Processed genes may be inserted into any part of the genome (chloroplast, mitochondrial, or nuclear), but insertion into a different compartment is especially problematic because the new sequence will evolve at a different rate. Plant nuclear sequences evolve much faster than plant chloroplast or mitochondrial sequences (Wolfe, Li, and Sharp 1987). The original, unedited sequence may remain as a functional or a silent copy, be deleted, or become a pseudogene (see the above references).

Methods

coxIII and *coxII*

We used 11 green plant sequences of a 381-nt portion of the mitochondrial *coxIII* gene from Hiesel, Combettes, and Brennicke's (1994) study: *Asplenium nidus, Cycas revoluta, Ginkgo biloba, Picea abies, Osmunda claytoniana, Equisetum arvense, Psilotum nudum, Lycopodium squarrosum, Marchantia polymorpha, Physcomitrella patens,* and *Oenothera berteriana* (GenBank accession numbers X76270–X76282). *Triticum aestivum* (Gualberto, Weil, and Grienenberger 1990; X52539) and *Zea mays* (X53055) mitochondrial *coxIII* sequences were also obtained from GenBank. The nuclear *Selaginella elegans* sequence (Hiesel, Combettes, and Brennicke 1994) was used as our *coxIII* processed paralog. Since the actual edit sites are shown in Hiesel, Combettes, and Brennicke (1994) or in the GenBank files, both DNA and cDNA sequences are available for each taxon.

Eleven *coxII* DNA sequences (792 base pairs long) were obtained from GenBank: *Beta vulgaris, Petunia hybrida, Glycine max, Pisum sativum, Vigna unguiculata, Daucus carota, Oenothera biennis, Oryza sativa, Zea mays, Triticum aestivum,* and *Psilotum nudum* (X55297, X17394, X04825, X02433, S48624, X63625, X00212, X01088, X52865, X01108, and X74310, respectively). The *Vigna* sequence is nuclear-encoded (Nugent and Palmer 1991), and there is both a nuclear (active) and a mitochondrial (silent) *Glycine* sequence. Sequences were aligned manually after removing one intron from *Zea, Triticum, Oryza, Petunia,* and *Beta,* and two from *Daucus.* cDNA sequences and editing sites were available for four taxa: *Pisum, Oenothera, Zea, Vigna,* and *Triticum.* We duplicated and changed the remaining sequences into "cDNAs" by comparing them to the known cDNA sequences and changing bases from C to T at all sites where editing is known to occur in the other taxa, with regard to differences between plant families (e.g., if grasses but not legumes are edited, an unknown grass C will be changed, but not an unknown legume C).

coxIII and *coxII* parsimony analyses were performed using PAUP's heuristic algorithm, with TBR branch swapping and 10 random stepwise addition sequence algorithms (Swofford 1993). To test for strength of the hypothesis, Bremer Support analyses (BRS; Bremer 1988) and 100 bootstrap replicates (Felsenstein 1985) were performed on each data set. In accordance with Davis (1995), we prefer the term "Bremer Support" to "decay index" because it more accurately reflects the meaning: higher scores mean higher support. BRS analyses were performed initially by obtaining trees up to five steps longer than the most parsimonious trees. Decay of the remaining nodes was found using inverse constraints in PAUP (Swofford 1993) as described by Johnson and Soltis (1994). To compare trees of the same taxa, we took the sum of the BRS values across all taxa as a measure of total Bremer Support, abbreviated here as TBS.

Results

Predictions

Because the edited sites in a DNA sequence represent potentially informative characters (the "edited" sites are allowed to mutate from T to C), we might predict that DNA trees will be more resolved than cDNA

Table 2
Characteristics of *coxIII* and *coxII* Sequences and PAUP Trees

Data Set	No. of Variable Positions	Figure	No. of Informative Characters	Analysis	No. of Trees	No. of Steps	CI	RI	TBS
coxIII....	367	2*a*	121	All genomics	1	344	0.683	0.668	57
		2*b*	97	All cDNAs	3	289	0.699	0.680	44
		2*c*	116	2 cDNAs and 11 DNAs	1	333	0.685	0.660	53
		2*d*	111	5 cDNAs and 8 DNAs	2	333	0.688	0.649	44
		2*e*	143	DNAs and *Selaginella*	2	447	0.667	0.610	47
			118	cDNAs and *Selaginella*	5	390	0.682	0.616	31
coxII	268	3*a*	77	All genomics	3	273	0.846	0.720	23
		3*b*	78	All cDNAs	3	272	0.846	0.731	24
		3*c*	127	DNA + Vign + Glyc/nuc	4	410	0.812	0.678	35
			85	2 cDNAs and 8 DNAs	1	279	0.835	0.712	27
			81	5 cDNAs and 5 DNAs	5	293	0.795	0.663	18
			99	DNA and *Vigna*	10	379	0.807	0.616	17
			95	cDNAs and *Vigna*	22	371	0.817	0.644	18

NOTE.—Abbreviations: CI = consistency index, calculated without autapomorphies; RI = retention index (Swofford 1993); TBS = total Bremer Support, the sum of BRS values across all nodes of a given tree.

trees. Also because of the information in the edited sites, the alternate forms of sequence (cDNAs, DNAs) may group together if they are mixed in an analysis. Plant nuclear processed paralogs are also expected to disrupt phylogeny because they will have evolved at a faster rate due to their presence in the nuclear genome (Wolfe, Li, and Sharp 1987). Increased rates of evolution in processed paralogs may result in their attraction to other long branches in phylogenetic analyses (Felsenstein 1978).

Effects of RNA Editing on Phylogeny Reconstruction—*coxIII*

With *coxIII,* we found only one most parsimonious tree using the genomic DNA sequences, and three most parsimonious trees with the cDNA data (fig. 2*a* and *b*). The loss of phylogenetic information in the cDNA sequences is also demonstrated by the number of informative characters—121 for the genomic data and 97 for the cDNA data (see table 2)—and the number of steps in each tree—344 vs. 289. A comparison of nodal support between the two trees reveals that the nodes that were upheld in the cDNA consensus tree had similar bootstrap values (fig. 2*a* and *b*) to the same nodes in the DNA tree. BRS values (Bremer 1988) were somewhat higher in the DNA tree, and TBS was much higher for the DNA tree (57 vs. 44) because unresolved nodes (in the cDNA tree) have a BRS of zero.

When one cDNA sequence was included in an analysis of DNA sequences or one DNA sequence was included in an analysis of cDNA sequences, the resulting trees were identical in topology to the genomic or cDNA trees, respectively, regardless of taxon choice (trees not shown). Inclusion of two cDNA sequences with the rest of the DNA data, however, did change the topology of the trees (fig. 2*c*). This result varied with taxon choice: if the two taxa were in the same original clade on the DNA tree, the cDNA sequences tended to group together. If the two taxa were originally in distant clades, such as one angiosperm and one gymnosperm, the cDNA sequences rarely grouped together. When cDNAs and DNAs were mixed arbitrarily in phylogenetic analyses, the resulting trees were fairly unresolved and differed from the genomic tree (fig. 2*d*; not all analyses are shown). The trees from mixed analyses also had lower TBS indices than the genomic tree (53 and 44 vs. 57; table 2).

Inclusion of the nuclear *Selaginella* sequence also resulted in lower resolution and unexpected tree topologies (fig. 2*e*). Since the nuclear sequence might contain the "edited" form of DNA, it was analyzed with the cDNAs as well as the genomic DNAs. We found two most parsimonious trees using the DNAs and five using the cDNAs. In both cases, *Selaginella* and *Asplenium* formed a clade. If the lycopsids are monophyletic, one would expect *Selaginella* to form a clade with *Lycopodium*; if not, one would expect it to be either basal to the rest of the vascular plants or basal to the vascular plants excluding *Lycopodium.* The phylogram (fig. 2*e*) shows that the *Asplenium* and *Selaginella* nodes are on long branches, suggesting a long branch attraction effect (Felsenstein 1978).

Results from *coxII*

The *coxII* data set consisted of nine angiosperms and *Psilotum* (the outgroup). Because mitochondrial genes evolve so slowly (Wolfe, Li, and Sharp 1987) and this was a more limited taxon set than that of *coxIII,* there were only 268 (out of 792) variable positions (see table 2). We found three most parsimonious trees each for DNAs and cDNAs (fig. 3*a* and *b*). The numbers of variable characters and support measures were similar between the two trees, but the dicot clade was better supported in the DNA tree than in the cDNA tree (bootstrap of 92, BRS of 5 vs. bootstrap of 86, BRS of 3).

As with *coxIII, coxII* trees did not differ in topology when only one cDNA was included in an analysis of mostly DNAs (and vice versa), but when two *coxII*

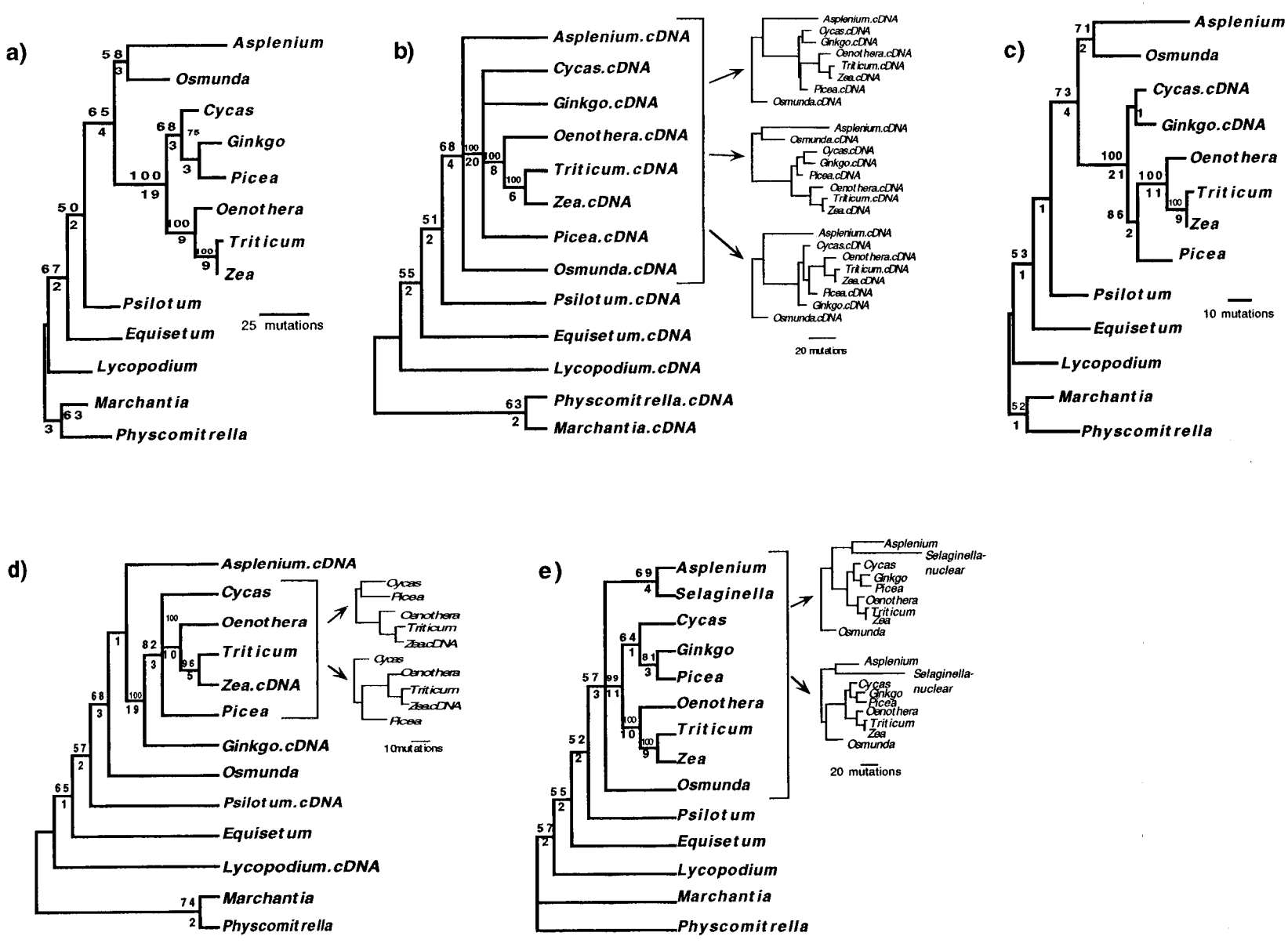

FIG. 2.—Parsimony analyses of *coxIII*. Bootstrap values are given above each branch and BRS below. Refer to table 2 for more information about each tree. *a*, Genomic DNAs. *b*, cDNAs: strict consensus of three trees on the left; the three alternate topologies with branch lengths on the right. *c*, Two cDNAs with genomic DNAs. *d*, An arbitrary mixture of DNAs and cDNAs: strict consensus of two trees on left; two alternate topologies with branch lengths on right. *e*, A processed paralog (*Selaginella*) in genomic DNA tree: strict consensus of two trees on left, the two alternate topologies with branch lengths on right.

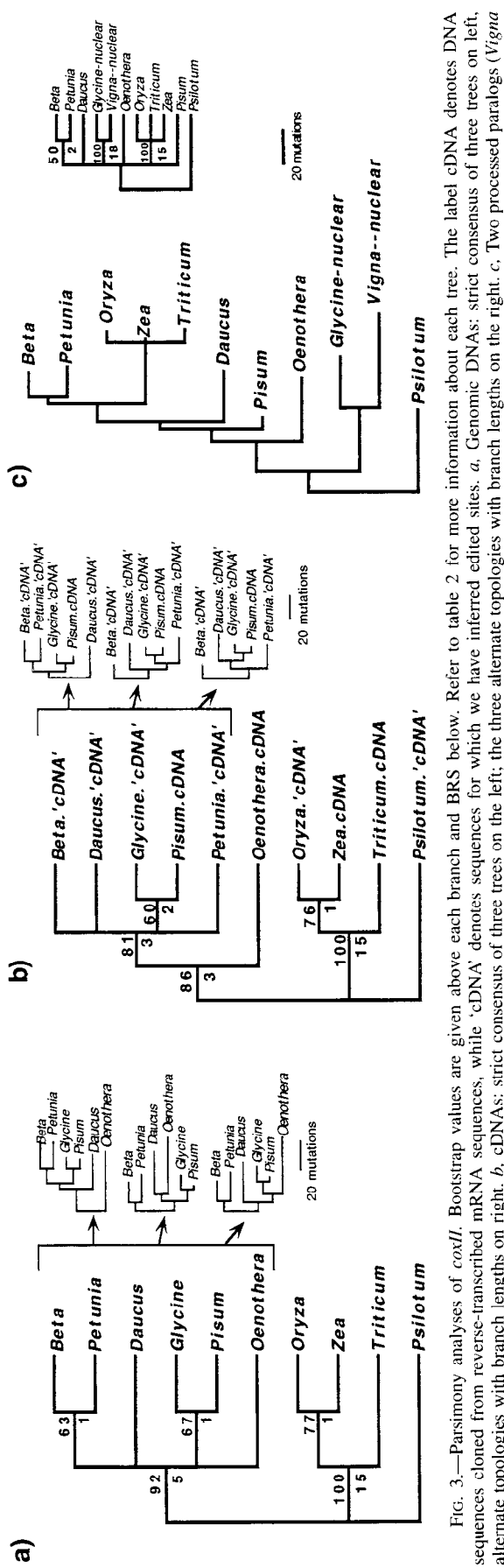

FIG. 3.—Parsimony analyses of *coxII*. Bootstrap values are given above each branch and BRS below. Refer to table 2 for more information about each tree. The label cDNA denotes DNA sequences cloned from reverse-transcribed mRNA sequences, while 'cDNA' denotes sequences for which we have inferred edited sites. *a*, Genomic DNAs: strict consensus of three trees on left, alternate topologies with branch lengths on right. *b*, cDNAs: strict consensus of three trees on the left; the three alternate topologies with branch lengths on the right. *c*, Two processed paralogs (*Vigna* and *Glycine*) in genomic DNA tree: one of four trees on left; strict consensus of four trees on right.

cDNAs were included in an analysis of mostly DNAs, the result varied with the taxa chosen: if the two were from the same main clade, they came together; however, if they were in different main clades (one monocot and one dicot, for example), the taxa did not form a clade. Also, when an arbitrary mixture of cDNAs and DNAs were analyzed, cDNAs grouped together.

Ten and 22 most parsimonious trees were found when the nuclear *Vigna* sequence was included in the DNA data set and the cDNA data set, respectively. The same result was found when the *Glycine* nuclear sequence was analyzed alone, and when both *Vigna* and *Glycine* were included, they formed a clade outside the legumes (fig. 3*c*).

Discussion

Most of our predictions were supported by the data: (1) DNA trees tend to be better supported than cDNA trees; (2) mixing the two types of sequences can wreak havoc on phylogenetic analyses; (3) plant nuclear processed paralogs tend to cause additional problems because of their long branches.

With the *coxIII* DNA data, we found a monophyletic gymnosperm and a monophyletic angiosperm clade as well as a fern clade sister to the seed plants, *Psilotum* sister to that clade, and then *Equisetum* (fig. 2*a*). Although one of the three most parsimonious cDNA trees fits this description, none was identical to the genomic DNA tree because the *Picea-Ginkgo* clade was not recovered. Our DNA results were similar to Hiesel, von Haeseler, and Brennicke's (1994) parsimony results from the *coxIII* cDNA data, but we felt that our choice of outgroup (the two bryophytes, *Marchantia* and *Physcomitrella*) was more appropriate because it was closer to the ingroup. Although our genomic DNA result was congruent to that presented by Hiesel, von Haeseler, and Brennicke (1994), we found three most parsimonious trees using the cDNA data.

Hiesel, von Haeseler, and Brennicke (1994) postulated that genomic DNA may not be "phylogenetically reliable" because RNA editing changes the sequence information during mRNA processing, but that cDNAs are useful because they are sequenced from mRNAs and predict the true protein sequence. However, editing may actually be a source of variation because certain substitutions may be allowed in otherwise conserved genomic DNA sequences, making edited sites effectively twofold degenerate. Consequently, because of the potential historical information carried in edited sites, cDNAs may provide fewer informative sites and therefore less resolution in a phylogenetic tree. Since we found little difference between cDNA and DNA trees other than greater resolution in DNA trees, our results appear to refute Hiesel, von Haeseler, and Brennicke's (1994) hypothesis that DNA sequences should not be used in phylogenetic analyses because they are "unreliable." Trypanosome sequences are extensively edited in a different way than plant sequences (Blum, Bakalara, and Simpson 1990), and we have not determined if their mRNA and genomic DNA phylogenies are the same. However,

Landweber and Gilbert (1994) describe trypanosome editing as "a novel source of frameshift mutations over evolutionary time," and more studies like theirs are needed to dissect the phylogenetic qualities of each type of sequence.

Mixing cDNAs and DNAs in phylogenetic analysis, however, can give confusing results. cDNAs will have T's at certain sites as a result of RNA editing, and these might be incorrectly interpreted as synapomorphies or symplesiomorphies in a phylogenetic analysis, depending on whether the sites are edited in the ingroup or in the outgroup. Our results varied with which taxa were chosen to be mixed in an analysis, indicating that the information contained in the edited sites can in some, but not all, cases outweigh the rest of the phylogenetic signal. In general, the more editing, the greater the chance that the artificially synapomorphic signal between two cDNAs will override the true phylogenetic signal. When possible, we recommend using either DNAs or cDNAs *but not both* in a given analysis. An alternative approach, if all editing sites are known, is coding edited sites as a fifth character, allowing the editing itself to contribute to the phylogenetic information. However, our experiments (trees not shown) revealed that trees produced from such five-character data are identical to genomic DNA trees. Another approach, especially when the available sequences are a combination of cDNAs and DNAs, is to leave out the sites that are known or suspected to be edited. Using only cDNAs should give approximately the same effect because many of the edited positions will be invariant.

Showing that DNA is reliable for uncovering phylogenetic relationships leads us to ask: What if some species lack editing capabilities, or some contain more editing sites than others? In protein-coding genes, edited codons (i.e., the amino acids) are highly conserved across many taxa and a position that "requires" editing will have either the conserved T, a C that is edited (Gray and Covello 1993), or a totally new codon. Therefore, a species without editing capabilities will most likely have a T at a given editing position, and in taxa with editing capabilities, a C may be substituted for T. There appears to be no reason that this type of substitution is any less reliable than any other synonymous substitution. Furthermore, as Hiesel, Combettes, and Brennicke (1994) demonstrated, most, if not all, vascular plants have editing capabilities, and gymnosperm sequences contain the most editing sites (table 1).

We hypothesized that if a single plant processed paralog was inserted into the nuclear genome from the mitochondrial, it might appear as a long branch and disrupt phylogeny, as Nugent and Palmer (1991) found, because plant nuclear sequences evolve faster (Wolfe, Li, and Sharp 1987) than chloroplast and mitochondrial sequences. In our analyses, the general topology of the *coxII* tree remained intact when the nuclear sequence from *Vigna* or *Glycine* was included, but *Vigna* (as a taxon) was not in its expected position—with the other legumes—and its branch is the longest on the tree. Nugent and Palmer (1991) hypothesized that this reflected a gene duplication event that preceded angiosperm diversification. In this case, the processed paralog was easy to identify because the tree topology did not reflect what is known of the phylogenetic history of the organisms; however, usually, tree topologies are not known at the outset, and a processed paralog may not be so obvious.

Processed paralogs might be detected by the following suite of characteristics: (1) absence of an intron in a gene that usually has an intron; (2) most known edited sites are already in edited form; (3) sequence has evolved at a different evolutionary rate, signaling insertion into a different genomic compartment; and (4) conflicting gene trees. If a processed paralog is a result of a recent event, or if reinsertion was into the same genomic compartment as the original sequence, its detection may be difficult if possible at all. In these cases, merely looking for long branches or unusual tree structure in a phylogenetic analysis is not sufficient, and at this time, it is unknown how often this has occurred. Although processed paralogy is an interesting phenomenon, it can be a major problem for phylogeneticists, and the only way to solve the problem may be to exclude the suspected sequences from the data set to determine their effects on tree topology.

While leaving a sequence out of an analysis may seem drastic, the consequences of including them are probably worse. If the reinsertion of a given sequence has occurred recently, the hazards of including that sequence in a phylogenetic analysis are similar to those of combining cDNAs and DNAs in the same analysis: the problem may be solved by either using only cDNA sequences in the analysis or by excluding edited sites. However, if the reinsertion event was relatively ancient and/or was into a faster- or slower-evolving genetic compartment, the sequence may have accrued more substitutions, potentially giving it a long branch and adding homoplasy to the data set.

In summary, we found no significant differences between using DNAs or cDNAs for phylogenetic reconstruction and no indication that genomic DNA is misleading in phylogenetic analyses. However, using a mixture of the two types of sequences can be troublesome. Although these *coxII* and *coxIII* data sets were limited in number of characters and taxa, we agree with Hiesel, von Haeseler, and Brennicke's (1994) conclusion that mitochondrial sequences will be useful for phylogenetic analyses, but we add that genomic DNAs are at least as useful as cDNAs for reconstructing phylogenetic events.

Acknowledgments

For helpful comments and discussion, we thank the following members of the Molecular Evolution Seminar group at Vanderbilt: Ned Young, Jim Leebens-Mack, Billie Swalla, Ölle Pellmyr, Andi Wolfe, and Gerry Moore. This research was partially supported by dissertation enhancement award (to L.M.B.), an N.S.C grant from Vanderbilt University and NSF grant DEB-9120258 (to C.W.D.). We thank Steve Wolfe for computer programming, and two anonymous reviewers for their comments and insights.

LITERATURE CITED

ARAYA, A., D. BEGU, and S. LITVAK. 1994. RNA editing in plants. Physiologia Plantarum **91**:543–550.

ARTS, G. J., H. VAN DER SPEK, D. SPEIJER, J. VAN DER BURG, H. VAN STEEG, P. SLOOF, and R. BENNE. 1993. Implications of novel guide RNA features for the mechanism of RNA editing in *Crithidia fasciculata*. EMBO J. **12**(4):1523–1532.

BALDAUF, S. L., and J. D. PALMER. 1990. Evolutionary transfer of the chloroplast *tuf*A gene to the nucleus. Nature **344**: 262–265.

BLUM, B., N. BAKALARA, and L. SIMPSON. 1990. A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. Cell **60**:189–198.

BOWE, L. M., and C. W. DEPAMPHILIS. 1995. Phylogenetic significance of RNA editing. Am. J. Bot. **82**:116.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution **42**:795–803.

COLLURA, R. V., and C.-B. STEWART. 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominids. Nature **378**:485–489.

COVELLO, P. S., and M. W. GRAY. 1990. Differences in editing at homologous sites in messenger RNAs from angiosperm mitochondria. Nucleic Acids Res. **18**:5189–5196.

———. 1992. Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome *c* oxidase (*cox2*) in soybean: evidence for RNA-mediated gene transfer. EMBO J. **11**(11): 3815–3820.

DAVIS, J. I. 1995. A phylogenetic structure for the monocotyledons, as inferred from chloroplast DNA restriction site variation, and a comparison of measures of clade support. Syst. Bot. **20**:503–527.

EMS, S., C. W. MORDEN, C. DIXON, K. WOLFE, C. W. DEPAMPHILIS, and J. D. PALMER. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant, *Epifagus virginiana*. Plant Mol. Biol. **29**(4):721–733.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. **19**:99–113.

GEISS, K. T., G. M. ABBAS, and C. A. MAKAROFF. 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.

GLAUBITZ, J. C., and J. E. CARLSON. 1992. RNA editing in the mitochondria of a conifer. Curr. Genet. **22**:163–165.

GRAY, M. W., and P. S. COVELLO. 1993. RNA editing in plant mitochondria and chloroplasts. FASEB J. **7**:64–71.

GUALBERTO, J. M., J.-H. WEIL, and J.-M. GRIENENBERGER. 1990. Editing of the wheat *coxIII* transcript: evidence for twelve C to U and one U to C conversions and for sequence similarities around editing sites. Nucleic Acids Res. **18**: 3771–3776.

HAJDUK, S. L., M. E. HARRIS, and V. W. POLLARD. 1993. RNA editing in kinetoplastid mitochondria. FASEB J. **7**:54–63.

HIESEL, R., B. COMBETTES, and A. BRENNICKE. 1994. Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. Proc. Natl. Acad. Sci. USA **91**:629–633.

HIESEL, R., A. VON HAESELER, and A. BRENNICKE. 1994. Plant mitochondrial nucleic acid sequences as a tool for phylogenetic analysis. Proc. Natl. Acad. Sci. USA **91**:634–638.

JOHNSON, L. A., and D. E. SOLTIS. 1994. *matK* DNA sequences and phylogenetic reconstruction in Saxifragaceae *s. str.* Syst. Bot. **19**(1):143–156.

LANDWEBER, L. F., and W. GILBERT. 1994. Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. Proc. Natl. Acad. Sci. USA **91**:8670–8674.

LI, W.-H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

MAIER, R. M., K. NECKERMANN, B. HOCH, N. B. AKHMEDOV, and H. KÖSSEL. 1992. Identification of editing positions in the *ndh*B transcript from maize chloroplasts reveals sequence similarities between editing sites of chloroplasts and plant mitochondria. Nucleic Acids Res. **20**(23):6189–6194.

NIXON, K. C., W. L. CREPET, D. STEVENSON, and E. M. FRIIS. 1994. A reevaluation of seed plant phylogeny. Ann. Mo. Bot. Gard. **81**(3):484–533.

NUGENT, J. M., and J. D. PALMER. 1991. RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. Cell **66**:473–481.

SCHUSTER, W., and A. BRENNICKE. 1994. The plant mitochondrial genome: physical structure, information content, RNA editing, and gene migration to the nucleus. Ann. Rev. Plant Phys. Plant Mol. Biol. **45**:61–78.

SUTTON, C. A., P. L. CONKLIN, K. D. PRUITT, and M. R. HANSON. 1991. Editing of pre-mRNAs can occur before *cis*- and *trans*-splicing in *Petunia* mitochondria. Mol. Cell. Biol. **11**: 4274–4277.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign, Ill.

WILSON, K. G., F. G. DONG, C. A. MAKAROFF JR., M. F. RABBI, A. BROCKHURST, and M. SABAT. 1994. Abstract. Evolution of the cox2 gene and introns in higher plants: can introns be used as biosystematic markers? Am. J. Bot. **81**(6):84.

WOLFE, K. H., W.-H. LI, and P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**:9054–9058.

YOKOBORI, S., and S. PÄÄBO. 1995. tRNA editing in metazoans. Nature **377**:490.

ZISCHLER, H., H. GEISERT, A. VON HAESELER, and S. PÄÄBO. 1995. A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. Nature **378**:489–492.

PAUL SHARP, reviewing editor

Accepted July 12, 1996