A Comparative Analysis of *numt* Evolution in Human and Chimpanzee

*Einat Hazkani-Covo**[1] *and Dan Graur**†

*Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel; and †Department of Biology and Biochemistry, University of Houston

Mitochondrial DNA sequences are frequently transferred into the nuclear genome, giving rise to *numt*s (nuclear DNA sequences of mitochondrial origin). So far, the evolutionary history of *numt*s has largely been studied by using single genomes. Here, we present the first attempt to study *numt* evolution in a comparative manner by using a pairwise genomic alignment. The total number of *numt*s was estimated to be 452 in human and 469 in chimpanzee. numts that were found in both genomes at identical loci were deemed to be orthologous; 391 *numts* (>80%) were classified as such. The preponderance of orthologous *numt*s is due to the very short divergence time between the 2 hominoids. The rest of *numts* were deemed to be nonorthologous. Nonorthologous *numt*s were subdivided into 1) ancestral *numt*s that have lost an ortholog in one species through deletion (12 in human and 11 in chimpanzee), 2) new *numt*s acquired by the insertion of a mitochondrial sequence after the divergence of the 2 species (34 in human and 46 in chimpanzee), and 3) paralogous *numt*s created by the tandem duplication of a preexisting *numt* (2 in human). This approach also enabled us to reconstruct the *numt* repertoire in the common ancestor of humans and chimpanzees (409 *numts*). Our comparative approach is also useful in identifying the exact boundaries of *numts*.

Mitochondrial DNA sequences are frequently transferred into the nuclear genome, giving rise to *numt*s (nuclear DNA sequences of mitochondrial origin, Lopez et al. 1994). *numt*s have been described in more than 80 species (Bensasson et al. 2001). For most species, the estimate of *numt* content and abundance is still incomplete. However, with fully sequenced genomes, it is possible to obtain an accurate estimate of *numt* abundance (Richly and Leister 2004). There is no correlation between the fraction of noncoding DNA and *numt* abundance (Richly and Leister 2004). The reason for the variation in *numt* abundance among genomes is not known. Conceptually, the differences might be due to 1) different rates of *numt* insertion, 2) different rates of *numt* deletion, and 3) different rates of *numt* postinsertional duplication.

All mammalian *numt* studied to date were found to be functionless, and it is thought that they became pseudogenized on arrival into the nucleus because of the differences between the nuclear and mitochondrial genetic codes (Gellissen and Michaelis 1987; Perna and Kocher 1996). In yeast, *numt*s are transferred under natural conditions during the repair of double-strand breaks (Ricchetti et al. 1999), and it was suggested that this is the cause for the ongoing colonization of different genomes by *numt*s. The continuing process of *numt* integration into the nuclear genome is evidenced by the finding of *numt*s that have been inserted into the human genome after the human–chimpanzee divergence (Ricchetti et al. 2004). Some of these *numts* are variable with respect to genomic presence or absence, indicating that they have only arisen recently in the human population. Transposition of *numt*s into genes has also been associated with human diseases (Willett-Brozick et al. 2001; Turner et al. 2003; Goldin et al. 2004).

From human genome data, different estimates of the number of *numt*s have been put forward in the literature (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Bensasson et al. 2003; Richly and Leister 2004). Additionally, phylogenetic methods have been suggested for dating the insertion of *numt*s into the nuclear genome (Mourier et al. 2001; Woischnik and Moraes 2002). Initial results indicated a fairly rapid process of *numt* insertion, however, some studies ignored the possibility of postinsertional nuclear duplication (e.g., Bensasson et al. 2000) resulting in overestimation of *numt* insertion rates. Hazkani-Covo et al. (2003) suggested a methodology for dating the insertion of *numt*s into the nuclear genome by using a single nuclear genome sequence and a mitochondrial phylogenetic tree. This methodology had the advantage of being able to detect *numt* duplication events. We discovered that the rate of *numt* insertion on the branch leading to humans was much lower than previously reported (Mourier et al. 2001; Woischnik and Moraes 2002). Most *numt*s turned out to be paralogs of preexisting *numt*s, rather than new insertions.

Two *numt*s are defined as orthologous if they are derived from a speciation event, but as paralogous if they are derived from a duplication event. So far, the evolutionary history of *numt*s has largely been studied by means of paralogous comparisons within single genomes (Mourier et al. 2001; Woischnik and Moraes 2002; Hazkani-Covo et al. 2003). The availability of closely related completely sequenced genomes has enabled us to use comparative methods to study directly orthologous *numt* evolution. We note that by using the methodology of Hazkani-Covo et al. (2003), the existence of orthologous *numt* in species other than humans was inferred indirectly. That inference, however, yielded a testable prediction. Thus, for example, a *numt* that was inferred to have been inserted in the common ancestor of human and chimpanzee should possess orthologs in both species. However, this prediction could be wrong if the mitochondrial phylogenetic tree is not the true tree. In addition, this methodology is only applicable to long *numt*s that have sufficient phylogenetic signal. With 2 or more genomes, the presence of orthologous *numt*s can be inferred directly, even when the *numt*s are short.

In the following, we suggest a protocol based on genome alignment to estimate the number of *numt*s in closely related species. We apply this approach to the genomes of human (Lander et al. 2001) and chimpanzee (*Pan troglodytes*;

[1] Present address: National Evolutionary Synthesis Center, Durham, North Carolina, USA.

Key words: *numts*, comparative evolution, promiscuous DNA, human genome, chimpanzee genome, mitochondrial DNA, genome evolution, pseudogenes.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 24(1):13–18. 2007
doi:10.1093/molbev/msl149
Advance Access publication October 20, 2006

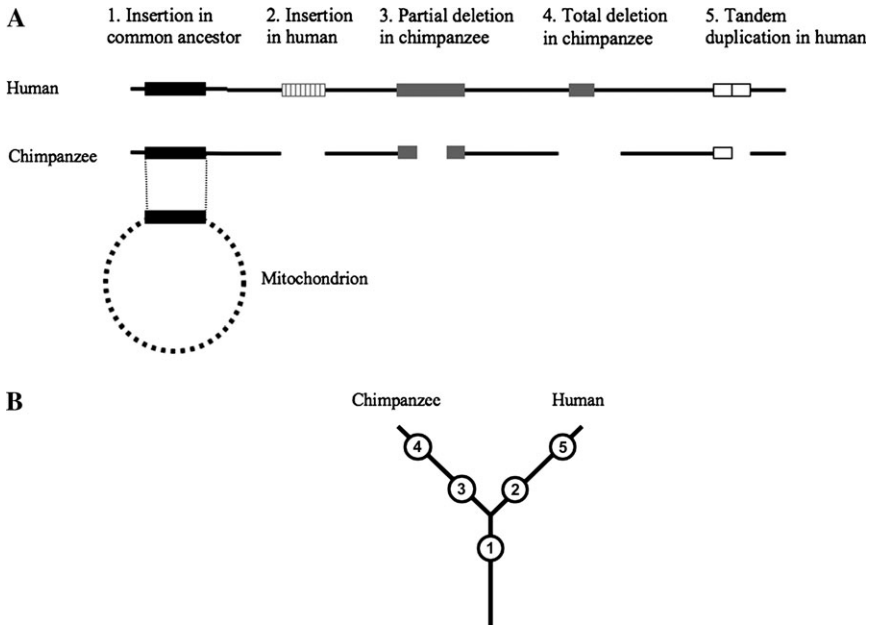

FIG. 1.—(*A*) *numt* classification based on genome alignment of homologous loci between human and chimpanzee. (*B*) Each evolutionary event is positioned on the inferred branch on the phylogenetic tree.

Mikkelsen et al. 2005), and use the alignment to identify evolutionary events that may have affected *numt* composition in each genome, as well as to reconstruct the *numt* makeup in the common ancestor of human and chimpanzee.

Because there are no hot spots for *numt* insertion (Zischler 2000), the presence of a *numt* at a particular locus in both genomes was taken to imply orthology (fig. 1). Nonorthologous *numt*s that are present in only one genome are further classified into insertions, partial or total deletions, or tandem duplications (fig. 1). Each such event can take place in either lineage. Nonorthologous *numt*s are identified by a gap in the alignment. The distinction between insertions and deletions is based on the fact that there exists no known mechanism for the precise excision of *numt*s. Thus, if the gap coincides precisely with the boundaries of the *numt*, an insertion is inferred. If the gap is smaller or larger than the *numt* in the other genome, we infer the occurrence of a partial or total deletion, respectively. Tandem *numt* duplications are characterized by adjacent homologous *numts* and a gap coinciding perfectly with the boundaries of the homolog from the other species. The assumptions used for *numt* classifications here were also used in PCR-based *numt* recognition (e.g., Lopez et al. 1994; Zischler et al. 1998; Herrnstadt et al. 1999).

Our analyses were based on genomic sequences and annotations from the University of California at Santa Cruz (Karolchik et al. 2004) Genome Center. First, Blast was used to search each of the human and chimpanzee genomes for regions of similarity with conspecific mitochondrial sequences (fig. 2, frame 1). Closely spaced mitochondrial hits were concatenated (fig. 2, frame 2). The distinction between orthologous and nonorthologous *numt*s (as described in fig. 1) was accomplished through a comparison of human and chimpanzee *numt* preliminary datasets. The comparison was based on the University of California-Santa Cruz genome alignment between human and chimpanzee. The analysis was performed in a reciprocal manner: comparing the human genome to the chimpanzee genome and comparing the chimpanzee genome to the human genome. For a detailed description of the methodology, see Supplementary Material online.

We found a similar number of *numts* in both genomes: 452 *numt*s in human and 469 *numt*s in chimpanzee (table 1). The total number of *numt*s in the 2 genomes was found to be similar to previous estimates in the literature. Unsurprisingly, because of the short time that has passed since the divergence of the 2 hominoids, 391 *numts* (87% in human and 84% in chimpanzee) were classified as orthologous, that is, were inserted into the nuclear genome before the divergence between the 2 lineages (table S1 in Supplementary Material online).

We identified 46 previously undescribed postspeciation *numt*s in the chimpanzee. These ranged in size between 37 and 3,076 bp. In addition, we identified 34 *numt*s in human. Our study, thus, increases the number of known human-specific *numts* (Ricchetti et al. 2004) by 26%, and identifies the shortest (29 vs. 47 bp) and the longest (5,219 vs. 1,323 bp) new *numt*s. Human and chimpanzee postspeciation *numts* that were found in this study are listed in table 2. The common ancestor of human and chimpanzee is estimated to have lived about 6 Myr ago (Goodman et al. 1998). Thus, the average rates of *numt* insertion are 5.7 insertions per 1 Myr in human and 7.7 *numt* insertions per 1 Myr in chimpanzee. The difference is not statistically significant ($P < 0.179$).

From among the postspeciation *numt*s, only 2 cases of tandem duplication were found (both in the human genome). In the first case, an internal segment of 30 bp within a *numt* located on chromosome 10 was duplicated once. The second case, in chromosome 12, includes 18 tandem duplications of a 47-bp sequence (fig. 3).

The number of events in which *numt*s were deleted from the genome is fairly similar between the 2 species.

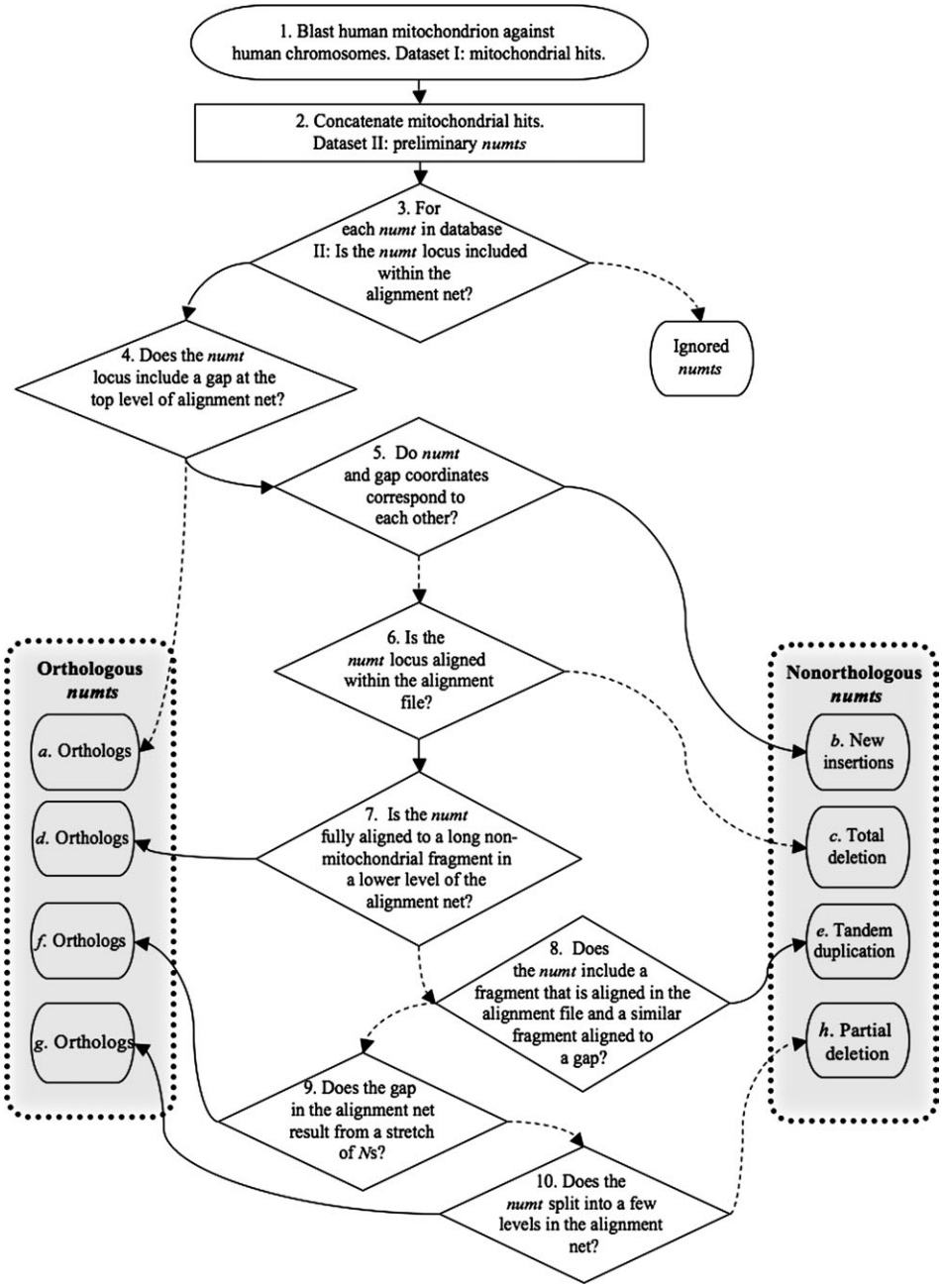

FIG. 2.—Flowchart of data collection and *numt* classification in human. Two types of UCSC files were used in the analysis: the nucleotide pairwise alignment file and the alignment net file. The final *numt* classification is determined after comparison with the chimpanzee genome (for details see Supplementary Methods, Supplementary Material online).

There are 12 deletion events in human, of which 11 are total deletions and 1 is a partial one. In chimpanzee, there are 11 deletion events, of which 7 are total deletions and 4 are partial. As far as the total deletions are concerned, one can distinguish between 2 separate groups: most of the *numt*s seem to have been deleted from the genome as part of a much larger segment. However, in a few cases, the *numt* deletion included only a limited flanking region.

The number of nonorthologous *numt*s is not large enough to be able to detect differences in *numt* evolutionary dynamics (insertion, deletion, or tandem duplication) between the 2 lineages. Still, we are now able to reconstruct the *numt*s constitution in the common ancestor of the 2 hominoids. The number of *numt*s in the common ancestor of human and chimpanzee is estimated at 409. This number includes 391 *numt*s that are still found in the 2 genomes, and a total of 18 *numt*s that were lost from 1 of the 2 genomes. Given the very low rate of *numt* deletion, the possibility that a *numt* has been lost in both genomes seems negligible.

We suggest that in comparison to single genome analyses, our methodology resulted not only in a more accurate estimate of the number of *numts* but also in a more precise identification of their boundaries. First, this protocol

Table 1
Numbers and Total Sizes (in Parentheses) of Different *numt* Types within the Genomes of Human and Chimpanzee

		Nonorthologous *numts*				
	Orthologous *numts*	New insertions	Tandem duplications	Evidence for *numt* deletions[b]	Ignored *numts*	Total Number of *numts*
Human	391 (395,530–437,048 bp)	34 (10,536 bp)	1 (1) (846 bp)	7 (4) (1,005 bp)	20 (25,834 bp)	452 (433,751–475,269 bp)
Chimpanzee	391 (395,530–437,048 bp)[a]	46 (8,442 bp)	0	11 (1) (2,620 bp)	21 (11,691 bp)	469 (418,283–459,801 bp)

[a] The total size of orthologous *numts* in chimpanzee was calculated according to human coordinates (see Supplementary Methods, Supplementary Material online). In order not to run into the risk of classifying the same *numt* twice, the size of partially deleted *numts* and very small tandem duplications (whose number appears in parentheses) was added to orthologous size. In addition, tandem duplications and evidence for partial deletion are not counted in the total number of *numts*. Underlined *numts* were used to estimate the repertoire of the common ancestor.

[b] Human *numts* are listed as evidence for deletions in the chimpanzee genome; chimpanzee *numts* are listed as evidence for deletions in the human genome.

distinguishes between orthologous and nonorthologous *numt*s. Second, by using genome alignment, we identified orthologous *numt*s that escaped detection by the usual Blasting of mitochondrial sequences against the nuclear genome. In 145 out of 391 cases, *numt*s were identified in only one of the genomes when the Blast analysis was used. However, in the majority of cases, alignment of those *numt*s to the corresponding fragment in the second genome revealed

Table 2
Postspeciation (New) *numt*s in the Human and Chimpanzee Genomes. Coordinates of the *numts* within the Chromosomes and the Mitochondria Are Shown as well as the *numt* size. Chromosome Names That Contain "Random" Include Unmapped Sequences from the Chromosome

	Chromosome	*Numt* Start	*Numt* End	Mitochondria Start	Mitochondria End	Size
Human						
1	1	37505010	37505083	8935	9008	74
2	1	212729637	212729675	9564	9602	39
3	2	33967073	33967125	1768	1820	53
4	2	81868148	81868389	7863	8104	242
5	2	149850064	149850195	613	744	132
6	3	25483960	25483998	10986	11024	39
7	3	68652747	68652775	12613	12641	29
8	3	97656933	97658255	1398	2720	1323
9	4	12392801	12393142	9339	9680	342
10	4	47689831	47689923	14982	15074	93
11	4	56109869	56109999	964	1094	131
12	4	79388079	79388310	2227	2458	232
13	4	163920153	163920320	12251	12418	168
14	5	73155790	73155830	10803	10843	41
15	5	134335215	134340433	10270	15488	5219
16	5	165938322	165938361	12148	12187	40
17	7	67613632	67613737	12962	13067	106
18	7	145086167	145086262	1615	1710	96
19	8	100464681	100464764	14862	14945	84
20	11	72948014	72948176	6643	6805	163
21	11	122411966	122412037	14661	14732	72
22	12	40043704	40043792	3792	3880	89
23	13	39140488	39140558	9524	9594	71
24	13	54343769	54343891	5109	5231	123
25	13	107774473	107774728	984	1239	256
26	17	42550249	42550316	10144	10211	68
27	17	51657732	51658384	6819	7471	653
28	17	79291501	79291541	6904	6944	41
29	18	2832230	2832352	14382	14504	123
30	18	43631604	43631795	7976	8167	192
31	20	9144571	9144612	2182	2223	42
32	20	13142959	13143001	3501	3543	43
33	20	56324532	56324601	12963	13032	70
34	22	34553532	34553578	6182	6228	47
Chimpanzee						
1	1	94875580	94875730	2368	2518	151
2	1	167557915	167557996	14438	14519	82
3	1	178753526	178753594	309	377	69
4	1	212351158	212351245	8419	8506	88
5	2	53678271	53678368	3042	3139	98
6	2	82799067	82799512	15690	16132	446
7	2	146514792	146514841	6710	6759	50
8	2	168762978	168763014	15223	15259	37
9	2	193113054	193113149	8920	9015	96
10	2	198849924	198849997	1753	1826	74

Table 2
Continued

	Chromosome	*Numt* Start	*Numt* End	Mitochondria Start	Mitochondria End	Size
11	2_random	50499996	50500065	11120	11189	70
12	3	113480884	113483978	7168	14547	3095
13	3	186259921	186259958	13880	13917	38
14	4	88229437	88229473	13473	13509	37
15	4_random	36403746	36403863	2142	2259	118
16	5	28122132	28122213	1268	1349	82
17	5	68511847	68511970	14670	14793	124
18	6_random	21154919	21154985	15098	15164	67
19	6_random	28204787	28204866	2175	2254	80
20	7	126455472	126455640	8260	8428	169
21	7	137974486	137974549	15611	15674	64
22	8	138224221	138224364	10786	10929	144
23	9	42240545	42240626	7950	8031	82
24	9	93119352	93119472	1828	1948	121
25	10	72885336	72885405	10088	10157	70
26	10	104435941	104436168	3540	3767	228
27	10_random	22373731	22374046	6709	7020	316
28	12	103361584	103361668	14582	14666	85
29	12_random	30969217	30969432	7824	8039	216
30	13	15766920	15767029	7433	7542	110
31	13	17844524	17844652	10950	11078	129
32	13	103738537	103738696	9982	10141	160
33	13	109897352	109897539	12157	12344	188
34	13_random	14479481	14479533	13246	13298	53
35	14	18106476	18106545	11189	11258	70
36	14	95292747	95292931	3795	3979	185
37	15	23943932	23944001	1328	1397	70
38	15	42692145	42692448	4778	5081	304
39	16	62752970	62753052	14596	14678	83
40	17	65338553	65338608	15361	15416	56
41	18	14965901	14966085	14448	6415	185
42	18_random	29158839	29158872	11943	11976	34
43	18_random	32760196	32760238	3393	3435	43
44	19_random	23568693	23568762	16025	16094	70
45	20	15505542	15505603	9864	9925	62
46	23	32703341	32703583	14372	14614	243

a cryptic or quasi-cryptic ortholog. In 15 cases, the existence of orthologous *numt*s in chimpanzee was inferred on the basis of a small stretch of *Ns* similar in size to the human *numt* in the homologous position. Finally, our protocol enables a more precise identification of the genomic coordinates of *numt*s. The comparative method allows concatenation of fragments that may otherwise be identified as independent *numts*.

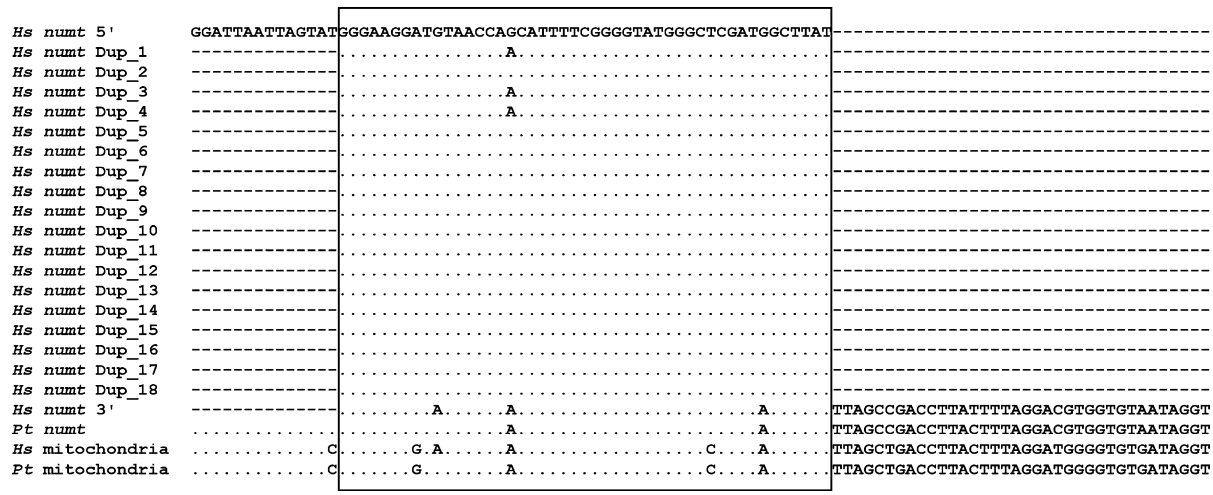

FIG. 3.—Multiple sequence alignment of 18 tandemly repeated *numt*s in human chromosome 12 (positions 125, 420, 954–125, 422, 037) and the homologous locus on chimpanzee chromosome 10. The alignment to human and chimpanzee mitochondria is also shown. Each repeat is 47 bp in length and aligns to mitochondrial coordinates 4418–4464 (box). The flanking regions of the human internally repeated *numt* align to human mitochondrial coordinates 4478–4382 and can also be aligned to a single chimpanzee *numt*. Duplications (Dup_) are numbered in order of their appearance from 5′ to 3′. Identical nucleotides in the alignment columns are indicated by a dot; dashes indicate gaps. Hs, *Homo sapiens*; Pt, *Pan troglodytes*.

Supplementary Material

Supplementary data and tables are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Shay Covo and Tal Dagan for their help. This work was supported in part by a grant (DBI-0543342) from the National Science Foundation.

Literature Cited

Bensasson D, Feldman MW, Petrov DA. 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J Mol Evol. 57:343–354.

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

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

Gellissen G, Michaelis G. 1987. Gene transfer. Mitochondria to nucleus. Ann N Y Acad Sci. 503:391–401.

Goldin E, Stahl S, Cooney AM, Kaneski CR, Gupta S, Brady RO, Ellis JR, Schiffmann R. 2004. Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. Hum Mutat. 24:460–465.

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.

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

Herrnstadt C, Clevenger W, Ghosh SS, Anderson C, Fahy E, Miller S, Howell N, Davis RE. 1999. A novel mitochondrial DNA-like sequence in the human nuclear genome. Genomics 60:67–77.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:493–496.

Lander ES, Linton LM, Birren B, et al. (256 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 39:174–190.

Mikkelsen TS, Hillier LW, Eichler EE, et al. (67 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

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.

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

Ricchetti M, Fairhead C, Dujon B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature. 402:96–100.

Ricchetti M, Tekaia F, Dujon B. 2004. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2:E273.

Richly E, Leister D. 2004. *NUMTs* in sequenced eukaryotic genomes. Mol Biol Evol. 21:1081–1084.

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

Turner C, Killoran C, Thomas NS, Rosenberg M, Chuzhanova NA, Johnston J, Kemel Y, Cooper DN, Biesecker LG. 2003. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Hum Genet. 112:303–309.

Willett-Brozick JE, Savul SA, Richey LE, Baysal BE. 2001. Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. Hum Genet. 109: 216–223.

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

Zischler H. 2000. Nuclear integrations of mitochondrial DNA in primates: inference of associated mutational events. Electrophoresis. 21:531–536.

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

William Martin, Associate Editor

Accepted October 11, 2006