J Mol Evol (2003) 57:343–354
DOI: 10.1007/s00239-003-2485-7

JOURNAL OF MOLECULAR EVOLUTION

Rates of DNA Duplication and Mitochondrial DNA Insertion in the Human Genome

Douda Bensasson, Marcus W. Feldman, Dmitri A. Petrov

School of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA

Received: 7 October 2002 / Accepted: 21 April 2003

Abstract. The hundreds of mitochondrial pseudogenes in the human nuclear genome sequence (numts) constitute an excellent system for studying and dating DNA duplications and insertions. These pseudogenes are associated with many complete mitochondrial genome sequences and through those with a good fossil record. By comparing individual numts with primate and other mammalian mitochondrial genome sequences, we estimate that these numts arose continuously over the last 58 million years. Our pairwise comparisons between numts suggest that most human numts arose from different mitochondrial insertion events and not by DNA duplication within the nuclear genome. The nuclear genome appears to accumulate mtDNA insertions at a rate high enough to predict within-population polymorphism for the presence/absence of many recent mtDNA insertions. Pairwise analysis of numts and their flanking DNA produces an estimate for the DNA duplication rate in humans of 2.2×10^{-9} per numt per year. Thus, a nucleotide site is about as likely to be involved in a duplication event as it is to change by point substitution. This estimate of the rate of DNA duplication of noncoding DNA is based on sequences that are not in duplication hotspots, and is close to the rate reported for functional genes in other species.

Correspondence to: Douda Bensasson, Evolutionary Genomics Department, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA; *email:* douda@pseudogene.net

Key words: Numt — numtDNA — Segmental duplication — Human population genetic markers

Introduction

There is a remarkable lack of information about mutations that involve more than a few nucleotides but are not visible at the chromosomal level. These are only now becoming accessible for study and the mutation rate at this level is surprisingly high (Lynch and Conery 2000; International Human Genome Sequencing Consortium 2001; Bailey et al. 2002a; Samonte and Eichler 2002). Duplications arising in the last 40 million years contribute to a significant proportion (> 5%) of the human genome (Bailey et al. 2002a; Samonte and Eichler 2002). Studies of yeast, *Drosophila*, and *C. elegans* also reveal many newly arisen gene duplicates, and rates of gene duplication have been estimated from these by dating duplications using generalized point substitution rates (Lynch and Conery 2000, 2001; Long and Thornton 2001; Gu et al. 2002). In this study we develop another approach and estimate the rate of DNA duplication, by using some of the hundreds of mitochondrial pseudogenes (numts) in the human nuclear genome (Fukuda et al. 1985; Bensasson et al. 2001; Mourier et al. 2001; Woischnik and Moraes 2002) to identify and date DNA duplication events.

There has also been a lot of recent interest in how and when the many numts in the human genome arose (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Hazkani-Covo et al. 2003). Human numts are associated with the fast-evolving and well-

characterized primate mtDNA lineage, and through that with a good fossil record, and so can be used to estimate the rate of mtDNA insertion into the nuclear genome and the rates of other mutational events involving numts. Numts are evenly distributed within and among chromosomes (Woischnik and Moraes 2002), and because in animals they have lost their function (Bensasson et al. 2001), they are likely to reflect the general DNA duplication rate of unconstrained sequence.

Each numt DNA sequence arose either as a new insertion of mtDNA from the mitochondrion or by DNA duplication of another previous insertion in the nuclear genome. It is clear from phylogenetic analysis of human numts and extant primate mtDNA that fragments of mtDNA have inserted into the nuclear genome multiple times in recent primate history (Fukuda et al. 1985; Hu and Thilly 1994; Mourier et al. 2001; Hazkani-Covo et al. 2003). Unfortunately, phylogenetic analyses cannot always firmly resolve whether a pair of numts arose from independent mitochondrial insertions. If two numts appear as sister taxa in a phylogenetic analysis, this may suggest that one arose by duplication of the other in the nucleus (Hazkani-Covo et al. 2003). However, independently arising numts can also appear as sister taxa if they arose in the same evolutionary period and so were very similar in sequence, or if they arose from a mitochondrial lineage (an ancestral polymorphism) that is now extinct, or due to imperfect phylogenetic reconstruction (Bensasson et al. 2001).

Here we develop a pairwise sequence comparison approach that uses the differences between numt and mtDNA evolution to distinguish mtDNA insertions from nuclear DNA duplication events. Since most, if not all, animal numts are noncoding, they apparently evolve without the selective constraints to which their mitochondrial progenitors are subject (Gellissen and Michaelis 1987; Perna and Kocher 1996). In brief, evidence of selective constraint on the differences between two numts is evidence that they arose from different mtDNA insertions (e.g., Fig. 1, numts A and H), whereas numts that arose by nuclear duplication (e.g., Fig. 1, numts A and B) should show no evidence of selective constraint in their nucleotide differences (Bensasson et al. 2000; Mundy et al. 2000). Using this distinction, we can exclude many numt pairs from a more labour intensive search for DNA duplications.

Because numts have no self-replicating mechanism, each numt duplication is part of a larger region of duplicated DNA, so numt pairs that arose by nuclear DNA duplication are characterized by DNA sequence homology that extends beyond numt regions. We use this characteristic to identify numts that arose by duplication, to estimate the rate at which mtDNA has been inserted into the nuclear genome, and to estimate the rate of DNA duplication.

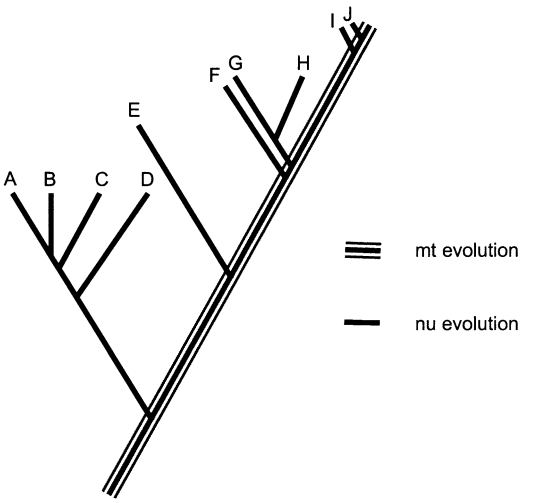

Fig. 1. The evolution of numt and mtDNA lineages. When DNA evolves in the absence of selective constraint, nucleotide changes are expected to accumulate at an equal rate at the 1st, 2nd, or 3rd positions of codons and this mode of evolution is illustrated by single black lines. The pattern of evolution expected in mitochondrial evolution is one of selective constraint and this is represented by *triple lines*.

Methods

Identifying and Characterizing Numts

Build 28 (the February 8, 2002, release) of the human genome project DNA sequence database was downloaded in FASTA format from NCBI (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/) and formatted and queried using the formatdb and blastn programs from the BLAST 2.1.3 package (from ftp://ncbi.nlm.nih.gov/blast/executables). Mitochondrial pseudogenes in the human nuclear genome (numts) were identified by using the whole human mtDNA genome sequence (accession No. NC_001807.3) as query with the "expect" threshold set to 0.0001, "word size" set at 7, and default settings. Chromosomal assignments for each numt were read from the title of each contig accession hit.

Most of the analysis described was automated in PERL (programs are available on request or at http://www.pseudogene.net).

BLAST Output Parsing and Numt Sequence Alignment. Hits less than 2.5 kb apart in the same contig were treated as part of the same numt. The local alignment generated by BLAST, as shown with the "query anchored without identities" output option for BLAST, was converted into FASTA format for use as the numt DNA sequence alignment. The BLAST local alignment tool does not return a complete alignment and large insertions or diverged segments, which are difficult to align, will not be included. This is an advantage for detecting selective constraints whose signal will be strongest in the most conserved DNA regions and which are best tested using the most strongly supported parts of an alignment. However, numt length and divergence from other sequences in the alignment will be underestimated. Tandem repetition of nuclear DNA sequence homologous to mtDNA and small insertions were removed from the alignment, thus maintaining the protein-coding reading frames of the mtDNA sequence against which numt sequences were aligned. Alignments were checked by eye using Bioedit Sequence Alignment Editor (Hall 1999; available at http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

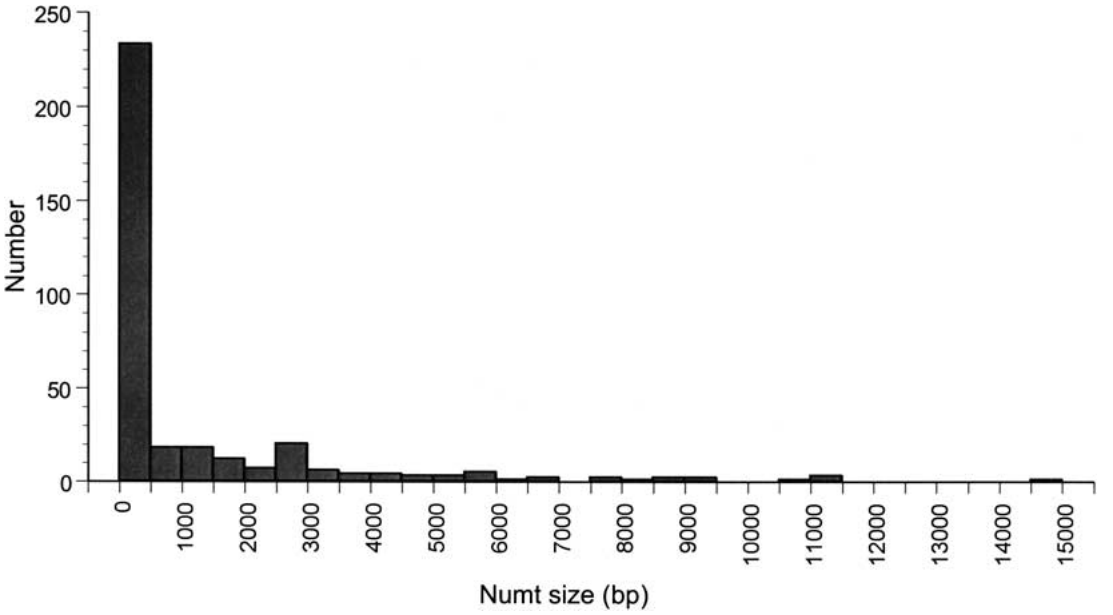

Fig. 2. The size distribution of numts identified in the human genome project sequence. These lengths include insertions, deletions, and tandem repeats.

Table 1. Summary of human numt use in this study

Analysis	Criterion	Number
Numt size, chromosomal distribution, analysis of divergence from mtDNA	All numts found using BLAST	348
Dating mtDNA insertions by phylogenetic reconstruction	Numts paralogous to >500 bp mtDNA alignment, minus the 19 numts that arose by DNA duplication	82
Analysis of selective constraints	Numts paralogous to >30 bp protein-coding DNA	236
Identification of DNA duplications by numt flanking DNA analysis	Numts paralogous to >200 bp protein-coding or >500 bp rRNA-coding DNA	127

Numt Sequence Analysis. Two lengths were estimated for each numt sequence. The first is the absolute length a numt spans in a contig sequence entry and therefore includes insertions, tandem repeats, and regions that are not easily aligned (lengths summarized in Fig. 2). The second represents the (ungapped) number of nucleotides used in the DNA sequence alignment and subsequent analyses. This second length was used to summarize and group data where sequence length may affect statistical power (Table 1).

Phylogenetic Reconstruction of the Age Distribution of Numts

Mitochondrial DNA Alignment. Whole mtDNA genome sequences from nine primates (human, *Homo sapiens*: NC_001807; chimp, *Pan troglodytes*: NC_001643; gorilla, *Gorilla gorilla*: NC_001645; orangutan, *Pongo pymaeus*: NC_001646; gibbon, *Hylobates lar*: NC_002082; baboon, *Papio hamadryas*: NC_001992; capuchin, *Cebus albifrons*: NC_002763; tarsier, *Tarsius bancanus*: NC_002811; loris, *Nycticebus coucang*: NC_002765) and two other mammals (treeshrew, *Tupaia belangeri*: NC_002521; mouse, *Mus musculus*: NC_001569) were aligned using ClustalW (1.81). To maintain the frame in which numt sequences were aligned against human mtDNA, nucleotide insertions in mtDNA relative to the human mtDNA sequence were removed in Bioedit Sequence Alignment Editor. Alignments were cropped to include only protein and rRNA coding regions, which represent the most conserved and better-aligned regions.

Dating Individual mtDNA Insertions by Phylogenetic Reconstruction. The numts used in this analysis were the 82 numts of 348 numts that had over 500 bp of homology to the mtDNA alignment and that did not arise by DNA duplication (Table 1; DNA duplications were identified as described under identification of DNA duplifications by Numt Flanking DNA Analysis, below). Shorter numts were excluded because their phylogenetic placement cannot readily be resolved.

The point at which each numt diverged from the mtDNA lineage was approximated by aligning each numt against the mammalian mtDNA sequence alignment described above and reconstructing its phylogenetic position relative to these sequences using PAML (Yang 1997). The PAML control file was referred to nine user trees for the analysis of each numt. These nine trees followed the taxonomic topology shown in Fig. 3 for mtDNA, which is in agreement with the phylogeny expected for these taxa (Goodman et al. 1998; Schmitz et al. 2001, 2002), with the numt falling out along one of branches 1–9 (Fig. 3) for each tree. The most likely tree was saved for each numt.

This PAML analysis of 82 numts was automated and repeated using three different molecular models, referred to here as HKY, HKY + G, HKY + G + gamma. In each case the model used was

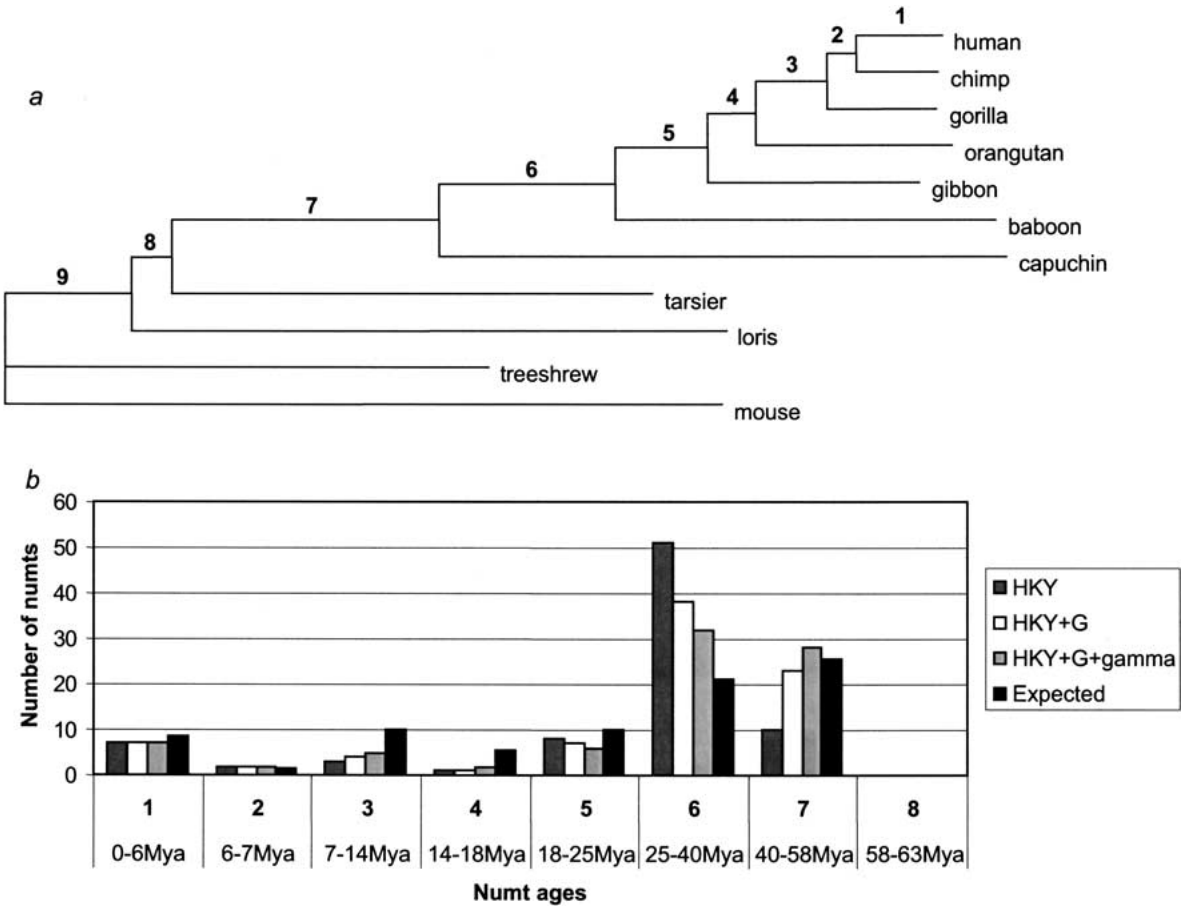

Fig. 3. Dating mtDNA insertions by phylogenetic reconstruction. Maximum likelihood trees were reconstructed for each numt relative to the mtDNA sequences shown in the tree in **a**; **a** shows branches 1–9, from which each numt may have diverged from the mtDNA lineage. The bars in each histogram bin 1–8 in **b** show the number of numt trees, for which numts diverged from branches 1–8, respectively (no numts diverged from branch 9). Analyses of each numt were made using PAML and the models described as HKY, HKY + G, and HKY + G + gamma under Methods. The "expected" values in the histogram are those that would be expected if mtDNA inserted into the nuclear genome at a uniform rate and if the primate divergence dates are accurate. The topology of the tree in **a** was estimated by analysis of 13.9 kb of mtDNA sequences in PAML using the HKY + G + gamma model and given a user tree with the branching order shown. Maximum likelihood analysis in PAUP, with a heuristic search and the HKY model, resulted in a tree with the branching order that is shown here and is consistent with the phylogeny expected for these taxa (Goodman et al. 1998; Schmitz et al. 2001, 2002).

HKY85; that of Hasegawa et al. (1985) and positions with missing or ambiguous data were removed from the alignment. For HKY, kappa was set at 8.3. For HKY + G, kappa was estimated for each tree by maximum likelihood and option G was used to estimate substitution rates separately for first, second, and third positions of codons and for rRNA positions. For HKY + G + gamma, the same settings were used as for HKY + G except a gamma distribution was applied to each of the four types of nucleotide position (1st, 2nd, 3rd and rRNA). The alpha parameter for each gamma distribution was set at 0.59 with four categories (variable ncatG in PAML = 4). The value (0.59) for the alpha parameter of the gamma distribution was estimated by maximum likelihood, using the HKY + G + gamma model and the mtDNA tree shown in Fig. 3.

A mitochondrial and not a nuclear mode of evolution was modeled because nucleotide changes sustained in the nucleus are restricted only to the numt lineage and we expect few changes in the nucleus because of its slower rate of mutation (Brown et al. 1982). The numt lineage forms a terminal branch and would therefore not be phylogenetically informative.

The best topology estimated using PAML was confirmed for eight numts, of various predicted ages, by maximum likelihood analysis and a heuristic search of all possible trees using Paup 4.0 (Swofford 2002) (data not shown).

The dates used to describe the bins in the histograms in Fig. 3 were those estimated by Goodman et al. (1998, Table 5). Expected numbers of numts falling on a particular branch in the histogram were based on these estimates of the relative length of time between each clade, assuming that numts accumulated at a uniform rate over the last 58 million years.

Detecting Selective Constraints in the Divergence Between Two Numts

Pairwise comparisons were made between the 236 of 348 numts that had at least 30 bp of sequence paralogous to protein-coding mtDNA (Table 1). The regions in the full numt alignment with homology to each mitochondrial protein-coding gene were concatenated into a single alignment that could be read in a continuous reading frame. The numbers of differences occurring at 1st, 2nd, or 3rd positions of codons were summed for every pairwise comparison. For every numt pair with more than 25 nucleotide differences, a chi-square test was used to test whether the number of differences

at 1st, 2nd, or 3rd positions of codons was significantly different from the 1:1:1 ratio expected under noncoding DNA divergence. For pairs with fewer than 25 differences in their sequence, the numbers of differences at 1st and 2nd positions were pooled, and a binomial exact probability was calculated to test whether the ratio of differences at 1st and 2nd positions to those at 3rd positions was significantly different from 2:1. Pairs with fewer than four differences were not tested and numt pairs that overlapped by less than 30 bp in the numt alignment were not compared.

Identification of DNA Duplications by Numt Flanking DNA Analysis

Pairs of numts and their flanking DNA were compared to identify which numts arose by DNA duplication. If one numt arose from another by DNA duplication in the nucleus, homology between the numts should extend into the nuclear DNA that flanks them and the degree of similarity between the numts and between flanking DNA regions should be the same. Whole contig sequence entries were downloaded for each numt pair examined, or 1 Mbp from either side of the numt if contig entries were longer than 5 Mbp. The length and percentage similarity of numt and flanking DNA homology (if present) were determined using Owen 1.2 (ftp://ncbi.nlm.nih.gov/pub/kondrashov/[Ogurtsov et al. 2002]). Similar results were obtained using BLAST.

As the number of numt comparisons possible is large (60,378 pairs; i.e., $348 \times 347 \times 0.5$), we reduced the number of comparisons by analyzing the flanking DNA of only the longest 127 numts (Table 1). These 127 (of 348) numts were the 113 numts that had >200 bp homology to protein-coding mtDNA and the 14 numts with >500 bp of homology to non-protein-coding mtDNA but <200 bp of homology to protein-coding mtDNA (Table 1). The longest numts were used because most are long enough to be dated by phylogenetic reconstruction.

The number of these possible pairwise comparisons (8001 pairs; $127 \times 126 \times 0.5$) was also substantially reduced by restricting analysis to the pairs of numts that overlapped in their homology to mtDNA. At the time of duplication, at least, this "overlap" should be complete. The number of pairs of numts for which flanking DNA was analyzed was still further reduced by excluding the pairs of numts that showed evidence of selective constraint in their differences. That is, significant bias in the number of nucleotide differences at 1st, 2nd, or 3rd positions of codons (at the 0.05 level without Bonferroni correction) or with an excess of differences at the 3rd position ($>40\%$ of at least 15 differences).

These restrictions (length, overlapping homology to mtDNA, and absence of selective constraints in differences) greatly reduce the number of comparisons of flanking DNA.

The expected length and degree of similarity between numts were confirmed for all numt pairs examined using Owen 1.2, and the chromosomal positions of each DNA segment involved in a duplication were estimated using the Map View tool for the human genome at NCBI.

Results

The Mitochondrial Pseudogenes Used in this Study

We identified 348 numts (Table 1) and their size distribution, shown in Fig. 2, is consistent with past observations of numts in the human genome project sequence (296 [Mourier et al. 200l] and 354 [Bensasson et al. 200l]) and with past estimates of hundreds of human numts from hybridization studies (Fukuda et al. 1985). Many more than 348 numts can be identified in the human genome, as illustrated by a recent study by Woischnik and Moraes (2002) that describes 612 mtDNA-like sequences. The number of numts identified by Woischnik and Moraes is probably much higher than other estimates because they used a BLAST program (tblastn) that is better suited to the identification of diverged protein-coding sequences and may identify numts that are more diverged from the modern mtDNA sequence. Woischnik and Moraes also used lower thresholds for statistical significance (a BLAST expect threshold of 10). The 348 numts studied here probably represent those that are less diverged from the modern mtDNA sequence and therefore arose more recently in primate history. Some numts are also missing from the human genome project sequence because it is incomplete. Our search covered approximately 84% of the human genome sequence (2860 Mbps from approximately 3400 Mbps [Li 1997]).

In agreement with past observations (Woischnik and Moraes 2002; Hazkani-Covo et al. 2003), the numts in this study appear to be evenly distributed among chromosomes. The 346 numts that have been mapped to chromosomes are distributed among chromosomes in approximately the proportions expected from the length of chromosome sequence represented in build 28 of the human genome sequence (G test on all chromosomes: $p = 0.22$, df = 23; G test on autosomes, X and Y: $p = 0.98$, df = 2).

The Oldest Numts in this Dataset Arose 58 Million Years Ago

Phylogenetic analysis suggests that the most ancient numts in our dataset predate the divergence of Old World and New World monkeys (40 Mya) but arose since the tarsier–Anthropoidea split (58 Mya) (Fig. 3) (Goodman et al. 1998).

The 82 numts that were used for phylogenetic reconstruction represent the longest numts in our dataset (Table 1). To get an indication of whether the oldest numts are represented in this subset, the nucleotide divergence from human mtDNA was calculated for each of the 348 numts used in this study (insertions and deletions were ignored). The oldest numts should be most diverged from human mtDNA. The 82 numts used for phylogenetic reconstruction showed a mean nucleotide divergence from mtDNA of 16% and no significant correlation between numt length and nucleotide divergence (Spearman's r_s, −0.15; $p = 0.19$). The remaining 266 showed a significantly lower divergence from mtDNA (Mann–Whitney U test: $p < 1 \times 10^{-6}$), with a mean of 12% and a significant positive correlation between numt length and divergence (Spearman's r_s, 0.59; $p < 1 \times$

10^{-6}). This would suggest that by estimating the age of only the longest numts by phylogenetic reconstruction, we are not underrepresenting the oldest numts.

A low divergence from mtDNA is expected for small numts if our identification of numts is limited by the divergence of numts from the human mtDNA, which we used as the BLAST query. This is because short numts would still get a high BLAST score if they are not very diverged from the query, whereas an ancient (and therefore diverged) numt would only have the same BLAST score, and therefore the same chances of being included in the BLAST results, if it were longer. We expect long numts to be less vulnerable to this limitation of the BLAST criteria. In support of this, long numt lengths are not correlated with divergence, so the ages of these 82 numts should be represented in the proportions found in the genome. If divergence from human mtDNA were not a limit to numt identification by BLAST, we might expect short numts to be older on average, because non-functional DNA is gradually lost as it sustains nucleotide deletions, though this effect may be very weak in mammals (Ophir and Graur 1997; see below).

The positive correlation between length and divergence suggests that our identification of numts is limited by their divergence from mtDNA. In agreement with this conclusion, use of old numts as BLAST queries reveals many more numts that were not among the 348 numts analyzed here (data not shown). The lack of mtDNA insertions that predate the tarsier–Anthropoidea split probably reflects our numt detection limit and not a real absence of such insertions.

The Rate of mtDNA Insertion is Approximately Uniform

The analysis described in Fig. 3 can also help to determine whether mtDNA insertion has been continuous. A large proportion of mtDNA insertions appears to have arisen during the millions of years between the New World monkey–Old World monkey split (25–40 Mya; branch 6 in Fig. 3. [Hazkani-Covo et al. 2003]). One reason for the large number of numts that diverge from the mtDNA lineage at branch 6 is that if the reported dates for primate divergences (Goodman et al. 1998) are relatively accurate, branch 6 represents a long time (15 million years) relative to most other branches in this analysis (e.g., branch 2 represents 1 million years). To account for this effect (that we may expect 15 times as many numts in branch 6 as in branch 2), we assess the continuity of mtDNA insertion by comparison to an ''expected'' number. The ''expected'' numbers given in Fig. 3 are those that would be expected if the mtDNA insertions studied here arose at a uniform rate since the tarsier–Anthropoidea split, given the estimated dates of primate divergences from Goodman et al. (1998) and that no numts have been subsequently lost.

Three increasingly realistic substitution models were used for the reconstruction of numt trees by maximum likelihood; HKY, HKY + G, and HKY + G + gamma (see Methods for full descriptions). HKY + G is a much better model than HKY for reconstructing the tree shown in Fig. 3 ($2 \times \Delta \ln L = 16102$, df = 4, $p = 0$) and HKY + G + gamma is by far the best of the three (HKY + G + gamma vs. HKY + G, $2 \times \Delta \ln L = 4900$, df = 4, $p = 0$; HKY + G + gamma vs. HKY, $2 \times \Delta \ln L = 21002$, df = 5, $p = 0$).

When numt trees were reconstructed using HKY or HKY + G models there still seemed to be a significant excess of mtDNA insertions 25–40 Mya (HKY model, chi-square test: $p = 2 \times 10^{-12}$, df = 6; HKY + G model, chi square test: $p = 6 \times 10^{-4}$, df = 6). However, as the substitution model is improved, this excess is reduced to the point where it is not statistically significant with the application of the better HKY + G + gamma model (chi-square test: $p = 0.0504$, df = 6).

Though not significant, there are still a few more mtDNA insertions arising 25–40 Mya than expected, but considering that even when using the HKY + G + gamma model the substitution model could still be improved, the rate of mtDNA insertion appears to be very close to uniformity (HKY + G + gamma is close to the expected distribution in Fig. 3).

Selective Constraint on Differences Between Most Numts

Evidence for selective constraint in the nucleotide differences between two numts is evidence that they have different coding progenitors and therefore arose by different insertions from mitochondrion to nucleus (Fig. 1). In the absence of selective constraint, nucleotide differences are expected to accumulate at equal rates at 1st, 2nd, and 3rd positions of codons. Pairs of numt sequences that differ from this expectation are said to have evidence of ''selective constraint'' in their divergence (see Methods).

This test could only be applied to the 236 numts that had at least 30 bp of homology to protein-coding mtDNA. Of the 27,730 ($236 \times 235 \times 0.5$) possible pairwise comparisons, only 16%, 4568, overlap by at least 30 bp in their homology to protein-coding DNA (Fig. 4). Of these 4568 pairwise comparisons 4127 (90%) show more nucleotide substitutions at the 3rd positions of codons (the least selectively constrained site in functional codons) than at the 1st or 2nd positions. Of these 4127 comparisons, 76% (3130) are statistically significant ($p < 0.05$), though up to 5%

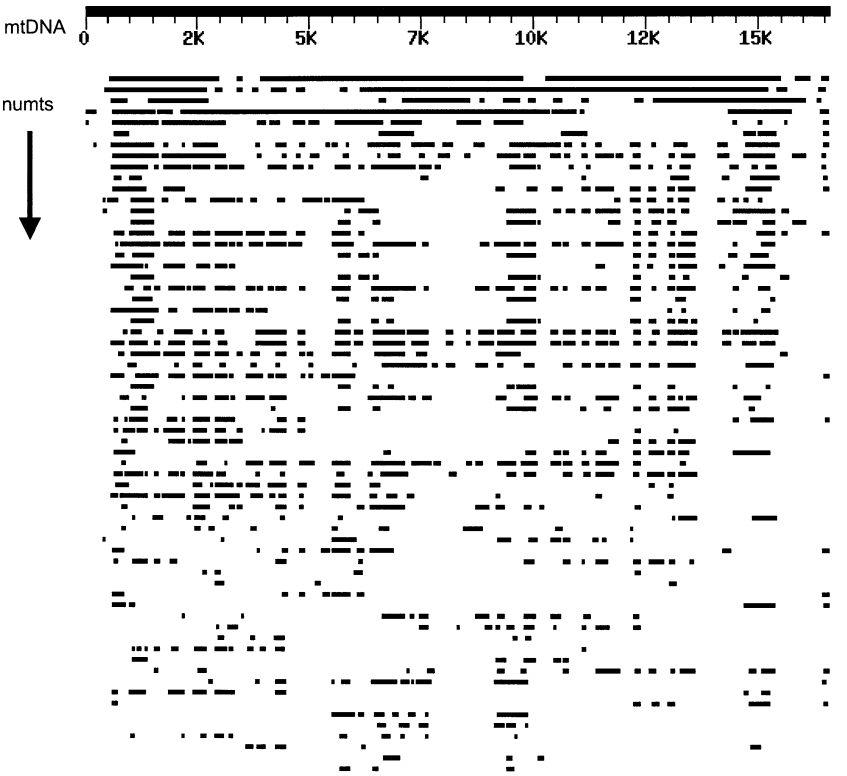

Fig. 4. The numt DNA sequences studied aligned against the human mtDNA genome. More than one numt may be represented on each line. Only 16% of possible pairwise numt comparisons overlap in their homology to mtDNA.

could be false positives, as no correction was made for multiple hits. The mean ratio of changes at 1st, 2nd, and 3rd codon positions is 2:1:4 as estimated from all pairwise comparisons of numts. Such a lack of uniformity is as expected from selectively constrained molecular evolution. The mean ratio from all pairwise comparisons of the mammalian mtDNA sequences shown in Fig. 3a is 2:1:5.

Few pairwise comparisons showed the neutral pattern of nucleotide substitution (1:1:1), yet many of these probably did not arise by duplication. Numt pairs that do not show evidence of selective constraint may have arisen by DNA duplication, but could also have arisen as insertions of similar mtDNA molecules or as ancient insertions whose pattern of selective constraint has since been obscured by neutral substitutions (e.g., numts F and G or numts I and J in Fig. 1).

The many pairwise comparisons, showing evidence of selective constraint, suggest that numts have inserted multiple times in the human genome from diverged mtDNA molecules, as also shown by phylogenetic analysis (Fig. 3) (Fukuda et al. 1985; Hu and Thilly 1994; Mourier et al. 2001). That the vast majority of numt comparisons showed evidence of selective constraint is surprising because not all the remaining comparisons involve duplication events, but also because a necessary restriction for this analysis is that numts "overlap" in their homology to mtDNA, though most numts do not (Fig. 4). The proportion of duplicated numts exhibiting this overlap is probably much higher than the proportion of duplicated numts in the total dataset, because numt duplicates would have been identical at the time of duplication. These data suggest that most numts arise as separate mtDNA insertions from mitochondrion to nucleus and not by duplication of existing numts.

Numt pairs that did not show an excess of differences at the 3rd positions of codons were studied to determine whether they arose by nuclear DNA duplication, and to estimate the rate at which the nuclear genome accumulates mtDNA insertions "*de novo*" from the mitochondrion.

A 13-Copy Family of Numts

The analysis of selective constraints in the 4568 pairwise comparisons above revealed 78 pairwise comparisons that each showed no significant evidence of selective constraint (chi-square test, $p > 0.05$) and represent every possible comparison among 13 numts. These 13 numts share an unusual arrangement of homology to noncontiguous mtDNA regions. Analysis of the flanking DNA of these numts also showed these 13 numts arose as duplications of 1–195-kb DNA segments. This is probably the same 13-copy family of numts as that recently described by Tourmen et al. (2002) and occurs in nine different chromosomal regions (Table 2). The length and divergence of the duplicated regions in which this

Table 2. Summary of seven duplication events

	Numt overlap (bp)	Chromosomal locations	mtDNA % D	Numt % D	Dup. % D	Duplication size (kb)
Duplication 1	5758	2q21.1, 2q21.1	20	4	3–5	~26
Duplication 2	2312	7q34, 9p13.2	15	7	5–8	91
Duplication 3	668	9q21.31, 9q22.31	19	0.8	4	10.5
Duplication 4	625	1q36.33^{T}, 6p25.3^{T}	1.8	0.3	0–2	~70
Duplication 5	1554	8q11.1^{C}, Yp11.1	17	15	13–19	77
Family of 3	504	6p21.31, 6p21.31, 6p21.31	21	0.3 & 3	1–3	69 & 16
Family of 13	496[a]	2p11.1^{C}, 3^{N}, 4p11^{C}, 9^{N}, 10^{N}, 10^{N}, 11^{N}, 13^{N}, 13^{N}, 13^{N}, 13^{N}, 17^{N}, 21^{N}	13[a]	3.5[a]	3–6	Up to 195

Note. Numt divergence from mtDNA (mtDNA % D); numt divergence from numt duplicate (numt % D); divergence across the entire duplicated region (Dup. % D). ^{T} Telomeric; ^{C} Centromeric; ^{N} Not mapped to an approximate cytogenetic position on chromosome.
[a] Mean values.

family of numts occurs suggest that its numts arose by multiple duplication events, but because the dynamics of duplicated regions may differ from those of unduplicated regions (Samonte and Eichler 2002), it is counted as only a single duplication event. A single numt from this family, the longest one (NT_024225.4 71692...74256), was chosen to represent this family in the analyses below.

Segmental DNA Duplications That Include Numt DNA Sequences

Our analysis of 127 numts of sufficient length for this analysis (see Methods and Table 1) revealed the 13-copy numt family described above and another 138 candidate pairs of numts whose sequences suggest that they may have arisen by DNA duplication. Fifty-three of these pairs are candidates for duplication because they overlap in their homology to protein-coding DNA but do not show evidence of selective constraint in their sequences. Pairwise comparison of the 14 longest numts with homology to non-protein-coding mtDNA resulted in a further 85 pairs of numts that overlap in these regions. The numt and flanking DNA for each of the 138 candidate pairs were analyzed.

Analysis of whether homology between numts extended into the flanking DNA of these 138 candidate numt pairs revealed a three-copy family of numts and five more duplication events, each represented by a single pair of numts (Table 2). As for the 13-copy numt family, the 3-copy numt family was counted as only a single independent duplication event, because of the higher rate of duplication expected for sequences that have already been duplicated. (Samonte and Eichler 2002). In total, 26 numts of 127 were involved in DNA duplications (Table 2). There appear to have been seven independent duplication events, so 7 of the 26 numts must have arisen as mtDNA insertions from the mitochondrion and 19 must have arisen by segmental DNA duplication.

The size of the seven duplication events described in Table 2 is typical of segmental duplications (International Human Genome Sequencing Consortium 2001). Bailey et al. (2002d) have reported that chromosomes 7, 9, 15, 16, 17, 19, 22, and Y are significantly enriched for both inter- and intrachromosomal duplications, while chromosomes 2, 3, 4, 5, 8, and 14 have fewer duplications than expected. Almost as many numts that are involved in duplications map to chromosomes for which we expect low duplication rates (six numts on chromosomes 2, 3, 4, and 8; Table 2) as those mapping to chromosomes with higher expected rates of duplication (seven numts on chromosomes 7, 9, 17, and Y; Table 2). Most duplication events in Table 2 (four of seven) do not involve centromeric or telomeric DNA, for which elevated duplication rates are also expected (Samonte and Eichler 2002).

Our analysis does not include two possible duplications that showed less than 0.5% sequence divergence, were not firmly resolved by flanking nonhomologous DNA sequence, and were not assigned to chromosomes, because these might have resulted from sequence misassembly. Such candidate duplications have also been excluded from other analyses of human genome duplications (International Human Genome Sequencing Consortium 2001). As a result, the duplication rate we report may be underestimated.

High Rates of mtDNA Insertion and DNA Duplication

Duplication analysis of 127 numts reveals that 19 of these arose by duplication from seven duplication events (Table 2). Using the estimated oldest numt age of 58 million years, and assuming a constant rate of mtDNA insertion, we can estimate the number of nuclear DNA duplications per numt per year as

$$\text{duplication rate} = \frac{N_d}{N_i \times 0.5 \times \text{oldest numt age}}$$

where N_d represents the number of duplication events, N_i represents the number of numts arising by mtDNA insertion and therefore the number available for duplication in the last 58 million years. It is calculated as the number of numts studied for duplication identification − the number of numts that arose by duplication (108 = 127 − 19; see Tables 1 and 2). Duplications of already duplicated numts are not counted in this analysis because of the changed dynamics of DNA regions that have already been duplicated (Samonte and Eichler 2002). The ''oldest numt age'' in the equation above is halved because only the oldest numts arose 58 Mya; under the assumption that the other numts arose continuously since then, the average duration that each numt was available for duplication in the nucleus is half that time (29 My). The duplication rate is estimated as 2.2×10^{-9} per numt per year.

Although this assumes that numts arose at a constant rate over 58 million years, changing this assumption has little effect on our conclusions. If the oldest numts arose 70 million or 30 million years ago, the duplication rate becomes 1.9×10^{-9} or 4.3×10^{-9} duplications per numt per year, respectively. Even if numts did not arise at a constant rate but all arose 58 million years ago (although this would contradict our results; see Fig. 3), the duplication rate would be 1.1×10^{-9}. The 95% confidence interval on the proportion of numts that were duplicated (7 of 108, and assuming a binomial distribution) suggests that the duplication rate is between 0.6×10^{-9} and 3.8×10^{-9} per numt per year.

Most pairs of numts do not appear to have arisen from duplication events. From this analysis, it appears that 85% of numts (108 of 127) arose by different mtDNA insertions. Applied to the total of 348 numts identified in this study, this would suggest that approximately 296 numts arose by mtDNA insertion. If the rate of mtDNA insertion is approximately constant, then from this analysis it can be estimated as 5.1 mtDNA insertions per genome per million years.

Discussion

This study reveals surprisingly frequent mtDNA insertion events in the human genome and a high rate of nuclear DNA duplication. These results are not explained by errors in the human genome sequence draft. In general the difficulties of genome assembly are likely to lead to an underestimated rate of DNA duplication (International Human Genome Sequencing Consortium 2001). In addition, very recently arising mtDNA insertions could be mistaken for mtDNA contamination and be excluded from the human genome project sequence. Analysis of numts in previous drafts of the human genome sequence, from August 2001 (build 26) and December 2001 (build 27), gives almost exactly the same results as those described here (data not shown). Though the human genome sequence is still being assembled, this has no obvious effect on our analysis or interpretation of the numt data.

An assumption of our estimation of the rates of mtDNA insertion and duplication is that the oldest numts in this study arose around the time of the tarsier–Anthropoidea split, approximately 58 million years ago (Goodman et al. 1998), and that the rate of mtDNA insertion has been approximately uniform. In general, the process of mtDNA insertion into the primate nuclear genome is thought to be a largely continuous one (Hu and Thilly 1994; Mourier et al. 2001; Hazkani-Covo et al. 2003). This assumption is supported by our analysis of the age distribution of mtDNA insertions. However, phylogenetic analysis of human numts with mtDNA appears to be sensitive to the type of substitution model used, and this may explain why a peak in numt accumulation between the Platyrrhini–Catarrhini and the Catarrhini–Hominidae [splits were observed by a different phylogenetic analysis of most of the same human numts (Hazkani-Covo et al. 2003]. We also observe such a peak but it disappeared when significantly better substitution models were used in the analysis. Even if our assumptions were wrong, and mtDNA insertions did not arise at a uniform rate and the oldest numts in this analysis arose at quite a different time than our data suggest, our estimate of the rate of DNA duplication, at least, would be affected very little (see Results).

The analyses of numts presented here, and elsewhere (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Hazkani-Covo et al. 2003), do not consider numts that have been lost by DNA deletion. The published estimates of the rate of DNA loss by small deletions in mammals (Graur et al. 1989; Ophir and Graur 1997) suggest that it is too low to affect this study. Using the estimated size and frequency of small insertions and deletions in human and murid pseudogenes in Ophir and Graur (1997), we estimate that the rate of DNA loss is of the order of 0.06 times the rate of point substitution in mammals. Applying an estimated rate of point substitution in mammals of 2×10^{-9} per nucleotide site per year to this, we would expect only 0.7% of a pseudogene to be deleted in 58 million years. This estimate is in agreement with the lack of significant negative correlations found between numt divergence from mtDNA (a proxy for numt age) and numt length.

Because no consideration has been made of duplications or insertions that may have once been fixed in the nuclear genome but have since been deleted by large DNA deletions, both the duplica-

tion rate and the rate of mtDNA insertion that we estimate may be underestimates. The duplications and mtDNA insertions discussed here are only those that reached fixation in the human genome, and so only reflect the germ-line mtDNA insertions that are not so deleterious that they cannot reach fixation. In addition, the rate of mtDNA insertion may also be underestimated because we extrapolate the age distribution that was estimated using the 82 longest numts, to all 348 numts in this study. Older numts are likely underrepresented in the dataset of 348 numts (see the discussion of the limitation of divergence from BLAST query for short numts in the Results). This is unlikely to be a problem for our estimate of the DNA duplication rate. The 26 numts (108 numts arising by mtDNA insertion − 82 phylogenetically analyzed) for which phylogenies were not directly estimated showed divergences from mtDNA that were not significantly different from those of the 82 numts (Mann–Whitney U test: $p = 0.18$) and showed no significant correlation between length and divergence (Spearman's $r_s = -0.09$, $p = 0.67$).

Although it may be an underestimate, our estimate of the rate of mtDNA insertion ($\mu_{numt} = 5.1 \times 10^{-6}$ mtDNA insertions per genome per year) is of an order that suggests that humans should vary with respect to the presence or absence of mtDNA insertions at nuclear positions. Assuming that these insertions are selectively neutral (strongly deleterious insertions are unlikely to have been included in this estimate of the mutation rate), we predict that on average any two haploid human genomes will differ in the presence or absence of mtDNA insertions at at least two loci ($\pi_{numt} \approx 2.07$). This is estimated from μ_{numt} by using observations of human nucleotide diversity ($\pi_{ps} \approx 0.00081$) (Przeworski et al. 2000) and the point substitution mutation rate ($\mu_{ps} \approx 2 \times 10^{-9}$) (Li 1997) and by assuming that under standard neutral theory $\pi \approx \theta = 4N_e\mu$ (Li 1997; Przeworski et al. 2000), so that π_{numt} per haploid genome $= \pi_{ps} \times (\mu_{numt}/\mu_{ps})$. Under the same assumptions, for five individuals (10 haploid genomes), at least five loci are expected to be variable in their presence or absence of mtDNA insertions (from $E[S] = \theta \times (1 + 1\sqrt{2} + 1\sqrt{3} + \ldots + 1\sqrt{10})$ (Watterson 1975), where θ is estimated as $\pi_{numt} = 2.07$). There is already experimental evidence to support these predictions as hybridization studies of human numts of different individuals revealed differences in numt presence or absence at different loci, even among siblings (Yuan et al. 1999). One such human mtDNA insertion polymorphism has already been characterized and utilized as a human population genetic marker (Thomas et al. 1996). Such cases could also be used to investigate the population dynamics of noncoding DNA insertions.

The human mitochondrial genome does not code for anything that would suggest it could actively promote its insertion and persistence in another (the nuclear) genome. The insertion and persistence of such large numbers of noncoding mtDNA fragments imply that it is the lack of foreign DNA availability in sequestered germ cells that limits horizontal gene transfer to the human genome.

The duplications observed in this study are 10–195 kb in size, are inter- or intrachromosomal, and occur in duplicate pairs and larger families (Table 2). They are therefore typical of the duplication class (segmental duplication) that was recently found to occur at a surprisingly high rate in the human genome (International Human Genome Sequencing Consortium 2001; Samonte and Eichler 2002). They differ in their degrees of divergence (0–19%; see Table 2), which suggests that duplications have been accumulating at least for as long as this study is able to detect them (approximately 58 million years); previous studies have focused on duplicates that are less than 10% diverged and so are younger than duplication 5 (Table 2) (International Human Genome Sequencing Consortium 2001; Bailey et al. 2002a, b; Samonte and Eichler 2002). The duplications reported here are not all close to centromeres or telomeres, or on chromosomes that may have elevated rates of DNA duplication (Table 2) (Bailey et al. 2002a). We did not count secondary duplications as separate duplication events because these are expected to occur at a higher rate than for unduplicated DNA (Samonte and Eichler 2002). We therefore expect rates of numt duplication to be typical of most of the human genome as the mtDNA insertions available for duplication are distributed evenly within and among chromosomes (see Results and Woischnik and Moraes [2002]).

The duplication rate we estimate, 2.2×10^{-9} per numt per year, is similar to the estimated rates of functional gene duplication reported for *Drosophila*, yeast, *C. elegans*, and *Arabidopsis*, $2–20 \times 10^{-9}$ per gene per year (Lynch and Conery 2000). As most numts are noncoding, the duplication rate reported here probably better reflects that of noncoding DNA. However, as close to 30% of human noncoding DNA is thought to occur in introns (International Human Genome Sequencing Consortium 2001), a large proportion of duplications that span thousands of nucleotides is likely to be associated with genes, even if these duplications were identified by analysis of noncoding DNA. Even so, it is perhaps surprising that our estimate of the rate of duplication using noncoding DNA is of the same order and perhaps less than that for functional gene duplication in humans according to a preliminary estimate of this rate (Lynch and Conery 2001). Our estimate of the rate of noncoding DNA duplication therefore supports the recent observation that there have been more recent

duplications in gene-rich regions than in gene-poor regions (Bailey et al. 2002a). Such a rate difference may suggest that functional gene duplicates most commonly reach fixation because they confer a selective advantage.

The duplication rate reported here and those reported for gene duplication in other organisms are high and similar to the rate of point substitution estimated for noncoding DNA in humans, 1.6×10^{-9}–2.5×10^{-9} per site per year (Li 1997). In other words, a nucleotide site is about as likely to be involved in a large (1–200-kb) duplication as it is to sustain a point mutation. If gene duplicates accumulated over the last 6 million years at the same rate as those of the numts studied here, we would expect humans to differ from chimps in the presence or absence of 792 gene duplicates (assuming 30,000 genes and 6 million years since humans and chimps diverged). Such a high rate of DNA duplication has profound implications for our understanding of molecular evolution.

Accessions

Accessions involved in duplications with numt start and end positions. Duplication 1: NT_022140.8, 86663…98139; NT_028068.5, 578756…583451. Duplication 2: NT_030040.3, 1217650…1220877; NT_023640.7, 962859…965351. Duplication 3: NT_008387.8, 395500…398161; NT_029369.4, 1270883…1271521. Duplication 4: NT_007412.8, 4088112…4093953; NT_004525.8, 5360872…5361497. Duplication 5: NT_023678, 409660…413224; NT_011896, 5697978…5705866. Family of 3: NT_007592.8, 188976…189711; NT_007592.8, 15097623…15098358; NT_023407.6, 296261…296717. Family of 13: NT_023943.8, 2398410…2400966; NT_026996.5, 30244…30916; NT_026996.5, 175125…177800; NT_026996.5, 215725…218400; NT_024563.3, 45191…47758; NT_024060.7, 5947…6593; NT_024060.7, 51589…52235; NT_024060.7, 83687…86359; NT_022110.7, 339859…342478; NT_030164.1, 119975…122655; NT_024225.4, 71692…74256; NT_005011.8, 285471…287877; NT_029489.1, 106894…107540.

Acknowledgments. Many thanks for helpful discussion, advice, and comments on the manuscript go to Aviv Bergman, Casey M. Bergman, Krista K. Ingram, and Dennis P. Wall, and to Jeffrey L. Boore for support in the later stages of this work. The comments of two anonymous reviewers substantially improved the manuscript. This work was partly funded by the Center for Computational Genetics and Biological Modeling.

References

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwatz S, Adams MD, Myers EW, Li PW, Eichler EE (2002a) Recent segmental duplications in the human genome. Science 297:1003–1007

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Shchwartz S, Rocchi M, Eichler EE (2002b) Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. Am J Hum Gene 70:83–100

Bensasson D, Zhang D-X, Hewitt GM (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol Biol Evol 17:406–415

Bensasson D, Zhang D-X, 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

Fukuda M, Wakasugi S, Tsuzuki T, Nomiyama H, Shimada K, Miyata T (1985) Mitochondrial DNA-like sequences in the human nuclear genome. J Mol Biol 186:257–266

Gellissen G, Michaelis G (1987) Gene transfer: Mitochondria to nucleus. Ann NY Acad Sci 503:391–401

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598

Graur D, Shuali Y, Li W-H (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol 28:279–285

Gu Z, Cavalcanti A, Chen F-C, P. B, Li W-H (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol Biol Evol 19:256–262

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis (program for windows 95/98/NT). Nucleic Acids Symp Ser 41:95–98

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 postinsertion duplications. J Mol Evol 56:169–174

Hu G, Thilly WG (1994) Evolutionary trail of the mitchondrial genome as based on human 16S rDNA pseudogenes. Gene 147:197–204

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Long M, Thornton K (2001) Gene duplication and evolution. Science 293:1551a

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M, Conery JS (2001) Gene duplication and evolution. Science 293:1551a

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

Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS (2002) OWEN: Aligning long colinear regions of genomes. Bioinformatics 18:1703–1704

Ophir R, Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205:191–202

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

Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. Trends Genet 16:296–302

Samonte RV, Eichler EE (2002) Segmental duplications and the evolution of the primate genome. Nature Rev Genet 3:65–72

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 (2002) The complete mitochondrial sequence of *Tarsius bancanus:* Evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol Biol Evol 19:544–553

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, MA

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

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

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Woischnik M, Moraes CT (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res 12:885–893

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13

Yuan JD, Shi JX, Meng GX, An LG, Hu GX (1999) Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. Cell Res 9:281–290