Using Secondary Structure to Identify Ribosomal Numts: Cautionary Examples from the Human Genome

Link E. Olson and Anne D. Yoder[1]

Department of Cell and Molecular Biology, Northwestern University Medical School, Chicago; and Division of Mammals, Field Museum of Natural History, Chicago

The identification of inadvertently sequenced mitochondrial pseudogenes (numts) is critical to any study employing mitochondrial DNA sequence data. Failure to discriminate numts correctly can confound phylogenetic reconstruction and studies of molecular evolution. This is especially problematic for ribosomal mtDNA genes. Unlike protein-coding loci, whose pseudogenes tend to accumulate diagnostic frameshift or premature stop mutations, functional ribosomal genes are not constrained to maintain a reading frame and can accumulate insertion-deletion events of varying length, particularly in nonpairing regions. Several authors have advocated using structural features of the transcribed rRNA molecule to differentiate functional mitochondrial rRNA genes from their nuclear paralogs. We explored this approach using the mitochondrial 12S rRNA gene and three known 12S numts from the human genome in the context of anthropoid phylogeny and the inferred secondary structure of primate 12S rRNA. Contrary to expectation, each of the three human numts exhibits striking concordance with secondary structure models, with little, if any, indication of their pseudogene status, and would likely escape detection based on structural criteria alone. Furthermore, we show that the unwitting inclusion of a particularly ancient (18–25 Myr old) and surprisingly cryptic human numt in a phylogenetic analysis would yield a well-supported but dramatically incorrect conclusion regarding anthropoid relationships. Though we endorse the use of secondary structure models for inferring positional homology wholeheartedly, we caution against reliance on structural criteria for the discrimination of rRNA numts, given the potential fallibility of this approach.

Introduction

Mitochondrial-derived nuclear pseudogenes (numts: Lopez et al. 1994) represent both boon and bane to molecular biologists (see reviews in Zhang and Hewitt 1996; Sorenson and Quinn 1998; Bensasson et al. 2001). They can provide valuable insight into ancestral mtDNA nucleotide states, relative substitution rates in the mitochondrial and nuclear genomes, and the history of major intergenomic translocation events (Fukuda et al. 1985; Lopez et al. 1994; Arctander 1995; Zischler et al. 1995). However, their unwitting inclusion in molecular studies can lead to spurious conclusions regarding phylogenetic relationships, molecular evolutionary dynamics, or both (Zhang and Hewitt 1996; Bensasson et al. 2001). Preferential polymerase chain reaction amplification of numts can occur (Arctander 1995; Collura and Stewart 1995; Sorenson and Fleischer 1996; Zhang and Hewitt 1996; Sorenson and Quinn 1998) and has even been reported in ancient DNA studies (van der Kuyl et al. 1995). This problem is particularly subtle for noncoding mtDNA sequences (Arctander 1995). For example, unlike protein-coding mtDNA loci, whose nuclear pseudogenes tend to accumulate diagnostic frameshift or premature stop mutations (e.g., Collura and Stewart 1997), functional ribosomal genes are not constrained to maintain a reading frame and can accumulate insertion-deletion (indel) events of varying length, particularly in nonpairing regions. Consequently, the identification of sequenced rRNA numts is less straightforward. Using examples from the human genome, we explore the effectiveness of using secondary structure models to identify rRNA numts and show that even relatively ancient numts that negatively affect phylogenetic inference can escape detection.

Several criteria have been proposed for discriminating between mtDNA and numt sequences (see Zhang and Hewitt 1996; Bensasson et al. 2001). Those applicable to rRNA genes involve either interpretation of the phylogenetic topology, the inferred secondary structure of the transcribed RNA molecule, or a combination of the two. (We assume sequence ambiguities on chromatograms and multiple bands on autoradiograms are automatically considered suspect and therefore limit our discussion here to clean sequences.) Phylogenetic indicators include ''unusual or contradictory'' (Zhang and Hewitt 1996, p. 250) or ''aberrant'' (Arctander 1995, p. 18) results. However, the absence of prior knowledge of phylogenetic relationships renders this criterion difficult to apply in some situations (Arctander 1995), particularly as the objective of many phylogenetic studies is the inference of the phylogenetic position of a given taxon (or taxa) in the first place. Significantly shorter branches may also indicate nuclear introgressions under the assumption that numts evolve more slowly than their mtDNA paralogs (Sorenson and Fleischer 1996; Sorenson and Quinn 1998). These same authors point out, however, that not only would recently introgressed copies fail detection by this criterion, but that the absence of selective constraints would eventually allow numts to accumulate substitutions at selectively conserved mtDNA sites such that branch lengths leading to *older* numts may be *longer* than those leading to their mtDNA paralogs (e.g., see Lopez et al. 1994, 1997).

A second widely advocated method for discriminating rDNA numts involves aligning sequences to sec-

[1] Present address: Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut

Key words: pseudogene, numt, 12S rRNA.

Address for correspondence and reprints: Link E. Olson, Department of Cell and Molecular Biology, Northwestern University Medical School, 303 E. Chicago Avenue, Chicago, Illinois 60611.
E-mail: lolson@fmnh.org.

Mol. Biol. Evol. 19(1):93–100. 2002

ondary structure models to determine whether structurally disruptive mutations or changes at phylogenetically conserved positions have occurred (Hickson et al. 1996; Sorenson and Fleischer 1996; Houde et al. 1997; Sorenson and Quinn 1998). Unfortunately, no explicit criteria have been proposed along these lines, leaving the underlying question of what constitutes a permissible change with regard to secondary structure or conserved sites unanswered. Secondary structure itself evolves, and novel structural features involving wholesale losses of otherwise highly conserved elements have been reported (e.g., Janke et al. 1994), as has the acquisition of novel stem structures (Olson 1999). Also, as Sorenson and Fleischer (1996) noted, numts will not necessarily be identifiable by structurally incompatible changes or substitutions in conserved sequence motifs owing, again, to their presumably slower substitution rates (see also Hickson et al. 1996).

Curiously, despite the apparent popularity of the secondary structure method of detection, few studies have used it to identify a suspected rRNA numt. Noor and Larkin (2000) employed structural criteria in their examination of suspiciously variable Drosophila 12S rRNA sequences reported by previous authors. These authors used three separate tests, two of which compared the number of changes in conserved regions (conserved sites and sites in stems) with those in nonconserved regions within and between several species. A third test compared minimal free energy values among taxa for a portion of the 12S rRNA molecule folded according to a structural model. Results of each of these tests suggested either errors in the sequences in question or the unintentional sequencing of one or more numts, although additional sequencing under similar conditions by Noor and Larkin (2000) failed to support a pseudogene explanation.

We are unaware of any study that has critically examined the effectiveness of using secondary structure to identify known numts. For example, none of the studies reporting nuclear insertions of mitochondrial rRNA or tRNA genes in table 1 of Zhang and Hewitt (1996) compared their resulting sequences with secondary structure models to ascertain the extent to which predicted structure was affected. Thus, the oft-cited utility of secondary structure models in discriminating rDNA numts remains largely speculative. We explore this approach using the 12S rRNA gene from the human mitochondrial genome and the recently completed Human Genome Project. Given the known genomic origin and status of these sequences in humans (functional mitochondrial gene or nuclear pseudogene), the availability of mtDNA sequences for the same locus from several additional primates, a well-supported phylogenetic hypothesis for the chosen taxa, and rigorously tested secondary structure models for the rRNA subunit of interest, we are afforded an unprecedented opportunity for examining this issue in detail. We consider a hypothetical scenario in which a species of uncertain phylogenetic affinity (*Homo sapiens*, in this case) is surveyed for a mitochondrial gene (12S) whose status (nuclear paralog vs. functional mitochondrial gene) we wish to determine using secondary structure. We consider two structural models and a number of numts in an attempt to avoid erroneous conclusions based on the potential idiosyncrasies of any single pseudogene or structural hypothesis.

Table 1
Stem Nomenclature Synonymy

SD96[a]	H96[b]
26	32
27	33
28	34
29	35
30	36
31	38
32	39
33	40
34	42
35	45
36	47
37	48

[a] Springer and Douzery (1996) model.
[b] Hickson et al. (1996) model.

Methods

We performed a BLAST search (Altschul et al. 1997) of the draft Human Genome using the complete human mitochondrial 12S rRNA gene (GenBank accession V00662) as the query sequence. We selected all matches for which the beginning and end of the original mitochondrial 12S were unambiguously present in order to limit our study to reasonably complete pseudogenes. Sequences were aligned by eye to the core secondary structure model for mammalian 12S rRNA of Springer and Douzery (1996) (SD96 hereafter). A second alignment for the third domain region of the gene was performed with reference to the model of Hickson et al. (1996) (H96 hereafter), which differs slightly from the corresponding portion of the SD96 model. To date, these differences with respect to mammalian 12S have not been investigated critically, and the relative accuracy of each model remains unknown. We therefore consider both models separately but use the stem nomenclature of Springer and Douzery (1996) for consistency (see table 1). These models were chosen over alternatives (e.g., Van de Peer et al. 2000) because they either focus on mammalian 12S rRNA structure (SD96) or make explicit hypotheses regarding conserved sequence motifs and nucleotides (H96). Bases involved in pairing were identified for both stem (both models) and tertiary (SD96 model only) interactions. The following additional primate mtDNA 12S sequences were included in these alignments and all subsequent analyses (GenBank accession numbers in parentheses): *Lemur catta* (AF038013), *Tarsius bancanus* (AF153001), *Papio hamadryas* (Y18001), *Nasalis larvatus* (AF069970), *Hylobates lar* (X99256), *Pongo pygmaeus* (D38115), *Gorilla gorilla* (X93347), *Pan paniscus* (D38116), and *Pan troglodytes* (X93335). Alignments are available as supplementary material on this journal's web page. We assume all GenBank mtDNA sequences used in this

study are themselves mitochondrial and not numts. We follow Groves' (1993) classification of primates in our references to higher taxa and employ the age estimates for primate taxa given in Goodman et al. (1998).

To investigate the extent to which secondary structure can be used to identify numts, we first inspected the correspondence of each human sequence to the SD96 and H96 models visually to identify any major indels occurring in either pairing regions or encompassing the conserved sites identified by Hickson et al. (1996) for Domain III. Nucleotides at these latter sites were also compared with the conserved states of the H96 model (their Appendix 1). Second, we surveyed each primate sequence for the total number of disruptive mutations, calculated as follows. For a given sequence, the absence of a pair bond otherwise present in *Bos* and at least two primate species at the homologous position was counted as one, whether it was because of nonpairing bases at that position (noncanonical T–G bonds were allowed, but A–C bonds were not) or a deletion encompassing one of the positions. Similarly, an uncompensated insertion within any stem resulting in a bulge was counted as one. Finally, we calculated minimal free energy values for Domain III for all sequences (mitochondrial and pseudogene), according to the predicted folding implied by both the SD96 and H96 models using the online version of mfold Version 3.1 (Mathews et al. 1999; Zuker, Mathews, and Turner 1999). Predicted stability calculations were limited to Domain III in order to obtain comparable results for both structural models. The additional two bases hypothesized to engage in pairing at the beginning of stem 26 in the SD96 model, which are not implicated in pairing in the H96 model, were not included for this same reason.

Parsimony analyses were performed using PAUP* 4.0b8 (Swofford 2001). The branch and bound search algorithm was employed with all characters equally weighted. Trees were rooted using *Lemur* and *Tarsius* as outgroups. Bootstrap support was calculated using 100 branch and bound search replicates. Tree searches were conducted for the alignment based on the SD96 model only, as the H96 model is limited to the third domain of the 12S molecule. Ambiguously aligned regions were excluded from parsimony searches. Separate analyses were conducted for each human sequence (three nuclear and one mitochondrial), and their respective phylogenetic positions were compared with the expected phylogeny of hominoids (Ruvolo 1997; Goodman et al. 1998). A fifth analysis in which all four human sequences were included was also carried out in an attempt to determine the relative age of each introgression event as well as to explore the possibility of gene duplication with respect to the pseudogenes sampled, i.e., whether one of the pseudogenes considered gave rise to one or both of the remaining numts.

Results and Discussion

A BLAST search of the human nuclear genome using the human mitochondrial 12S sequence resulted in three matches meeting our previously listed criteria. These are located on chromosome 10 (952 bases spanning positions 151857–152793 in GenBank accession NT024128.3, 88% identical under BLAST criteria), chromosome 11 (952 bases, positions 1023270–1024221 in NT009243.3, 95% identical), and chromosome 5 (957 bases, positions 730591–731547 in NT006961.3; 94% identical). These are hereafter referred to as numt1, numt2, and numt3, respectively. Each of these possesses a seemingly intact 5′ and 3′ region, and the lengths fall within, or correspond closely to, the range observed in hominoid mitochondrial 12S (951–956 bases), suggesting that no indel events longer than a few bases have occurred in these copies. Finally, it warrants mention that additional shorter numts were recovered in our BLAST search that were not included in this study. On the basis of our brief examination of several of these, it is obvious that numerous partial 12S numts of varying length, antiquity, and preservation reside in the human nuclear genome, many of which would undoubtedly be instantly recognized as nonfunctional pseudogenes. We limited our investigation to those copies that could theoretically escape detection through the sequencing stage of a study employing complete 12S rRNA gene sequences.

All three numts were readily alignable to both secondary structure models (fig. 1; H96 model and an expanded version of fig. 1 are available online as supplementary material). The only indels larger than one nucleotide occurring within a putative stem structure in any pseudogene are both found in numt3. These include a 4-base deletion encompassing portions of either stem 1 or stem 3 (see fig. 1 for alternative placements of this indel), which would result in the loss of three pair-bonds in either case. No comparable indels were observed in any mitochondrial sequence.

Single-nucleotide insertions resulting in bulges within stems are observed in both mitochondrial and numt sequences (fig. 1). A single-nucleotide deletion in stem 28 (both models) of *Papio* results in either the loss of an otherwise ubiquitous bulge or the loss of an adjacent pair bond, depending on the inferred positional homology. Several stems (e.g., 7, 19, 20, 34, 36) are found to vary with respect to the number of pairing bases in both mitochondrial and numt sequences. Pairing in both positions of stem 20 is lost entirely in *Papio* (unless A–C bonds are considered), and only four of the 10 conserved bonds in stem 40 occur in numt 1.

Of the 56 conserved positions in the H96 model for Domain III, only three were found for which one or more sequences possess an alternative base (fig. 1). Two of these involve mutations to C's in positions otherwise conserved for T's in stem 31; one occurs in numt1 and one in *Lemur*. The mutation in numt1 occurs in the fourth position of stem 31 and is not compensated for in the corresponding nucleotide in 31′, indicating the loss of a T–A pair bond in this stem. Noncanonical C–A bonds are generally discounted, but they have nonetheless been suggested for this stem in other mammals (Springer and Douzery 1996) and may be relatively common among vertebrates in Domain III (Hickson et al. 1996). In addition, other mammals possess a C at

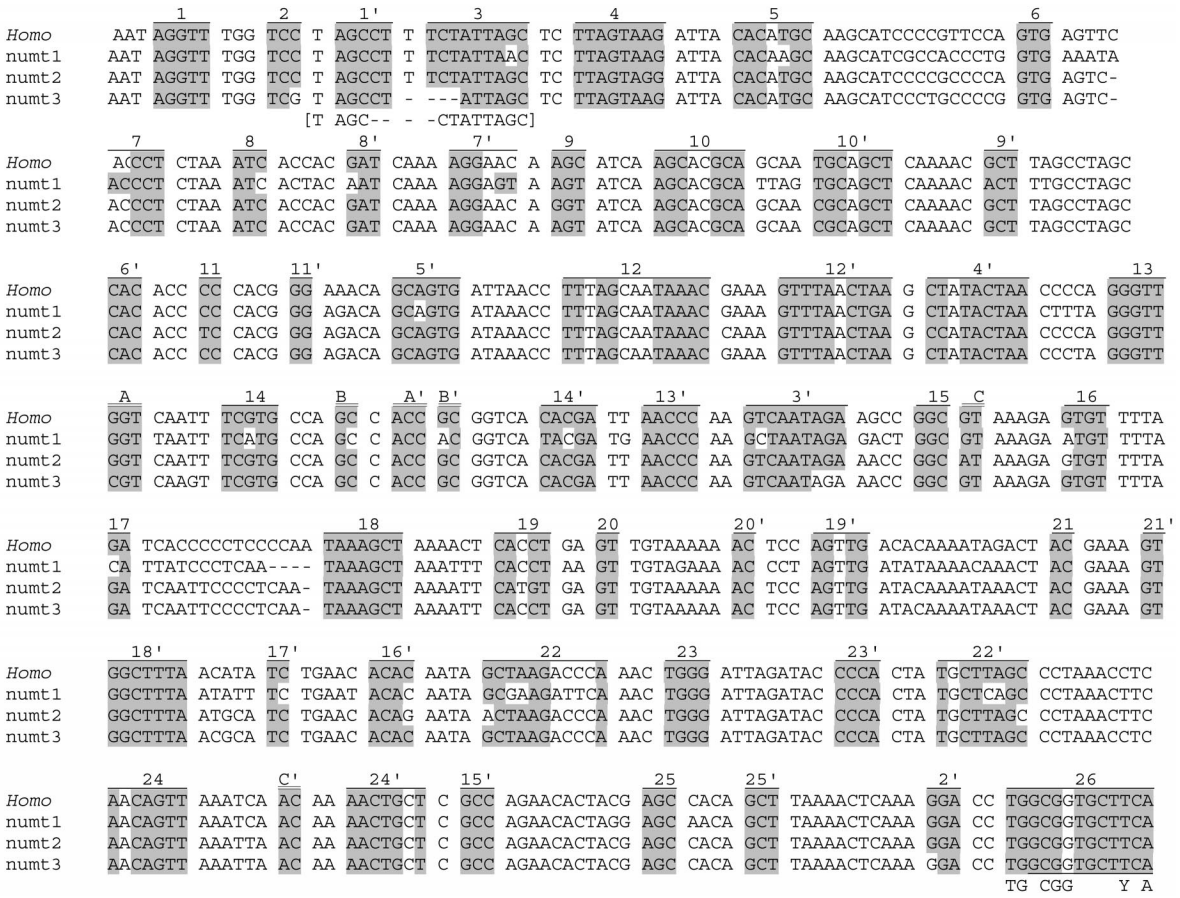

FIG. 1.—Human mitochondrial 12S rDNA sequence and three nuclear pseudogenes from the human genome aligned to the core mammalian secondary structure model of Springer and Douzery (1996). Stem boundaries are denoted above with single lines and are numbered according to the SD96 model. Boundaries for bases involved in tertiary interactions are indicated with double lines above and labeled with upper-case letters. For both types of interactions, canonical and G–T bonds are shaded (C–A bonds were not considered). An alternative alignment for the region spanning stems 1 and 3 in numt3 is shown below in brackets. The portion of Domain III corresponding to the region considered in the H96 model and used in the free energy calculations (see table 2) is underlined. Letters below the alignment in this region represent the conserved states in the H96 model (from their Appendix I; Y = C or T, R = A or G); those for which one or more sequences in this study were found to exhibit an alternative state are shown in bold. Note that these conserved positions may not correspond structurally in each model, i.e., some pairing sites in the SD96 model do not pair in the H96 model and vice versa. An expanded version of this figure showing all additional primate 12S sequences and ambiguously aligned sites excluded from phylogenetic analyses is available online as supplementary material or from the senior author.

this position (Springer and Douzery 1996; personal observation). Similarly, the C in what is otherwise the last pairing position of stem 31 in the SD96 model observed in *Lemur* is uncompensated at the corresponding position in 31′, resulting in the loss of a T–A bond. The third site found to possess a nonconserved base in any sequence occurs in the nonpairing region connecting stems 36 and 27. Normally constrained to a purine, a cytosine is observed in *Papio*, *Nasalis*, and numt1. With respect to the highly conserved positions proposed by Hickson et al. (1996), then, only one of the three pseudogenes (numt1) exhibits any deviation, and for every aberrant conserved site in numt1, one or more of the mitochondrial sequences are found to be similarly variable, either at the same position or within the same stem.

The total number of disruptive mutations for each sequence is given in table 2. Not surprisingly, a much greater range is encountered when the entire molecule is considered (SD96 only), in which case numt1 exhibits more than twice as many disruptive mutations as any other sequence. Values for numts 2 and 3 do not fall outside the range observed in other primate (mtDNA) sequences. When only Domain III is considered, no numt exceeds all other primate sequences in the number of disruptive mutations observed.

Free energy values for the predicted folding of Domain III are given in table 2. In all but one case (*Lemur*), sequences were more stable when folded according to the H96 model than when folded to the SD96 model, suggesting the former may be more accurate for this region of 12S in primates if thermal stability is the primary target of selection. Numts 1 and 3 were less stable than all other primate sequences according to the SD96 model, but both fell within the range of primate values for the H96 model. Numt 2, on the other hand, fell within the range of nonhuman hominoid values for the SD96 model and had the lowest free energy of *any* sequence according to the H96 model.

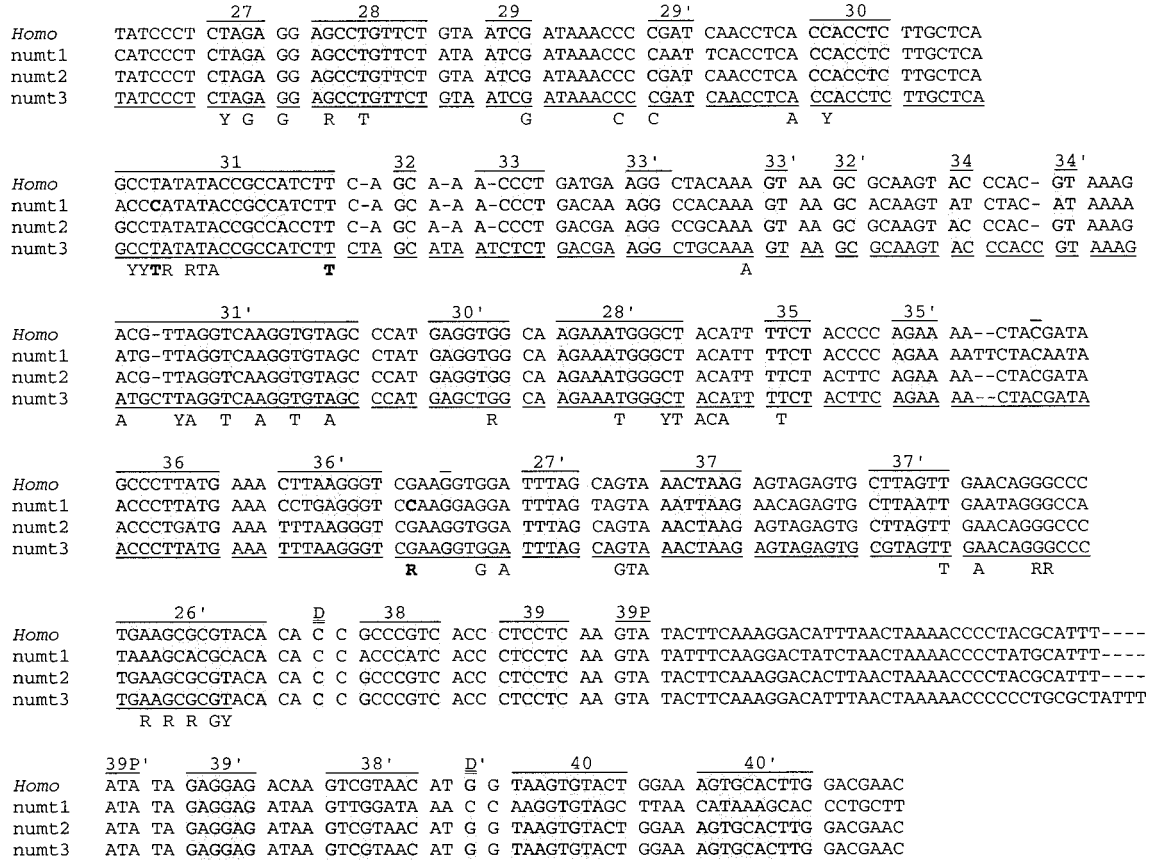

FIG. 1 *(Continued)*

Results of parsimony analyses are shown in fig. 2. When only mitochondrial sequences are analyzed (fig. 2*A*), a single most-parsimonious topology is recovered. Although the expected human-chimp sister relationship is not obtained, bootstrap support for the conflicting gorilla-chimp clade is weak; all other nodes receive ≥80% bootstrap support and conform to the well-corroborated phylogenetic hypothesis relating these taxa. Analyses in which individual numts were substituted for the human mitochondrial sequence are shown in fig. 2*B–D*, and the results of an analysis with all sequences included are summarized in fig 2*E*. The well-supported phylogenetic position of numt1 (fig. 2*B* and *E*) suggests that it introgressed from the mitochondrial genome prior to the hominoid radiation. A 4-base deletion occurring in all hominoid sequences (including all numts) in the nonpairing region (in both models) between stems 30 and 31 serves to establish a maximum age of origin for numt1. A survey of 44 additional primate 12S sequences in GenBank (see *Appendix*) suggests this 4-base deletion is unique to catarrhines, and that the introgression of numt1 therefore likely postdated the platyrrhine-catarrhine divergence. Collectively, this evidence indicates that numt1 originated between 18 and 25 MYA. Numts 2 and 3, on the other hand, appear to be of much more recent origin, apparently having diverged from the human mitochondrial genome sometime after the divergence of orangutans from the remaining hominids (fig. 2*C–E*). The poorly resolved phylogenetic position of numts 2 and 3 relative to the *Gorilla*, *Pan*, and *Homo* mitochondrial sequences limits inferences of the relative timing of these introgressions. The possibility that either numt2 or numt3 gave rise to the other cannot be discounted based on our analyses, given the recovery of this arrangement in two of the nine most-parsimonious trees.

Table 2
Disruptive Mutations and Free Energy Values

	Disruptive Mutations			$\Delta G°$ (kcal/mol), Domain III	
	Entire Gene	Domain III			
Species	SD96	SD96	H96	SD96	H96
Bos	—	—		−29.4[b]	−43.0[b]
Lemur	7	3	5	−30.8	−30.4
Tarsius	6	2	3	−36.4	−37.3
Papio	10	3	3	−32.8	−40.6
Nasalis	≥8[a]	2	3	−39.5	−45.2
Hylobates	5	2	4	−39.0	−45.1
Pongo	5	1	2	−38.8	−53.3
Gorilla	5	0	0	−48.4	−55.5
Pan paniscus	5	0	0	−50.1	−57.2
Pan troglodytes	5	0	0	−48.2	−55.3
Homo	4	0	0	−48.8	−55.9
numt1	21	3	2	−27.1	−34.7
numt2	6	1	1	−43.5	−58.0
numt3	10	3	3	−28.6	−38.8

[a] Value represents a minimum estimate due to missing data at the 5′ region.

[b] Free energy values shown for *Bos* are based on the model as taken directly from each respective paper.

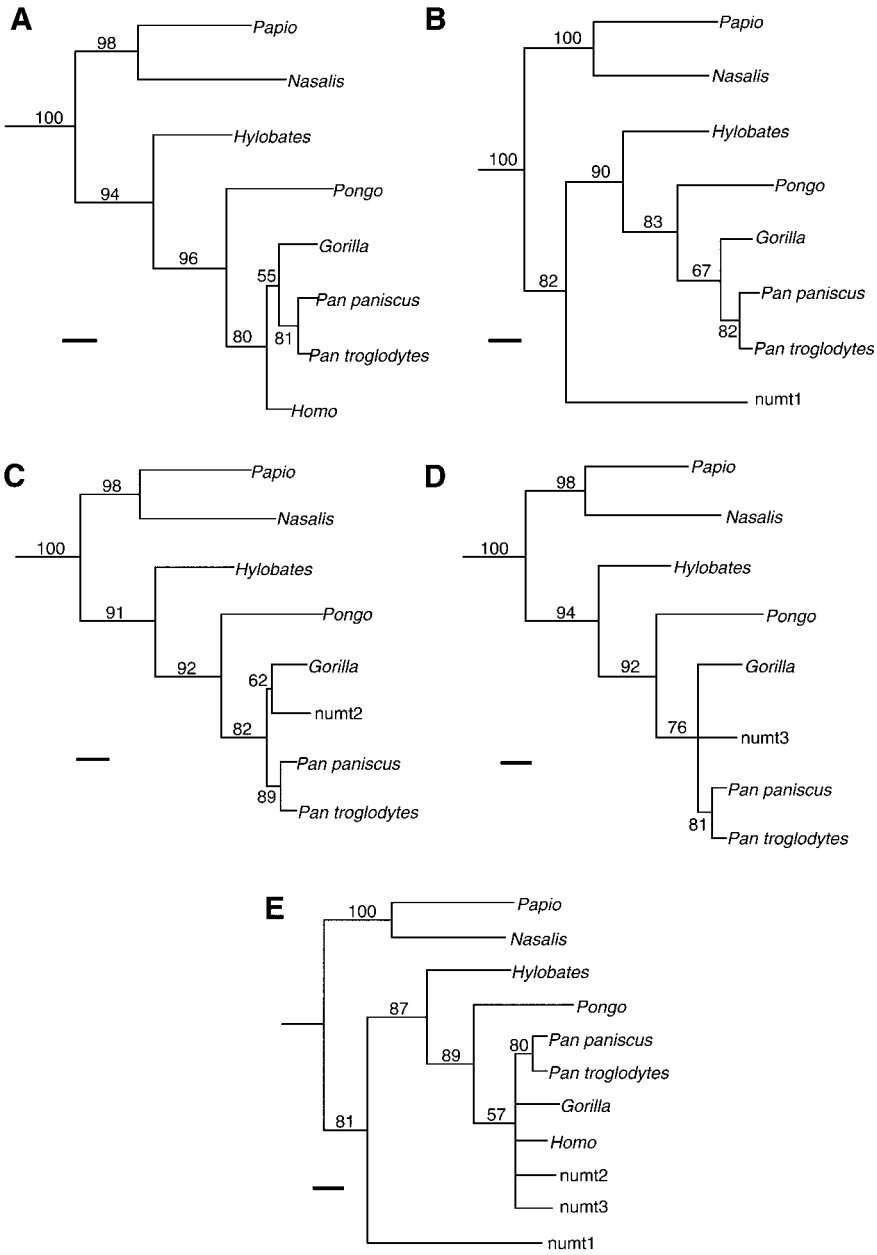

FIG. 2.—Results of parsimony analyses. Outgroup taxa (lemur and tarsier) were used to root the trees but are not shown here. Values represent bootstrap support, scale bar represents 10 changes. *A*, mitochondrial sequences only, single most-parsimonious tree. *B&C*, nonhuman mitochondrial sequences and numts 1 and 2, respectively, single most-parsimonious trees. *D*, nonhuman mitochondrial sequences and numt3, strict consensus of two equally parsimonious trees. *E*, All sequences included, strict consensus of nine equally parsimonious trees.

If nodes that receive less than 70% bootstrap support are collapsed, the topologies in fig. 2*A*, *C*, and *D* become identical. This suggests that the unintentional inclusion of two of the three numts would probably not result in a topology discordant with the true phylogeny of hominids. The unwitting analysis of numt1, however, would result in a dramatically different hypothesis of relationships. In none of the trees in fig. 2 do branch lengths leading to any numt appear disproportionate, and relative-rates tests (Tajima 1993) failed to single out any pseudogene sequences with respect to the mitochondrial sequences (results not shown).

In our exploration of the inferred structural characteristics of three 12S rRNA pseudogenes from the human genome, we found only two structurally disruptive indel events in stem regions, both of which, ironically, occur in what is likely to be the most recently introgressed copy (numt3). The more disruptive of the two—suggesting the loss of three pair bonds in one or possibly two stems—is well outside the region amplified by the popular Kocher et al. (1989) primers and is hence not included in the H96 model; the same is true for the inordinate loss of pair bonds in stem 40 of numt1. Neither of the other numts exhibits any immediately sus-

picious indels. A simple count of disruptive mutations would suggest numt 1 (but not numts 2 or 3) to be an outlier, but only when the entire gene is assayed; none of the numts appear inordinately modified when Domain III alone is considered. This suggests that studies employing only partial 12S sequences, which may be more prone to numt contamination to begin with (e.g., van der Kuyl et al. 1995), may be even further compromised if secondary structure alone is used to check for numt amplification.

Inspection of conserved positions in the H96 model similarly fails to elicit alarm. Two of the three numts (not surprisingly, the two youngest) are invariant at all of these sites. Numt1 varies at only two positions, but so do other mammalian mitochondrial sequences at both positions for the same nucleotide. Broadening the taxonomic sample to include other primates would likely assuage suspicion. All 25 cercopithecid (Old World monkeys) and 16 out of 17 platyrrhine (New World monkeys) 12S sequences from GenBank listed in the *Appendix* possess either a C or a T at this position (rather than the A or G proposed in the H96 model), suggesting that the selective constraints that have maintained a purine at this position in other animal taxa are not acting similarly in primates.

Thus, with the possible exception of numt3, a thorough visual inspection of these sequences against a structural model would likely fail to identify the pseudogenes. Numt1, by far the oldest copy, appears questionable when all disruptive mutations are counted along the entire molecule, but not when shorter pieces are considered. Comparisons of predicted stability for Domain III are perhaps more sophisticated, but their results are equally equivocal and cast as much suspicion on some mitochondrial sequences (e.g., *Lemur*) as on some pseudogenes, whereas one pseudogene surpasses all other sequences in estimated thermal stability. As would be predicted based on its age, numt1 is the least stable of the human sequences, falling outside the range of free energies observed in other primates in the SD96 model but *not* under the H96 model. This potential for model-dependent conclusions has implications beyond our study, particularly for studies investigating the evolution of rRNA. We suggest that either alternative models of secondary structure need to be considered or that such models be fine-tuned for the taxa of interest. With respect to our hypothetical scenario, we believe that, in the absence of additional evidence, numts1, 2, and possibly 3 would be accepted as mitochondrial if inadvertently sequenced and subjected to a cursory check against secondary structure. This is particularly sobering in the case of numt1, given its well-supported phylogenetic placement, which differs substantially from that of humans and could lead to myriad erroneous conclusions if accepted as a mitochondrial sequence. Whereas a more exhaustive investigation of these sequences might very well identify a pseudogene signature for these numts, casual inspection of secondary structure can clearly be insufficient for discriminating both recently introgressed as well as relatively ancient mitochondrial rRNA pseudogenes.

Reference to secondary structure is often claimed to be or advocated (e.g., Houde et al. 1997) as a means of confirming the cytoplasmic origin of sequenced mitochondrial rRNA genes. The expectation that rRNA pseudogenes will eventually accumulate structurally disruptive mutations in the form of excessive point substitutions or indel events or both is certainly a reasonable one, and reference to secondary structure is likely a relatively rapid method of numt identification *if such mutations have occurred.* Criteria for accepting a given rDNA sequence as mitochondrial based on structural characteristics alone are often vague, at best, and probably for a good reason—namely, uncertainty as to what constitutes structurally and selectively tenable mutations. Although we strongly advocate the use of secondary structure for inferring positional homology (see Kjer 1995) and studying character evolution (regardless of the dizzying tedium involved), we caution against relying on concordance to secondary structure as a means of identifying pseudogenes, as this may engender false confidence. Rather, we suggest that researchers who suspect such numts employ more direct methods such as cloning or singe-stranded conformation polymorphism (Bensasson et al. 2001).

Acknowledgments

We thank Michael Alfaro for discussion and comments on an earlier draft of this paper, Mohammed Noor and Karl Kjer for their insightful and constructive reviews, and Keith Crandall for his efficient handling of the manuscript. This research was supported by NSF grant DEB-9985205 to A.D.Y.

APPENDIX

Forty-four additional primate 12S sequences surveyed for the deletion between stems 30 and 31 and the conserved position discussed in the text. GenBank accession numbers in parentheses.

Tarsius syrichta (AF069976), *Callithrix pygmaea* (AF069983), *Callithrix jacchus* (AF069982), *Callimico goeldii* (AF069981), *Saguinus geoffroyi* (F069972), *Saguinus oedipus* (AF069973), *Leontopithecus rosalia* (AF069969), *Cebus apella* (AF069965), *Saimiri sciureus* (AF069974), *Aotus trivirgatus* (AF069977), *Alouatta palliata* (AF069964), *Alouatta seniculus* (AF069975), *Ateles* sp. (AF069978), *Brachyteles arachnoides* (AF069979), *Callicebus moloch* (AF069980), *Chiropotes satanas* (AF069966), *Lagothrix lagotricha* (AF069968), *Pithecia pithecia* (AF069971), *Cercocebus aterrimus* (L35192), *Cercocebus torquatus* (L35204), *Cercopithecus aethiops* (L35185, L35187, L35189, L35190, L35194, L35207), *Cercopithecus ascanius* (L35202), *Cercopithecus cephus* (L35191), *Cercopithecus diana* (L35193), *Cercopithecus galeritus* (L35208), *Cercopithecus mitis* (L35197), *Cercopithecus mona* (L35198), *Cercopithecus neglectus* (L35182), *Cercopithecus nictitans* (L35199), *Cercopithecus patas* (L35186), *Colobus guereza* (L35195), *Macaca mulatta* (L35203), *Macaca sylvanus* (L35188), *Mandrillus sphinx* (L35196), *Miopithecus talapoin* (L35205), *Papio*

cynocephalus (L35184), *Papio ursinus* (L35206), *Presbytis cristatus* (L35200), and *Gorilla gorilla* (L35209).

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ARCTANDER, P. 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proc. R. Soc. Lond. **262**:13–19.

BENSASSON, D., D.-X. ZHANG, D. L. HARTL, and G. M. HEWITT. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. **16**:314–321.

COLLURA, R. V., and C.-B. STEWART. 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature **378**:485–492.

FUKUDA, M., S. WAKASUGI, T. TSUZUKI, H. NOMIYAMA, K. SHIMADA, and T. MIYATA. 1985. Mitochondrial DNA-like sequences in the human nuclear genome. Characterization and implications in the evolution of mitochondrial DNA. J. Mol. Biol. **186**:257–266.

GOODMAN, M., C. A. PORTER, J. CZELUSNIAK, S. L. PAGE, H. SCHNEIDER, J. SHOSHANI, G. GUNNELL, and C. P. GROVES. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9**:585–598.

GROVES, C. P. 1993. Order Primates. Pp. 243–277 *in* D. E. Wilson and D. M. Reeder, eds. Mammal species of the world. A taxonomic and geographic reference. 2nd edition. Smithsonian Institution Press, Washington.

HICKSON, R. E., C. SIMON, A. COOPER, G. S. SPICER, J. SULLIVAN, and D. PENNY. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12SrRNA. Mol. Biol. Evol. **13**:150–169.

HOUDE, P., A. COOPER, E. LESLIE, A. E. STRAND, and G. A. MONTAÑO. 1997. Phylogeny and evolution of 12S rDNA in Gruiformes (Aves). Pp. 121–158 *in* D. P. MINDELL, ed. Avian molecular evolution and systematics. Academic Press, San Diego, Calif.

JANKE, A., G. FELDMAIER-FUCHS, W. K. THOMAS, A. VON HAESELER, and S. PÄÄBO. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. Genetics **137**:243–256.

KJER, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. **4**:314–330.

KOCHER, T. D., W. K. THOMAS, A. MEYER, S. V. EDWARDS, S. PÄÄBO, F. X. VILLABLANCA, and A. C. WILSON. 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA **86**:6196–6200.

LOPEZ, J. V., M. CULVER, J. C. STEPHENS, W. E. JOHNSON, and S. J. O'BRIEN. 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mol. Biol. Evol. **14**:277–286.

LOPEZ, J. V., N. YUHKI, R. MASUDA, W. MODI, and S. J. O'BRIEN. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. **39**:174–190.

MATHEWS, D. H., J. SABINA, M. ZUKER, and D. H. TURNER. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**:911–940.

NOOR, M. A., and J. C. LARKIN. 2000. A re-evaluation of 12S ribosomal RNA variability in *Drosophila pseudoobscura*. Mol. Biol. Evol. **17**:938–941.

OLSON, L. E. 1999. Systematics, evolution, and biogeography of Madagascar's tenrecs (Mammalia: Tenrecidae). Doctoral dissertation, University of Chicago, Chicago, Ill.

RUVOLO, M. 1997. Genetic diversity in hominoid primates. Annu. Rev. Anthropol. **26**:515–540.

SORENSON, M. D., and R. C. FLEISCHER. 1996. Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus. Proc. Natl. Acad. Sci. USA **93**:15239–15243.

SORENSON, M. D., and T. W. QUINN. 1998. Numts: a challenge for avian systematics and population biology. Auk **115**: 214–221.

SPRINGER, M. S., and E. DOUZERY. 1996. Secondary structure and patterns of evolution among mammalian mitochondrial 12S rRNA molecules. J. Mol. Evol. **43**:357–373.

SWOFFORD, D. L. 2001. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

VAN DE PEER, Y., P. DE RIJK, J. WUYTS, T. WINKELMANS, and R. DE WACHTER. 2000. The European small subunit ribosomal RNA database. Nucleic Acids Res. **28**:175–176.

VAN DER KUYL, A. C., C. L. KUIKEN, J. T. DEKKER, W. R. K. PERIZONIUS, and J. GOUDSMIT. 1995. Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. J. Mol. Evol. **40**:652–657.

ZHANG, D.-X., and G. M. HEWITT. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol. Evol. **11**:247–251.

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.

ZUKER, M., D. H. MATHEWS, and D. H. TURNER. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. Pp. 11–43 *in* J. BARCISZEWSKI and B. F. C. CLARK, eds. RNA biochemstry and biotechnology. Kluwer Academic, Boston, Mass.

KEITH CRANDALL, reviewing editor

Accepted September 17, 2001