

Letter to the Editor

The Human Genome Project Reveals a Continuous Transfer of Large Mitochondrial Fragments to the Nucleus

Tobias Mourier, Anders J. Hansen, Eske Willerslev, and Peter Arctander

Department of Evolutionary Biology, Zoological Institute, University of Copenhagen, Copenhagen, Denmark

Mitochondrial genomes are believed to gradually transfer DNA fragments (numts) into the nuclear chromosomes of eukaryotic cells during evolution (reviewed in Zhang and Hewitt 1996). This assumption relies on hybridization studies of mitochondrial DNA sequences (mtDNA) (Tsuzuki et al. 1983), sequencing of numts (e.g., Lopez et al. 1994; Arctander 1995; Zischler et al. 1995; Herrnstadt et al. 1999), and similarity searches in sequence databases (Blanchard and Schmidt 1996; Bensasson et al. 2001). Here we present the first extensive analysis of numts in the human nuclear genome. Through a combination of conventional BLAST alignment (Altschul et al. 1997) and a DNA block aligning (DBA) algorithm (Jareborg, Birney, and Durbin 1999), we searched roughly 93.5% of the human genome (<http://www.ncbi.nlm.nih.gov/genome/seq/>) for numts. This approach revealed three notable findings. First, several numts exceed the size of the longest human numt reported to date (Herrnstadt et al. 1999). Second, all parts of the mitochondrial DNA are represented in the nuclear genome. Finally, the integration of mtDNAs into the nucleus is a continuous evolutionary process, thereby verifying previous beliefs (Zhang and Hewitt 1996; Wallace et al. 1997; Herrnstadt et al. 1999).

Through the web service provided by NCBI (<http://www.ncbi.nlm.nih.gov/>), we compared the complete human mitochondrial DNA and the working draft of the human nuclear genome (as of mid-April 2001) using BLAST. This procedure was followed by alignment using the DBA algorithm (Jareborg, Birney, and Durbin 1999), which found collinear blocks of conserved sequence allowing for indels between blocks. The rationale for this twofold alignment procedure stems from the assumption that two mechanisms may obscure the BLAST alignment. First, the extant mtDNA will have diverged from the ancestral sequence. Second, as the numts are presumably released from selection, larger deletions and insertions may take place.

Hits from the BLAST search (default settings) in the same sense and within the vicinity (4–6,128 bp) of each other were assessed to potentially stem from a single insertion event. If such a group of hits involved more than 100 identical positions, the genomic sequence covering all the hits and their intervening sequences were retrieved. This sequence was aligned to the correspond-

ing mtDNA sequence using the DBA algorithm. The sequences were considered a result of a single insertion event if the DBA algorithm was able to align more than 80% of the mtDNA sequence in a collinear way.

Following the above criteria, we found 296 numts ranging between 106 and 14,654 bp in size (table 1). Fifteen of these were found to be longer than 5,842 bp, previously reported by Herrnstadt et al. (1999) as the length of the longest human numt.

Furthermore, we found that all positions of the mitochondrial genome are represented in the nuclear DNA, with the domain comprising the control region being relatively underrepresented (fig. 1). As this could be an artifact caused by the distal position of the control region in the linear mtDNA sequence, we constructed an alternative representation in which the control region was central. Neither this nor the removal of the low-complexity filter of BLAST produced additional hits to this region (not shown). The deficiency of numts from the control region probably results from the significantly higher evolutionary rate of extant mtDNA in this region (Saccone, Pesole, and Sbisá 1991). This hypothesis is further supported by the increased number of numts in the region comprising the central conserved domain (fig. 1).

Interestingly, we found 4 numts covering the complete control region (table 1), signifying that these are at least the result of a DNA-based transfer (for a discussion see Shay and Werbin [1992] and references therein).

To estimate the time of insertion of the numts, we collected all numt-mitochondria alignments longer than 2,000 bp (i.e., either complete numts, if they were com-

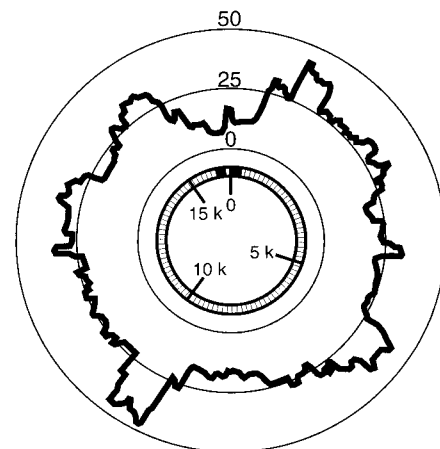


FIG. 1.—Circular diagram of the number of numts descending from a given position in the mitochondria (thick line). The inner hatched circle depicts the mitochondria, with the two hypervariable segments of the control region (encompassing the central conserved domain) highlighted (black).

Key words: mitochondrial DNA, nuclear insertions, human genome.

Address for correspondence and reprints: Tobias Mourier, Department of Evolutionary Biology, Zoological Institute, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark. E-mail: tmourier@zi.ku.dk.

Mol. Biol. Evol. 18(9):1833–1837. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
The 60 Longest Human Numts

No. ^a	Accession ^b	Chromosome	Length ^c	Position ^d	Mit. Pos. ^e	DBA cover ^f	Id. ^g	Phylo. ^h
1	NT_006129.2	4	14,654	530,845–545,549	672–15,325	0.99	0.75	2
2	NT_008413.2	9	12,281	73,386–91,990	1,294–13,574	0.99	0.75	3
3	NT_004836.2	1	10,541	724,546–735,399	12,218–6,189	0.95	0.72	3
4	NT_005102.2	2	10,474	466,390–478,391	8,312–2,216	0.94	0.70	1
5	NT_006721.2	5	9,067	444,304–453,412	6,117–15,183	1.00	0.89	1
6	NT_024128.2	10	8,974	89,285–98,234	2,139–11,112	1.00	0.83	2
7	NT_009151.2	11	8,914	574,889–1,033,967	752–9,665	0.97	0.74	
8	NT_007712.2	7	8,764	14,611–22,554	3,117–11,880	0.89	0.69	2
9	NT_024667.2	15	8,764	49,178–57,662	3,117–11,880	0.96	0.73	1
10	NT_005164.2	2	8,663	17,548–26,131	6,692–15,354	0.99	0.75	2
11	NT_0196112.2	16	6,928	278,992–431,122	8,688–15,615	0.90	0.68	1
12	NT_022127.2	2	6,721	149,031–155,565	3,799–10,519	0.96	0.75	
13	NT_005445.2	2	6,567	1,334,852–1,341,636	16,434–6,431	0.94	0.71	1
14	NT_008224.2	8	6,150	523,487–529,110	9,176–15,325	0.85	0.63	
15	NT_010289.2	15	6,117	1,119,636–1,125,355	9,786–15,902	0.93	0.68	1
16	NT_024346.2	11	5,424	133,184–138,604	9,820–15,243	0.99	0.73	1
17	NT_011618.2	X	5,317	760,316–765,591	576–5,892	0.97	0.78	
18	NT_022121.2	2	5,317	163,829–170,802	575–5,891	0.97	0.77	
19	NT_022208.2	2	5,058	176,917–182,316	12,220–708	0.88	0.65	1
20	NT_007995.2	8	4,724	301,945–307,342	636–5,359	0.95	0.75	
21	NT_025827.1	10	4,513	120,512–125,898	3,821–8,333	0.97	0.73	1
22	NT_025657.1	2	4,379	50,342–55,439	9,196–13,574	0.93	0.70	
23	NT_007583.2	6	4,321	972,351–976,677	5,435–9,755	—	0.98	1
24	NT_005229.2	2	4,275	2,719,563–2,724,104	6,966–11,240	0.98	0.74	1
25	NT_024640.1	14	3,959	103,603–107,756	11,367–15,325	0.97	0.72	
26	NT_011896.3	Y	3,882	5,666,802–5,672,422	596–4,477	0.83	0.64	
27	NT_004328.2	1	3,812	387,888–391,662	9,782–13,593	0.98	0.72	1
28	NT_007772.1	7	3,761	23,528–27,595	2,793–6,553	0.99	0.76	
29	NT_023678.2	8	3,551	219,881–228,579	16,498–3,479	0.87	0.70	
30	NT_009859.2	13	3,422	209,439–213,269	13,051–16,472	0.96	0.71	
31	NT_006654.2	5	3,380	437,998–441,377	12,662–16,041	—	0.87	1
32	NT_011574.1	X	3,362	364,748–368,477	1,054–4,415	0.93	0.71	
33	NT_022290.1	2	3,267	139,811–143,078	11,801–15,067	0.99	0.73	
34	NT_025076.2	18	3,267	63,799–67,066	11,801–15,067	0.99	0.73	
35	NT_025259.2	X	3,188	583,624–590,336	12,136–15,323	0.90	0.66	
36	NT_022790.2	4	3,177	430,560–434,014	16,498–3,105	0.86	0.70	
37	NT_008421.2	9	3,083	1,252,046–1,255,273	11–3,093	0.91	0.76	1
38	NT_005496.2	3	2,996	500,209–503,607	9,345–12,340	1.00	0.72	
39	NT_022852.2	4	2,979	846,486–849,454	13,046–16,024	1.00	0.78	1
40	NT_022127.2	2	2,939	194,666–197,564	10,657–13,595	0.99	0.74	
41	NT_021990.2	1	2,867	81,390–84,148	11,367–14,233	0.97	0.72	
42	NT_024891.2	17	2,793	1,009,574–1,013,294	2,087–4,879	0.94	0.76	
43	NT_024633.1	14	2,723	110,583–113,284	6,594–9,316	0.98	0.74	
44	NT_022497.2	3	2,713	572,868–575,560	6,604–9,316	0.96	0.72	
45	NT_005229.2	2	2,692	1,127,113–1,129,798	10,440–13,131	1.00	0.76	1

Table 1
Continued

No. ^a	Accession ^b	Chromosome	Length ^c	Position ^d	Mit. Pos. ^e	DBA cover ^f	Id. ^g	Phylo. ^h
46	NT_025295.2	X	2,588	129,449–132,250	14,685–703	0.81	0.62	
47	NT_006322.2	4	2,521	534,125–536,928	9,781–12,301	0.98	0.73	
48	NT_007653.2	7	2,496	222,299–224,791	598–3,093	1.00	0.85	1
49	NT_009243.2	11	2,452	244,075–246,524	521–2,972	—	0.94	1
50	NT_008387.2	9	2,397	1,196,615–1,199,276	9,202–11,598	0.99	0.73	
51	NT_024535.2	13	2,384	114,983–117,781	11,032–13,415	0.95	0.69	
52	NT_005164.2	3	2,365	256,359–258,716	3,799–6,163	0.98	0.76	
53	NT_007664.2	7	2,305	125,183–127,480	13,065–15,369	1.00	0.75	1
54	NT_006961.2	5	2,277	686,336–688,608	421–2,697	—	0.94	1
55	NT_008136.2	8	2,269	601,918–604,576	1,013–3,281	0.84	0.64	
56	NT_024891.2	17	2,202	993,589–995,958	16,434–2,066	0.84	0.67	
57	NT_008524.2	9	2,100	116,089–118,186	4,773–6,872	1.00	0.77	1
58	NT_024128.2	10	2,098	150,954–153,089	16,376–1,904	0.95	0.78	
59	NT_010577.2	16	2,045	382,708–385,073	2,427–4,471	1.00	0.79	
60	NT_023290.2	5	1,963	170,306–172,278	5,934–7,896	1.00	0.79	

NOTE.—See Supplementary Material on the MBE website for a complete list of human numts.

^a Annotation corresponds to the numbering in figure 2.

^b GenBank accession number of working draft sequence.

^c Length of mitochondrial sequence aligned to numt.

^d Position of numt in GenBank sequence.

^e Position of corresponding sequence in human mitochondria (GenBank accession number NC_001807).

^f Percentage of mitochondrial sequence aligned by DBA. — = stem from a single BLAST hit that subsequently did not undergo alignment using DBA.

^g Percentage of identity between numt and mtDNA as determined by either BLAST or DBA.

^h Blocks of alignment used for phylogeny (fig. 2).

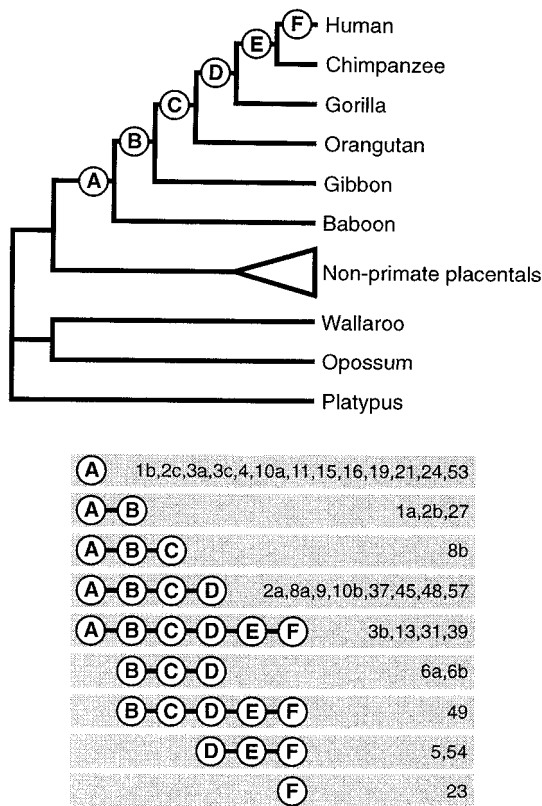


FIG. 2.—Consensus tree of the phylogenetic positions of human numts, based on 35 individual bootstrap analyses of all blocks from the DBA alignment longer than 2,000 bp. The trees were constructed in PAUP*, version 4.0b4a (Swofford 1998), using the neighbor-joining algorithm based on maximum-likelihood (ML) distance measures. The shape parameters of the gamma distributions, α (0.28–0.43), and the transition-transversion rates (1.6–2.9) were estimated using ML. Six branching points are depicted on the tree (A–F). On the basis of 100% support with 100 bootstrap replicates and Platypus as the outgroup, numts could be confined to one or more of the branching points, as shown below the tree. For example, numts listed in the gray box (A) have 100% bootstrap support positioned at branching point A, whereas numts listed in the box (A–B) with the same support only can be confined to either branching point A or branching point B. Needless to say, numts in the box covering all positions (A–F) are restricted to the primate clade, but their exact position is undetermined. If two or more alignment blocks come from the same numt, these have letter suffixes (see table 1 for details). The following mtDNA sequences were used (GenBank accession numbers in parentheses): human (*Homo sapiens*; NC_001807), chimpanzee (*Pan troglodytes*; NC_001643), gorilla (*Gorilla gorilla*; NC_001645), Orangutan (*Pongo pygmaeus*; NC_001646), gibbon (*Hylobates lar*; NC_002082), baboon (*Papio hamadryas*; NC_001992), wallaroo (*Macropus robustus*; NC_001794), opossum (*Didelphis virginiana*; NC_001610), and platypus (*Ornithorhynchus anatinus*; NC_000891). Nonprimate placentals: alpaca (*Lama pacos*; NC_002504), armadillo (*Dasypus novemcinctus*; NC_001821), bat (*Chalinolobus tuberculatus*; NC_002626), cat (*Felis catus*; NC_001700), cow (*Bos taurus*; NC_001567), European hedgehog (*Eriaceus europaeus*; NC_002080), flying fox (*Pteropus scapulatus*; NC_002619), guinea pig (*Cavia porcellus*; NC_000884), Madagascar hedgehog (*Echinops telfairi*; NC_002631), rabbit (*Oryctolagus cuniculus*; NC_001913), squirrel (*Sciurus vulgaris*; NC_002369), and tree shrew (*Tupaia belangeri*; NC_002521).

pletely alignable, or subsets of numts of which DBA blocks exceeded 2,000 bp) and aligned these with the corresponding mtDNA sequences from a variety of mammals. The phylogenetic analysis supported the gen-

eral conviction that numt DNAs are continually integrated into the nuclear genome as a result of several independent evolutionary events (fig. 2).

Since we used the working draft of the human nuclear genome for analysis, we cannot exclude that some of the recent integration events are simply due to erroneous sequencing of mitochondrial contamination. However, this will not change the above conclusions. On the contrary, the above findings may be an underestimate, since recently transferred numts may not have reached fixation (e.g., Zischler et al. 1995) and therefore may not be present in the available human genome draft.

This study presents the first extensive large-scale survey of human numts based on the human genome project—an initial step on the way to a complete catalog of human numts.

As previously stated (Perna and Kocher 1996), human numts may serve as both obstacles and tools in understanding the evolution of the human mitochondria. For example, the large number of long numts can confound studies on mitochondrial heteroplasmy as well as phylogenetic and population studies using mtDNA markers. For these studies, decisive knowledge of human numts may be crucial in detecting erroneous results due to false amplification of nuclear homologs.

On the contrary, since numts may be regarded as “molecular fossils” of mtDNA (Zischler, Geisert, and Castresana 1998), they may provide fruitful insight into the evolution of modern human mitochondria and help to uncover the evolutionary basis of contemporary human diseases related to the genetics of the mitochondria.

Supplementary Materials

A table of all 296 human numts is provided on the *Molecular Biology and Evolution* web site.

Acknowledgments

We thank Douda Bensasson, Kasi B. Desfor, Sylvia Mathiasen, and Seirian Sumner for help and discussions. A.J.H. and E.W. were supported by the VELUX foundation of 1981, Denmark. A.J.H. and E.W. contributed equally to this work and should be regarded as joint authors.

LITERATURE CITED

- ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- ARCTANDER, P. 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc. R. Soc. Lond. B Biol. Sci.* **262**:13–19.
- BENSASSON, D., D.-X. ZHANG, D. HARTL, and G. HEWITT. 2001. Mitochondrial pseudogenes: evolution’s misplaced witnesses. *Trends Ecol. Evol.* **16**:314–321.
- BLANCHARD, J. L., and G. W. SCHMIDT. 1996. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol. Biol. Evol.* **13**:537–548.

- HERRNSTADT, C., W. CLEVENGER, S. S. GHOSH, C. ANDERSON, E. FAHY, S. MILLER, N. HOWELL, and R. E. DAVIS. 1999. A novel mitochondrial DNA-like sequence in the human nuclear genome. *Genomics* **60**:67–77.
- JAREBORG, N., E. BIRNEY, and R. DURBIN. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**:815–824.
- LOPEZ, J. V., N. YUHKI, R. MASUDA, W. MODI, and S. J. O'BRIEN. 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.
- PERNA, N. T., and T. D. KOCHER. 1996. Mitochondrial DNA: molecular fossils in the nucleus. *Curr. Biol.* **6**:128–129.
- SACCONE, C., G. PESOLE, and E. SBISÁ. 1991. The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. *J. Mol. Evol.* **33**:83–91.
- SHAY, J. W., and H. WERBIN. 1992. New evidence for the insertion of mitochondrial DNA into the human genome: significance for cancer and aging. *Mutat. Res.* **275**:227–235.
- SWOFFORD, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Mass.
- TSUZUKI, T., H. NOMIYAMA, C. SETOYAMA, S. MAEDA, and K. SHIMADA. 1983. Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. *Gene* **25**:223–229.
- WALLACE, D. C., C. STUGARD, D. MURDOCK, T. SCHURR, and M. D. BROWN. 1997. Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations. *Proc. Natl. Acad. Sci. USA* **94**:14900–14905.
- ZHANG, D.-X., and G. M. HEWITT. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol. Evol.* **11**:247–251.
- ZISCHLER, H., H. GEISERT, A. VON HAESELER, and S. PÄÄBO. 1995. A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* **378**:489–492.
- ZISCHLER, H., H. GEISERT, and J. CASTRESANA. 1998. A hominoid-specific nuclear insertion of the mitochondrial d-loop: implications for reconstructing ancestral mitochondrial sequences. *Mol. Biol. Evol.* **15**:463–469.

PEKKA PAMILO, reviewing editor

Accepted June 4, 2001