Mitochondrial Insertions into Primate Nuclear Genomes Suggest the Use of numts as a Tool for Phylogeny

Einat Hazkani-Covo
National Evolutionary Synthesis Center, Durham, NC

Homoplasy-free characters are a valuable and highly desired tool for molecular systematics. Nuclear sequences of mitochondrial origin (numts) are fragments of mitochondrial DNA that have been transferred into the nuclear genome. numts are passively captured into genomes and have no transposition activity, which suggests they may have utility as phylogenetic markers. Here, five fully sequenced primate genomes (human, chimpanzee, orangutan, rhesus macaque, and marmoset) are used to reconstruct the evolutionary dynamics of recent numt accumulation in a phylogenetic context. The status of 367 numt loci is used as categorical data, and a maximum parsimony approach is used to trace numt insertions on different branches of the taxonomically undisputed primate phylogenetic tree. The presence of a given numt in related taxa implies orthologous integration, whereas the absence of a numt indicates the plesiomorphic condition prior to integration. An average rate of 5.65 numts per 1 My is estimated on the tree, but insertion rates vary significantly on different branches. Two instances in which the presence–absence pattern of numts does not agree with the phylogenetic tree were identified. These events may be the result of either lineage sorting or reversal. Using the numts reported here to reconstruct primate phylogeny produces the canonical primate tree topology with high bootstrap support. Moreover, numts identified in gorilla Supercontigs were used to test the human–chimp–gorilla trichotomy, yielding a high level of support for the sister relationship of human and chimpanzee. These analyses suggest that numts are valuable phylogenetic markers that can be used for molecular systematics. It remains to be tested whether numts are useful at deeper phylogenetic levels.

Homoplasy-free characters such as short interspersed elements (SINEs) are powerful tools for phylogenetic and population analysis, providing information that is independent of sequence or morphology (Shedlock and Okada 2000; Xing et al. 2005; Kriegs et al. 2006; Ray et al. 2006). Other rare genomic changes (RGCs) such as organelle gene order, changes in the genetic code, and protein insertions/deletions are also useful for molecular systematics (Rokas and Holland 2000). However, identifying informative RGCs for a given phylogeny is challenging, and new RGCs are needed.

Nuclear DNA sequences of mitochondrial origin, or numts (Lopez et al. 1994), are fragments of mitochondrial genome found in most eukaryote nuclear genomes (Bensasson, Zhang, et al. 2001; Richly and Leister 2004). The transfer of these fragments into the nuclear genome has been ongoing since free-living z-proteobacteria were acquired as mitochondria (Timmis et al. 2004). numts appear in markedly different abundances among species. Two processes can contribute to this difference: integration rates and postinsertions processes as duplications and deletions. Despite the growing catalog of numts from new genome sequences, our knowledge of recent numt insertions is based solely on analysis of single genomes such as human (Bensasson, Petrov, et al. 2001; Hazkani-Covo et al. 2003), vole (Triant and DeWoody 2007), or honeybee (Behura 2007) and on one pair of genomes (human–chimpanzee) (Ricchetti et al. 2004; Hazkani-Covo and Graur 2007). With the availability of multiple primate genomes, it is now possible to identify numt insertions based on genome alignment and to estimate their insertion rates in a phylogenetic context. Estimation of the frequency of mitochondrial transfer in closely related species should improve our understanding of the source of numt variation among species.

Zischler (2000) originally proposed that numts would provide useful phylogenetic markers. numts are passively captured into double-strand breaks in the nuclear genome via non-homologous end joining mechanism (Ricchetti et al. 1999; Hazkani-Covo and Covo 2008). There is no evidence for active mechanisms of numt excision, integration, or duplication (Bensasson et al. 2003). They have homology to different mitochondrial regions and thus have little similarity in length or sequence. Because integration of identical numts in the same nuclear site in different taxa is intrinsically unlikely, the presence of a numt at the same locus in two species suggests descent from a single insertion event, whereas the absence of a numt suggests that the species diverged prior to numt integration. The known ancestral condition of numt insertion (absence of numt) also provides unambiguous tree rooting, similar to the situation with SINEs (Shedlock et al. 2004). Whether numts can serve as RGCs for phylogenetic studies has not been tested in a comprehensive way. Here, numt insertions are identified and dated on an undisputed primate phylogenetic tree (Goodman et al. 1998), focusing on recent numt insertions in Catarrhini (Old-World monkeys, apes, and humans) after their divergence from the Platyrrhini (New-World monkeys). The results demonstrate the potential utility of numts as markers for inferring phylogenies.

The genomes of five primates were analyzed: humans (Homo sapiens), two great apes—the chimpanzee (Pan troglodytes) and orangutan (Pongo pygmaeus abelii), an Old-World monkey—rhesus macaque (Macaca mulatta), and a New-World monkey—the marmoset (Callithrix jacchus), which was used as an outgroup (IHGS 2001; CSAC 2005; Gibbs et al. 2007). The full numt complement in these genomes was identified by using Blast to search the nuclear genomes for regions of similarity with their own mitochondrial sequences (supplementary table 1, Supplementary Material online). numts were then identified that are present in some of the four ingroup species, but absent...
in others (present/absent-informative numts), by performing a pairwise comparison of each of the four species against the others (described in Hazkani-Covo and Graur 2007) based on alignments from the University of California Santa Cruz Genome Center (Karolchik et al. 2004). A numt is scored as absent in a species where there is a gap in the pairwise genome alignment that coincides with the boundaries of the numt in other species (fig. 1). For each numt that is absent in at least one of the four genomes, the corresponding locus in all genomes was determined. Possible character status include 1) presence of a numt, 2) absence of a numt 3) other (appearance of sequence other than numt or a gap in the alignment that is much longer than numt size), and 4) unknown (sequence not available or no detectable synteny).

Three hundred twenty-eight present/absent-informative numts from the primate genome set were then placed on the primate phylogenetic tree (fig. 2). Maximum parsimony (MP) character state reconstruction was used to infer when the insertion of each numt occurred. In an unweighted parsimony analysis, every change has the same cost. Because precise excision of a numt should be rare, weighted parsimony analysis was also performed, in which the penalty for numt removal (a → b) is two-times higher than the penalty of the other steps. Of the 328 numts, 235 show the predicted and perfect synteny patterns as indicated (below branches, fig. 2). Eighty-five of the remaining 93 numts were dated to a single branch, and eight were dated on two branches in an equally parsimonious manner using the weighted parsimony. The unweighted parsimony classifies the 93 numts with 64 on a single branch, 10 on two branches, and 19 on three or more branches (supplementary fig. 1, Supplementary Material online).

Numt insertions in close proximity to other inserted sequences might be missed in examination of only numt-size gaps in alignments, so alignments with gaps up to 1-kb flanking a given numt were also examined (e.g., see fig. 1c). Numts associated with larger gaps could reflect coinsertion with other sequences or postinsertion processes as deletions occurring as part of genome evolution. Because of the homoplasy concern, in the case of deletions, only loci that show perfect synteny patterns, where all genomes but the one with the numt show the same gap boundaries (fig. 2, below branches), were counted as new insertions. An additional 39 insertions were identified in this way, giving a total of 367 new numt insertions on the primate tree.

I found two cases in which the presence–absence pattern does not agree with the phylogenetic tree. In the first case, a numt present in human, chimpanzee, and rhesus, is missing from orangutan (fig. 3). In the second case, a numt appears in chimpanzee, orangutan, and rhesus but not in human and marmoset. The first event might be the result of a reversal because multiple Alu elements (that might have mediated the deletion) are present near the numt. It seems that numt reversal causing homoplasy is rare, as is precise deletion of Alu elements (36 events, van de Lagemaat et al. 2005). Lineage sorting might be a better explanation of these events if an ancestral polymorphism regarding the absence/presence of numts becomes fixed in different lineages in a form inconsistent with speciation (Shedlock et al. 2004).

Weighted parsimony results yielded 40, 68, 90, and 101 species-specific numts for human, chimpanzee, orangutan, and rhesus macaque, respectively (fig. 2). Similar numbers were obtained by unweighted parsimony (supplementary fig. 1, Supplementary Material online). The number of insertions on inner branches was also determined. For consistency, numts showing equally parsimonious assignments to two branches are counted on the more basal one. The branch leading to the common ancestor of human and chimpanzee after the divergence of orangutan has 29 numt insertions (22/29 are unique to the branch). Insertion of 39 numts is reconstructed in the common ancestor of the apes and humans after its divergence from the Old World monkeys (38/39 are unique to this branch). Most of the 367 numts correspond to a single mitochondrial sequence. However, 25 numts are composed of several mitochondrial fragments, up to five mitochondrial fragments per numt.

Numt insertion rates were calculated by dividing the number of numts (by weighted parsimony) by the length of their respective branches: 6, 13, and 23 Ma for human–chimpanzee, human–orangutan, and human–rhesus, respectively (Glazko and Nei 2003). The average rate on the entire tree is 5.65 numts per 1 My (367 numts/65 My), in agreement with previous estimates of 5.1–5.7 numts per 1 My (Bensasson et al. 2003; Ricchetti et al. 2004; Hazkani-Covo and Graur 2007). Insertion rate varies significantly among the six branches ranging from 3.9 numts in 1 My on the human–chimp–orangutan branch and 11.3 numts in 1 My on the chimpanzee branch (fig. 2). The null hypothesis that numt sample came from a population having a 6:6:13:23:7:10 ratio in human:chimpanzee:orangutan: rhesus:human–chimpanzee:human–chimpanzee–orangutan in corresponding to their branch lengths was rejected ($\chi^2$ goodness of fit, df = 5, $P < 2 \times 10^{-16}$). The trend toward lower rates on more basal branches might result from insufficient synteny to classify recent numts or from losses on subsequent branches. However, differences in
FIG. 2.—Placement of 367 numt branch–specific insertions on the primate phylogenetic tree. The total number of numts in genomes (regardless of their insertion time, supplementary table 1, Supplementary Material online) is shown above branches. The number of branch–specific numts is shown below branches. The format is $A + B$, where $A$ is the number inferred under weighted MP and $B$ the number that are associated with large gaps. The most common form of genome alignment is shown below branches; arrows point to the possible branches. Numbers in open circles represent the number of numt loci with equally parsimonious placements on two branches; gray circle diameter is proportional to the rate of numt insertion. The rate values, calculated as the number of weighted species-specific numts (summarized above circles) divided by time (Glazko and Nei 2003), appear inside the circles. Dollo parsimony resulted in the same tree and the numbers in parentheses indicate the percentage of bootstrap replicates that support each node. Tree is not scaled.
mitochondria can be Blasted against the nuclear genome of the species in a phylogenetic study. In these cases, the nuclear and mitochondrial genome is available for one numt genetic branching order of species (Bensasson, Zhang, et al. 2001; see discussion in Salem et al. 2003). The 137 numt that were previously placed on human, chimpanzee, and human–chimpanzee branches were used to interrogate the gorilla Supercontigs using BLAT. The criterion that were used for phylogeny reconstruction. Using Dollo parsimony (PAUP*, Swofford 2003) produces the same canonical primate tree topology (fig. 2) with 100% bootstrap support. The recent Gorilla gorilla genome draft now available from the Sanger center provides the opportunity to test the human (H), chimpanzee (C), and gorilla (G) trichotomy using numts. Although the consensus topology is ((H,C),G), some nuclear and mitochondrial analysis support other phylogenies (Satta et al. 2000; Chen and Li 2001; see discussion in Salem et al. 2003). The 137 numts that were previously placed on human, chimpanzee, and human–chimpanzee branches were used to interrogate the gorilla Supercontigs using BLAT. The criterion that both numt nuclear flanking regions (100 bp) are present on a single Supercontig resulted in 82 numts that were used for phylogeny reconstruction. Using Dollo parsimony results in the verification of sister grouping of human and chimpanzee with 100% bootstrap support.

These results confirm that the presence or absence of numts at specific loci is useful in determining the phylogenetic branching order of species (Bensasson, Zhang, et al. 2001). numts might be particularly beneficial in cases where the nuclear and mitochondrial genome is available for one of the species in a phylogenetic study. In these cases, the mitochondria can be Blasted against the nuclear genome to search for numts, and these numts can be tested in the other species using polymerase chain reaction (Ricchetti et al. 2004). The availability of many mitochondrial genomes has an additional advantage in organisms like metazoans, where the mean evolutionary rate in mitochondrial genomes is about 10 times higher than in the nuclear genome (Brown et al. 1982). In these cases, the DNA distances between numts and mitochondria can be used to make a rough estimate of numt insertion time (Hazkani-Covo et al. 2003; Ricchetti et al. 2004). Thus, when searching for markers to distinguish between several tree topologies, one could specifically amplify numts that date to the branches of interest (hence with a better chance of being informative).

Supplementary Material

Supplementary table 1 and supplementary figure 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

numt insertion rates also appear between the sister taxa human and chimpanzee (P < 0.007), suggesting that integration rate may be an important factor in the variable numt coverage in genomes. Given that reversals are rare, it seems likely that numts classified as new insertions are indeed new. Using the numts reported here to reconstruct primate phylogeny with Dollo parsimony (PAUP*, Swofford 2003) produces the same canonical primate tree topology (fig. 2) with 100% bootstrap support. The recent Gorilla gorilla genome draft now available from the Sanger center provides the opportunity to test the human (H), chimpanzee (C), and gorilla (G) trichotomy using numts. Although the consensus topology is ((H,C),G), some nuclear and mitochondrial analysis support other phylogenies (Satta et al. 2000; Chen and Li 2001; see discussion in Salem et al. 2003). The 137 numts that were previously placed on human, chimpanzee, and human–chimpanzee branches were used to interrogate the gorilla Supercontigs using BLAT. The criterion that both numt nuclear flanking regions (100 bp) are present on a single Supercontig resulted in 82 numts that were used for phylogeny reconstruction. Using Dollo parsimony results in the verification of sister grouping of human and chimpanzee with 100% bootstrap support.

These results confirm that the presence or absence of numts at specific loci is useful in determining the phylogenetic branching order of species (Bensasson, Zhang, et al. 2001). numts might be particularly beneficial in cases where the nuclear and mitochondrial genome is available for one of the species in a phylogenetic study. In these cases, the mitochondria can be Blasted against the nuclear genome to search for numts, and these numts can be tested in the other species using polymerase chain reaction (Ricchetti et al. 2004). The availability of many mitochondrial genomes has an additional advantage in organisms like metazoans, where the mean evolutionary rate in mitochondrial genomes is about 10 times higher than in the nuclear genome (Brown et al. 1982). In these cases, the DNA distances between numts and mitochondria can be used to make a rough estimate of numt insertion time (Hazkani-Covo et al. 2003; Ricchetti et al. 2004). Thus, when searching for markers to distinguish between several tree topologies, one could specifically amplify numts that date to the branches of interest (hence with a better chance of being informative).

Acknowledgments

Research support was provided by the National Evolutionary Synthesis Center (NSF #EF-0423641). I would like to thank Todd Vision, Trina Roberts, Brian Sidlauskas, David Swofford, Brian O’Meara, and Dorothée Huchon for discussion and comments. The marmoset and orangutan genomes were produced and kindly provided by The Genome Center at Washington University School of Medicine in St Louis and can be obtained from ftp://genome.wustl.edu/pub/. The gorilla sequence data were produced by the gorilla Sequencing Group at the Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/sequences/gorilla.

Literature Cited


<table>
<thead>
<tr>
<th>Human</th>
<th>TTGAACCTCCTAAAGCCCCCTTA</th>
<th>/230TAAAAACCTTTGATGCTCAAAATTATC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimp</td>
<td>TTGAACCTCCTAAAGCCCCCTTA</td>
<td>TAAAAACCTTTGATGCTCAAAATTATC</td>
</tr>
<tr>
<td>Orangutan</td>
<td>TTGAACCTCCTTTAAAGCCCCCTTA</td>
<td>GGGCCTCAAGTATTATC</td>
</tr>
<tr>
<td>Rhesus</td>
<td>TTGAATTTCTCAAAGCCCCCTTA</td>
<td>TAAAAACCTTTGATGCTCAAAATTATC</td>
</tr>
<tr>
<td>Human_Mi</td>
<td>-----------------------------</td>
<td>---------------------------------</td>
</tr>
</tbody>
</table>