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; Woischinka 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 (Henz and Martin 2001; Mourier et al. 2001; Woischink and Moraes 2002). Moreover, it is likely that the accumulation of NUMTs is a continuous evolutionary process (Mourier et al. 2001; Woischink 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. 2001a): 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; Woischink 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., 2001a). 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. 1a): 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. 1a).

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 1b. 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. 1b).

Our data extend the previous observations of Bensasson et al. (2001a) 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. 2001a) or mammals as human, mouse, and rat (this study), differ substantially in their NUMT content (table 1 and fig. 1a). 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. 1b 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
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; 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

<table>
<thead>
<tr>
<th>Species</th>
<th>mtDNA Total Size [bps]</th>
<th>Transferred [%]</th>
<th>Nuclear Genome (Mbps)</th>
<th>NUMTs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>Intergenic Regions</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>16,571</td>
<td>98</td>
<td>2910.0</td>
<td>2183.0 (75%)</td>
</tr>
<tr>
<td>M. musculus</td>
<td>16,299</td>
<td>99</td>
<td>2500.0</td>
<td>nd</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>16,300</td>
<td>25</td>
<td>2800.0</td>
<td>nd</td>
</tr>
<tr>
<td>F. rubripes</td>
<td>16,447</td>
<td>34</td>
<td>320.0</td>
<td>212.0 (66%)</td>
</tr>
<tr>
<td>C. elegans</td>
<td>13,794</td>
<td>1</td>
<td>97.0</td>
<td>45.6 (47%)</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>19,517</td>
<td>3</td>
<td>122.7</td>
<td>71.3 (58%)</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>15,363</td>
<td>0</td>
<td>278.2</td>
<td>215.0 (78%)</td>
</tr>
<tr>
<td>C. intestinalis</td>
<td>14,788</td>
<td>74</td>
<td>116.7</td>
<td>nd</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>5,967</td>
<td>&lt;1</td>
<td>22.9</td>
<td>9.5 (42%)</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>85,779</td>
<td>1</td>
<td>12.5</td>
<td>3.6 (29%)</td>
</tr>
<tr>
<td>S. pombe</td>
<td>19,431</td>
<td>8</td>
<td>12.5</td>
<td>5.0 (40%)</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>366,924</td>
<td>45</td>
<td>115.4</td>
<td>64.1 (56%)</td>
</tr>
<tr>
<td>O. sativa</td>
<td>490,520</td>
<td>38</td>
<td>420.0</td>
<td>232.3 (55%)</td>
</tr>
</tbody>
</table>

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.

Nuclei 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; 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**


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**


William Martin, Associate Editor

Accepted January 27, 2004