Heredity (2013) 111, 314–320

www.nature.com/hdy

ORIGINAL ARTICLE

Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization

M Michalovova[1,2], B Vyskot[1] and E Kejnovsky[1,2]

We analysed the size, relative age and chromosomal localization of nuclear sequences of plastid and mitochondrial origin (NUPTs-nuclear plastid DNA and NUMTs-nuclear mitochondrial DNA) in six completely sequenced plant species. We found that the largest insertions showed lower divergence from organelle DNA than shorter insertions in all species, indicating their recent origin. The largest NUPT and NUMT insertions were localized in the vicinity of the centromeres in the small genomes of *Arabidopsis* and rice. They were also present in other chromosomal regions in the large genomes of soybean and maize. Localization of NUPTs and NUMTs correlated positively with distribution of transposable elements (TEs) in *Arabidopsis* and sorghum, negatively in grapevine and soybean, and did not correlate in rice or maize. We propose a model where new plastid and mitochondrial DNA sequences are inserted close to centromeres and are later fragmented by TE insertions and reshuffled away from the centromere or removed by ectopic recombination. The mode and tempo of TE dynamism determines the turnover of NUPTs and NUMTs resulting in their species-specific chromosomal distributions.
Heredity (2013) **111,** 314–320; doi:10.1038/hdy.2013.51; published online 29 May 2013

Keywords: promiscuous DNA; NUPT; NUMT; chromosome; plant

INTRODUCTION

Plant cells contain three genomes: the nuclear genome and the genomes of mitochondria and plastid. During evolution, organellar genomes have been partly reduced in size and DNA sequence fragments have been transferred to nucleus (Bensasson *et al.,* 2001; Selosse *et al.,* 2001). Some DNA sequences, referred to as 'promiscuous DNA' are still moving between these three genomes (Ellis, 1982). The most frequent DNA transfer is from organelles to nucleus (Lewin, 1984; Martin and Herrmann, 1998). After integration into the nucleus, organellar sequences are called NUPTs (nuclear plastid DNA) or NUMTs (nuclear mitochondrial DNA). NUPTs and NUMTs have been found in almost all plant species studied (Ayliffe *et al.*, 1998; Hazkani-Covo *et al.,* 2010). About 18% of *Arabidopsis* genes were acquired from a cyanobacterial ancestor of plastid (Martin *et al.,* 2002). A systematic analysis of NUPTs that assessed their age, size, structure and chromosomal localization has been conducted only in rice (Matsuo *et al.,* 2005; Guo *et al.*, 2008) and the *Arabidopsis* genome (Shahmuradov *et al.,* 2003). In rice, large NUPTs were found preferentially in pericentromeric regions of most chromosomes (Matsuo *et al.,* 2005), but general conclusions about their localization, length and relative age could not be drawn because a systematic study of NUPTs in other species had not been done yet.

There are still debates about the mechanism of DNA transfer from organelles to nucleus and the mode of integration (Leister, 2005; Kleine *et al.,* 2009). It was suggested that the insertion mechanism involves double-strand break repair via NHEJ (Hazkani-Covo and Covo, 2008). The availability of whole-genome sequences of more species has enabled a more detailed study of genomic structure of NUPTs and NUMTs. Early analysis of only four species detected no correlation between the abundance of NUMTs and the size of nuclear or mitochondrial genomes (Richly and Leister, 2004a). However, recent systematic studies of dozens of species showed that the number and cumulative length of NUPTs and NUMTs correlate with both genome size and the number of organelles in cell (Hazkani-Covo *et al.,* 2010; Smith *et al.,* 2011). The least diverged NUPTs and NUMTs are the largest, which implies that original insertions were large and have decayed over time (Yuan *et al.,* 2002; Richly and Leister, 2004b; Huang *et al.,* 2005). However, many NUMTs probably originate not by de novo insertions but by post-insertion duplications (Hazkani-Covo *et al.,* 2003). The majority of large NUPTs in rice seem to have been eliminated from the nuclear genome within 1 million years (Matsuo *et al.,* 2005), indicating that the turnover and accumulation of NUPTs and NUMTs is influenced by forces controlling the expansion and contraction of nuclear DNA. The great variation in chromosomal localization of NUPTs between maize inbred lines also suggests that plastid insertions in nucleus are frequent and recent (Roark *et al.,* 2010). In addition to large NUPTs and NUMTs, short recent insertions were also found in plant nuclear genomes. Some of these short sequences were organized as clusters, which might indicate that they were concatemerized before integration (Richly and Leister, 2004b; Noutsos *et al.,* 2005). Although occasionally, NUPTs and NUMTs become functional genes in the nucleus (Stegemann and Bock, 2006), the majority of organellar DNA

[1]Department of Plant Developmental Genetics, Institute of Biophysics ASCR, Brno, Czech Republic and [2]Laboratory of Genome Dynamics, CEITEC—Central European Institute of Technology, Masaryk University, Brno, Czech Republic
Correspondence: Professor E Kejnovsky, Laboratory of Plant Developmental Genetics, Institute of Biophysics ASCR, Kralovopolska 135, CZ 612 65, Brno, Czech Republic.
E-mail: kejnovsk@ibp.cz
Received 19 September 2012; revised 26 March 2013; accepted 15 April 2013; published online 29 May 2013

insertions in nucleus is mutated, partially deleted and expanded by insertions of nuclear DNA (Noutsos *et al.,* 2005).

Here we show, based on an analysis of six sequenced plant species, that the largest NUPT and NUMT insertions are the most recent. They are localized preferentially at the centromeres in species with small genomes (*Arabidopsis* and rice), but they are more fragmented and localized along whole chromosomes in species with larger genomes (soybean and maize). We suggest that the dynamism of transposable elements (TEs) together with recombination rates are responsible for the turnover and distribution patterns of promiscuous DNA.

MATERIALS AND METHODS

The sources of sequence data of the analyzed plant genomes are presented in Supplementary Table S1. Organellar insertions were identified using BLAST 2.2.25 (Altshul *et al.,* 1990) with an e-value threshold set to 0.01, filtering switched off (-dust no), a mismatch penalty of -2 and a word size of 9. All hits shorter than 100 bp were removed in order to eliminate repetitive sequences not associated with the true insertions. NUPTs and NUMTs were in all figures and supplements calculated as follows: every two insertions identified by BLAST that fulfilled the three criteria were considered to be part of one insertion event and the length of this merged insertion site was calculated as the merged length of the individual organellar insertions. The criteria were: (i) the distance between insertions must be less than the maximum distance within the nucleus, which was set to 0.03% of the chromosome length, (ii) the maximum distance within the plastid genome must be $<3\%$ of the organellar genome length, (iii) the difference in nucleotide divergence between insertions must be $<20\%$ (assumes that a greater difference in divergence points to two insertion events instead of one). The orientation of neighboring insertions was not taken into account, allowing merging of all insertions that fulfilled our criteria and not only those that preserved the orientation of cpDNA in the nucleus. These merged insertions were then considered as one NUPT or NUMT. Their genomic coordinates are listed in Supplementary Table S2.

In all figures the sequence divergence was assessed as nucleotide divergence of NUPT/NUMT from their organellar counterparts. For illustrative purposes, only insertions longer than 100 bp, 400 bp, 300 bp and 500 bp for NUPTs and 100 bp, 400 bp, 300 bp and 300 bp for NUMTs were visualized for *A. thaliana*, *Z. mays*, *O. sativa* and *G. max,* respectively (Figure 2).

When the distance of two organellar sequences homologous to NUPTs/NUMTs was calculated in organellar genomes, the shorter distance from two possibilities was taken into account due to the circular character of genomes. In case of overlaps, NUPTs/NUMTs were evaluated as distant. Exact coordinates of NUPTs homologous to IR regions could not be determined because of their repetitive character and were treated as they would have matched the first IR region.

Data processing was performed using C#, which is part of Microsoft.NET Framework and Perl and subsequently visualized using gnuplot, a command-line driven graphing utility available from http://www.gnuplot.info/. Transposon density was calculated using a sliding window of length 50 000 with RepeatMasker 3.3.0 (www.repeatmasker.org) outputs at divergence 24 with the exception of maize for which the sliding window was set to 100 000.

RESULTS

The size, relative age and chromosomal localization of NUPTs and NUMTs

We analysed plastid and mitochondrial DNA sequences localized in nucleus (NUPTs and NUMTs) in six completely sequenced plant species—*Arabidopsis* (*A. thaliana*), rice (*O. sativa*), grapevine (*Vitis vinifera*), sorghum (*Sorghum bicolor*), soybean (*G. max*) and maize (*Z. mays*). We focused on the size, chromosomal localization and relative age of NUPTs and NUMTs. In Figure 1, the NUPTs and NUMTs sizes correspond to the size of the vertical line and the color of the line indicates the nucleotide divergence of NUPTs from cpDNA or NUMTs from mtDNA. In all five chromosomes of *Arabidopsis,* NUPTs and NUMTs were localized close to centromere (up to a maximum distance of 6 Mbp from centromere). The NUPTs in *Arabidopsis* were mostly short, up to 600 bp in most cases (Figure 1). The maximal NUPT length was about 5 kb and was found on chromosome 3 (Figure 1c). The frequency of NUMTs was higher than NUPTs with evidence of accumulation around the centromeres of all chromosomes. Although the majority of NUMTs reached a maximal length of 2 kb, 16 longer NUMTs gathered around the centromere of chromosome 2. Collectively, the total length of these longer insertions was more than 286 kb. The divergence of the largest NUMTs from mtDNA was very low (maximum 4%) suggesting their recent origin, but many other NUMTs exhibited divergence values of about 20% (green lines in Figure 1).

The NUPTs in rice were often longer than in *Arabidopsis* (Supplementary Figure S1). In total, 151 kb of organellar DNA was found on chromosome 10 in a cluster consisting of 5 insertions. A 143 kb cluster consisting of an assemblage of 14 insertions was found on chromosome 4. Again, the largest NUPTs were localized mostly close to pericentromeres and were of recent origin (divergence from cpDNA max 4%, Supplementary Figure S1). The length of NUMTs was similar to the length of NUPTs, and the largest NUMTs were most recent. The largest NUMTs were found on chromosome 12 where a cluster of 21 smaller insertions with a cumulative length of 223 kb was located (Supplementary Figure S1). Similar clusters of NUMTs were found on chromosome 1 (totaling about 28 kb), chromosome 4 (totaling 23 kb), chromosome 6 (totaling 25 kb) and chromosome 9 (almost 43 kb).

In grapevine, the majority of NUPT and NUMT insertions were short (up to 600 bp) with some examples reaching 13 kb in NUPTs and cluster of 21 kb in NUMTs (Supplementary Figure S2). In sorghum, all insertions were highly diverged (20% and more, green lines in Supplementary Figure S3), quite long (frequently up to 5 kb), and distributed mainly around regions of putative centromeres. The NUPTs in soybean were mostly short (up to 600 bp), with maximal insertion length of 10 kb (Supplementary Figure S4). The abundance of short (up to 600 bp) NUPTs was diminished in the centromeres in the majority of chromosomes (Supplementary Figure S4), but the largest insertions were spread along whole chromosomes with only few present in centromeric regions. The divergence of the largest NUPTs from cpDNA was up to 15% (Supplementary Figure S4). In maize the majority of NUPTs and NUMTs were shorter than 10 kb (Supplementary Figure S5). However, many insertions were large, reaching lengths of 34 kb (chromosome 2). Similarly, some NUMTs reached a length of 108 kb (chromosome 1). A cluster of large NUMTs in the centromeric region of chromosome 9 totaled 225 kb in length (Supplementary Figure S5). All large insertions were recent, with a maximum 4% divergence from cpDNA or mtDNA.

These results showed that patterns of both chromosomal distribution and relative age in NUPTs and NUMTs were similar in rice (Supplementary Figure S1), grapevine (Supplementary Figure S2) and maize (Supplementary Figure S5) having similar density of insertions occurrence, similar divergence and length range. In contrast, these patterns differed between NUPTs and NUMTs in *Arabidopsis* (Figure 1) and sorghum (Supplementary Figure S3) having higher number of NUMTs that were more diverged and lower number of NUPTs that were relatively recent. Taken together, these data demonstrate that mode of turnover could be specific for each type of organellar DNA sequences in these species.

In order to visualize data from all chromosomes of one species clearly in one picture (Figure 2), we decided to represent the length of NUPTs and NUMTs by the area of circles. These are plotted against

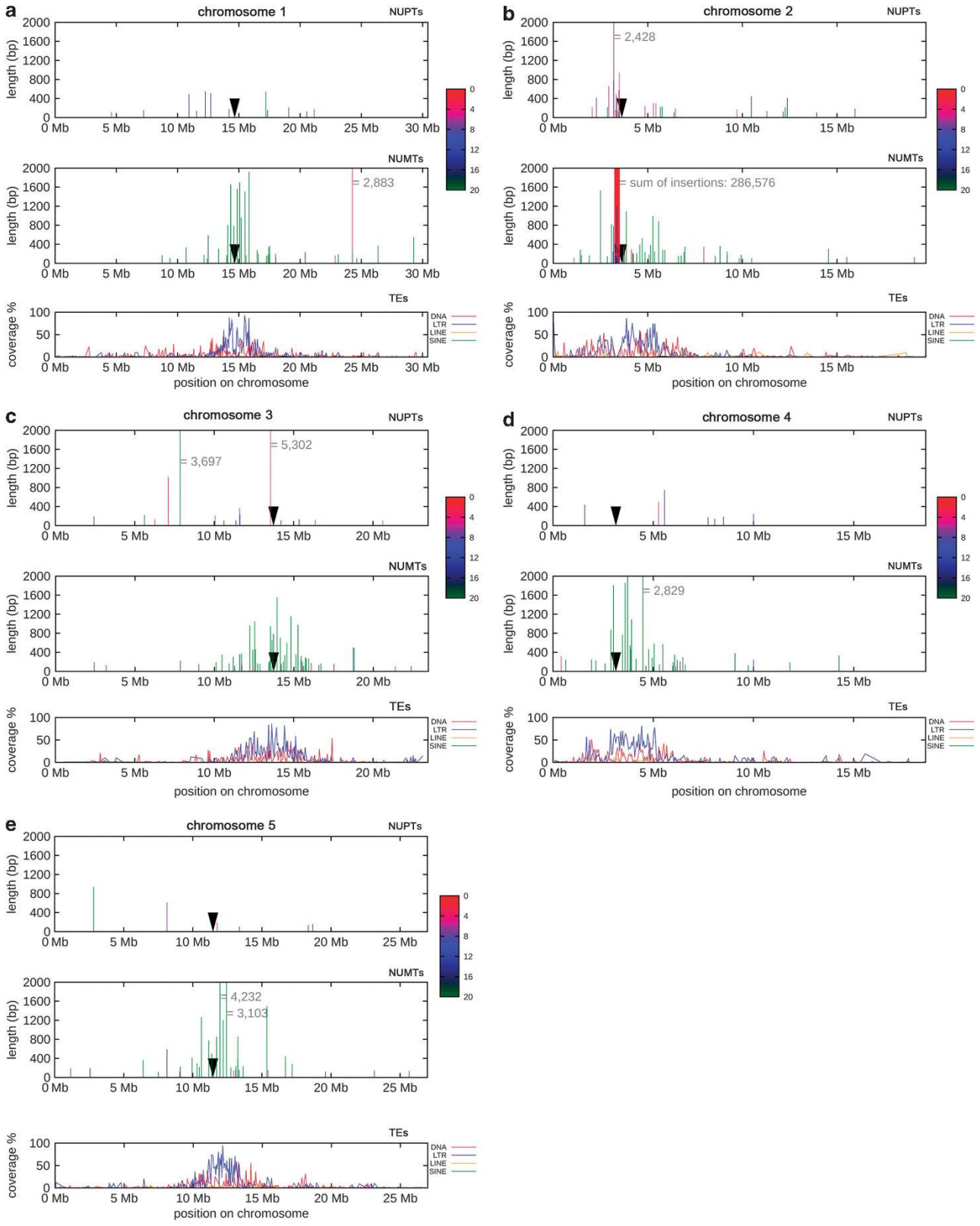

Figure 1 Distribution of NUPTs and NUMTs on five chromosomes of *Arabidopsis*. The length of vertical lines indicates the length of the insertion, the color indicates the nucleotide divergence from organellar DNA in percentage. The coverage of the chromosome by different TE families measured as the proportion of TEs in sliding window is visualized for each chromosome. Individual chromosomes are in separate panels (**a**–**e**).

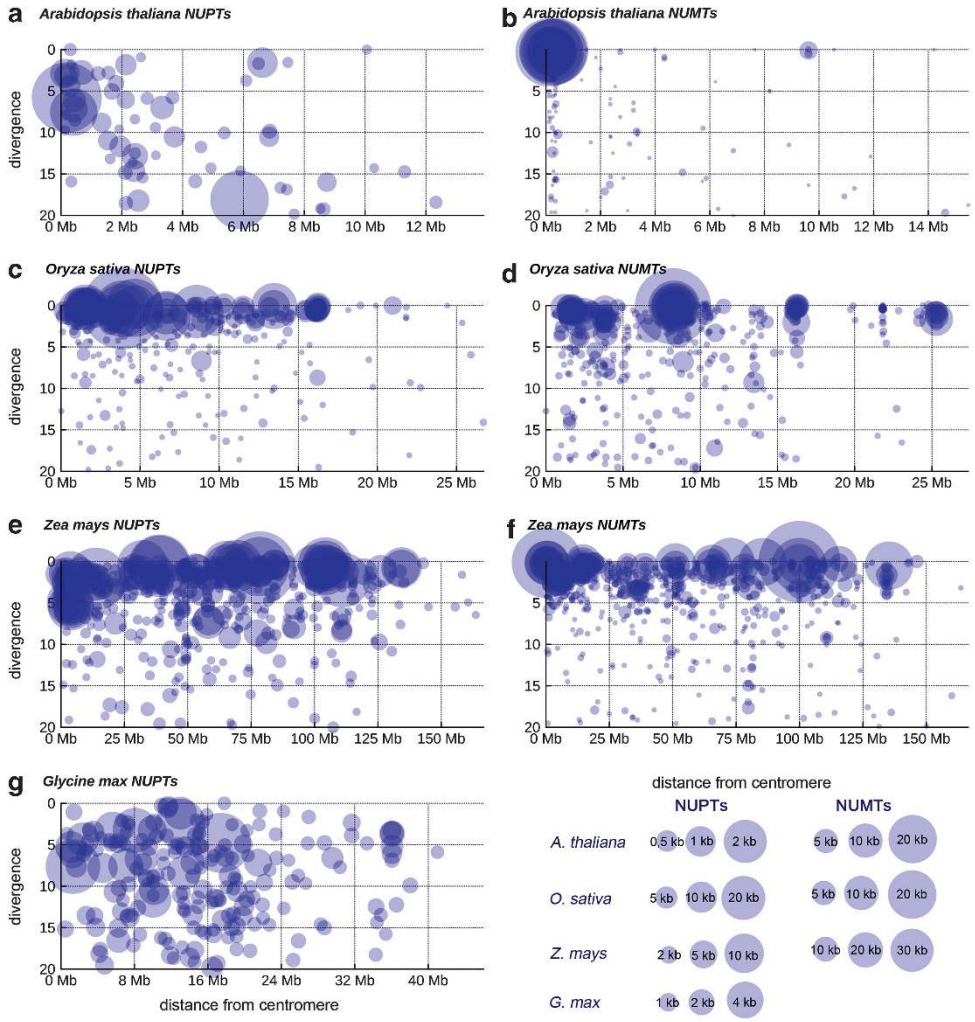

Figure 2 The size, age and chromosomal localization of promiscuous DNA. The size of NUPTs and NUMTs, their distance from the centromere and nucleotide divergence from organellar DNA in *Arabidopsis* (**a** and **b**), rice (**c** and **d**), maize (**e** and **f**) and soybean (**g**). The area of the circles corresponds to the size of NUPT/NUMT, but the scale is different for each species.

the divergence from organelle DNA (Y-axis) and distance from the centromere (X-axis). This figure represents another mode of visualization and contains only species for which the position of the centromere is known and the distance of NUPTs or NUMTs from the centromere could be measured—that is, *Arabidopsis*, rice, maize and soybean. The mtDNA sequence in soybean is not available and therefore the analysis of NUMTs in this species is not included. Again, from this representation, it is evident that the largest insertions exhibit the lowest divergence from organellar DNA, suggesting their recent insertion. The divergence (relative age) of NUPTs was higher only in soybean (Figure 2). The largest NUPTs, as well as NUMTs were often localized in the vicinity of centromeres in the small genomes of *Arabidopsis* and rice, but were localized both in pericentromeres and in other regions of chromosomes in the large genomes of soybean and maize (Figure 2).

We calculated the total length of all NUPTs and NUMTs in genomes in relation to the genome size (Figure 3). We found that the sum of insertions correlates with genome size but that the genomic proportion (size of circle in Figure 3) represented by promiscuous DNA did not. The length of the largest individual NUPT/NUMT insertions in specific species did not correlate with its genome size. The large NUPTs were found in the small genome of rice (up to 151 kb) and in the large genome of maize (up to 61 kb). The same was true for NUMTs; the large NUMTs were found in *Arabidopsis* (286 kb), rice (223 kb) and maize (225 kb), while the longest NUMTs in sorghum were only 11 kb long. These patterns indicate that the number of insertions rather than their length composes the cumulative length of promiscuous DNA in genome and that insertions are more fragmented in larger genomes.

In summary, we found that the largest organellar insertions showed lower divergence from cpDNA or mtDNA than smaller insertions in all six studied species indicating that larger insertions are of more recent origin. Second, we revealed a correlation of total NUPTs/NUMTs with genome size but no correlation between the genome proportion formed by NUPTs/NUMTs and genome size. Third, we found that centromeres are preferential niches for promiscuous DNA localization, especially in species with small genomes.

Correlation of NUPTs/NUMTs with the distribution of TEs

We studied the relationship between the number of NUPT/NUMT insertions and the distribution of TEs. We visualized coverage of individual chromosomes with TEs, distinguishing between various

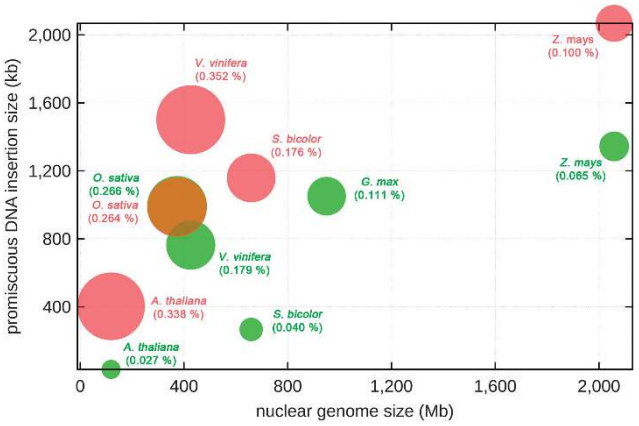

Figure 3 Promiscuous DNA and genome size. The promiscuous DNA size (cumulative length of NUPTs or NUMTs) plotted against nuclear genome size. NUPTs are in green, NUMTs are in red. The size of the circle corresponds to the proportion (indicated in percent, %) of promiscuous DNA in the nuclear genome of specific species.

types of TE families (Figures 1 and 4, Supplementary Figures S1–S5). The contrasting patterns of NUPT/NUMT insertions and TE distributions of the six studied species are summarized in Figure 4. In this Figure, we show typical representatives of each of the studied plant species. All chromosomes can be examined in Supplementary Material (Supplementary Figure S1–S5). We found that TEs were specifically localized or enriched in centromeres and pericentromeres in *Arabidopsis* and soybean (Figures 1, 4a and e, Supplementary Figure S4). Localization of NUPTs and NUMTs correlated positively with the distribution of TEs in *Arabidopsis* and sorghum (Figures 4a and d) but negatively in grapevine and soybean (Figures 4c and e). No correlations between the localization of NUPTs/NUMTs and TE were observed in rice and maize (Figures 4b and f). These results show that similar patterns of promiscuous DNA distribution were observed in species differing in genome size and in species with different TE content—for example, in rice (20–40% coverage, Figure 4b) or maize (70% coverage, Figure 4f). It appears that the distribution pattern of promiscuous DNA does not depend on the abundance or localization of TEs in genome but rather reflects the dynamism of TEs.

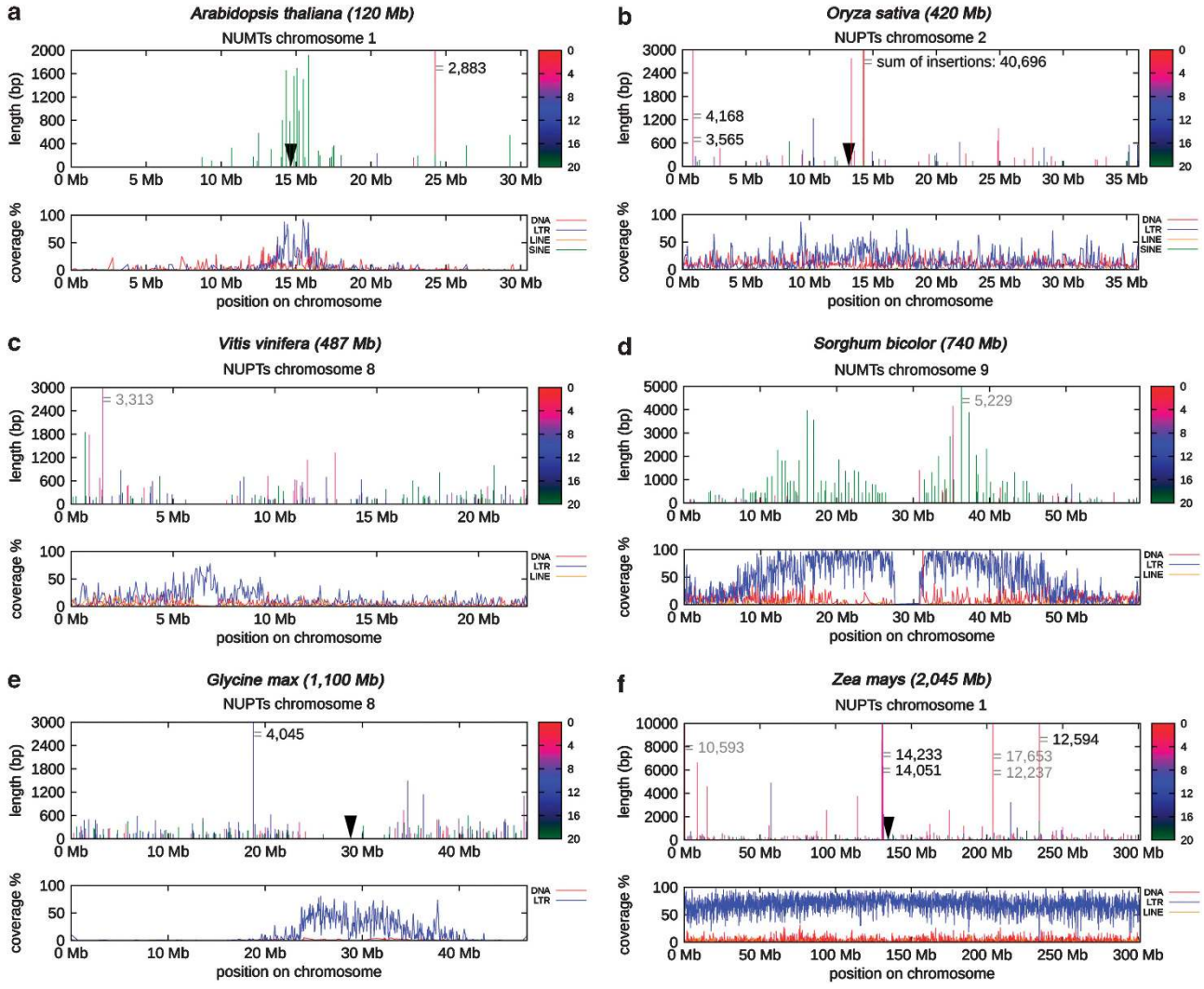

Figure 4 Contrasting patterns of promiscuous DNA and TEs localization. Distribution, size and nucleotide identity (with organellar DNA) of NUPTs or NUMTs along chromosome 1 of *Arabidopsis* (**a**), chromosome 2 of rice (**b**), chromosome 8 of grapevine (**c**), chromosome 9 of sorghum (**d**), chromosome 8 of soybean (**e**) and chromosome 1 of maize (**f**). In parallel, the abundance of TEs along the chromosomes is marked. Genome sizes of analysed species are indicated.

A Model of NUPT/NUMT and TE dynamics

Our finding that the largest NUPTs showed the lowest divergence from cpDNA indicates that they are most recent. Therefore, our data support the view that bulk DNA is transferred from the plastid or mitochondria to the nucleus (Henze and Martin, 2001). We found that older (more diverged) NUPTs/NUMTs were shorter, suggesting that some fragmentation processes are in action.

We propose a model (Figure 5) where new plastid and mitochondria DNA sequences are inserted in the vicinity of centromeres, later are fragmented by TE insertions and, finally, partially or completely removed by ectopic recombination between homologous regions, such as provided by TEs (Figure 5a). The mode and tempo of TE dynamism, which could be species-specific, determines the turnover of NUPTs/NUMTs and leads to their specific chromosomal distribution. As a result NUPTs and NUMTs are found in centromeres and pericentromeres in species with small genomes like *Arabidopsis* and rice that have a low abundance of TEs and probably lower genome dynamism, that is, lower activity of TEs and lower recombination rate (Figures 2a–d). In contrast, NUPTs or NUMTs are also present in other chromosomal regions in plants with large genomes and higher dynamism. This is especially obvious in maize, which has the largest genome analysed in this study (Figures 2e and f).

In order to test the validity of our model we measured the distance of all neighboring NUPTs or NUMTs in both nuclear and organellar genomes in maize (Supplementary Figure S6). This species was chosen because of its large genome size and high proportion of TEs and we therefore expected that the fragmentation of NUPTs/NUMTs by TEs, if present, should be visible. We restricted our analysis to neighboring organellar insertions up to 100 kb mutual distance what in our opinion best reflects the situations of recent fragmentations.

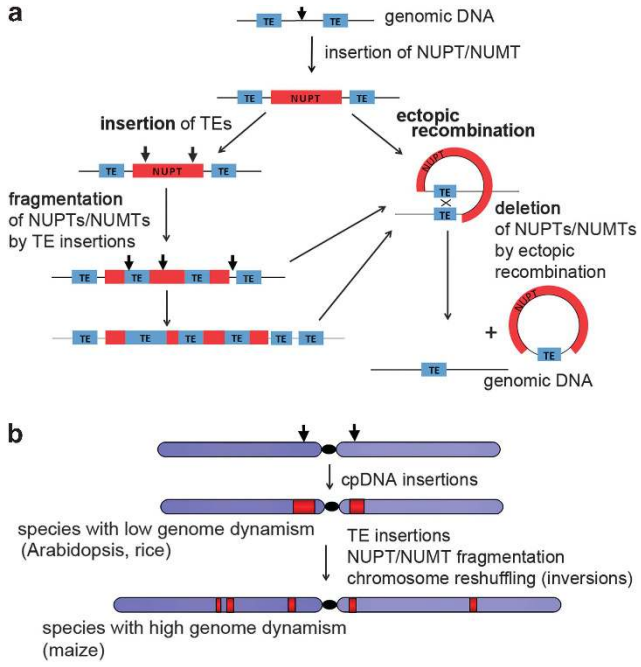

Figure 5 Model of dynamism of promiscuous DNA and TEs. (**a**) Insertions of TEs lead to fragmentation of NUPTs/NUMTs. Older NUPTs/NUMTs accumulate mutations causing their divergence from cpDNA/mtDNA. Ectopic recombination between homologous TEs results in deletions of intervening regions. (**b**) Both processes of TE insertion and chromosome reshuffling result in NUPT/NUMT fragmentation and their movement away from centromeric regions.

We calculated only situations when intervening sequences between two neighboring NUPTs or NUMTs was formed by TEs (at least 90% coverage by TEs) and we took into account only NUPTs or NUMTs of the same orientation and maximal 5% mutual divergence (Supplementary Figures S6A, E). Under these conditions neighboring NUPTs or NUMTs represented originally continual organellar sequence that was interrupted by TEs. We found that majority of neighboring organellar insertions showed no or minimal distance in organellar genome irrespective of their distance in nuclear genome (abundant crosses distributed along horizontal axis). These results clearly support our model of NUPT and NUMT fragmentation by TEs. Results of modified calculation where orientation of neighboring NUPTs or NUMTs was not considered and their mutual divergence was allowed to be higher (10%) are also summarized in Supplementary Figure S6.

DISCUSSION

NUPTs and NUMTs have been studied in plants for many years and a large number of cases have been characterized, especially in *Arabidopsis* and rice where genome-wide identification studies of NUPTs and NUMTs have been carried out (Shahmuradov *et al.*, 2003; Matsuo *et al.*, 2005). In this study, we analysed size, divergence from organellar genome and chromosomal localization of NUPTs and NUMTs in six plant species in the largest survey to date. This enabled us to look for patterns of NUMT and NUPT evolution on a larger scale. Our results confirm the findings of previous genome-wide surveys of NUPTs and NUMTs demonstrating that NUPTs as well as NUMTs are very common, despite the pattern of variations in the species studied.

We found that the numbers of NUPTs and their divergence from organellar sequence were similar to the number of NUMTs and their divergence in rice, grapevine and maize. In contrast, NUMTs were more frequent and diverged more from organellar genome than NUPTs in *Arabidopsis* and sorghum. This suggests that the number of insertions and their age do not correlate with the genome size of the species. However, we observed differences in localization of organellar insertions between species having small and those having large genomes—NUPTs and NUMTs were gathered close to centromeres in small genomes while they were present along whole chromosomes in larger genomes. The novelty of our report consists in proposing a model of organellar DNA turnover in nucleus where fragmentation of originally long organellar insertions by TEs, represents an important process. This fragmentation is more obvious in large genomes, like maize, probably because TEs represent high-genomic proportion here and are highly active.

Our model proposes that organellar DNA sequences are preferentially inserted in the vicinity of centromeres. Centromeres and pericentromeres are both better able to engulf large cpDNA or mtDNA and they represent a more stable genomic environment for integrated NUPTs/NUMTs as suggested by Matsuo *et al.* (2005). This phenomenon is visible in small genomes but the situation is probably the same for larger genomes too. Our model predicts that especially in species with higher genome dynamism, not only fragmentation of NUPTs and NUMts by TEs, but also TE-based recombination can result in reshuffling of genomic regions and shifting of NUPTs/NUMTs away from the (peri)centromeres (Figure 5b). However, it is also possible that NUPTs and NUMTs are gathered near the centromeres in species having small genomes because there is little intergenic space outside the centromeric regions. In this way, the amount of intergenic space would be an additional factor to the activity of TEs. However, the accumulation of NUPTs and NUMTs in

the centromeres of the rye B chromosomes that are more gene-poor (Martis *et al.*, 2012) supports an explanation based on the attractivity of centromeres for organellar insertions than preferential insertions into gene-poor regions. TEs probably have a significant role in reshuffling of chromosomal regions and removing NUPTs and NUMTs.

Previous studies on maize suggested that the doubling of its genome over as little as 3 million years is due to TE accumulation (SanMiguel *et al.,* 1996, 1998). In maize, intact retroelements outnumber solo LTRs by >5:1 (SanMiguel *et al.,* 1996), while this ratio is 1:1 in *Arabidopsis* and 2:3 in rice (Devos *et al.,* 2002). This pattern suggests that the maize genome is in an evolutionary phase of expansion while the genomes of *Arabidopsis* and rice are probably under stronger evolutionary constraints and deletions by ectopic recombination are more frequent (Kejnovsky *et al.,* 2009, 2012). Indeed, the decay of TEs in rice is rapid; the half-life of LTR retrotransposon sequences in rice was found to be <6 million years (Ma *et al.,* 2004). Similarly, 80% of NUPTs are eliminated from the rice nuclear genome within a million years (Matsuo *et al.,* 2005). The rate of DNA loss differs between DNA fragment sizes and among species (Petrov *et al.,* 2000) and can potentially contribute to the size-dependent filtering of NUPTs or NUMTs (Richly and Leister, 2004a,b). In this way, the nuclear genome absorbs, fragments, reshuffles and eliminates plastid and mitochondrial sequences with a species-specific mode and tempo.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research has been supported by the Grant Agency of the Czech Republic (grants P305/10/0930, P501/10/0102, P501/12/2220, P501/12/G090), grant AV0Z50040702 from the Academy of Sciences of the Czech Republic, by the project 'CEITEC—Central European Institute of Technology' (CZ.1.05/1.1.00/02.0068) from European Regional Development Fund and by the project OPVK (CZ.1.07/2.3.00/20.0045). We highly acknowledge the access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme 'Projects of Large Infrastructure for Research, Development and Innovations' (LM2010005).

Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
Ayliffe MA, Scott NS, Timmis JN (1998). Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. *Mol Biol Evol* **15**: 738–745.
Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* **16**: 314–321.
Devos KM, Brown JKM, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res* **12**: 1075–1079.
Ellis J (1982). Chloroplast genes inside plant mitochondria. *Nature* **299**: 678–679.
Guo X, Ruan S, Hu W, Cai D, Fan L (2008). Chloroplast DNA insertions into nuclear genome of rice: the genes, sites and ages of insertion involved. *Func Integr Genomics* **8**: 101–108.
Hazkani-Covo E, Sorek R, Graur D (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.
Hazkani-Covo E, Covo S (2008). Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet* **4**: e1000237.
Hazkani-Covo E, Zeller RM, Martin W (2010). Molecular Poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* **6**: e1000834.
Henze K, Martin W (2001). How mitochondrial genes get into the nucleus? *Trends Genet* **17**: 383–387.
Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol* **138**: 1723–1733.
Kejnovsky E, Leitch I, Leitch A (2009). Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol Evol* **24**: 572–582.
Kejnovsky E, Hawkins JS, Feschotte C (2012). Plant transposable elements: biology and evolutionWendel JF, Greilhuber J, Dolezel J, Leitch IJ (eds). *Plant Genome Diversity*, Volume 1. Springer: Wien, Heidelberg, NewYork, Dortrecht, London.
Kleine T, Maier UG, Leister D (2009). DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Ann Rev Plant Biol* **60**: 115–138.
Leister D (2005). Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet* **21**: 655–663.
Lewin R (1984). No genome barriers to promiscuous DNA. *Science* **224**: 970–971.
Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14**: 860–869.
Martin W, Herrmann RG (1998). Gene transfers from organelles to the nucelus: how much, what happens, and why? *Plant Physiol* **118**: 9–17.
Martin W, Rujan T, Richly E, Hansen A, Cornelson S, Lins T *et al.* (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* **99**: 12246–12251.
Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FK, Macas J, Schmutzer T *et al.* (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci USA* **109**: 13343–13346.
Matsuo M, Ito Y, Yamauchi R, Obokata J (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell* **17**: 665–675.
Noutsos C, Richly E, Leister D (2005). Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res* **15**: 616–628.
Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000). Evidence for DNA loss as a determinant of genome size. *Science* **287**: 1060–1062.
Richly E, Leister D (2004a). NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol* **21**: 1081–1084.
Richly E, Leister D (2004b). NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol Biol Evol* **21**: 1972–1980.
Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ (2010). Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. *Cytogenet Genome Res* **129**: 17–23.
SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998). The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**: 43–45.
SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A *et al.* (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**: 765–768.
Selosse M-A, Albert B, Godelle B (2001). Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends Ecol Evol* **16**: 135–141.
Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003). Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Plant Mol Biol* **52**: 923–934.
Smith DR, Crosby K, Lee RW (2011). Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biol Evol* **3**: 365–371.
Stegemann S, Bock R (2006). Experimental reconstruction of functional gene tarsnefer from tobacco plastid genome to the nucleus. *Plant Cell* **18**: 2869–2878.
Yuan Q, Hill J, Hsiao K, Moffat K, Ouyang S, Cheng Z *et al.* (2002). Genome sequencing of a 239-kb region of rice chromosome 10 L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. *Mol Gen Genet* **267**: 713–720.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)