NUMTs in Sequenced Eukaryotic Genomes

Erik Richly and Dario Leister

Max-Planck-Institut für Züchtungsforschung, Köln, Germany

Mitochondrial DNA sequences are frequently transferred to the nucleus giving rise to the so-called nuclear mitochondrial DNA (NUMT). Analysis of 13 eukaryotic species with sequenced mitochondrial and nuclear genomes reveals a large interspecific variation of NUMT number and size. Copy number ranges from none or few copies in *Anopheles*, *Caenorhabditis*, *Plasmodium*, *Drosophila*, and *Fugu* to more than 500 in human, rice, and *Arabidopsis*. The average size is between 62 (baker's yeast) and 647 bps (*Neurospora*), respectively. A correlation between the abundance of NUMTs and the size of the nuclear or the mitochondrial genomes, or of the nuclear gene density, is not evident. Other factors, such as the number and/or stability of mitochondria in the germline, or species-specific mechanisms controlling accumulation/loss of nuclear DNA, might be responsible for the interspecific diversity in NUMT accumulation.

In eukaryotes, nuclear DNA exists that is homologous to mitochondrial DNA (mtDNA). These sequences, which originate from the invasion of nuclear DNA by mtDNA, are designated nuclear mtDNA (NUMT) (Lopez et al. 1994). NUMTs exhibit different degrees of homology to their mitochondrial counterparts; are variable in size; evenly distributed within and among chromosomes, and, in cases, are highly rearranged and/or fragmented (Zhang and Hewitt 1996; Ricchetti, Fairhead, and Dujon 1999; Woischnik and Moraes 2002). Available data indicate a predominant descent of NUMTs from nonhomologous recombination of nuclear DNA with mtDNA fragments leaking out of damaged mitochondria (Henze and Martin 2001; Mourier et al. 2001; Woischnik and Moraes 2002). Moreover, it is likely that the accumulation of NUMTs is a continuous evolutionary process (Mourier et al. 2001; Woischnik and Moraes 2002; Bensasson, Feldman, and Petrov 2003). However, the duplication of NUMTs preexisting in the nuclear genome should have also contributed to increase their number (Lopez et al. 1994; Bensasson, Zhang, and Hewitt 2000; Tourmen et al. 2002; Hazkani-Covo, Sorek, and Graur 2003). NUMTs have been so far detected in more than 82 species (Bensasson et al. 2001*a*): 30 in baker's yeast (Ricchetti, Fairhead, and Dujon 1999) and between 296 and 612 in the human genome (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Bensasson, Feldman, and Petrov 2003). In *Caenorhabditis elegans*, *Plasmodium falciparum*, *Drosophila melanogaster*, and other species they are rare or even absent (Bensasson et al., 2001*a*). Here, an updated inventory of NUMTs in the sequenced eukaryotic genomes and causes for the specificity of NUMT accumulation are discussed.

Both the mitochondrial and the nuclear genome sequence are known for 13 eukaryotic species. Employing BLASTN searches at a range of different threshold levels allowed us to identify diverged and/or small NUMTs, as well as conserved and/or long ones. The result is that dramatic differences in the content of NUMTs in the different genomes are evident (fig. 1*a*): at a threshold of 10^{-4} they range from less than ten in *Fugu*, *Drosophila*, *Plasmodium*, and *Caenorhabditis* to more than 500 in human, rice, and *Arabidopsis*. Between 10 and 100 NUMTs are present in rat, *Ciona*, *Neurospora*, and in the yeast species. No NUMTs at all have been detected in *Anopheles*. For *N. crassa* only a preliminary estimate is possible because sequence information of its mtDNA is still incomplete; nevertheless, between 11 (threshold $<10^{-50}$) and 22 (10^{-4}) NUMTs exist in this fungal species (fig. 1*a*).

In human, mouse, and *Ciona*, most or almost all mtDNA sequences were transferred to the nucleus (table 1), indicating that in principle all mitochondrial sequences are transferable, as argued by Allen (1993). Size distributions of NUMTs in species with more than 10 copies are shown in figure 1*b*. The longest NUMTs are present in *Neurospora*, *Arabidopsis*, and *Ciona*, whereas the yeast species and rat contain, in average, the smallest ones. When the length of NUMTs is normalized based on the size of the mitochondrial chromosome, *Ciona*, human and mouse were the species with the longest NUMTs (fig. 1*b*).

Our data extend the previous observations of Bensasson et al. (2001*a*) concerning the NUMT content across species. The content is highly variable, ranging from 0 in *Anopheles* to more than 400 kbp in rice (table 1). Relative to the genome size, the two plant species contain the largest fraction of NUMTs (around 0.10% and 0.17% in rice and *Arabidopsis*, respectively). Organisms of related genera, such as different insect species (Sunnucks and Hales 1996; Bensasson et al. 2001*a*) or mammals as human, mouse, and rat (this study), differ substantially in their NUMT content (table 1 and fig. 1*a*). It is known that the fraction of noncoding nuclear DNA varies among eukaryotes, even among related species (Hartl 2000; Petrov 2001). If the frequency of NUMTs in noncoding regions of the genome were similar among species, their number should increase in species with more noncoding nuclear DNA, based on the assumption that transfers of mtDNA fragments into expressed regions of the genome is counterselected. This can explain the abundance of NUMTs in *Homo sapiens* but not their rarity or absence in species such as *Caenorhabditis* and *Anopheles* (table 1). Furthermore, the size of the mitochondrial chromosome does not correlate with NUMT's frequency or size distribution (fig. 1*b* and table 1).

Then what is the reason behind the variable abundance of NUMTs in different species? Two explanations can be suggested. (1) *The frequency of DNA transfer*

Key words: duplication, gene transfer, genome evolution, mitochondria, NUMT, pseudogene.

E-mail: leister@mpiz-koeln.mpg.de.

Mol. Biol. Evol. 21(6):1081–1084. 2004
DOI:10.1093/molbev/msh110
Advance Access publication March 10, 2004

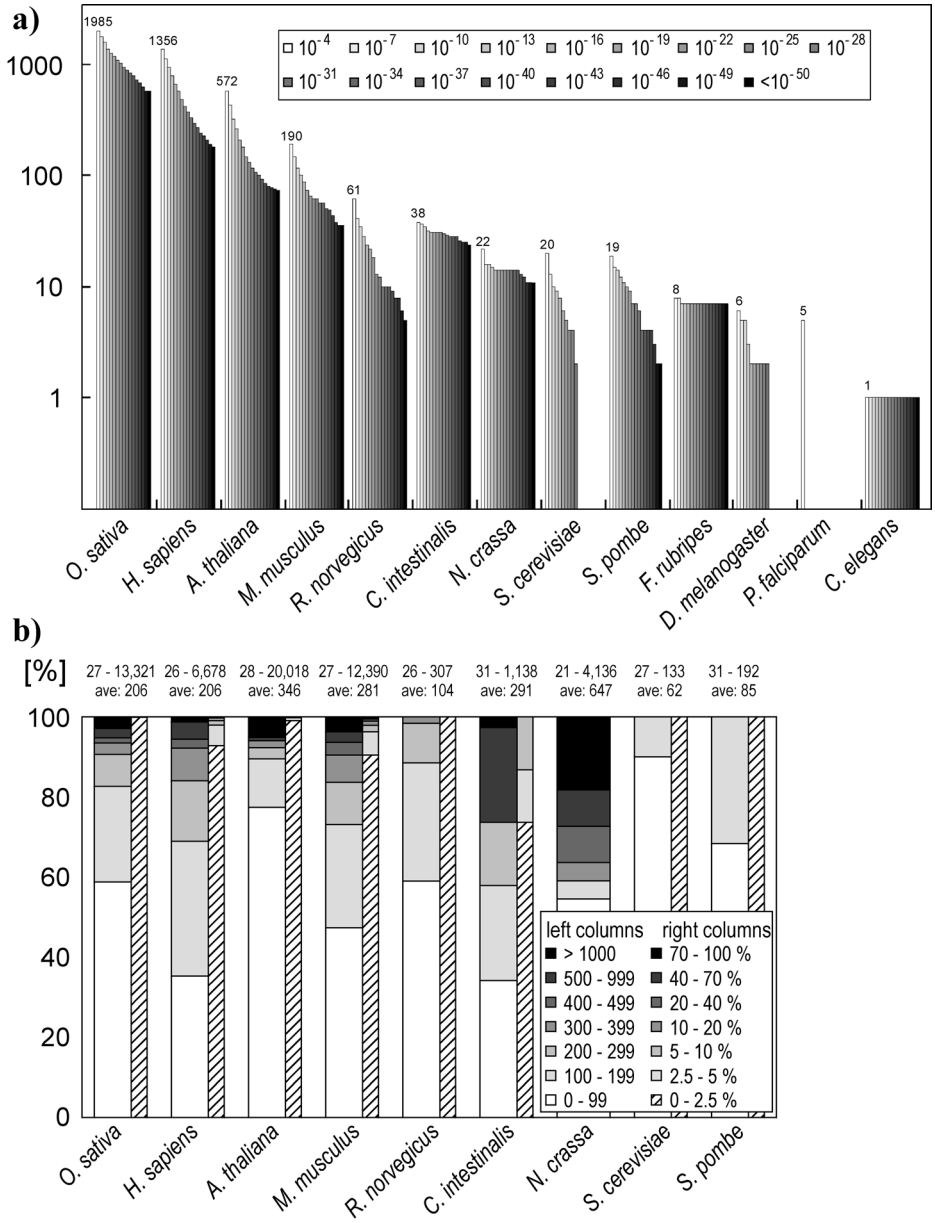

FIG. 1.—Number and length distribution of NUMTs in different nuclear eukaryotic genomes. (*a*) Frequency of NUMTs detected by BLASTN with thresholds from 10^{-4} to $<10^{-50}$ indicated by different shading and presented at a logarithmic scale. In *Anopheles*, no NUMT was detected. (*b*) Length distribution of NUMTs detected at a threshold of 10^{-4}. Only species with more than 10 NUMTs were considered. Left columns represent distributions of absolute lengths classes (in bps); right columns refer to the distribution of relative length of NUMTs considered as fractions (in %) of the complete size of mtDNA. Note that for *N. crassa* the complete mtDNA sequence is not known. On top of columns, minimum and maximum NUMT sizes and average values (ave), are reported.

from mitochondria to the nucleus differs between species. The mtDNA escape into the cytoplasm, and ultimately its transfer to the nucleus, can be influenced by the vulnerability of mitochondria to stress and other factors (Bensasson et al. 2001*a*; Woischnik and Moraes 2002), as well as by the number of mitochondria present in each cell—particularly of the germline. Accordingly, species-specific differences in the formation of the germline and/or the number of mitochondria per cell may account for some of the interspecific differences in NUMT abundance observed. The number of mitochondria per cell could, for instance, explain the low number of NUMTs in *Plasmodium*—an organism having only one mitochondrion per cell (Divo et al. 1985; Hopkins et al. 1999). Furthermore, the number of somatic cell divisions from zygote to meiosis (and the loss of the nuclear envelope during each division) should influence the frequency of mitochondrion-to-nucleus DNA transfer (Walbot and Evans 2003). This might be the reason for the high NUMT content in the plants rice and *Arabidopsis*. Also, the efficiency of nuclear import of mtDNA and/or of its integration into the nuclear genome might differ between species. (2) *The rate of loss of NUMTs is different among species*. The rate and spectrum of DNA loss from the

Table 1
Sizes of mtDNA and of Nuclear Genomes and of NUMTs Detected by BLASTN at a Threshold of 10^{-4}

	mtDNA		Nuclear Genome (Mbps)			NUMTs	
Species	Total Size [bps]	Transferred [%]	Total	Intergenic Regions	Noncoding Regions	bps	10^{-3}%
H. sapiens	16,571	98	2910.0	2183.0 (75%)	2852.0 (98%)	279,170	9.6
M. musculus	16,299	99	2500.0	nd	2375.0 (95%)	53,453	2.1
R. norvegicus	16,300	25	2800.0	nd	nd	6,340	0.2
F. rubripes	16,447	34	320.0	212.0 (66%)	285.0 (89%)	5,624	1.8
C. elegans	13,794	1	97.0	45.6 (47%)	71.8 (74%)	126	0.1
D. melanogaster	19,517	3	122.7	71.3 (58%)	99.4 (81%)	534	0.4
A. gambiae	15,363	0	278.2	215.0 (78%)	258.7 (93%)	0	0.0
C. intestinalis	14,788	74	116.7	nd	101.5 (87%)	11,051	9.9
P. falciparum	5,967	<1	22.9	9.5 (42%)	10.8 (47%)	152	0.7
S. cerevisiae	85,779	1	12.5	3.6 (29%)	3.8 (30%)	1,241	9.9
S. pombe	19,431	8	12.5	5.0 (40%)	5.4 (43%)	1,614	12.9
A. thaliana	366,924	45	115.4	64.1 (56%)	81.9 (71%)	198,105	171.7
O. sativa	490,520	38	420.0	232.3 (55%)	336.0 (80%)	409,104	97.4

NOTE.—"Transferred" mtDNA refers to the fraction (in %) of mtDNA that gave rise to NUMTs (repeated transfers of the same sequence are not considered). In the column "NUMTs (bps)" all nuclear sequences homologous to mtDNA are included, also when deriving from the same mtDNA sequences. "Noncoding regions" include all non-exon sequences as listed by Taft and Mattick (2003). Values in column "NUMTs (10^{-3} %)" refer to the ratio of NUMTs to the total size of the nuclear genome in the species concerned. nd, data not determined.

nucleus might shape the accumulation and size pattern of NUMTs. A specific spectrum of DNA loss could favor the deletion of NUMTs while still allowing the accumulation of massive amounts of noncoding DNA elements with different size. This type of DNA loss could lead to genomes with a large fraction of noncoding DNA but only with few NUMTs. Vice versa, a different control on DNA loss would allow more compact genomes to accumulate many NUMTs (such as in *Arabidopsis*). It is well known that the rate of DNA loss varies substantially for different fragment sizes and among species (Petrov et al. 2000; Bensasson et al. 2001*b*; Devos, Brown, and Bennetzen 2002), and this could explain the absence of a strict correlation between the abundances of noncoding nuclear DNA and NUMTs.

In conclusion, the causes for the interspecific diversity of NUMTs with respect to both copy number and length distribution remain obscure. The analysis of additional eukaryotic genomes to be completely sequenced in the future, in combination with the experimental analysis of the rates of mtDNA migration to the nucleus—particularly in related species that differ dramatically in their NUMT contents—should shed light onto the question how and to which extent eukaryotes deal with NUMTs and other pseudogenes in their genomes.

Materials and Methods

Sequence Analyses

Full-length mtDNA sequences were retrieved from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html). Nuclear DNA sequences were obtained from the NCBI (http://www.ncbi.nlm.nih.gov/genomes/static/EG_T.html; *Caenorhabditis elegans*, *Drosophila melanogaster*, *Anopheles gambiae*, *Homo sapiens*, *Mus musculus*, *Plasmodium falciparum*, *Rattus norvegicus*, *Schizosaccharomyces pombe*), Munich Information Center for Protein Sequences (http://mips.gsf.de/proj/thal/db/index.html; *Arabiopsis thaliana*), Joint Genome Institute (http://www.jgi.doe.gov/genomes/index.html; *Ciona intestinalis, Fugu rubripes*), Saccharomyces Genome Database (http://www.yeastgenome.org/; *Saccharomyces cerevisiae*), Center for Genomics Research (http://www.broad.mit.edu/annotation/fungi/neurospora/; *N. crassa*), and The Institute for Genomic Research (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/; *Oryza sativa*).

NCBI-BLASTN (Altschul et al. 1990) was carried out locally with standard settings and thresholds ranging from 10^{-4} to $<10^{-50}$. Whole mitochondrial genomes were BLASTed either against draft nuclear genome sequences (human, mouse, rice, and rat) or, in all other cases, against complete genomes.

Numbers for total genome sizes and the amount of intergenic sequences were extracted from the Web pages listed above. Values of non–protein-coding DNA listed in table 1 were extracted from Taft and Mattick (2003).

Web Site

When additional genome sequences become available, updated versions of figure 1 and table 1 will be made available at: http://www.mpiz-koeln.mpg.de/~leister/mbe_2004.html.

Acknowledgments

D.L. is supported by a Heisenberg stipend of the Deutsche Forschungsgemeinschaft (LE 1265/8). We thank Francesco Salamini for valuable comments on the manuscript.

Literature Cited

Allen, J. F. 1993. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. J. Theor. Biol. **165**:609–631.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

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

Bensasson, D., D. A. Petrov, D. X. Zhang, D. L. Hartl, and G. M. Hewitt. 2001b. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. **18**:246–253.

Bensasson, D., D. Zhang, D. L. Hartl, and G. M. Hewitt. 2001a. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. **16**:314–321.

Bensasson, D., D. X. Zhang, and G. M. Hewitt. 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. **17**:406–415.

Devos, K. M., J. K. Brown, and J. L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. **12**:1075–1079.

Divo, A. A., T. G. Geary, J. B. Jensen, and H. Ginsburg. 1985. The mitochondrion of *Plasmodium falciparum* visualized by rhodamine 123 fluorescence. J. Protozool. **32**:442–446.

Hartl, D. L. 2000. Molecular melodies in high and low C. Nat. Rev. Genet. **1**:145–149.

Hazkani-Covo, E., R. Sorek, and D. Graur. 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.

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

Hopkins, J., R. Fowler, S. Krishna, I. Wilson, G. Mitchell, and L. Bannister. 1999. The plastid in *Plasmodium falciparum* asexual blood stages: a three-dimensional ultrastructural analysis. Protist **150**:283–295.

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.

Mourier, T., A. J. Hansen, E. Willerslev, and P. Arctander. 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol. Biol. Evol. **18**:1833–1837.

Petrov, D. A. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. **17**:23–28.

Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. Science **287**:1060–1062.

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

Sunnucks, P., and D. F. Hales. 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. **13**:510–524.

Taft, R. J., and J. S. Mattick. 2003. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biology http://genomebiology.com/2003/5/I/PI.

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

Walbot, V., and M. M. Evans. 2003. Unique features of the plant life cycle and their consequences. Nat. Rev. Genet. **4**:369–379.

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

Zhang, D. X., and G. M. Hewitt. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol. Evol. **11**:247–251.

William Martin, Associate Editor

Accepted January 27, 2004