J Mol Evol (2003) 56:169–174
DOI: 10.1007/s-00239-002-2390-5

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Dynamics of Large *Numts* in the Human Genome: Rarity of Independent Insertions and Abundance of Post-Insertion Duplications

Einat Hazkani-Covo,[1] Rotem Sorek,[1,2] Dan Graur[1]

[1] Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[2] Compugen Ltd., 72 Pinchas Rosen St., Tel Aviv 69512, Israel

Received: 12 July 2002 / Accepted: 9 September 2002

Abstract. We determined the phylogenetic positions of 82 large nuclear pseudogenes of mitochondrial origin (*numts*) within the human genome. For each *numt*, two possibilities pertaining to its origin were considered: (1) independent insertion from the mitochondria into the nucleus, or (2) genomic duplication subsequent to the insertion. A significant increase in the rate of *numt* accumulation is seen after the divergence of Platyrrhini (New World monkeys) from the Catarrhini (Old World monkeys, apes and humans). By using pairwise phylogenetic analyses, we were able to demonstrate that this peak in *numt* accumulation is mostly the result of duplication of preexisting nuclear *numts* rather than the result of an increase in mitochondrial-sequence insertion. In fact, only about a third of all the *numt* repertoire in the human nuclear genome is due to insertions of mitochondrial sequences, the rest originated as duplications of preexisting *numts*. Hence, we conclude that *numt* insertion occurs at a much lower rate than previously reported. As expected under the assumption that genomic duplications occur at rates that are uninfluenced by content, older *numts* were found to be duplicated more times than recently inserted ones.

Key words: *Numts* — Human genome — Promiscuous DNA — Gene duplicaton — Pseudogenes — Primates

Introduction

Starting with the findings of Stern and Lonsdale (1982) on the transfer of genetic information among genomes, hundreds of studies have documented the ubiquity of genetic-information flow between organelles and between organelles and the nucleus (e.g., Blanchard and Schmidt 1995; Collura and Stewart 1995; Fukuda et al. 1985; Lopez et al. 1994). This type of "disrespect" for genomic barriers has been dubbed "promiscuous DNA" (Ellis 1982; Lewin 1983). To date, examples have been found for five out of the six possible types of gene transfer among genomes: chloroplast to mitochondria, mitochondria to chloroplast, chloroplast to nucleus, nucleus to mitochondria, and mitochondria to nucleus (Thorsness and Weber 1996).

While the transfer of functional mitochondrial genes into the nucleus has most probably ceased before the emergence of animals, approximately 1,000 million years ago (Boore 1999), the integration of functionless mitochondrial sequences into the nuclear genome has continued unremittingly, and nuclear pseudogenes of mitochondrial origin or *numts* (pronounced "*new-mights*", Lopez et al. 1994) have been described in numerous eukaryotes (Bensasson et al. 2001). All mammalian *numts* studied to date were found to be functionless, and it is thought that because of the differences between the nuclear and mitochondrial genetic codes, they became pseudogenes immediately on arrival into the nucleus. *Numts* have an uneven taxonomic and chromosomal distribution,

Correspondence to: D. Graur; *email:* graur@post.tau.ac.il

but so far no diagnostic features have been described for the regions flanking the *numt* integration sites (Bensasson et al. 2001). Gene transfer from the mitochondria to the nucleus most probably occurs through direct DNA transfer, rather than through cDNA-mediated transfer (Henze and Martin 2001).

Recently, the full *numt* repertoire of the human nuclear genome was published (Mourier et al. 2001; Woischnik and Moraes 2002). On the basis of phylogenetic analyses, both groups concluded that the evolutionary process *of numt* insertion is continuous (Bensasson et al. 2001), and that it occurs at fairly rapid rates. However, we find their phylogenetic analyses incomplete, especially since they regard all *numts* as independent mitochondrial insertions and neglect the possibility of post-insertion nuclear duplication (e.g., Bensasson et al. 2000). In this study, we attempted to reconstruct the evolutionary dynamics of *numt* accumulation. In particular, we determined for each *numt* whether it was derived independently through the integration of a mitochondrial sequence or whether it was created through the nuclear-genome duplication of a preexisting *numt*.

Materials and Methods

Data Collection. The FASTA algorithm (Pearson and Lipman 1988) was used to search each of the human chromosomes (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/) for regions of similarity with the human mitochondrial sequence (Genebank, NC_001807). Ninety-four hits that were longer than 1,000 bp were selected for further analysis. After filtering overlapping results and choosing the ones that had the longer hits, we used the Smith-Waterman algorithm (Smith and Waterman 1981) to join closely spaced (< 100 Kb) hits that were found on the same contig and in the same orientation. The algorithm was employed to ensure that each *numt* in our analysis appears only once, i.e., that it was not artificially divided into segments. These procedures reduced the number of hits to 82 *numts*.

Phylogenetic Analysis of numts. Ten full mitochondrial sequences were selected for phylogenetic analysis and were aligned using ClustalW (Higgins et al. 1996). A user tree (Fig. 1A) was built for fully sequenced mitochondrial genomes from eight primates, a sister taxon (*Tupaia belangeri*, Scandentia), and an outgroup (*Phoca vitulina*, Pinnipedia, Carnivora). The taxa were chosen on the basis of complete-mitochondrial sequence availability and the possibility of building a taxonomically undisputed phylogenetic tree (Goodman et al. 1998). Genebank accession numbers for the mitochondrial sequences are: NC_001807 for human (*Homo sapiens*), NC_001643 for chimpanzee (*Pan troglodytes*), NC_001645 for gorilla (*Gorilla gorilla*), NC_001646 for orangutan (*Pongo pygmaeus*), NC_002082 for white-handed gibbon (*Hylobates lar*), NC_002764 for Barbary macaque (*Macaca sylvanus*), NC_002763 for white-fronted capuchin (*Cebus albifrons*), NC_002765 for slow loris (*Nycticebus coucang*), NC_002521 for northern tree shrew (*Tupaia belangeri*), and NC_001325 for harbor seal (*Phoca vitulina*).

Branch lengths were calculated through maximum-likelihood methodology with the DNAML program in PHYLIP 3.573 (Felsenstein 1993). ClustalW was used to align each of the 82 *numts* to the 10 mitochondrial sequences. Each *numt* was added to each of the nine branches on the lineage leading to the human genome, and by using DNAML we computed the likelihood of each of the nine resulting trees.

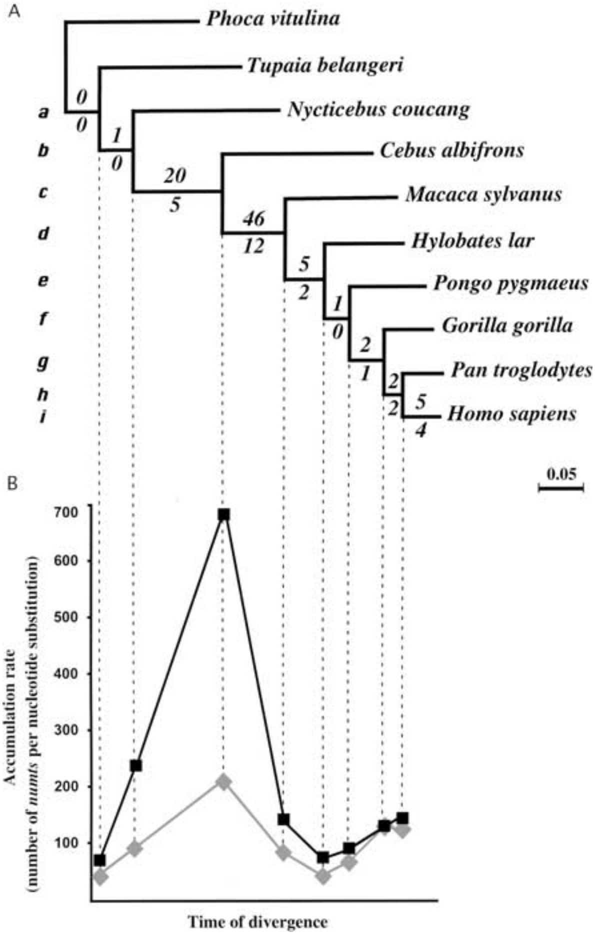

Fig. 1. **(A)** Maximum likelihood phylogenetic tree based on 10 complete mitochondrial sequences from primates and outgroups. Branch lengths were calculated with the DNAML program and are proportional to numbers of nucleotide substitutions in the mitochondria. Branch lengths are measured in units of nucleotide substitution per site (see bar). Numbers of *numts* that have originated at various evolutionary times (above branch), and number of separate insertions (below branch) are indicated. The notation for the tree branches (*a–i*) is also used in Table 1. **(B)** Temporal dynamics of *numt* accumulation in the nuclear genome (black line), and those of separate *numts* insertions (gray line), plotted on a time axis derived from the maximum likelihood phylogenetic tree in **A**. Time axis is measured in units of nucleotide substitution per site (see scale bar).

The nine trees were given two scores: (1) The unweighted score was the number of times that each of the nine trees emerged as the most likely tree. (2) The weighted score was calculated as follows: If the likelihood of the best tree was significantly different from the other trees, the tree was given a score of 1. If two trees could not be shown to differ from each other in a statistically significant manner ($p < 0.05$), each of the two trees was given a score of 0.5. If three trees could not be shown to differ from one another in a statistically significant manner, each of the three trees was given a score of 0.33, and so on. For each of the nine trees, we summed the scores over the 82 *numts*.

Phylogenic Analysis of Pairs of numts. We compiled a database of pairs of *numts*, in which each pair contains a short *numt* that is fully contained within a long one. We used the previously determined maximum-likelihood branch location for the longer *numt* to identify the phylogenetic position of the shorter *numt*. The maximum likelihood position for the shorter *numt* in a pair was identified with the user-tree option in the DNAML program. If the two *numts* emerged as sister taxa on the same branch, we concluded that the shorter *numt* represents a partial duplication of the longer one. In such a case, the longer *numt* is called the "father" and the shorter one is called the "son".

Inferrence of the Number of Independent numt *Insertions. Numts* that participate in pairs only as fathers but never as sons were deemed to have been created by insertion. *Numts* that did not appear in the database of pairs were also classified as independent insertions. All other *numts* were inferred to have been created by duplication of a preexisting *numt*.

Results

Eighty-two *numts* longer than 1,000 bp were identified in the human nuclear genome (Table 1). The chromosomal distribution of *numts* was found not deviate significantly from a random distribution (χ^2 = 22.85; df = 23; $p \ll 0.47$). This finding is in agreement with Mourier et al. (2001).

By using maximum likelihood methodology, it was possible to place each of the 82 *numts* in their temporal evolutionary context (Fig. 1A). When adjacent placements on the phylogenetic tree could not be distinguished from one another with sufficient statistical confidence, we assigned equal probability of *numt* origin on each of the indistinguishable branches. Numbers of *numts* were similar for both the weighted and the unweighted method.

We applied Grubbs' extreme studentized deviate test (Barnett and Lewis 1994) on the unweighted numbers of *numts* divided by the lengths of their respective branches. A statistically significant (p < 0.01) 30-fold increase in the rate of *numt* accumulation was observed to have occurred on the branch leading to Catarrhini (Old World monkeys, apes, and humans) after their divergence from the Platyrrhini (New World monkeys) approximately 40 million years ago (Fig. 1B).

The dramatic change in the rate of *numt* accumulation could be due to increase in the rate of independent sequence transfers from the mitochondria to nucleus or due to post-insertion duplications within the nuclear genome. In order to distinguish between the two possibilities, we analyzed 323 *numt* pairs. Nine of the 82 *numts* were found to have no relation to the other *numts* and were, thus, considered as independent insertions. The other 73 *numts* were inferred to have been created by the duplication of 17 ancestral *numts*. Thus, only 30% of all *numts* in the human nuclear genome have been created by insertion; the others have accumulated by subsequent duplication.

We placed each of the 26 independently inserted *numts* on the branches of the mitochondrial phylogenetic tree (Fig. 1A). Again, we found a relative excess of *numt* accumulation (this time attributed solely to insertion) on the branch leading to Catarrhini after its divergence from the Platyrrhini (Fig. 1B). Nevertheless, Grubbs' extreme studentized deviate test is no longer statistically significant.

The ratios between the number of *numts* and the number of *numt* insertions on the branches ranged from 4 to 1, with the higher values obtained for the older branches. This indicates, that older *numts* have been duplicated more times than younger ones.

Discussion

Recently, several papers analyzing the full *numt* repertoire reported a continuous evolutionary transfer of mitochondrial sequences into the human nuclear genome. Mourier et al. (2001) used a combination of BLAST (Altschul et al. 1997) and DNA-block alignment (Jareborg et al. 1999), and found that the human nuclear genome contains 296 *numts*, 94 of which were longer than 1,000 bp. In our survey, we have only identified 82 such *numts*, most probably because of our more conservative criteria for inclusion. Although the method of Mourier et al. (2001) is suitable for the identification of the human *numt* repertoire, their phylogenetic analysis is, to say the least, inconclusive. First, Mourier et al. (2001) ignored the possibility of *numt* duplication. Second, since many of their *numts* consisted of disjointed segments, in many cases *numts* were placed in more than one phylogenetic position on the tree. This is evolutionarily impossible and should be regarded as an artifact of their use of the block-alignment algorithm, which has yielded *numts* with varying degrees of similarity to the mitochondrial parent.

In the study by Woischnik and Moraes (2002), the authors searched for hits of single mitochondrial genes in the nuclear genome, and used their coordinates to combine them into longer *numts*. Woischnik and Moraes (2002) discovered 612 *numts*. The phylogenetic analysis in Woischnik and Moraes (2002) was carried out gene by gene, so that parts of the same *numt* were most probably positioned on different branches of the tree. And again, the possibility of *numt* duplication occurring subsequent to the insertion of the mitochondrial sequence was ignored.

Here, we performed an analysis on 82 long *numts*. We did not aim to identify the entire *numt* repertoire, a process that was most probably completed by Mourier et al. (2001), but to reconstruct *numt* evolutionary history by taking into account the possibility of genomic duplication. We found that the number of *numts* is positively correlated with branch length. For example, the longest branches, i.e., those representing the divergence between Platyrrhini and Catarrhini and between Strepsirhini and the rest of the Primates, show the higher number of *numts*. In other words, our analysis indicates that *numt* insertion into the nuclear genome is a continuous and largely monotonic evolutionary process. However, our analysis also indi-

Table 1. Human *numts* longer than 1,000 bp

Contig	Chromosome	Length	Mitochondria position	Contig position	%Similarity	Tree location[a]
NT_023115.7	5	8821	6390-15211	3721-12548	88.80%	g
NT_007412.7	6	5888	3912-9800	141462-135572	98.18%	i
NT_030001.2	7	5831	6149-11980	1337640-1331818	59.52%	d
NT_009184.7	11	5765	9821-15586	121983-116221	59.79%	d
NT_004836.7	1	5634	573-6207	1451784-1457418	75.15%	d
NT_010530.7	16	5302	8584-13886	2334047-2328751	70.91%	d
NT_007091.7	5	5222	10266-15488	588077-582855	93.99%	i
NT_024089.7	10	5177	3294-8543	678570-684680	61.50%	c
NT_022208.5	2	4654	11590-16244	178103-182758	71.30%	d
NT_028400.2	X	4177	2232-6409	179660-183822	74.01%	d
NT_023451.7	6	4027	7669-11696	1133421-1137446	54.89%	c
NT_022140.7	2	3827	8296-12123	91132-94949	75.70%	d
NT_006129.6	4	3712	8296-12008	427566-423870	74.47%	d
NT_024862.6	17	3697	2141-5838	2611-6307	84.66%	e
NT_006654.7	5	3464	12661-16125	1068972-1072433	86.56%	g
NT_006129.6	4	3379	663-4042	99606-96231	79.44%	d
NT_030719.1	7	3356	3293-6649	24338-27690	63.92%	d
NT_026437.5	14	3305	12417-15412	2347244-2343953	66.51%	d
NT_005151.7	2	3291	11802-15093	2651955-2648663	68.82%	d
NT_007995.7	8	3086	2009-5095	191614-188537	62.55%	d
NT_011362.7	20	2766	1045-3811	20988535-20985785	70.75%	d
NT_005229.7	2	2751	10409-13160	1250750-1248006	75.73%	d
NT_005129.7	2	2603	737-3340	2165807-2168408	71.33%	d
NT_023678.6	8	2547	1033-3580	412979-410448	74.39%	d
NT_007769.5	7	2545	576-3121	224815-222274	83.35%	d
NT_030040.2	9	2545	576-3121	2066695-2069239	83.56%	d
NT_009243.7	11	2451	523-2974	832183-829734	94.09%	h
NT_008541.7	9	2443	4548-6991	595250-597696	75.80%	d
NT_022852.7	4	2432	12890-15322	852792-850357	75.50%	e
NT_007884.7	7	2392	13066-15458	437318-434933	73.82%	d
NT_022790.7	4	2366	577-2943	689245-691612	68.46%	d
NT_011896.7	Y	2333	14237-16570	5697142-5699469	61.46%	c
NT_008583.7	10	2309	11645-13954	3338266-3340570	74.51%	d
NT_024814.5	16	2303	13905-16208	204841-202534	71.53%	c
NT_006961.7	5	2281	420-2701	103574-101297	93.59%	h
NT_008583.7	10	2173	1703-3876	2182212-2184389	77.92%	d
NT_008251.7	8	2107	13932-16039	858254-856146	76.97%	e
NT_023290.4	5	2099	5891-7990	178922-181031	70.80%	d
NT_009952.7	13	2048	13942-15990	205485-207526	72.84%	d
NT_006576.7	5	2034	14139-16173	1204488-1206530	71.44%	c
NT_005229.7	2	1995	9473-11468	1588832-1590824	64.54%	c
NT_011649.7	X	1993	14031-16024	1271724-1269724	74.75%	d
NT_019350.7	3	1984	8619-10603	936218-934237	73.07%	c
NT_011613.7	X	1893	14678-16571	56571-54679	65.36%	d
NT_008150.6	8	1868	4865-6733	812806-814667	75.75%	d
NT_006322.7	4	1861	9441-11302	515297-517150	73.51%	d
NT_015360.7	16	1833	2783-4616	18822-16998	76.03%	d
NT_026437.5	14	1810	575-2385	2931912-2933718	65.18%	d
NT_006936.7	5	1770	14274-16044	198918-197149	69.68%	c
NT_005638.6	3	1669	9105-10774	622432-624101	75.03%	d
NT_027070.4	8	1631	10772-12403	775266-773640	73.92%	d
NT_005112.7	2	1609	10995-12604	158769-157171	74.83%	d
NT_025395.6	X	1584	12995-14579	136603-135032	61.78%	c
NT_008218.7	8	1553	7024-8577	7163-8721	94.23%	i
NT_030590.1	2	1531	8272-9803	339514-341038	71.87%	c
NT_009151.7	11	1470	8309-9779	6648376-6646914	73.00%	c
NT_030828.1	15	1452	11683-13135	3728186-3729632	75.77%	d
NT_024862.6	17	1448	14335-15783	139720-138272	82.29%	f
NT_010441.7	16	1446	12196-13642	244421-242976	74.90%	d
NT_030761.1	9	1417	8474-9891	1556287-1554877	74.44%	d
NT_022475.4	3	1384	1387-2771	511841-510458	93.07%	i
NT_022066.1	1	1368	14671-16039	76000-74633	77.24%	e
NT_005151.7	2	1353	4855-6208	2698233-2699587	81.59%	e
NT_009184.7	11	1337	14505-15842	6619287-6617962	72.73%	c

Table I. Continued

Contig	Chromosome	Length	Mitochondria position	Contig position	%Similarity	Tree location[a]
NT_019350.7	3	1324	3501-4825	939172-940487	72.95%	d
NT_030846.1	17	1306	3178-4484	618593-617293	76.95%	d
NT_025766.6	7	1282	14748-16030	707948-709233	70.74%	c
NT_010289.7	15	1275	14465-15740	2202527-2201254	74.46%	c
NT_015326.7	12	1249	3923-5172	860080-858838	71.24%	c
NT_028068.4	2	1224	8398-9622	460959-462176	75.43%	d
NT_026965.2	2	1221	13007-14228	32515-33735	72.42%	d
NT_010530.7	16	1211	10474-11685	2401941-2403161	72.44%	d
NT_005445.7	2	1153	15418-16571	819365-818210	63.75%	c
NT_004836.7	1	1110	13693-14803	3956054-3957164	70.59%	c
NT_024296.3	11	1084	6449-7533	72596-73680	75.16%	d
NT_022497.7	3	1076	8275-9351	541783-542854	71.89%	d
NT_011512.4	21	1050	4910-5960	22840953-22839916	64.87%	c
NT_011574.3	X	1040	753-1793	969765-968737	62.78%	c
NT_019284.6	1	1037	2071-3108	225524-226553	67.43%	b
NT_019350.7	3	1035	1015-2050	941525-940486	77.36%	c
NT_010164.7	14	1025	5584-6609	6073254-6072232	92.98%	i
NT_006560.7	5	1008	6953-7961	1103416-1104421	78.12%	d

[a] Tree location as in Fig. 1.

cates that the process of *numt* insertion is less frequent than previously reported. The peak in *numt* accumulation that is found on the branch representing the divergence of Platyrrhini from the Catarrhini is mostly the result of duplication of preexisting nuclear *numts* rather than the result of an increase in mitochondrial-sequence insertion. In fact, this phenomenon is a general one; on average only one of every three *numts* in our genome is the result of an independent integration event, while the other two originate from duplications within the nuclear genome.

Under the assumption that genomic duplications occur at rates that are uninfluenced by content, older *numts* should appear in larger copy numbers than recently inserted *numts*. Let us consider, for instance, branch C in Figure 1. Twenty *numts* were located on this branch, yet these 20 *numts* are derived from only five independent insertions. In contrast, of the five *numts* inferred to have accumulated in the human genome after the *Homo-Pan* divergence, four are most probably independent insertions. These five *numts* are expected to have no homologues in non-human genomes. The fact that older *numts* were indeed found to be duplicated more times than younger *numts* strengthens the confidence in the reliability of our results.

Acknowledgments. We wish to thank the people of Compugen Ltd., especially Idit Azar, Gad S. Cojocaru, and Erez Levanon. E. H.-C. wishes to thank Shay Covo for his helpful advise.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley &Sons, New York

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

Blanchard JL, Schmidt GW (1995) Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol 41:397–406

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780

Collura RV, Stewart CB (1995) Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378:485–489

Ellis J (1982) Promiscuous DNA–chloroplast genes inside plant mitochondria. Nature 299:678–679

Felsenstein J (1993) PHYLIP: Phylogeny inference package and manual. Version 3.5 Department of Genetics, University of Washington, Seattle

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

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

Henze K, Martin W (2001) How do mitochondrial genes get into the nucleus? Trends Genet 17:383–387

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266:383–402

Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 9:815–824

Lewin R (1983) Promiscuous DNA leaps all barriers. Science 219:478–479

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

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

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444–2448

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol l47:195–197

Stern DB, Lonsdale DM (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. Nature 299:698–702

Thorsness PE, Weber ER (1996) Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. Int Rev Cytol 165:207–234

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