J Mol Evol (2005) 61:1–11
DOI: 10.1007/s00239-004-0293-3

JOURNAL OF MOLECULAR EVOLUTION

Forty Million Years of Independent Evolution: A Mitochondrial Gene and Its Corresponding Nuclear Pseudogene in Primates

Jürgen Schmitz,[1] Oliver Piskurek,[2] Hans Zischler[3]

[1] Institute of Experimental Pathology, ZMBE, University of Münster, Von-Esmarch-Str. 56, D-48149 Münster, Germany
[2] Faculty of Bioscience & Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku Yokohama 226-8501, Japan
[3] Institute of Anthropology, Johannes Gutenberg-University, D-55099 Mainz, Germany

Received: 30 September 2004 / Accepted: 25 February 2005 [*Reviewing Editor:* Dr. Rafael Zardoya]

Abstract. Sequences from nuclear mitochondrial pseudogenes (numts) that originated by transfer of genetic information from mitochondria to the nucleus offer a unique opportunity to compare different regimes of molecular evolution. Analyzing a 1621-nt-long numt of the rRNA specifying mitochondrial DNA residing on human chromosome 3 and its corresponding mitochondrial gene in 18 anthropoid primates, we were able to retrace about 40 MY of primate rDNA evolutionary history. The results illustrate strengths and weaknesses of mtDNA data sets in reconstructing and dating the phylogenetic history of primates. We were able to show the following. In contrast to numt-DNA, (1) the nucleotide composition of mtDNA changed dramatically in the different primate lineages. This is assumed to lead to significant misinterpretations of the mitochondrial evolutionary history. (2) Due to the nucleotide compositional plasticity of primate mtDNA, the phylogenetic reconstruction combining mitochondrial and nuclear sequences is unlikely to yield reliable information for either tree topologies or branch lengths. This is because a major part of the underlying sequence evolution model — the nucleotide composition — is undergoing dramatic change in different mitochondrial lineages. We propose that this problem is also expressed in the occasional unexpected long branches leading to the "common ancestor" of orthologous numt sequences of different primate taxa. (3) The heterogeneous and lineage-specific evolution of mitochondrial sequences in primates renders molecular dating based on primate mtDNA problematic, whereas the numt sequences provide a much more reliable base for dating.

Key words: Mitochondrial — rDNA — Nuclear mitochondrial pseudogene — numt — Primates — Papionini

Introduction

Transfer of genetic information and integration of mitochondrial DNA sequences into eukaryotic nuclear chromosomes is a frequent and apparently continuous process over evolutionary time scales. Mourier et al. (2001) queried the working draft sequence of the human genome to determine the frequency and distribution of nuclear mitochondrial pseudogenes (numts), the remnants of this transfer. Altogether 296 numts were identified, ranging between 106 and 14,654 nt in size and representing all positions of the mitochondrial genome. This is likely to be an underestimate since some integrations are too recent to have gained fixation and are therefore polymorphic with respect to their presence in humans (Thomas et al. 1996). Presumably, numt DNAs

Jürgen Schmitz and Oliver Piskurek contributed equally to the present paper.

Correspondence to: Jürgen Schmitz; *email:* jueschm@uni-muenster.de

evolve without functional constraints and accumulate mutational changes at a reduced rate compared to the corresponding mitochondrial sequence. Thus, especially for long evolutionary periods of separation, similarity between numts and extant mt sequences will be eroded.

In a phylogenetic analysis Mourier et al. (2001) pointed out the independent and continuing integration of numts on the lineage leading to humans. On the other hand, it is also argued that large amounts of existing numts represent the outcome of recurring nuclear duplication events rather than being predominantly generated by multiple but independent integration events (Hazkani-Covo et al. 2003). In humans, numts are apparently regularly dispersed over the chromosomal complement and represent more than 500,000 nt of nuclear sequences (Tourmen et al. 2002). To date, numts have been found in more than 82 eukaryotic species including Fungi, Viridiplantae, and Metazoa (Bensasson et al. 2001). Numts often contaminate sequence information from mitochondrial DNA sets used to infer phylogenetic and population genetic affiliations between taxonomic units. This demands a critical verification of the genuine origin of sequences analyzed (Zhang and Hewitt 1996). At the same time, numts provide a unique and powerful tool to compare evolutionary processes of homologous sequences at paralogous locations.

Based on nucleotide divergence between orthologous 896-nt-long mitochondrial DNA fragments of apes, Brown et al. (1982) propose that the evolutionary rate of mitochondrial protein coding genes in primates is on average 5 to 10 times higher than in nuclear protein coding genes. In addition, it is obvious that the evolutionary rate of mtDNA may vary substantially between genes and nucleotide positions (Arctander 1995; Smith and Eyre-Walker 2003). The heterogeneous mode of mitochondrial sequence evolution is best reflected in the often observable conflicting branching orders resulting from phylogenetic reconstructions based on either mitochondrial or nuclear DNA information. One of the most prominent examples is the phylogenetic position of tarsiers in the primate tree. Whereas mitochondrial sequences consistently cluster tarsiers and strepsirrhines together with high bootstrap or quartet puzzling support values (Schmitz et al. 2002a), retropositional evidence obtained from nuclear DNA analyses clearly indicates a sister group relationship of tarsiers and anthropoids (Schmitz et al. 2001). In a second example pertaining to the phylogenetic position of the flying lemur (*Cynocephalus variegatus*, Dermoptera), we were able to reject a close phylogenetic affiliation of flying lemurs with higher primates, that was proposed on the basis of complete mitochondrial genome comparisons (Arnason et al. 2002), by defining the presence/absence pattern of six retrotransposable markers (Schmitz et al. 2002b, 2003). As a possible reason for these conflicting interpretations between mtDNA- and nuclear DNA-based phylogenetic analyses, we proposed a lineage-specific directional mutation mechanism in mitochondrial DNA. Gibson et al. (2004) lend support to this hypothesis comparing a broader sample of complete eutherian mitochondrial genomes. While the different mitochondrial genes and noncoding regions are selectively forced according to their functionality, numts evolve inoperably in a nuclear specific manner and presumably without specific selection pressure. Fukuda et al. (1985) described a reduced evolutionary rate of numts after insertion into the nuclear DNA. As a result, numts appear "frozen" and represent ancestral properties of their corresponding mitochondrial sequences at the time point of transposition. Perna and Kocher (1996) describe numts as molecular fossils in the nucleus. The level of fossilization is more pronounced the higher the evolutionary rate of the mitochondrial source gene.

We specifically addressed the question if either heterogeneously evolving mitochondrial sequences or presumably inoperable mitochondrial pseudogenes residing in the nucleus are better suited for reconstructing the primate phylogenetic tree. Furthermore, we tested the reliability of results based on mixed data sets consisting of both mitochondrial and paralogous numt sequences. In the present study we retrace a 40-MY-old numt integration, charting its course of evolution and that of the corresponding mitochondrial gene in 18 primate species. To this end we applied numt flanking primers located in nuclear regions adjacent to the numt sequences of different primate taxa that are derived from a single nuclear integration event of mtDNA which took place in a common ancestor of the analyzed primate taxa. Long-range combined with nested PCRs enabled us to generate genuine mitochondrial sequences. In this way we assured a comparison of orthologous mitochondrial sequences with one defined paralogous numt, thus retracing their independent evolution during primate divergence and over a time span of 40 MY.

Materials and Methods

Database Searches

To identify mitochondrial rRNA pseudogenes in the human GenBank working draft sequences, we performed a BLAST search of mitochondrial rDNAs in genomic sequences. For technical simplicity, we selected mitochondrial pseudogenes that could be PCR-amplified from unique flanking nuclear sequences yielding fragments smaller than 2000 nt. In this way we retrieved and focused on a 1621-nt-long human mitochondrial pseudogene which is com-

Table 1. PCR primer sequences and amplified fragment length

Name	Primer sequence	Fragment size (nt)	
		mt	ψmt
lmt-f	5′-AGG TTT GGT CCT AGC CTT-3′	3828	
lmt-r	5′-GGG ATT AAT TAG TAC GGG AAG-3′		
mt-f	5′-CGT AAA RAC GTT AGG TCA AGG TGT A-3′	1753	
mt-r	5′-GAT CAC GTA GGA CTT TAA TCG TTG A-3′		
int-f	5′-CAA RCC TAC CAA GCC TGK TGA TAG C-3′	1067	1075
int-r	5′-TTK TTT CCT AGK GTC TAA AGA GCT G-3′	753	822
ψmt-f	5′-CCT CCG TGG TCT TAT GGT CTG T-3′		1677
ψmt-r	5′-GTA GAA CCC CGT CCC TAC TGC T-3′		

Note. lmt-f: long mitochondrial fragment forward primer. lmt-r: long mitochondrial fragment reverse primer. mt-f: mitochondrial fragment forward primer. mt-r: mitochondrial fragment reverse primer. ψmt-f: pseudogene mitochondrial fragment forward primer. ψmt-r: pseudogene mitochondrial fragment reverse primer. int-f: pseudogene and mitochondrial fragment forward primer. int-r: pseudogene and mitochondrial fragment reverse primer.

posed of 186 nt of 5′ partial 12S rDNA, 69 nt of the valine tRNA, and 1366 nt of the 16S rDNA. The human pseudogene sequence is available in GenBank, accession number X02226, and was first described by Nomiyama et al. (1985).

DNA Extraction

Primate tissues were either obtained from animals held in captivity at the German Primate Center or the primate GeneBank (C. Roos; http://www.dpz.gwdg.de/GENEbank/genebank.html). Using standard protocols (Sambrook et al. 1989) genomic DNA was isolated from the following primate tissue or blood samples: **Hominidae**—*Homo sapiens, Pan troglodytes, Gorilla gorilla*, and *Symphalangus syndactylus*; **Cercopithecidae**—*Macaca fascicularis, M. nemestrina, Papio hamadryas, Mandrillus sphinx, Cercopithecus aethiops, Colobus guereza, Pygathrix nemaeus*, and *Presbytis entellus*; **Platyrrhini**—*Callithrix jacchus, Callimico goeldii, Leontopithecus rosalia, Saguinus oedipus, Aotus azarae, Cebus apella, Saimiri sciureus, Callicebus cupreus, Chiropotes albinasus*, and *Lagothrix lagotricha*; **Tarsiidae**—*Tarsius bancanus*; and **Strepsirrhini**—*Lemur catta*.

PCR Procedure

PCR primers were chosen to hybridize in the nuclear flanks surrounding the pseudogene, to clearly define the integration by its flanking, unique sequences, and to avoid confusion with additional independent rDNA integrations at different loci. The human PCR fragment derived was 1677 nt long. In addition, we designed two internal PCR primers resulting in overlapping PCR fragments of 822 and 1075 nt, respectively. In order to retain comparable sequences of genes and pseudogenes, we also amplified the paralogous mitochondrial gene sequences for all investigated individuals. To avoid inadvertent PCR amplification of possible mt pseudogenes, we first performed a long-range PCR of an approximately 3800-nt mitochondrial fragment embedding the rDNA target sequence with one primer at the 5′ end of the 12S rRNA and the second primer at the 3′ end of the transfer RNA methionine gene and the beginning of the adjacent NADH2 gene. This fragment was used as a template for subsequent PCR amplification. The nested PCR amplification including the pseudogene analogous region was 1753 nt in length. Using the internal primers designed for the pseudogene in combination with the mitochondrial specific primers, we amplified a 753- and a 1067-nt fragment. All primer sequences and additional information are shown in Table 1. PCR reactions were carried out for 30 cycles, each consisting of 30 s at 94°C, 30 s at the primer specific annealing temperature, and 60 s per 1-kb fragment length at 72°C. The PCR fragments were purified by agarose gel electrophoresis, ligated into pGEM-T vector (PROMEGA), and electroporated into TOP10 cells (Invitrogen). Plasmid sequencing was performed with universal primers using an automated LI-COR DNA sequencer 4200. In our nested PCR approach to amplify the genuine mtDNA, the sequenced clones from all products display one sequence, thus we can exclude an inadvertent coamplification of pseudogenes.

Data Deposition

Hominidae: *Homo sapiens* (AF420032/AF420050), *Pan troglodytes* (AF420033/AF420051), *Gorilla gorilla* (AF420034/AF420052), *Symphalangus syndactylus* (AF420035/AF420053).

Cercopithecidae: *Macaca fascicularis* (AF420036/AF420054), *M. nemestrina* (AF420037/AF420055), *Mandrillus sphinx* (AF420039/AF420057), *Papio hamadryas* (AF420038/AF420056), *Cercopithecus aethiops* (AF420040/AF420058), *Colobus guereza* (AF420041/AF420059), *Pygathrix nemaeus* (AF420042/AF420060), *Presbytis entellus* (AF420043/AF420061).

Platyrrhini: *Leontopithecus rosalia* (AF420044/AF420062), *Aotus azarae* (AF420045/AF420063), *Cebus apella* (AF420046/AF420064), *Saimiri sciureus* (AF420047/AF420065), *Lagothrix lagotricha* (AF420049/AF420067), *Chiroptotes albinasus* (AF420048/AF420066); Tarsiidae: *Tarsius bancanus* (AF348159).

Strepsirrhini: *Lemur catta* (AJ421451), *Nycticebus coucang* (AJ309867); Scandentia, *Tupaia belangeri* (AF217811).

Papionini monophyly: *Theropithecus gelada* (AY603034); *Lophocebus aterrimus* (AY603035); *Cercopithecus aethiops* (AY603036); *Cercocebus agilis* (AY603037).

Remark: The *Cebus apella* mitochondrial pseudogene AF420064 includes an ~300-bp-long Alu retroposon.

Nucleotide Composition and Phylogenetic Reconstruction

Nucleotide frequencies of the 18 analyzed anthropoids and 4 outgroup species were compiled with the PAUP* program (Swofford 2000). To analyze single- and double-stranded rDNA sequence regions separately, we estimated a secondary structure model corresponding to Hixson and Brown (1986 [12S rRNA]), Sprinzl et al. (1998 [tRNA]), and Gutell and Fox (1988 [16S rRNA]) (see sup-

plemental Figs. 1A–C at http://zmbe2.uni-muenster.de/expath/addmat/numtl.htm).

Sequence alignments were carried out by CLUSTAL X (Thompson et al. 1997). To account for nucleotide compositional heterogeneity, phylogenetic reconstructions were performed using the LogDet distance transformation with 10,000 bootstrap steps (neighbor-joining search) as implemented in PAUP* (Swofford 2000). In addition, we performed a Bayesian inference using MrBayes version 3.0 (Huelsenbeck and Ronquist 2001). The specified model of sequence evolution has been set to GTR with four rate categories to approximate the gamma distribution of rates across sites. The Markov chain Monte Carlo (MCMC) process was set so that three chains were run independently for 1,000,000 generations and with trees sampled every 100 generations. The number of initial members in the chain to skip before starting sampling was set to 500 trees.

Upon initial numt sequence comparison we identified a 23-nt insertion in *Papio* and *Mandrillus*. To reconstruct the origin of the "integration in an integration" in Old World monkeys and make use of the molecular cladistic information of this insertion, we sequenced the orthologous regions of diverse representatives of cercopithecoids (see above).

Spectrum from LogDet Distances

For hominoids, cercopithecoids, and platyrrhines we performed separate phylogenetic reconstructions using members of the corresponding sister lineages as outgroups. For the well-accepted nodes labeled with Arabic letters (Fig. 4), we performed a spectral analysis as implemented in Spectrum 2.0 (http://taxonomy.zoology.gla.ac.uk/~mac/spectrum). To visualize supporting and conflicting splits, we draw a lento plot for the mtDNA and numtDNA sequences, respectively (Fig. 5) (for details of the method see Penny et al. 1999).

Molecular Dating

Average divergence times for three primate-splitting points were estimated by subsampling the sequences to a total of 505 quartets using the program QDate 1.11 (Rambaut and Bromham 1998). This method can accommodate rate heterogeneity between taxa. We calculated the splitting dates for the common ancestor of Cercopithecinae, Cercopithecoidea, and Anthropoidea. Reference points based on fossil evidence for the human–chimpanzee split (5 MYA), for the common ancestor of the subtribe Papioninae (8 MYA), the Cercopithecinae split (10 MYA), the common ancestor of Colobini (12 MYA), and the last common ancestor of platyrrhines and the Cercopithecoidea (26 MYA) were taken from Delson (1992) and MacFadden (1990).

We specified the REV model of substitution with the constrained two-rate model and tested it against the five-rate model. We incorporated four categories of site-specific rate heterogeneity and defined the respective shape of the gamma distribution determined by the alpha parameter from the ML reconstruction ($\alpha = 0.21$ for mt and $\alpha = 1.07$ for numt).

Relative Rate Test

To compare the relative rates of mt genes and the corresponding numts, the adequate outgroup is the branching point of mtDNA lineages that took place before the transposition event.

Accordingly, we chose *Tarsius bancanus* to test the molecular clock. However, it turned out that tarsiers are too distant to detect significant differences in molecular rates. To obtain a closer reference point, we estimated the hypothetical ancestral sequence at the splitting point of the numts from a user-defined tree using the ML method implemented in PAML (Yang 2000) and using MP method implemented in PAUP* (Swofford 2000). To avoid misleading results by using a hypothetical reconstructed sequence as an outgroup which might be problematic due to uncertainties in estimating the character state at the most variable positions, we restricted our analysis to the strict consensus nucleotide positions of ML and MP derived ancestral sequences (including 1573 of 1623 nts of the analyzed human sequence). RRTree version 1.1.11 was used to compare lineage-specific substitution rates (Robinson et al. 1998). A phylogenetic reconstruction of the mtDNA and numt DNA sequences under investigation is shown as supplemental Fig. 2 at http://zmbe2.uni-muenster.de/expath/addmat/numtl.htm).

Results

In a GenBank screening we identified and analyzed more than 20 human nuclear paralogues of the mitochondrial rDNA region. For the comparative analysis of sequence evolution, we chose numts displaying an appropriate size for PCR amplification with primers located in the flanking nuclear regions. We determined the time of transfer to the nucleus by relating the numts with the respective mt rDNA sequences in a phylogenetic tree reconstruction incorporating representatives of all primate infraorders (Catarrhini, Platyrrhini, Tarsiiformes, Lemuriformes). Finally, we chose a 1621-nt-long insertion of mitochondrial rDNA into the nuclear DNA that was transferred to the nucleus in the common ancestor of all extant anthropoid primates. This mitochondrial pseudogene comprises 186 nt, which are homologous to the 3′ end of the 12S rRNA gene, the valine tRNA gene, and an almost-complete 16S rRNA specifying region (1366 nt). The respective integration was described for the human genome in Nomiyama et al. (1985). We were able to localize the pseudogene and the nuclear flanking regions on human chromosome 3.

Analyzing the selected numt locus for 18 different anthropoid primates by PCR, we could verify the presence-state in all of these, however, no PCR signal was detectable in *Tarsius* and strepsirrhines. There are two possible reasons for this. One is that the transfer to the nucleus took place after the split of tarsiers and the absence locus is too small to be detectable by PCR. Alternatively, the flanking region diverged to a degree that prevents PCR amplification.

Nucleotide Composition

In Fig. 1A we depict the nucleotide composition of mt and numt sequences for all species under investigation. Furthermore, we separately analyzed single- and double-stranded regions (Figs. 1B and C). The nucleotide composition is relatively constant in numts and is expected to represent the distribution close to

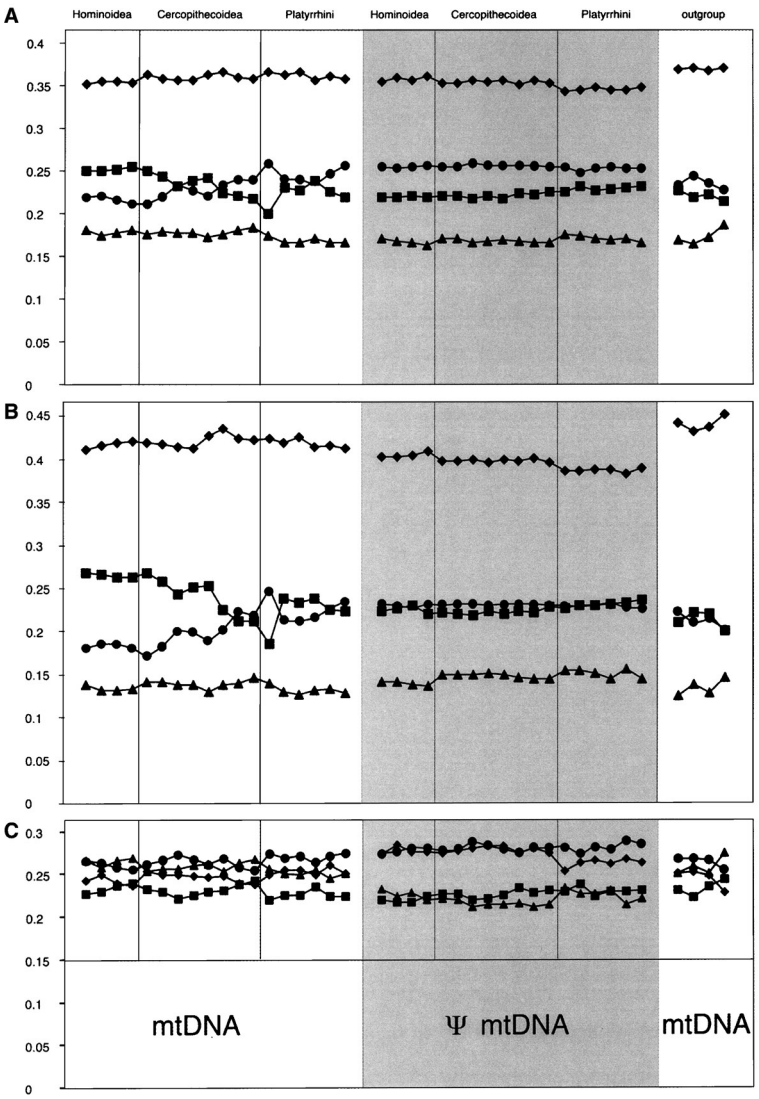

Fig. 1. Nucleotide frequencies of the investigated taxa. Frequencies for the pseudogenes are displayed on a gray background. Nucleotide frequencies for (**A**) the complete sequences and for (**B**) single-stranded-, and (**C**) double-stranded regions of the rDNA. The adenosine base frequency is represented by rhombs, cytosines by gray rectangles, guanines by triangles, and thymines by circles.

the time point of transposition. On the other hand, the mtDNA nucleotide composition with the CT distribution in particular is rather plastic. Plasticity is mainly expressed in the single-stranded regions of the RNA genes analyzed.

Phylogenetic Reconstruction

In Schmitz et al. (2002a) we described an extensive nucleotide compositional plasticity of mitochondrial DNA in primates. Figure 1 shows that this holds also for the analyzed mitochondrial rDNA sequences. At present, only the LogDet distance-based transformation takes variation of nucleotide composition into account. We therefore applied the LogDet distance transformation for phylogenetic reconstructions and verified the resulting trees by Bayesian estimations. The tree shown in Fig. 2 comprises the entire sampling of primates for both the mt and numt sequences including a Scandentian representative, *Tupaia belangeri*, as an outgroup. The resulting tree topology is almost identical upon excluding constant sites from the LogDet reconstruction. In addition, we performed separate LogDet distance analyses of the mt and numt data sets (see supplemental Figs. 3A and B at http://zmbe2.uni-muenster.de/expath/addmat/numtl.htm). The branching orders of those "subtrees" are congruent to the tree drawn from the combined data set.

The Bayesian phylogenetic reconstruction, allowing us to correct for rate heterogeneity across sites but not taking biases in nucleotide composition into account, also failed to find the widely accepted primate consensus tree and shows a similar topology compared to the LogDet reconstruction. A closer look into internal branches exhibits several additional misinterpretations generated by applying the Bayesian algorithm. This is exemplified for the mitochondrial part of the tree by displaying a paraphyly of the marmosets as well as a paraphyly of the Papionini

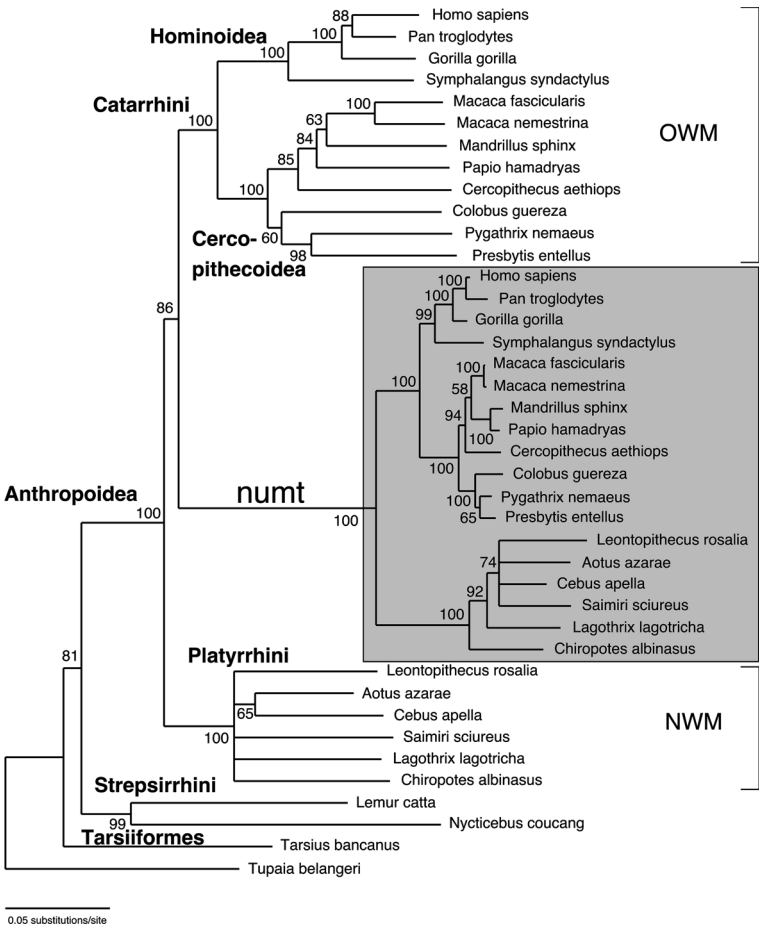

Fig. 2. Phylogenetic reconstruction of mitochondrial genes and pseudogenes of 18 anthropoids. The reconstruction is based on a LogDeterminant correction implemented in PAUP*. Bootstrap values are indicated at the corresponding nodes. Excluding constant sites from the LogDet reconstruction yields an almost-identical tree topology and corresponds to a ML tree. *Tupaia belangeri* was used as an outgroup. Branch lengths represent nucleotide substitutions per site. Mitochondrial pseudogene sequences cluster together and are shown in the boxed gray part of the tree. The Old World monkeys (OWM) include the Hominoidea and Cercopithecoidea. NWM, New World monkeys. Note that this tree topology is not congruent with the actual evolutionary history (see text).

(see supplemental Fig. 4 at http://zmbe2.uni-muenster.de/expath/addmat/numtl.htm). Therefore, the LogDet tree that takes varying nucleotide compositions into account is closest to the widely accepted primate phylogenetic tree topology.

Further support for the Old World monkey tree topology was obtained by a highly informative, so-called "rare genomic change" (RGC; see Springer et al. 2004). We were able to define an insertion of 23 nt in the numt sequence of all Papionini, thus clearly separating them from the remaining cercopithecoids (Fig. 3).

The tree presented in Fig. 2 shows several striking characteristics. (1) *Tarsius bancanus* is the descendant of the basal primate split. This artificial placement was described by Andrews et al. (1998) and was discussed for the entire mitochondrial genome by Schmitz et al. (2001). Using retrotranspositional evidence, we have been able to show that tarsiers are the appropriate sister group to anthropoid primates (Schmitz et al. 2001, 2005). (2) From the presence of numts in all platyrrhines we would expect that all numts split off before the separation of platyrrhines. This is not the case in Fig. 2. (3) There is an unexpected long branch connecting all numts. (4) There are some conflicting phylogenetic relationships between the mt and the numt branches. To account for the latter point, we performed detailed LogDet analyses of the phylogenetic relationships of each primate infraorder (Fig. 4).

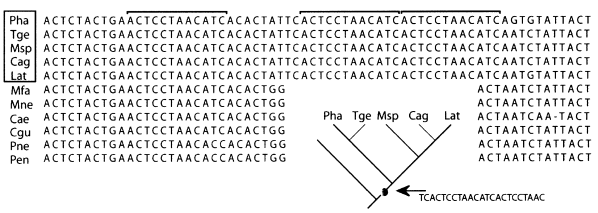

Fig. 3. Papionini monophyly supported by an insertion. The common ancestry of Papionini (boxed species) is shown by an orthologous insertion of a 23-nt sequence (arrow). Duplicated regions are shown by lines above the sequences. Pha (*Papio hamadryas*), Tge (*Theropithecus gelada*), Mfa (*Macaca fascicularis*), Msp (*Mandrillus sphinx*), Cag (*Cercocebus agilis*), Lat (*Lophocebus aterrimus*), Mfa (*Macaca fascicularis*), Cae (*Cercopithecus aethiops*), Cgu (*Colobus guereza*), Pne (*Pygathrix nemaeus*), and Pen (*Presbytis entellus*).

Based on Fig. 4 we performed a spectral analysis to visualize support and conflict of the branching points labeled a–i. Again, we used the LogDet transformation matrix to compute the distribution of phylogenetic signal and random noise. Although the supporting signal is strong in some splits, the con-

Table 2. Quartet dating analysis from fossil calibration points for mt and numt

LCA	Reference (MYA)	mt	numt
Cercopithecoidea	13	13.4 ± 0.4 (8/8/0)	15.5 ± 0.4 (0/0/16)
Catarrhini	25	20.5 ± 1.1 (0/0/25)	24.9 ± 1.8 (0/0/25)
Anthropoidea	40	38.1 ± 4.5 (29/23/412)	51.2 ± 4 (0/8/456)

Note. Reference dating of DNA sequences by Page and Goodman (2001). Numbers in parentheses indicate faulty quartets (first number), quartets with significance (second number), and nonsignificant (third number) log likelihood differences between the constrained two-rate model and the unconstrained five-rate model (see Materials and Methods), respectively. LCA, last common ancestor; MYA, million years ago.

flicting signals are prominent in the mitochondrial data set while being only marginal in the numt subset of sequences. In two cases (e and i) the conflicting signal of the mitochondrial data exceeds the phylogenetic signal. This leads to the collapse of the branches as seen in Fig. 4. In addition to phylogenetic reconstructions based on sequence information, we could recognize an insertion of 23 nt in the Papionini numts of *Papio, Theropithecus, Mandrillus, Cercocebus*, and *Lophocebus* that is absent in other primates. This presence/absence pattern can be used as a molecular cladistic marker to solve the highly controversial discussed monophyly of this subgroup of Papionini (Fig. 3).

Molecular Dating

In order to establish the effects of the different evolutionary histories of mt and numts on molecular dating, we performed a QDating based on five calibration points deduced from the fossil record (see Materials and Methods). We focused our analyses on the splitting date of the last common ancestors of Cercopithecoidea, Catarrhini, and Anthropoidea and compared this to the reference dating of Page and Goodman (2001) of 13, 25, and 40 MYA, respectively (Table 2). Comparing the dating results as computed from the mtDNA subset, it became obvious that the numt sequences tend to predate the mtDNA derived splitting dates. However, in 37 of a total of 505 quartets the mt sequences gave incorrect confidence intervals and were therefore excluded from the analyses. In 31 quartets the mt sequences gave significant differences between a two-rate and a five-rate model of evolution. In contrast, numt data include only eight quartets with lineages showing significantly more than two evolutionary rates. However, most erroneous quartets and significant rate heterogeneities arose from including platyrrhines in the molecular calibration. The rapid radiation of platyrrhines accompanying the colonization of South America some 40 MYA is the probable reason for a heterogeneous evolutionary pattern (Singer et al. 2003).

Relative Rate Test

It is obvious from the supplemental Fig. 2 (http://zmbe2.uni-muenster.de/expath/addmat/numt1.htm) that the terminal branches based on the mitochondrial data are almost longer than those leading to the pseudogene sequences. To obtain relative evolutionary rates it is absolutely essential to have an adequate outgroup. Because the outgroup to gene and pseudogene, *Tarsius bancanus*, is highly problematic and the remaining strepsirrhines are too distant to give meaningful results, we reconstructed a hypothetical sequence of the lineage before the transposition event took place to use as the closest synthetic outgroup. As a result, all hominoid and cercopithecid terminal branches are significantly longer when comparing the mtDNA subset to the numt sequences. However, in platyrrhines we could detect no significant differences in the relative rates comparing mt and numt sequences (data not shown).

Discussion

The study of genes and their corresponding pseudogenes allows a retracing of selective constraints as well as changes in the mutational pattern at the molecular level. In general, pseudogenes are probably free from functional constraints and evolve at higher rates than their paralogues. An entirely different situation emerges when fast-evolving mitochondrial genes transpose to chromosomal locations. Depending on which mitochondrial region emigrates, the evolutionary rate may slow down to 1/10, with the mitochondrial region continuing to evolve in a nucleus-like manner.

In this study we compare the evolutionary pathway of a pseudogene which separated from the corresponding mitochondrial gene more than 40 MYA. From the presence of the orthologous numt in all higher primates we expected the branch leading to the pseudogenes to be basal to higher primates. However, in Fig. 2 all numts cluster amid the higher primates. This artificial clustering is affected by rooting the tree with strepsirrhines, *Tarsius*, and the *Tupaia* inde-

pendent of their individual use or as a paraphyletic outgroup cluster. The presence of the numts in all analyzed NWMs clearly uncovers this reconstruction error.

It has to be stressed at this point that the example presented herein represents a system in which true orthology between the analyzed primate rDNA numts is assured. This is due to the fact that single-copy nuclear flanks are used as PCR anchors to amplify numts in different primate taxa that are derived from a single integration event of mtDNA into nuclear DNA that happened in a common ancestor of the respective taxa. It is therefore conceivable that two independent integrations of the same mitochondrial region which took place at around the same time point in primate evolution—and thus displaying similar nucleotide compositions—will be clustered together as sister groups in a phylogenetic tree reconstruction despite the fact that they are derived from different integration events. The hypothesis that numt sequences arose to a considerable part by duplication (Collura et al. 1996; Hazkani-Covo et al. 2003) should therefore be scrutinized again incorporating information of the numt flanking nuclear regions to verify the true orthology between the numt sequences analyzed.

In Schmitz et al. (2002a) we described the mitochondrial specific change of the mode of evolution on the lineage leading to higher primates. To avoid basing our comparison on inappropriate trees, we focus on detailed and well-supported parts of the tree with clear outgroup rooting (Fig. 4). For hominoids the numt sequences revealed the same tree topology as mt sequences and showed comparably stronger support for the human chimpanzee clade (Fig. 4A). In cercopithecids the branching order differs between mt and numt sequences. The Papionini clade merging *Papio* and *Mandrillus* is not supported by the mt data set. In the numt data set we found strong bootstrap support (100%) for the Papionini. This result was further supported by a 23-mt insertion whose presence could be traced back to the common ancestor of all Papionini (Fig. 3). However, the phylogenetic resolution in platyrrhines was very weak. Nevertheless, the numt data set revealed a split merging together *Leontopithecus*, *Aotus*, *Cebus*, and *Saimiri*. Recently, we found a molecular cladistic marker supporting this clade (Singer et al. 2003). Figure 5 indicates that although the phylogenetic signal of the mitochondrial data set is stronger compared to numt data, the phylogenetic noise is getting stronger as well, leading to the collapse of the widely accepted phylogenetic history in two cases (nodes e and i). The strong mitochondrial clustering of the mandrills with macaques is stable independent of sampling, rooting, and methods of reconstruction. We suppose that directional mutation pressure leading to nucleotide

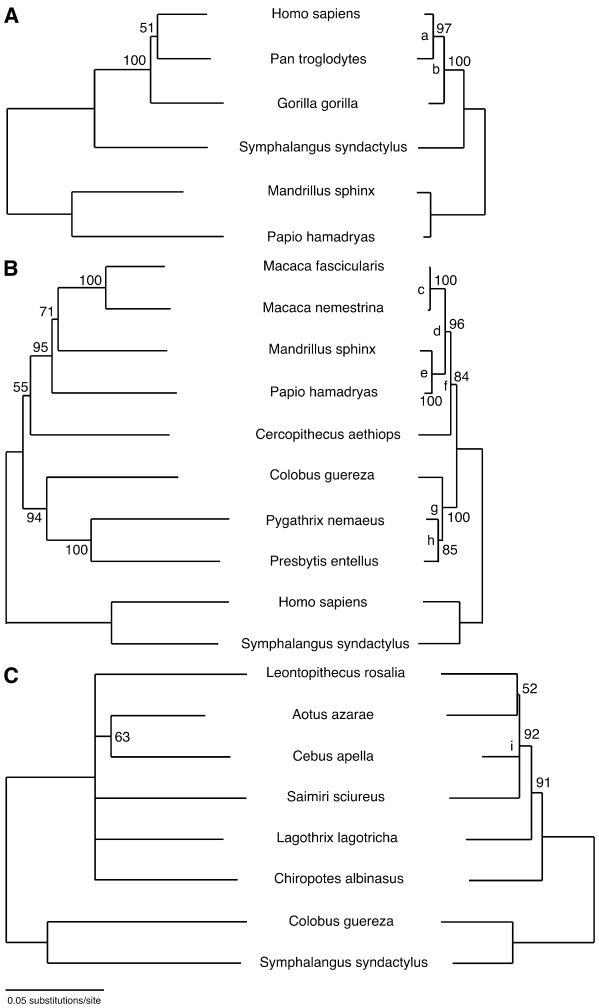

Fig. 4. Detailed phylogenetic reconstructions for mitochondrial genes (left) and mitochondrial pseudogenes (right). LogDet-based reconstruction for (**A**) Hominidae, (**B**) Cercopithecidae, and (**C**) Platyrrhini. Bootstrap values are indicated at the corresponding nodes. Branch lengths represent nucleotide substitutions per site. Lowercase letters indicate splits that were tested in a lento plot for support and conflicting signals (see Fig. 5).

compositional effects is the main source of inhomogeneity in mitochondrial data. In Fig. 1 we show the nucleotide composition for the investigated species for the complete data set (Fig. 1A), as well as for the single-stranded (Fig. 1B) and double-stranded (Fig. 1C) RNA-specifying regions. In comparison to the numt sequences, a high variation of the nucleotide composition is expressed in the mitochondrial DNA. However, the more constrained double-stranded region (Fig. 1C) represents a more homogeneous distribution of nucleotides and comes closer to the distribution as found in numts. Representing the sampling of Fig. 4B and restricting the analyses to the mt double-stranded regions, we found no further support for the erroneous clade merging the macaques and the mandrill in phylogenetic reconstructions (tree not shown). This indicates once more the effects of mitochondrial nucleotide compositional plasticity

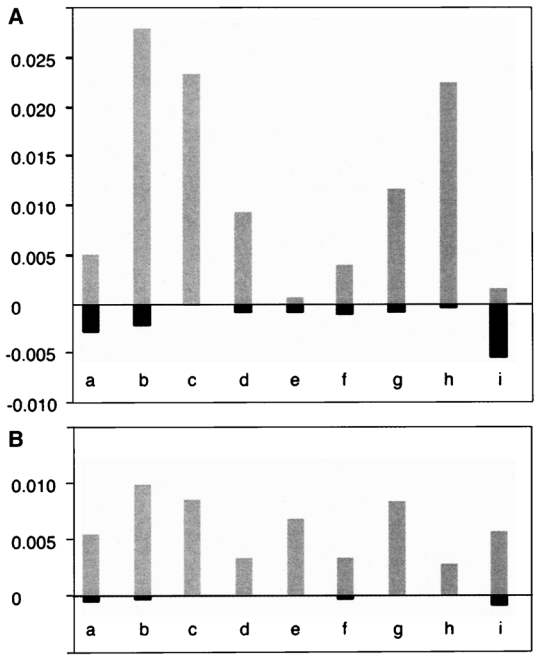

Fig. 5. Lento plot for the splits indicated as lowercase letters in Fig. 4 for (**A**) mitochondrial genes and (**B**) pseudogenes. Bars above the horizontal line denote the expected number of changes per site. Bars below the horizontal line characterize the relative conflict for incompatible splits.

on phylogenetic reconstructions. It has to be mentioned that nucleotide composition may be one, but not necessarily the only, parameter confounding phylogenetic tree reconstructions based on mitochondrial rDNA sequences. Restricting phylogenetic analyses to double-stranded rDNA sequence regions may be problematic because of a possible accumulation of compensatory changes that keep the secondary structures of rRNA intact. Thus, the assumption of independent evolution of single sites is likely to be violated. On the other hand, by interacting with different ribosomal proteins single-stranded rRNA regions may be particularly under functional constraints and may therefore show a heterogeneous conservation pattern.

Transitions/Transversions

Zhang and Hewitt (1996) mentioned that the ratio of transitions (Ti) versus transversions (Tv) is usually an order of magnitude lower in numts than in mt. One example is the Ti/Tv ratio of 10 and 1.7 for respective mt and numt cyt*b* sequences of the bird genus *Scytalopus* (Arctander 1995). The hypothetical translocation time is more than 5 MYA. On the other hand, Zischler et al. (1998) show mt and numt data of the mt control region present in all hominoids and absent in cercopithecoids. This implies a transposition time of more than 18 MY with Ti/Tv ratios of 0.9 and 4.8 for mt and numt, respectively. Our present mitochondrial RNA gene data of the more than 40-MY-old integration in primates give a Ti/Tv ratio of 1.59 and 2.2 for mt and numt, respectively. Mundy et al. (2000) found a Ti/Tv ratio of 4.25 and 2.57 for platyrrhine-specific cyt*b* mt and numts. They also found a respective value of 2.35 for two different numts and a value of 4.51 for an older numt. From this heterogeneity in the Ti/Tv ratios, we propose that there is no certain mechanism maintaining a characteristic Ti/Tv ratio in numts.

Molecular Dating

The accuracy of molecular dating depends on two basic requirements: first, a reliable fossil calibration of the molecular clock and, second, an accumulation of mutation over time in comparable stretches. The data presented indicate a high heterogeneity of the mitochondrial evolutionary rates in conjunction with a high plasticity of the nucleotide composition among different primate lineages. These cast doubt on the suitability of mitochondrial data for molecular dating in primates. On the other hand, numts represent homogeneity in both the rate of evolution and the nucleotide composition. For numts, quartet dating was therefore universally applicable while mitochondrial data produced 37 erroneous results of 505 possible quartets (Table 2). Finally, numt data tend to predate the currently established fossil splitting points (Table 2). Tavare et al. (2002) calculated an earlier divergence time for the first divergence of primates. Based on numt data our results support an earlier splitting point for Cercopithecoidea and Anthropoidea.

Rate and Pattern of Sequence Evolution in Mitochondrial rDNA and Corresponding Pseudogenes

There are two main features in the evolution of mitochondrial rDNA and corresponding nuclear pseudogenes that set them apart from each other. First, there is a pronounced difference in evolutionary rate heterogeneity over sites with alpha values, describing the gamma distribution of rates across the sites of 0.21 *vs.* 1.07 in mt and numt, respectively. Thus, nucleotide substitutions in the nuclear copies are distributed more randomly, whereas rate heterogeneity is more pronounced in mitochondrial rDNA. This most likely results from structural constraints exerted on the mitochondrial encoded rRNA molecule.

Second, the overall rate of evolution is lowered in the nuclear compartment, as can be seen upon comparing terminal branch lengths. Due to the differences in evolutionary rate for the different mitochondrial genes, with, e.g., rDNA evolving considerably slower than noncoding control regions,

Table 3. Nucleotide changes on the common branch of pseudogenes

	A	C	G	T
A		5	6	7
C	6		0	23
G	15	2		1
T	9	33[c]	1	

Note. The *X* axis indicates the nucleotide in the common ancestor of the genes and pseudogenes. The Y axis indicates the changes on the lineage leading to the pseudogenes [c]including three CpG sites.

this can be expected to a different extent for the various mtDNA–numt combinations. However, one of the sticking problems concerning the evolution of mitochondrial pseudogenes, an unexpected long branch leading to the most recent common ancestor (MRCA) of the numt sequences of single taxa, is expressed in the presented data set too. Extrapolating from a reduced evolutionary rate in the terminal branches of rDNA numts compared to the long branches leading to the mitochondrial rDNA sequences, the extended basal branch does not fit the hypothesis of numt evolution.

To take a closer look at the nucleotide compositional conditions that make up this branch, we analyzed all character changes on the lineage leading to the numts (see Table 3). From this substitution matrix it is obvious that C–T and T–C transitions dominate the changes that occurred on the branch leading to the common numt ancestor. However, of 33 C-to-T and 23 T-to-C changes, only 3 and 6, respectively, exhibit a consistency index (CI) exceeding 0.5. This indicates that the long branch leading to the numts is dominated by a majority of phylogenetically noninformative character changes.

Collura and Stewart (1995) argue that certain molecular processes during transposition may cause damage during the process of transposition entailing the unexpected long branch leading to the pseudogenes. In the face of the reduced numt evolutionary rate we would expect these changes to be reflected in high CI values, which apparently is not the case in our rDNA example. This is in line with the assumption that recent mitochondrial integrations should display a damage-induced increased branch length as well, which has not been confirmed by observation so far. Considering numt evolution on the nuclear level only, Zischler et al. (1998) suggested that changes caused by the methylation of CpG sites in the nuclear DNA could in part bring about the unexpected long branch leading to numts. However, in our sequences we could determine only few potential CpG methylation sites, which cannot fully explain the increased branch length.

Regarding sequence evolution on the mitochondrial site and taking into account the nucleotide compositions displayed in Fig. 1, we propose that the long branch leading to the numts is also brought about by the compositional plasticity of primate mitochondrial DNA. First, sequences of similar nucleotide compositions are likely to be clustered in phylogenetic tree reconstructions, and second, there is a possibility for reconstruction artifacts due to uneven nucleotide composition upon uniting mitochondrial genes and nuclear pseudogenes in a dataset.

We were able to show that nucleotide compositional changes are mostly expressed in the mt rDNA loop regions and that the phylogenetic signal is more robust in the conserved double-stranded region (Fig. 1). Restricting the analysis to double-stranded RNA specifying sequences, the branch leading to the pseudogenes is in fact only 2.6 times longer than the average of the terminal pseudogene branches. In comparison, the single-strand regions show a value of 10.2 (data not shown). We therefore propose that the longer the divergence time and the higher the evolutionary rate difference between genes and pseudogenes, the less convincing the phylogenetic tree on the basis of a mixed set of both mitochondrial and nuclear data. The long branch leading to the pseudogene is therefore predominantly not an expression of synapomorphic evolution of the nuclear pseudogenes. Rather, it can be explained by a divergence of the evolutionary mode mainly on the mitochondrial site. This is expressed in the nucleotide composition that is conserved upon integration into the nuclear genome, which causes difficulties when merging two different data sets together in one underlying reconstruction model.

Acknowledgments. We acknowledge the excellent technical assistance of Martina Ohme and Claudia Schwiegk. Thanks go to Christian Roos for primate DNA and to Kira Gee for editorial assistance. Two anonymous reviewers are gratefully acknowledged for their helpful comments.

References

Andrews TD, Jermiin LS, Easteal S (1998) Accelerated evolution of cytochrome b in simian primates: adaptive evolution in concert with other mitochondrial proteins? J Mol Evol 47:249–257

Arctander P (1995) Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proc R Soc Lond B Biol Sci 262:13–19

Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu XF, Janke A (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. Proc Natl Acad Sci USA 99:8151–8156

Bensasson D, Zhang D, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol 16:314–321

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Collura RV, Stewart CB (1995) Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378:485–489

Collura RV, Auerbach MR, Stewart CB (1996) A quick, direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes. Curr Biol 6:1337–1339

Delson E (1992) Evolution of Old World monkeys. In: Jones JS, Martin RD, Pilbeam D, Bunney S (eds) Cambridge encyclopedia of human evolution. Cambridge University Press, Cambridge, pp 217–222

Fukuda M, Wakasugi S, Tsuzuki T, Nomiyama H, Shimada K, Miyata T (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

Gibson A, Gowri-Shankar V, Higgs PG, Rattray M (2004) A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol 22:251–264

Gutell RR, Fox GE (1988) A compilation of large subunit RNA sequences presented in a structural format. Nucleic Acids Res 16:175–313

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hazkani-Covo E, Sorek R, Graur D (2003) Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J Mol Evol 56:169–174

Hixson JE, Brown WM (1986) A comparison of the small ribosomal RNA genes from the mitochondrial DNA of Great Apes and humans: sequence, structure, and phylogenetic implications. Mol Biol Evol 3:1–18

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754–755

MacFadden BJ (1990) Chronology of Cenozoic primate localities in South America. J Hum Evol 19:7–21

Mourier T, Hansen AJ, Willerslev E, Arctander P (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol 18:1833–1837

Mundy NI, Pissinatti A, Woodruff DS (2000) Multiple nuclear insertions of mitochondrial cytochrome b sequences in callitrichine primates. Mol Biol Evol 17:1075–1080

Nomiyama H, Fukuda M, Wakasugi S, Tsuzuki T, Shimada K (1985) Molecular structures of mitochondrial-DNA-like sequences in human nuclear DNA. Nucleic Acids Res 13:1649–1658

Page SL, Goodman M (2001) Catarrhine phylogeny: noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. Mol Phylogenet Evol 18:14–25

Penny D, Hasegawa M, Waddell PJ, Hendy MD (1999) Mammalian evolution: timing and implications from using the LogDeterminant transform for proteins of differing amino acid composition. Syst Biol 48:76–93

Perna NT, Kocher TD (1996) Mitochondrial DNA: molecular fossils in the nucleus. Curr Biol 6:128–129

Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442–448

Robinson M, Gouy M, Gautier C, Mouchiroud D (1998) Sensitivity of the relative-rate test to taxonomic sampling. Mol Biol Evol 15:1091–1098

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Schmitz J, Zischler H (2003) A novel family of tRNA-derived SINEs in the colugo and two new retrotransposable markers separating dermopterans from primates. Mol Phylogenet Evol 28:341–349

Schmitz J, Ohme M, Zischler H (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. Genetics 157:777–784

Schmitz J, Ohme M, Zischler H (2002a) The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol Biol Evol 19:544–553

Schmitz J, Ohme M, Suryobroto B, Zischler H (2002b) The colugo (*Cynocephalus variegatus*, Dermoptera): The primates' gliding sister? Mol Biol Evol 19:2308–2312

Schmitz J, Roos C, Zischler H (2005) Primate phylogeny: molecular evidence from retroposons. Cytogenet Genome Res 108:26–37

Singer SS, Schmitz J, Schwiegk C, Zischler H (2003) Molecular cladistic markers in New World monkey phylogeny (Platyrrhini, Primates). Mol Phylogenet Evol 26:490–501

Smith NGC, Eyre-Walker A (2003) Partitioning the variation in mammalian substitution rates. Mol Biol Evol 20:10–17

Springer MS, Stanhope MJ, Madsen O, de Jong WW (2004) Molecules consolidate the placental mammal tree. Trends Ecol Evol 19:430–438

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 26:148–153

Swofford DL (2000) PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA

Tavare S, Marshall CR, Will O, Soligo C, Martin RD (2002) Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726–729

Thomas R, Zischler H, Paabo S, Stoneking M (1996) Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. Hum Biol 68:847–854

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80:71–77

Yang Z (2000) Phylogenetic analysis by maximum likelihood (PAML), Version 30. University College London, London

Zhang D-X, Hewitt GM (1996) Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol Evol 11:247–251

Zischler H, Geisert H, Castresana J (1998) A hominoid-specific nuclear insertion of the mitochondrial D-loop: implications for reconstructing ancestral mitochondrial sequences. Mol Biol Evol 15:463–469