# NUPTs in Sequenced Eukaryotes and Their Genomic Organization in Relation to NUMTs

*Erik Richly and Dario Leister*

Abteilung für Pflanzenzüchtung und Ertragsphysiologie, and Abteilung für Pflanzenzüchtung und Genetik; Max-Planck-Institut für Züchtungsforschung, Köln, Germany

NUPTs (nuclear plastid DNA) derive from plastid-to-nucleus DNA transfer and exist in various plant species. Experimental data imply that the DNA transfer is an ongoing, highly frequent process, but for the interspecific diversity of NUPTs, no clear explanation exists. Here, an inventory of NUPTs in the four sequenced plastid-bearing species and their genomic organization is presented. Large genomes with a predicted low gene density contain more NUPTs. In *Chlamydomonas* and *Plasmodium*, DNA transfer occurred but was limited, probably because of the presence of only one plastid per cell. In *Arabidopsis* and rice, NUPTs are frequently organized as clusters. Tight clusters can contain both NUPTs and NUMTs (nuclear mitochondrial DNA), indicating that preNUPTs and preNUMTs might have concatamerized before integration. The composition of such a hypothetical preNUPT-preNUMT pool seems to be variable, as implied by substantially different NUPTs:NUMTs ratios in different species. Loose clusters can span several dozens of kbps of nuclear DNA, and they contain markedly more NUPTs or NUMTs than expected from a random genomic distribution of nuclear organellar DNA. The level of sequence similarity between NUPTs/NUMTs and plastid/mitochondrial DNA correlates with the size of the integrant. This implies that original insertions are large and decay over evolutionary time into smaller fragments with diverging sequences. We suggest that tight and loose clusters represent intermediates of this decay process.

## Introduction

In plants, plastid DNA (ptDNA) coexists with homologous sequences present in the nucleus. These sequences are designated "NUPTs" (nuclear ptDNA) and they originate—similar to the case of NUMTs (nuclear mitochondrial DNA)—from the infiltration of nuclear DNA by organellar DNA (reviewed in Timmis et al. [2004]). The generation of NUPTs is likely the result of nonhomologous recombination of nuclear DNA with ptDNA fragments leaking out from plastids. Because in flowering plants extranuclear DNA is usually maternally inherited (Corriveau and Coleman 1987; Zhang, Liu, and Sodmergen 2003), organelle-to-nucleus transfer of DNA should preferentially occur during pollen development when a programmed degeneration of plastids takes place.

NUPTs vary interspecifically in size and copy number, as well as intraspecifically in species such as pea, barley, and wheat (Ayliffe, Scott, and Timmis 1998). Hypervariability, in terms of unique spectra of NUPTs in individual plants or even in different tissues of the same individual, has been noted in spinach and beet (Ayliffe, Scott, and Timmis 1998). This variability could be caused by endopolyploidy with some, possibly heterochromatic, portions of the nucleus being underreplicated. In addition, the cited experiments used methylation-sensitive restriction enzymes to resolve cpDNA from nuclear DNA; therefore, differential methylation cannot be ruled out as the cause of the apparent tissue variation. Alternative explanations are either a high instability of NUPTs or that ptDNA is translocated to the nucleus at a very high frequency (Ayliffe, Scott, and Timmis 1998). The latter explanation is supported by experimental data obtained in tobacco (Huang, Ayliffe, and Timmis 2003; Stegemann et al. 2003), which implies that plastid-to-nucleus DNA transfer is an ongoing, highly frequent process in flowering plants. In the green alga *Chlamydomonas reinhardtii*, however, plastid-to-nucleus transfer could not be detected experimentally (Lister et al. 2003).

NUPTs have been found in almost all dicot and monocot species investigated (Timmis and Scott 1983; Scott and Timmis 1984; Ayliffe, Timmis, and Scott 1988; Pichersky and Tanksley 1988; Dujardin 1990; Pichersky et al. 1991; Ayliffe and Timmis 1992*a*, 1992*b*; Ayliffe, Scott, and Timmis 1998). Their organization in tandem arrays of fragments derived from disparate regions of the plastid DNA, as well as the existence of concatamers of NUMTs and NUPTs, implies that the fragments join together from an intracellular pool of organellar DNA before they integrate into the nuclear genome (Blanchard and Schmidt 1995). Recently, completely or partially sequenced plant genomes have been analyzed for the occurrence of NUPTs. In *A. thaliana*, only relatively few NUPTs were found, totaling between 11 (Arabidopsis Genome Initiative 2000) and 20 kbps (Shahmuradov et al. 2003), and corresponding to less than 16 % of the *A. thaliana* ptDNA. In *Oryza sativa*, Shahmuradov et al. (2003) detected a total of 218 kbps of NUPTs, representing 83% of the plastid chromosome of rice. Two very large NUPTs of 33 (Yuan et al. 2002) and 131 kbps (Rice Chromosome 10 Sequencing Consortium 2003) exist on rice chromosome 10. In the genome of the green alga *C. reinhardtii*, however, NUPTs have been not detected yet (Lister et al. 2003).

Previously, we have described the number and size distribution of NUMTs in 13 eukaryotic genomes (Richly and Leister 2004). In this paper, an inventory of NUPTs in *A. thaliana*, rice, *C. reinhardtii*, and the malaria parasite *Plasmodium falciparum*, and their genomic organization in *A. thaliana* and rice in relation to NUMTs is presented. Causes for the intraspecific diversity of NUPT accumulation, as well as for the evolutionary dynamics of the genomic organization of NUPTs and NUMTs, are discussed.

Key words: gene transfer, genome evolution, mitochondrion, NUMT, NUPT, plastid.

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

*Mol. Biol. Evol.* 21(10):1972–1980. 2004
doi:10.1093/molbev/msh210
Advance Access publication July 14, 2004

**Table 1**
**Sizes of ptDNA, of Nuclear Genomes, and of NUPTs Detected by BlastN at a Threshold of $10^{-4}$**

| | ptDNA | | Nuclear Genome (Mbps) | | | NUPTs | |
|---|---|---|---|---|---|---|---|
| Species | Total Size (bps) | Transferred (%) | Total | Intergenic Regions | Noncoding Regions | bps | $10^{-3}$% |
| *O. sativa* | 134,525 | 99 | 420.0 | 232.3 (55 %) | 336.0 (80 %) | 804,737 | 191.6 |
| *A. thaliana* | 154,478 | 19 | 115.4 | 64.1 (56 %) | 81.9 (71 %) | 35,235 | 30.5 |
| *C. reinhardtii* | 203,828 | 1 | 95.4 | nd | nd | 2,461 | 2.6 |
| *P. falciparum* | 34,682 | <1 | 22.9 | 9.5 (42 %) | 10.8 (47 %) | 125 | 0.5 |

NOTE.—"Transferred" ptDNA refers to the fraction (in %) of ptDNA that gave rise to NUPTs (repeated transfers of the same sequence were not considered). "Noncoding regions" include all nonexon sequences as listed by Taft and Mattick (2003). In the seventh column, all nuclear sequences homologous to ptDNA are included, also when deriving from the same ptDNA sequences. Values in column eight refer to the ratio of NUPTs to the total size of the nuclear genome in the species concerned; nd indicates data not determined.

## Materials and Methods

### Sequence Analyses

Full-length organellar nucleotide sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). Nuclear DNA sequences were obtained from MIPS (http://mips.gsf.de/proj/thal/db/index.html; *Arabidopsis thaliana*), JGI (http://www.jgi.doe.gov/genomes/index.html; *Chlamydomonas reinhardtii*), TIGR (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/; *Oryza sativa* ssp. *japonica*), and PlasmoDB (http://plasmodb.org/; *Plasmodium falciparum*). For the analysis of clusters of NUPTs or NUMTs in rice, only the subfraction of sequences belonging to the completely sequenced chromosomes 1 (Sasaki et al. 2002), 4 (Feng et al. 2002), and 10 (Rice Chromosome 10 Sequencing Consortium 2003) were considered.

NCBI-BlastN (Altschul et al. 1990) was carried out locally with standard settings and thresholds ranging from $10^{-4}$ to less than $10^{-50}$. Whole plastid genomes were Blasted either against draft nuclear genome sequences (rice and *Chlamydomonas*) or against complete chromosomes (*Plasmodium*, *Arabidopsis*, and rice chromosomes 1, 4, and 10).

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).

### Comparison of Random and Actual Genomic Distributions of NUPTs and NUMTs in *A. thaliana*

NUMT identification has been provided previously (Richly and Leister 2004). NUPTs and NUMTs separated by less than 5 kbps of DNA of nonorganellar origin were considered as tight clusters. The distribution of NUPTs and NUMTs in the genome of *A. thaliana* was compared by a randomization test to a random expectation. This was conducted by creating simulated genomes in which the chromosomal locations of NUPTs and NUMTs were randomly reassigned. In each simulated genome, the number and size of NUPTs and NUMTs were the same as in the real genome. The distribution of NUPTs and NUMTs in the actual genome was compared with that of 10,000 simulated genomes.

## Results

### Number and Size of NUPTs in Different Species

For the flowering plants *O. sativa* (Hiratsuka et al. 1989; Feng et al. 2002; Goff et al. 2002; Sasaki et al. 2002; Yu et al. 2002; Rice Chromosome 10 Sequencing Consortium 2003) and *A. thaliana* (Sato et al. 1999; Arabidopsis Genome Initiative 2000), as well as for the green alga *C. reinhardtii* (Maul et al. 2002; http://genome.jgi-psf.org/chlre1/chlre1.home.html), complete or draft sequences of both the plastid and the nuclear DNA are available. In addition, both the nuclear (Gardner et al. 2002) and the apicoplast (a relict plastid [Wilson et al. 1996]) genome of the human malaria parasite *P. falciparum* have been sequenced. Employing BlastN searches with different threshold levels allowed us to identify diverged and/or small NUPTs, as well as conserved and/or long ones (see *Supplementary Material*). The result is that dramatic differences in the content of NUPTs in the different genomes are evident (fig. 1): at a threshold of $10^{-4}$, they range from three in *P. falciparum* to more than 2,000 in *O. sativa*. Contradictory to Lister et al. (2003), who could not detect NUPT sequences in *C. reinhardtii*, we could list 41 NUPTs for this species.

In *O. sativa*, almost all regions of the ptDNA were transferred to the nucleus (table 1), indicating that, in principle, all plastid sequences are transferable, as suggested by Allen (1993). Size distributions of NUPTs are shown in figure 2. The longest NUPTs are present in *O. sativa* and *A. thaliana*, whereas *C. reinhardtii* and *P. falciparum* contain, on average, smaller ones. Because in our analysis each Blast hit was counted individually, integrants of ptDNA interrupted by stretches of genomic DNA without homology to ptDNA (e.g., the 33-kbps NUPT of Yuan et al. [2002] and the 131-kbps one [Rice Chromosome 10 Sequencing Consortium 2003] present on chromosome 10 of rice), were not counted as continuous NUPTs, but as complex loci containing several NUPTs.

When the length of NUPTs was normalized based on the size of the plastid chromosome, rice resulted as the species containing the longest NUPTs: 98.2% of rice NUPTs sized less than 2.5% of the length of the plastid chromosome in this species, but 1.8 % of rice NUPTs sized between 2.5 % and 10 % of the rice plastome. In the other three species, all NUPTs had a size less than or equal to 2.5 % of the size of the respective plastome. In cases,

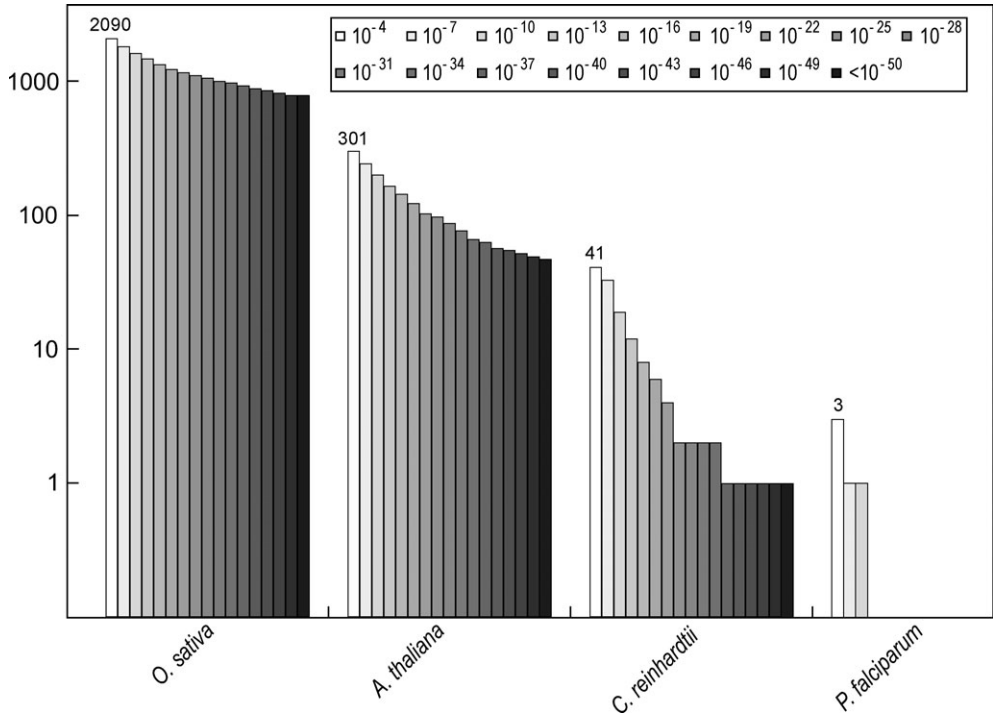


FIG. 1.—Number of NUPTs in different nuclear eukaryotic genomes. The frequency of NUPTs was detected by BlastN with thresholds from $10^{-4}$ to less than $10^{-50}$ (indicated by different shading and presented in logarithmic scale).

the larger NUPTs contained complete open reading frames, resulting into the presence of nuclear copies of two functional plastid genes in *A. thaliana* and of numerous ones in rice.

## Total NUPT Content in Different Genomes

The data presented in this paper extend the observations of Blanchard and Schmidt (1995) concerning the NUPT content across species. The content is highly variable, ranging from 125 bps in *P. falciparum* to more than 800 kbps in rice (table 1). Relative to the nuclear genome size, rice contains the largest fraction of NUPTs (around 0.2%). If the frequency of NUPTs in noncoding genomic regions were similar among species, their number should increase in species with more noncoding nuclear DNA, assuming that transfer of ptDNA fragments in expressed regions of the genome is counterselected. In *A. thaliana* and rice, NUPTs were, in fact, predominantly located in intergenic regions (table 2): only around 25% of NUPTs (and NUMTs) were found within genes, although genes make up 44% and 45%, respectively, of the genomes of *Arabidopsis* and rice (table 1). This bias towards integration in intergenic regions could explain the increase in NUPT abundance following the order *P. falciparum–C. reinhardtii–A. thaliana*–rice (table 1); however, no linear correlation between the size of noncoding regions and the total amount of NUPTs in these species was noted. Furthermore, the size of the plastid chromosome did not correlate with the frequency or size distribution of NUPTs (figs. 1 and 2 and table 1).

## Tight Clusters of NUPTs and NUMTs in *A. thaliana* and Rice

Of the four species investigated, precise physical mapping of DNA sequences, based on the complete sequence of the nuclear genome or of some of its chromosomes, was feasible in *A. thaliana* and in the rice chromosomes 1, 4, and 10. Inspection of the chromosomal location of NUPTs in *A. thaliana*, as well as on the three completely sequenced rice chromosomes, allowed recognition that they often clustered. NUPTs separated by less than 5 kbps were assigned to "tight" clusters. We found that 151, or around 50%, of the *Arabidopsis* NUPTs were organized in such clusters (table 2). In rice, 467 (69%) of the 677 NUPTs located on the three chromosomes analyzed were found in similar clusters, including the previously described 33-kbps and 131-kpbs integrations of ptDNA (Yuan et al. 2002; Rice Chromosome 10 Sequencing Consortium 2003), because they contained also stretches of DNA of nonorganellar origin. Moreover, in both species, NUPTs organized in tight clusters were frequently derived from different regions of the plastid DNA (data not shown), indicating that the major fraction of tightly clustered NUPTs derived from rearranged ptDNA or from concatamers of ptDNA fragments. When NUMT sequences were also included in the analysis, many of the tight clusters were shown to contain both NUMTs and NUPTs (heterogeneous clusters [table 2]). This suggests that either (1) tight clusters derive from large integrants of organellar DNA because of insertions of stretches of nonorganelle DNA and that a pool of both types of fragmented organellar DNA exists in the cell, which can concatamerize before their insertion in the nuclear DNA

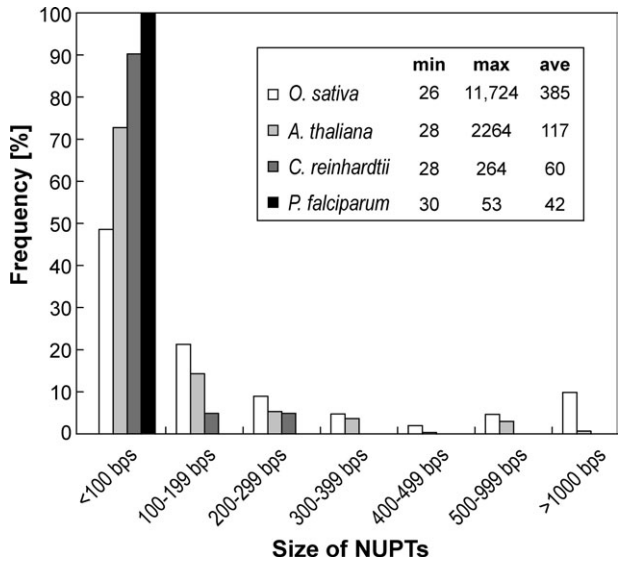


FIG. 2.—Length distribution of NUPTs in genomes of different model species. The distribution of the absolute lengths (in bps) of NUPTs detected at a threshold of $10^{-4}$. In the figure, minimum (min) and maximum (max) NUMT sizes, as well as average values (ave), are reported. Because each Blast hit was counted individually, integrants of ptDNA interrupted by stretches of genomic DNA without homology to ptDNA (e.g., the 33-kbps [Yuan et al. 2002] and 131-kbps [Rice Chromosome 10 Sequencing Consortium 2003] integrants present on chromosome 10 of rice), were not counted as continuous NUPTs, but as tight clusters containing several NUPTs.

or (2) that chromosomal hotspots of organelle-to-nucleus DNA transfer exists in which repeatedly, over evolutionary times, integration events occur. Such hotspots could be caused by specific attributes of the chromosomal region itself (for instance the condensation state) or by additional integrations of similar elements by homologous recombination or other mechanisms.

## Loose Clusters of NUPTs and NUMTs in *A. thaliana* and Rice

NUPTs and NUMTs have been reported to be evenly distributed among chromosomes and within individual chromosomes (Blanchard and Schmidt 1995; Bensasson et al. 2001*a*; Woischnik and Moraes 2002; Hazkani-Covo, Sorek, and Graur 2003). To test this assumption, we compared the genomic organization of NUPTs and NUMTs in *A. thaliana* and on the three completely sequenced rice chromosomes with a hypothetical chromosomal organization that would result from a completely random distribution of NUPTs and NUMTs. To this scope, each existing tight cluster was treated as a single locus, thereby reducing the number of loci considered (*A. thaliana*: 197 NUPT and 349 NUMT loci; rice: 309 NUPT and 277 NUMT loci). In both species, the genomic distribution of NUPTs and NUMTs differed markedly from a random distribution (fig. 3). In particular, NUPTs and NUMTs were more frequent in DNA regions of 100-kbps compared with their random distribution situation. In *Arabidopsis*, even 100 to 200 and 200 to 300 kbps distances were overrepresented (fig. 3).

**Table 2**
**Genomic Distribution of NUPTs and NUMTs in *A. thaliana* and Rice**

| | *Arabidopsis* | | Chromosomes 1, 4, and 10 of Rice | |
|---|---|---|---|---|
| | NUPTs | NUMTs | NUPTs | NUMTs |
| Total number | 301 | 572 | 677 | 566 |
| Genic regions | 79 | 166 | 177 | 138 |
| Intergenic regions | 222 | 406 | 500 | 428 |
| Total number of tight clusters | 47 (151) | 60 (288) | 101 (467) | 80 (367) |
| Homogenous clusters | 37 | 49 | 68 | 47 |
| Heterogeneous clusters | | 10 | | 33 |

NOTE.—In *Arabidopsis*, all five chromosomes were analyzed, and in rice, the chromosomes 1, 4, and 10. "Genic regions" refer to all gene sequences from the start to the stop codon (including introns), whereas "intergenic regions" refer to the size of the rest of the genome (see also table 1). "Tight clusters" refers to NUPTs or NUMTs not separated by more than 5 kbps; numbers in parentheses refer to the total number of NUPTs/NUMTs organized in tight clusters. "Homogenous clusters" include clusters containing either only NUPTs or NUMTs, whereas in "heterogeneous clusters," both types of nuclear organellar DNA are present.

## Evolutionary Relationship of Tight and Loose Clusters

The existence of the "loose" clusters described above can be, in principle, attributed to two possible causes. (1) Hotspots for the integration of organellar DNA in the nuclear genome. (2) Origin from tight clusters caused by the dispersion of NUPTs and NUMTs over a wider chromosomal region by translocation events or by insertion of other, unrelated DNA sequences. If the second scenario is real, tight clusters should have been derived from organelle-to-nucleus DNA transfer more recently than NUMTs or NUPTs organized as single loci or as loose clusters. As a measure of the divergence of NUPTs and NUMTs from their organellar counterparts, their level of sequence identity to plastid or mitochondrial DNA, respectively, was used. We found that the tight cluster of NUMTs making up the complex 620-kbps mtDNA insertion on chromosome 2 of *A. thaliana* (Stupar et al. 2001) was on average 99.0% identical to *Arabidopsis* mtDNA, whereas the genome-wide average of *Arabidopsis* NUMTs was only 95.4%. With respect to the 33-kbps (Yuan et al. 2002) and the 131-kbps ptDNA fragment (Rice Chromosome 10 Sequencing Consortium 2003) on rice chromosome 10, the corresponding tight clusters of NUPTs were, on average, 99.8% and 99.5% identical to rice ptDNA, respectively. Again, the genome-wide average of rice NUPTs was markedly lower (92.2%).

When the NUPTs and NUMTs of rice and *Arabidopsis* were tested for their sequence divergence from the original organellar DNA, it was found that large NUPTs and NUMTs, as well as large clusters, exhibited, in general, less sequence divergence, than small fragments or clusters (figs. 4 and 5). It was noted that, on average, larger NUPTs or NUMTs are less diverged than shorter fragments; however, also in the fraction of less diverged NUPTs or NUMTs (i.e., between 95% and 100% identity to cpDNA or mtDNA) the majority of nuclear organellar DNA is short (<250 bps) (fig. 4). When also tight clusters were considered in this analysis, the less diverged—and therefore relatively recently integrated—NUPTs and

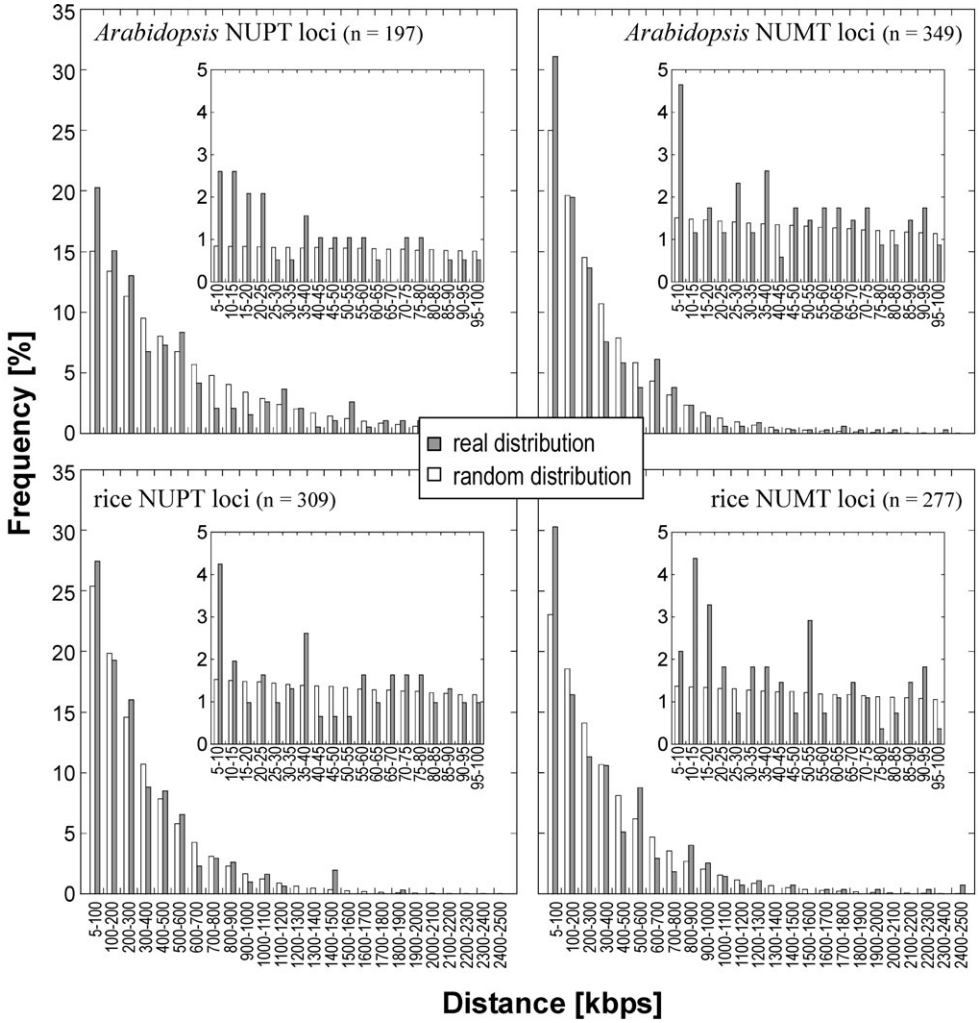


FIG. 3.—Distribution of NUPTs and NUMTs in the genome of *A. thaliana* and in three rice chromosomes. NUPTs (this study) and NUMTs (Richly and Leister 2004) identified by BlastN at a threshold of $10^{-4}$ were considered. NUPTs and NUMTs separated from each other by less than 5 kbps (tight clusters [see table 2]) were treated as one locus. A total of 197 NUPT and 349 NUMT loci in *Arabidopsis*, as well as 309 NUPT and 277 NUMT loci in rice, were analyzed. The actual distribution of insertions of nuclear organellar DNA is displayed as relative frequency of distances between adjacent NUPTs or NUMTs in 100-kbps intervals, ranging from a distance of between 5 and 100 kbps to a distance between 2,400 and 2,500 kbps. In the insets, for the 5-kbps to 100-kbps interval, the relative frequency of distances between adjacent NUPTs or NUMTs is resolved in 10-kbps intervals. The actual distribution of NUPTs or NUMTs was compared with a random expectation by creating 10,000 simulated genomes in which the chromosomal locations of NUPTs or NUMTs were randomly reassigned. In each simulated genome, the number and size of NUPTs and NUMTs were the same as in the real genome.

NUMTs were found to be organized in large pieces of nuclear organellar DNA or tight clusters of it (fig. 5). This suggests that tight clusters are most likely not caused by chromosomal hotspots for organelle DNA integrations, because independent and continuous integration events over evolutionary times are incompatible with the low level of sequence divergence observed for long NUPTs and tight clusters. Moreover, a plausible prediction is that large, tight clusters, such as the ones described above, will evolve over evolutionary time into less compact tight clusters and ultimately into loose clusters. However, it cannot be excluded that, independently of the large integration events, also relatively small pieces of organellar DNA can directly insert into the nuclear genome. A large set of relatively small NUPTs and NUMTs with high homology to ptDNA and mtDNA exists (fig. 4), which could have derived from such small-scale events.

## Discussion

Number and size distribution of NUPTs are different in the four sequenced eukaryotic genomes investigated. The spectrum ranges from three copies, totaling approximately 100 bps, in the malaria parasite *P. falciparum* to 2,090 copies, covering around 800 kbps, in rice. The latter species contains the largest NUPTs (~400 bps), whereas in *P. falciparum* and the green alga *C. reinhardtii*, relatively small NUPTs predominate. To explain this variability in abundance and size distribution of NUPTs, the same reasons as outlined for NUMTs (see Richly and Leister [2004]) can be proposed.

(1) The frequency of DNA transfer from plastids to the nucleus differs between species. The escape of organellar DNA into the cytoplasm, and ultimately its transfer to the nucleus, can be influenced by the vulnerability of the

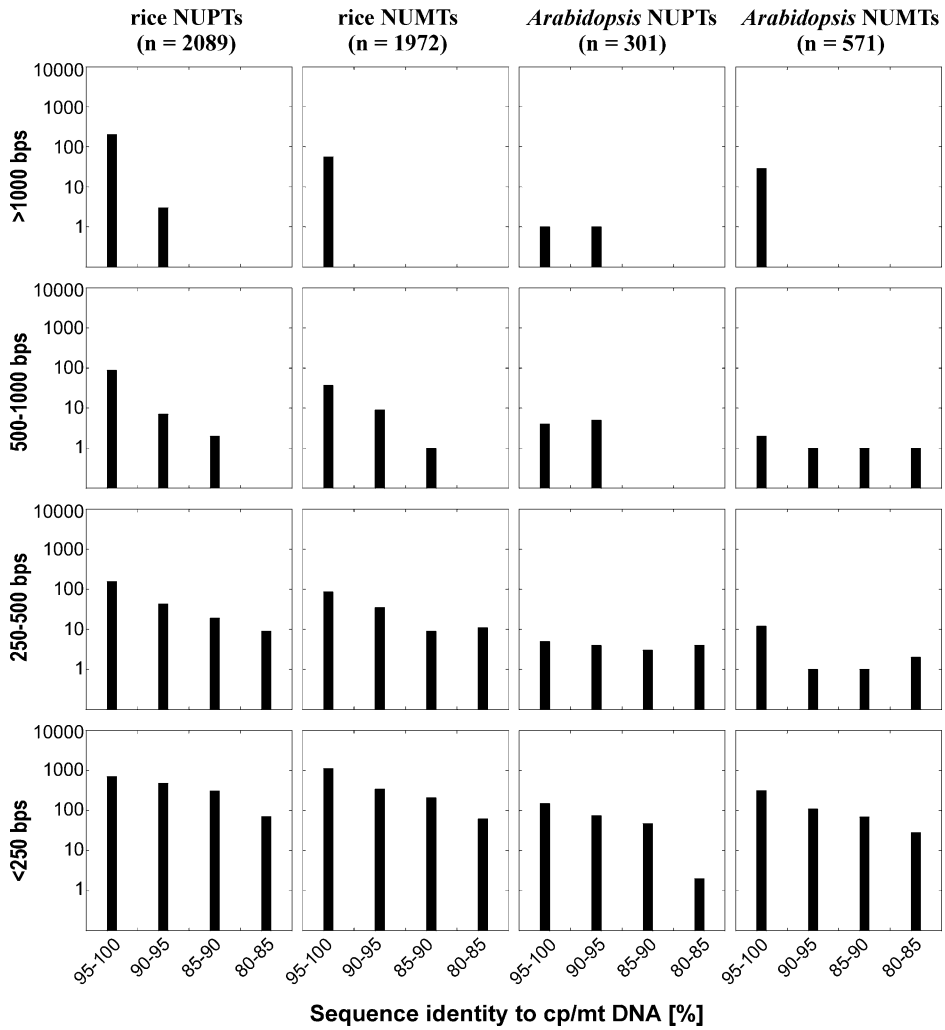


FIG. 4.—Number of *Arabidopsis* and rice NUPTs and NUMTs for four different size classes and for four different classes of sequence divergence (relative to cpDNA and mtDNA). NUPTs and NUMTs were detected by BlastN with a threshold of $10^{-4}$ (see figure 1), and their frequency is presented in logarithmic scale.

organelle to stress and other factors (Bensasson et al. 2001*a*; Woischnik and Moraes 2002), by the copy number of genomes present in each organelle, and the number of the respective organelle present in each cell—particularly of the germline. Accordingly, species-specific differences in the formation of the germline, and/or in the number of plastids per cell, and/or in the programmed degeneration of plastids during pollen development may account for interspecific differences in NUPT abundance. All four species analyzed have a maternal inheritance of ptDNA in common (Coleman 1984; Corriveau and Coleman 1987; Creasey et al. 1994). The number of plastids per cell could, however, explain the low number of NUPTs in *C. reinhardtii* (Lister et al. 2003; this study) and *P. falciparum* (this study)—both are organisms bearing only one plastid per cell (Rochaix 1995; Hopkins et al. 1999). In addition, the efficiency of nuclear import of ptDNA or of its integration into the nuclear genome might differ between species.

(2) The rate of loss of NUPTs is different among species. The rate and spectrum of DNA loss from the nucleus might also shape the accumulation and size pattern of NUPTs. It is well known that the rate of DNA loss from the nucleus varies substantially for different DNA fragment sizes and among species (Petrov et al. 2000; Bensasson et al. 2001*b*; Devos, Brown, and Bennetzen 2002). A specific spectrum of DNA loss could favor the deletion of NUPTs, while still allowing the accumulation of massive amounts of noncoding nuclear DNA elements with different size. This preferential NUPT deletion could lead to genomes with a large fraction of noncoding DNA but only with few NUPTs. Vice versa, a different control on DNA loss would allow more compact genomes to accumulate many NUPTs. This possibly explains the discrepancy in the abundance of NUPTs and NUMTs in the genomes of rice and *A. thaliana*. Whereas in *Arabidopsis*, approximately 200 kbps of NUMT sequences exist (Richly and Leister 2004), the same species contains only 35 kbps of NUPTs (this study). On the contrary, rice contains around 400 kbps of NUMTs (Richly and Leister 2004) but double the amount of NUPTs (this study). This difference might be indicative of a size-specific elimination of noncoding DNA in the two species. *Arabidopsis* NUPTs are relatively small (on average 117 bps) and might be more efficiently deleted than larger fragments,

such as the NUMTs in this species (with an average size of 346 bps [Richly and Leister 2004]). Accordingly, in rice, NUPTs are markedly longer than NUMTs (385 versus 206 bps) and more abundant than NUMTs. If this scenario of a size-dependent filtering of NUPTs and NUMTs in *Arabidopsis* and rice is real, the question is why the size distribution of NUPTs and NUMTs differs a priori in the two species. The size of the organellar genomes cannot be made responsible for this difference: both NUPT and NUMT fragments are very small relative to the complete organelle genome, and in both species, the mitochondrial genome is larger than the respective ptDNA. Moreover, once released into the cytoplasm, the two types of organellar DNA should have the same fate and result into similar types of fragments. However, examples of an independent control of inheritance of mitochondria and plastids in the same species are known (Sodmergen et al. 2002), which indicates that the transfer of plastid and mitochondrial DNA to the nucleus may also occur with a different rate.

In conclusion, no clear explanation exists for the interspecific diversity of NUPTs and NUMTs in copy number and length distribution. It seems possible that large, and often concatamerized, DNA fragments represent original insertions of organelle DNA. In this context, the complex 620-kbps mtDNA insertion on chromosome 2 of *A. thaliana* (Stupar et al. 2001) and the 131-kbps ptDNA fragment on rice chromosome 10 (Rice Chromosome 10 Sequencing Consortium 2003) should be considered as relatively recent events. Interspersion with subsequent insertions of nonorganelle DNA of the original large NUPTs or NUMTs, in combination with local rearrangements, should result in tight clusters, which then ultimately evolve into loose clusters. Similar genetic mechanisms operate in the evolution of the NBS-LRR multigene family in *A. thaliana*, where closely related NBS-LRR genes deriving from tandem duplications are separated by insertions of unrelated DNA, as well as by duplication of NBS-LRR genes to nearby loci, or to other chromosomes (Leister 2004). Our analysis, however, cannot decide whether loose clusters are caused by chromosomal hotspots of organellar DNA integration or by the decay of tight clusters. To clarify this, additional types of evolutionary analyses, such as the one described recently by Hazkani-Covo, Sorek, and Graur (2003), are required to determine the relative ages of the individual NUPTs and NUMTs contained in loose clusters. Tight clusters are, nevertheless, less diverged from organellar DNA than short NUPTs and NUMTs (see figure 5), tempting us to speculate that they are relatively young insertions of organellar DNA that either derive from single insertions of concatamerized cpDNA and mtDNA or from multiple, and more or less simultaneous, insertions in a chromosomal hotspot. Future studies will have to focus on the detailed analysis of these large, recent integrants, which was not the scope of our global, genomic-type analysis.

The now available inventories of NUPTs and NUMTs in *A. thaliana* and rice should allow a systematic characterization of the junctions between concatamerized NUPTs and/or NUMTs, as well as between such elements and nuclear DNA, to deduce information about mechanisms of integration of organellar DNA into nuclear DNA. The analysis of additional plastid-bearing species to be completely sequenced in the future and, in particular, the availability of high-quality chromosomal sequences allowing in silico physical mapping, should shed further light onto the question of how NUPTs and NUMTs are organized in the genome and how they evolve. Moreover, the analysis of angiosperm species with biparental cytoplasmic inheritance should elucidate the role of programmed degeneration of plastids and mitochondria in organelle-to-nucleus DNA transfer. There is considerable nuclear genome sequence available for *Medicago truncatula* (May and Dixon 2004), which inherits plastids (and plastid DNA) biparentally (Lilienfeld 1962). This should, in the near future, make it possible to test whether the anticipated lower rate of breakdowns of the plastome in male gametes of this species also leads to a lower rate of plastid-to-nucleus DNA transfer. In this context, characterization of NUMTs and NUPTs in species that form sperm cells in which only one of the two types of organelle DNA is degraded (such as in *Musella lasiocarpa* [Zhang, Liu, and Sodmergen. 2003]) are of special interest.

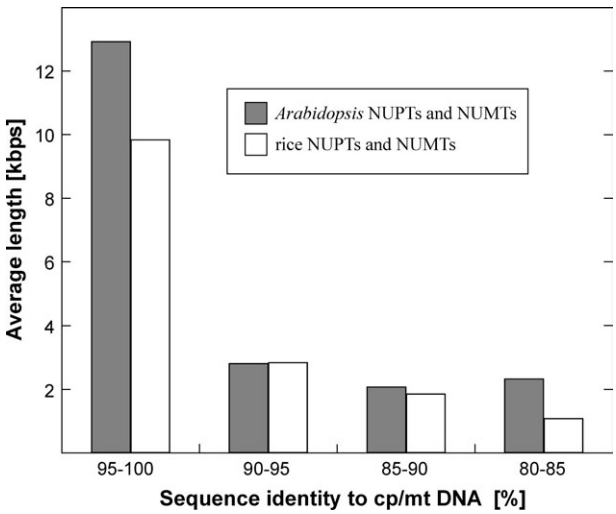


FIG. 5.—Relationship of length and sequence divergence of NUPTs and NUMTs in single and complex loci. Single-locus NUPTs or NUMTs with a size greater than 500 bps, as well as tight clusters containing in total more than 500 bps of NUPTs or NUMTs sequence, were considered. The sequence identity between NUPTs or NUMTs and ptDNA or mtDNA, respectively, was assessed by BlastN analysis. Note that for fragments smaller than 500 bps, no such clear correlation between length and sequence divergence exist, indicating that small organelle DNA fragments might be also transferred to the nucleus independently of the large integrants.

In a state of eukaryote evolution, where almost all transferable genes of plastids and mitochondria have been relocated to the nucleus, future analyses have to show whether present organelle-to-nucleus DNA transfer is a futile mechanism or an important mutagen that drives evolution.

## Supplementary Material

Compilation of all BlastN hits that identify NUPTs in *Arabidopsis*, rice, *Chlamydomonas* and *Plasmodium* is available at the MBE Web site and http://mpiz-koeln.mpg.de/~leister/mbe_2004b.html.

## Acknowledgments

D.L. is supported by a Heisenberg stipend of the Deutsche Forschungsgemeinschaft (LE 1265/8). We thank Francesco Salamini and Nakao Kubo 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.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Ayliffe, M. A., and J. N. Timmis. 1992*a*. Plastid DNA sequence homologies in the tobacco nuclear genome. Mol. Gen. Genet. **236**:105–112.

———. 1992*b*. Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. Theoret. Appl. Genet. **85**: 229–238.

Ayliffe, M. A., N. S. Scott, and J. N. Timmis. 1998. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol. Biol. Evol. **15**:738–745.

Ayliffe, M. A., J. N. Timmis, and N. S. Scott. 1988. Homologies to chloroplast DNA in the nuclear DNA of a number of chenopod species. Theoret. Appl. Genet. **75**:282–285.

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

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

Blanchard, J. L., and G. W. Schmidt. 1995. Pervasive migration of organellar DNA to the nucleus in plants. J. Mol. Evol. **41**: 397–406.

Coleman, A. W. 1984. The fate of chloroplast DNA during cell fusion, zygote maturation and zygote germination in *Chlamydomonas reinhardtii* as revealed by DAPI staining. Exp. Cell Res. **152**:528–540.

Corriveau, J. L., and A. W. Coleman. 1987. Detection and analysis of organelle DNA changes during pollen development. Am. J. Bot. **74**:629–629.

Creasey, A., K. Mendis, J. Carlton, D. Williamson, I. Wilson, and R. Carter. 1994. Maternal inheritance of extrachromosomal DNA in malaria parasites. Mol. Biochem. Parasitol. **65**:95–98.

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.

Dujardin, P. 1990. Homologies to plastid DNA in the nuclear and mitochondrial genomes of potato. Theoret. Appl. Genet. **79**: 807–812.

Feng, Q., Y. Zhang, P. Hao et al. (74 co-authors). 2002. Sequence and analysis of rice chromosome 4. Nature **420**:316–320.

Gardner, M. J., N. Hall, E. Fung et al. (45 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature **419**:498–511.

Goff, S. A., D. Ricke, T. H. Lan et al. (55 co-authors). 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296**:92–100.

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.

Hiratsuka, J., H. Shimada, R. Whittier et al. (16 co-authors). 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome—intermolecular recombination between distinct transfer-RNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. **217**:185–194.

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.

Huang, C.Y., M. A. Ayliffe, and J. N. Timmis. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature **422**:72–76.

Leister, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet. **20**:116–122.

Lilienfeld, F. A. 1962. Plastid behavior in reciprocally different crosses between two races of *Medicago truncatula* Gaertn. Seiken Ziho **13**:3–38.

Lister, D. L., J. M. Bateman, S. Purton, and C. J. Howe. 2003. DNA transfer from chloroplast to nucleus is much rarer in *Chlamydomonas* than in tobacco. Gene **316**:33–38.

Maul, J. E., J. W. Lilly, L. Cui, C. W. dePamphilis, W. Miller, E. H. Harris, and D. B. Stern. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell **14**:2659–2679.

May, G. D., and R. A. Dixon. 2004. *Medicago truncatula*. Curr Biol. **14**:R180–181.

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.

Pichersky, E., J. M. Logsdon Jr, J. M. McGrath, and R. A. Stasys. 1991. Fragments of plastid DNA in the nuclear genome of tomato: prevalence, chromosomal location, and possible mechanism of integration. Mol. Gen. Genet. **225**:453–458.

Pichersky, E., and S. D. Tanksley. 1988. Chloroplast DNA sequences integrated into an intron of a tomato nuclear gene. Mol. Gen. Genet. **215**:65–68.

Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300**:1566–1569.

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

Rochaix, J. D. 1995. *Chlamydomonas reinhardtii* as the photosynthetic yeast. Annu. Rev. Genet. **29**:209–230.

Sasaki, T., T. Matsumoto, K. Yamamoto et al. (80 co-authors). 2002. The genome sequence and structure of rice chromosome 1. Nature **420**:312–316.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res. **6**:283–290.

Scott, N. S., and J. N. Timmis. 1984. Homologies between nuclear and plastid DNA in spinach. Theoret. Appl. Genet. **67**:279–288.

Shahmuradov, I. A., Y. Y. Akbarova, V. V. Solovyev, and J. A. Aliyev. 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol. Biol. **52**: 923–934.

Sodmergen, Q. Zhang, Y. Zhang, W. Sakamoto, and T. Kuroiwa. 2002. Reduction in amounts of mitochondrial DNA in the sperm cells as a mechanism for maternal inheritance in *Hordeum vulgare*. Planta **216**:235–244.

Stegemann, S., S. Hartmann, S. Ruf, and R. Bock. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc. Natl. Acad. Sci. USA **100**:8828–8833.

Stupar, R. M., J. W. Lilly, C. D. Town, Z. Cheng, S. Kaul, C. R. Buell, and J. Jiang. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc. Natl. Acad. Sci. USA **98**:5099–5103.

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 Biol. http://genomebiology.com/2003/5/I/PI.

Timmis, J. N., M. A. Ayliffe, C. Y. Huang, and W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. **5**:123–135.

Timmis, J. N., and N. S. Scott. 1983. Sequence homology between spinach nuclear and chloroplast genomes. Nature **305**:65–67.

Wilson, R. J., P. W. Denny, P. R. Preiser et al. (11 co-authors). 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. **261**:155–172.

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

Yu, J., S. Hu, J. Wang et al. (100 co-authors). 2002. A draft sequence of the rice genome *Oryza sativa* L. ssp. *indica*. Science **296**:79–92.

Yuan, Q., J. Hill, J. Hsiao, K. Moffat, S. Ouyang, Z. Cheng, J. Jiang, and C. R. Buell. 2002. Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. Mol. Genet. Genomics **267**:713–720.

Zhang, Q., Y. Liu, and Sodmergen. 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. Plant Cell Physiol. **44**:941–951.

William Martin, Associate Editor

Accepted July 9, 2004