Multiple Nuclear Insertions of Mitochondrial Cytochrome b Sequences in Callitrichine Primates

Nicholas I. Mundy,* Alcides Pissinatti,† and David S. Woodruff‡

*Institute of Biological Anthropology, University of Oxford, Oxford, England; †Centro de Primatologia do Rio de Janeiro, Fundação Estadual de Engenharia do Meio Ambiente, Rio de Janeiro, Brazil; and ‡Department of Biology and Center for Molecular Genetics, University of California at San Diego

We report the presence of four nuclear paralogs of a 380-bp segment of cytochrome b in callitrichine primates (marmosets and tamarins). The mitochondrial cytochrome b sequence and each nuclear paralog were obtained from several species, allowing multiple comparisons of rates and patterns of substitution both between mitochondrial and nuclear sequences and among nuclear sequences. The mitochondrial DNA had high overall rates of molecular evolution and a strong bias toward substitutions at third codon positions. Rates of molecular evolution among the nuclear sequences were low and constant, and there were small differences in substitution patterns among the nuclear clades which were probably attributable to the small number of sites involved. A novel method of phylogenetic reconstruction based on the large difference in rates of evolution at different codon positions among mitochondrial and nuclear environments (Arctander 1995; Sunnucks and Hales 1996; Lopez et al. 1997) was used to determine whether different nuclear paralogs represent independent transposition events or duplications following a single insertion. This method is generally applicable in cases where differences in pattern of molecular evolution are known, and it showed that at least three of the four nuclear clades represent independent insertion events. The insertion events giving rise to two of the nuclear clades predate the divergence of the callitrichines, whereas those leading to the other two nuclear clades may have occurred in the common ancestor of marmosets.

Introduction

Since the first reports of transpositions of mitochondrial DNA into the nuclear DNA of vertebrates (Fukuda et al. 1985; Zullo et al. 1991; Smith, Thomas, and Patton 1992), further cases have been described at an increasing rate (reviewed in Zhang and Hewitt 1996). Several mitochondrial DNA insertions into the nucleus have been reported for primates (Fukuda et al. 1985; Collura and Stewart 1995; Zischler et al. 1995; van der Kuyl et al. 1995; Zischler, Geisert, and Castresana 1998), but to date these studies have been confined to humans, apes, and Old World monkeys. Such transpositions pose a problem for phylogenetic analyses if they remain undetected (e.g., Collura and Stewart 1995) but are sometimes useful as outgroups (Zischler et al. 1995).

The greatest significance of nuclear paralogs of mitochondrial DNA, however, may be their utility in comparing molecular evolution of homologous sequences in mitochondrial and nuclear environments (Arctander 1995; Sunnucks and Hales 1996; Lopez et al. 1997). Although data on comparison of modes of evolution in mitochondrial and nuclear compartments are accumulating, there is currently little information on whether there are differences in mode of evolution among different nuclear paralogs.

In this study, we report on multiple nuclear insertions of a segment of cytochrome b in callitrichine primates (marmosets and tamarins) which were discovered during a project on the molecular phylogeny of marmosets (*Callithrix*). These data allow the comparison of molecular evolution of mitochondrial DNA and four nuclear paralogs, along with comparison among the nuclear paralogs. We also present a novel method of phylogenetic reconstruction based on different modes of molecular evolution in the mitochondrial and nuclear compartments and use it to infer multiple independent insertion events of the nuclear paralogs.

The callitrichine primates comprise five genera (*Callithrix*, *Cebuella*, *Callimico*, *Saguinus*, and *Leontopithecus*) that form a sister clade with squirrel monkeys (*Saimiri*) and capuchins (Schneider et al. 1996; von Dornum and Ruvolo 1999). The marmosets consist of three species groups (Rylands et al. 1993), the *Callithrix jaccus* group in the Atlantic forest and eastern Brazil (including *C. jaccus*, *Callithrix penicillata*, *Callithrix kuhli*, *Callithrix geoffroyi*, and *Callithrix aurita* in this study), the *Callithrix argentata* group in central Brazil (including *C. argentata*, *Callithrix melanura*, *Callithrix humeralifer*, and *Callithrix maues* in this study) and the monospecific pygmy marmoset, *Cebuella*, in western Amazonia.

Materials and Methods

Samples

Blood or tissue samples were used for the following New World monkey individuals (all numbers are lab ID numbers): *C. jaccus* (Cjaccus 292), *Cebuella pygmaea* (Cebuella 592, Cebuella 25), *Saguinus oedipus* (Saguinus 296), *Leontopithecus rosalia* (Leontopithecus 22), *Callimico goeldii* (Callimico 1), *Saimiri oerstedii* (Soerstedtii 128), and Alouatta sara (Alouatta 584) (all from the Zoological Society of San Diego). Hair samples were used for the following marmosets, all from the Rio de Janeiro Primate Center: *C. jaccus* (Cjaccus 2), *C. argentata* (Cargentata 1), *C. melanura* (Cmelanura 3), *C. humeralifer* (Chumeralifer 3), *C. mauesi* (Cmauesi 2), *C. aurita* (Caurita 2), *C. geoffroyi* (Cgeof-
froyi 2), *C. penicillata* (Cpenicillata 4), and *C. kuhli* (Ckuhli 3). For one individual of *C. melanura* (Cmelanura 2), a museum skin from the Field Museum of Natural History was used. Hair samples from Brazil were imported under appropriate Brazilian and U.S. CITES permits.

**DNA Extraction, Amplification, Cloning, and Sequencing**

Genomic DNA was isolated from liver, skeletal muscle, and blood using standard procedures (Sambrook 1989). Total DNA was extracted from hairs using 5% Chelex 100 (Biorad, Hercules) following the method of Garza and Woodruff (1992). Total DNA was extracted from the museum skin using a commercial kit (Qiagen Tissue Kit; Qia-gen) according to the manufacturer’s instructions. Polymerase chain reaction (PCR) amplifications were performed in a thermal cycler (Hybaid) in a 25-μl total volume containing 0.5–1.0 U Taq polymerase (PE Applied Biosystems), 1 × PCR buffer, 50 mM each dNTP, 1.5 mM MgCl₂, and 25 μg bovine serum albumin (Fraction V, Sigma, St. Louis, Mo.). Cycling parameters were 94°C for 3 min, followed by 40 cycles of 94°C for 30 s, 50–60°C for 60 s, and 72°C for 90 s, followed by 72°C for 10 min. Direct cloning of PCR products was carried out using TA cloning kits (Invitrogen) according to the manufacturer’s instructions.

Primers and dNTPs were removed from PCR products using GeneClean (BIO 101). Sequencing was performed on both strands manually with Sequenase (USB) and 35S-dATP; for direct sequencing of PCR products, Sequenase (USB) was used (Garza and Woodruff 1992). Products of sequencing reactions were separated on 8% polyacrylamide gels (Long Ranger, J. T. Baker). Dried gels were exposed to X-ray film (X-Omat or Biomax; Kodak Eastman) for 1–7 days. Sequences were read by eye. In later experiments, sequencing was performed by cycle sequencing using dRhodamine terminators (PE Applied Biosystems) and run on an ABI 377 sequencing apparatus. Where possible, sequences were verified from multiple clones. Sequences obtained by TA cloning were given a transformation number and a clone number (e.g., Cjacchus 2.16.4 represents clone 4 from transformation 16, which was of specimen *C. jacchus* 2).

**Primers**

The original primers used to amplify a portion of cytochrome *b* were L15375 (5′-GGCTCAAGTAACCATT-CAGG-3′) and H5 (5′-TACTGGTTGTCTCCGATTC-3′). Primers used to amplify a 2.2-kb segment containing cytochrome *b* and control region were L14724 (5′-CGAAGCTTGATATGAAAAACCATCGTTG-3′) and CCRH2 (5′-CAGAAGGCTAGGACCCAAACCT-3′). Clade-specific primers were CBAF/CBR for clade A (CBAF: 5′-TGACGCCCATAATACCTATCAGGGG-3′; CBAE: 5′-GACCGGTTCTCCGAGTTTAGC-3′) and CBEF/CBER for clade E (CBEF: 5′-CCATCCACATCCTAGCATTGAA-3′; CBER: 5′-GCCCTCTGAAATGATATTGTCCTCA-3′).

**Sequence Analysis**

Sequences were unambiguously aligned by eye. As sequences obtained with different primer pairs were of slightly different lengths, the ends of some sequences were removed to create a 354-bp alignment for further analysis. Phylogenetic analyses were performed using PAUP* (Swofford 1999). For maximum-parsimony analyses, the transition-to-transversion ratio was set to 4:1 (taken from the data set), codon position weighting varied from first: second: third of 1:1:1 or 3:12:1 (taken from the data set), and random branch addition was used. For neighbor-joining trees, Kimura two-parameter distances were used. All bootstrap resamplings were performed with 500 replicates.

The significance of the ratio of interclade pairwise distances by codon position was determined using a randomization test. In each iteration of the test, the sequences belonging to the pair of clades under consideration were randomly assigned to two new groups with the same sizes as the clades. The difference in distance ratios by codon position for interclade comparison to the maximum intraclade comparison was then compared with the same difference obtained from the actual data. Results were considered significant at the 1% level if the difference from the randomization exceeded the difference from the data on 1% or fewer of the iterations (one-tailed test). All randomizations were performed using 500 iterations.

**Results and Discussion**

**Multiple Cytochrome b-like Sequences from Individual Marmosets**

In our early experiments on marmoset samples, unreadable sequencing ladders were generally obtained when PCR products with primers L15375 and H5 were directly sequenced. Cloning of PCR products revealed the presence of two or more cytochrome *b*-like sequences per individual, with between 11% and 25% uncorrected sequence divergence between the different sequences from the same individual. There were also occasional cases in which single clones from the same individual differed by a single base pair from other clones. Such instances were presumed to be due to PCR errors and were ignored.

Across-species comparisons showed that the sequences fell into five different groups, A–E. In order to increase the number of sequences obtained from each individual, primers specific to sequences from groups A and E were designed and used to obtain group A and E sequences from most individuals, although in most cases these sequences were still determined by cloning of PCR products (except for group A sequences from Cpenicillata 4 and Ckuhli 3). In total, two to five cytochrome *b*-like sequences were obtained from each individual marmoset sample (fig. 1), with five sequences being obtained from Chumeralifer 3. Relative to published primate cytochrome *b* sequences (e.g., those included in fig. 2), all sequences in group D have two deletions of 1 and 2 bp and all sequences in group E have independent deletions of 3 and 4 bp. Inspection of the sequences...
showed that none of the five groups of sequences could be explained by in vitro recombination of the other sequences during the PCR.

In a neighbor-joining tree (fig. 1), the five groups of sequences form five clades with high bootstrap support. Within each clade, the structure is generally consistent with known marmoset phylogeny (Tagliaro et al. 1997; unpublished data), e.g., the three species groups of marmosets, the jacchus group, the argentata group, and Cebuella, are monophyletic in every clade, with the exceptions of the jacchus and argentata groups in clade D and Cebuella in clade E. Similar results, including high bootstrap support for the five clades, were obtained by maximum parsimony (bootstrap values shown in fig. 1).

Identification of Clade A as the Mitochondrial Clade

The ratio of the mean number of substitutions for codon position 3 to codon position 2 within clades is much greater in clade A than in clades B–E (table 1), strongly suggesting that clade A contains functional mitochondrial sequences, whereas clades B–E consist of nuclear pseudogenes derived from mitochondrial DNA. Support for this assumption comes from the following:

1. All sequences in clade A are translatable, whereas all members of clades D and E and one of the clade C sequences (Chumeralifer 3.21.1) have mitochondrial stop codons. (2) Using conserved primers L14724 and CCRH2, a 2.2-kb segment of mitochondrial DNA containing cytochrome b and control region was amplified from samples Cjachus 292 and Cebuella 592 and directly sequenced. The clean sequences thus obtained matched the clade A sequences obtained from these samples using primers CBAF and CBAR.

Multiple nuclear paralogs of mtDNA have been found in a few cases (e.g., aphids [Sunnucks and Hales 1996] and birds [Sorensen and Fleischer 1996]) but our data set is unusual in comprising multiple sequences from multiple clades.
The failure to obtain sequences from all clades from all individuals is more likely to be for technical reasons (e.g., mismatches in primer sites, sequencing of a small number of clones) than the actual absence of the sequences in those individuals. We have no evidence that nuclear paralogs of mtDNA are more likely to be obtained from hair samples than from blood samples, as has been described for elephants (Greenwood and Pääbo 1999), since members of all four nuclear clades B–E were obtained from blood samples as well as hair samples.

Comparisons of Substitution Pattern Among Clades

As expected, the mitochondrial sequences evolve more rapidly and more unevenly with respect to codon position than do the nuclear clades: mean pairwise intracodon distances were almost five times as high for the mitochondrial clade as for the nuclear clades, and, in addition to the high codon position 3:2 ratio, the mean number of substitutions for codon position 3 to codon position 1 was higher for the mitochondrial than for the nuclear clades. The transition/transversion (Ts/Tv) ratio for the mitochondrial clade (4.25) was lower than expected, probably because of saturation—pairwise comparisons among closely related species gave much higher ratios.

Our data provide an unusual opportunity to compare evolution among multiple nuclear mtDNA paralogs. The overall rates of evolution of the clades are very similar (mean intracodon distances are presented in table 1), but there is an almost twofold variation in the Ts/Tv ratio. Additionally, there is some variation in rates of evolution by codon position away from the expected value of 1:1:1. As nuclear paralogs of mtDNA are unlikely to be functional, the most likely explanation for these variations is the small number of substitutions involved. For example, the relatively low Ts/Tv ratio and relatively high ratios of change by codon positions 3:1 and 3:2 for clade C is largely explained by the contribution from the four variable sites (at nucleotide positions 34, 113, 325, and 334) which split the jacchus and argentata species groups: of these four sites, three are at codon position 3 (two transversions and one transition) and one is at codon position 1 (transition). Also, the largest ratios of change by codon positions 3:1 and 3:2 occur among nuclear clade D, which has the smallest number of sequences (three).

### Table 1

Patterns of Substitution: Average Pairwise Differences Within Clades

<table>
<thead>
<tr>
<th>Clade</th>
<th>Transitions (Ts)</th>
<th>Transversions (Tv)</th>
<th>Total Changes</th>
<th>K2P Distance</th>
<th>Stop Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Total</td>
<td>1</td>
</tr>
<tr>
<td>A (N = 9)</td>
<td>8.03</td>
<td>2.34</td>
<td>20.28</td>
<td>30.65</td>
<td>0.94</td>
</tr>
<tr>
<td>B (N = 7)</td>
<td>0.86</td>
<td>2.86</td>
<td>1.90</td>
<td>5.62</td>
<td>0.86</td>
</tr>
<tr>
<td>C (N = 8)</td>
<td>2.46</td>
<td>0.50</td>
<td>1.64</td>
<td>4.54</td>
<td>0.00</td>
</tr>
<tr>
<td>D (N = 3)</td>
<td>1.34</td>
<td>1.33</td>
<td>3.33</td>
<td>5.96</td>
<td>0.67</td>
</tr>
<tr>
<td>E (N = 11)</td>
<td>1.60</td>
<td>3.49</td>
<td>0.87</td>
<td>6.00</td>
<td>1.30</td>
</tr>
</tbody>
</table>

**Age of Insertion Events**

In order to determine when the nuclear insertion events occurred, TA cloning from three primer pairs (L15375/H5, CBAF/CBAR, and CBEF/CBER) was performed for the remaining callitrichine genera (Callimico, Saguinus, and Leontopithecus) and members of two outgroups to the callitrichines (Saimiri and the more distant Alouatta). In all callitrichine genera and Saimiri, multiple cytochrome b–like sequences were found, whereas a single sequence was found for Alouatta. The following were identified as mitochondrial sequences: Saguinus 296.52.3, Soerstedti 128.63.2, and Alouatta 584.4mt (all confirmed by direct sequencing of L14724/CCHR2 product) and Leontopithecus 22.64.2 and Callimico 1.66.2 (these were obtained with primers CBAF/CBAR and do not contain stop codons, and their ratio of change by codon position is consistent with mtDNA).

All of these sequences, combined with published cytochrome b sequences of Saimiri sciureus and several Old World monkeys and apes (Collura and Stewart 1995), were used in phylogenetic reconstructions. In maximum-parsimony and neighbor-joining analyses, the five clades of marmoset sequences were conserved, and certain of the new sequences consistently grouped with one of the five clades. However, relationships of many of the new sequences and relationships among many of the clades were unstable.

Figure 2 shows a 50% consensus tree of 500 bootstrap replicates using maximum parsimony that is consistent with other trees generated. The five clades of marmoset sequences from the tree in figure 1 are still present and well supported. Sequences from the three other callitrichine genera (Saguinus, Leontopithecus, and Callimico) are present in an expanded nuclear clade E, and one sequence from Saguinus is present in nuclear clade D. Interestingly, none of the other callitrichine D or E sequences contain deletions like those in the marmoset sequences, showing that the deletions occurred in the common ancestor of marmosets (clade E) or since the split with Saguinus (clade D). None of the sequences from the other taxa were grouped with marmoset mitochondrial clade A or nuclear clade B or C.

Three further independent nuclear paralogs were identified: Callimico 2.66.2, which contains a stop codon and has a sister relationship with Callimico 2.66.1, the putative mitochondrial sequence from Callimico, and Soerstedti 128.55.5 and 128.55.3.
Nuclear transpositions of the same segment of cytochrome \( b \) have been described in Old World monkeys and apes (Collura and Stewart 1995). When these additional nuclear pseudogenes were included in phylogenetic reconstructions (not shown), they grouped with the other catarrhine (Old World monkey and ape) sequences and thus represent insertion events independent of those reported here in New World monkeys.

Do Clades B–E Represent Independent Nuclear Insertion Events or Duplication Following Insertion?

As standard methods of phylogenetic reconstruction had failed to resolve this issue, an alternative approach was taken based on the different substitution rates by codon position expected in functional mtDNA and nonfunctional nuclear inserts. The rationale for this approach is shown in figure 3, which shows the three possible phylogenies for mitochondrial clade A and nuclear clades B and C. In the two cases where nuclear insertion occurred independently (trees 1 and 2 in fig. 3), the interclade difference between clades B and C includes a period of mitochondrial evolution corresponding to the time between the two insertion events. In contrast, if clades B and C represent duplication from a single insertion event (tree 3 in fig. 3), the comparison between them does not include a period of mitochondrial evolution. The greatly elevated rate of third-position substitution in mitochondrial DNA and the overall slow rate of evolution of nuclear DNA mean that the signature of a period of mitochondrial DNA evolution may be detectable in interclade comparisons.

Pairwise interclade distance ratios by codon position (position 3/position 1 and position 3/position 2) comparisons are shown in table 2. As expected, the comparisons between clade A and all other clades are intermediate between the values for the two respective intraclade comparisons, as ratios by codon position are not expected to vary outside the extremes represented by evolution in the mitochondria or the nucleus. In contrast, many of the pairwise interclade comparisons among nuclear clades show elevated ratios of genetic distance by codon position. In particular, all pairwise comparisons involving clades B and C and other nuclear clades show greater ratios for both positions 3/1 and 3/2 compared to the corresponding intraclade values from table 1 for the relevant clades.

Table 2

<table>
<thead>
<tr>
<th>Clade</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>—</td>
<td>2.47</td>
<td>2.16</td>
<td>2.49</td>
<td>2.19</td>
</tr>
<tr>
<td>B</td>
<td>5.30</td>
<td>—</td>
<td>2.94*</td>
<td>3.04*</td>
<td>2.98*</td>
</tr>
<tr>
<td>C</td>
<td>(11.44, 0.69)</td>
<td>6.07*</td>
<td>(1.44, 1.21)</td>
<td>(1.44, 1.99)</td>
<td>(1.44, 0.30)</td>
</tr>
<tr>
<td>D</td>
<td>10.83</td>
<td>(0.69, 2.89)</td>
<td>—</td>
<td>2.83*</td>
<td>2.53*</td>
</tr>
<tr>
<td>E</td>
<td>(11.44, 3.02)</td>
<td>6.51*</td>
<td>(2.89, 3.02)</td>
<td>—</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Note.—Above diagonal: codon position 3/codon position 1; below diagonal: codon position 3/codon position 2. The numbers in parentheses below each value are the corresponding intraclade values from table 1 for the relevant clades.

* Significant at the 1% level.
positions 3/2 than do the respective intraclide comparisons, which is in the direction expected if a period of mitochondrial evolution was involved, and these differences are all significant. It should be noted that the variations in intraclide substitution patterns among nuclear clades (table 1) are not a significant confounding factor in this analysis, since the ratios for both positions 3/1 and positions 3/2 in the intraclide comparisons involving clades B and C are all greater than all of the intraclide comparisons. We conclude that clades B and C represent independent insertion events.

In contrast, ratios by codon position between clades D and E did not exceed maximum intraclide values. This suggests either that clades D and E represent duplications after a single nuclear insertion event, or that they arose by independent insertion events that occurred in a relatively short time interval. Analysis of the nuclear sequences flanking these inserts would be useful for distinguishing these possibilities.

Thus, knowledge of mode of sequence evolution can be a useful tool in phylogenetic analysis and can give better resolution for certain specific questions than traditional methods of phylogenetic reconstruction. The use of this method is not restricted to nuclear pseudogenes; it can be applied in any case in which different modes of sequence evolution are involved. The power of the method will increase when several sequences from different clades are available and when transposition events occurred at well-spaced time intervals.

**Supplementary Material**

Sequences have been deposited in GenBank under accession numbers AF245046–AF245095. The full sequence alignment is available from the journal website, or directly from N.I.M.

**Acknowledgments**

We thank the Center for the Reproduction of Endangered Species, the Zoological Society of San Diego, Oliver Ryder, the Field Museum, and Bruce Patterson for samples. Two anonymous reviewers made useful comments on the manuscript. This work was partly supported by the U.S. National Science Foundation.

**LITERATURE CITED**


B. FRANZ LANG, reviewing editor

Accepted March 27, 2000