Analysis of nuclear mitochondrial DNAs and factors affecting patterns of integration in plant species

Takanori Yoshida, Hazuka Y. Furihata and Akira Kawabe*

Faculty of Life Sciences, Kyoto Sangyo University, Kamigamo Motoyama, Kita-ku,
Kyoto, Kyoto 603-8555, Japan

(Received 19 July 2016, accepted 7 November 2016; J-STAGE Advance published date: 22 February 2017)

Sequences homologous to organellar DNA that have been integrated into nuclear genomes are referred to as nuclear mitochondrial DNAs (NUMTs) and nuclear plastid DNAs (NUPTs). NUMTs in nine plant species were analyzed to reveal the integration patterns and possible factors involved. The cumulative lengths of NUMTs in two-thirds of species analyzed were greater than those of NUPTs observed in a previous study. The age distribution of NUMTs was similar to that of NUPTs, suggesting similar mechanisms for integration and degradation of both NUPTs and NUMTs. Nuclear genome size and the cumulative length of NUMTs showed a significant positive correlation for older but not younger NUMTs. The same correlation was also found between nuclear genome size and older NUPTs in 17 species. These results suggested that genome size is a key factor to determine the cumulative length of relatively older NUPTs/NUMTs. Although the factor(s) determining the cumulative length of younger NUPTs/NUMTs is unclear, these sequences may be more deleterious, which could explain the different manner of determining the cumulative length of younger NUPTs/NUMTs in nuclear genomes. In addition, a relationship between the cumulative length of integrated NUMTs and complexity of mitochondrial genomes (i.e., the number of repeats) was found. The results indicate that the structural complexity of both NUMTs and their original mitochondrial sequences affects integration and degradation processes.

Key words: mitochondrion, NUMTs, organellar genome, plants

INTRODUCTION

Mitochondria are cytoplasmic organelles that have their own distinct genomes, which originated from the endosymbiosis of ancient bacteria (Mereschkowsky, 1905; Margulis, 1981). In comparison with those in the genomes of current, free-living bacteria, the number of genes in mitochondrial genomes is small, and the genes encode only a small fraction of the proteins that function in mitochondria (Kleine et al., 2009). During the process of endosymbiosis, a large number of genes were transferred from the organellar genomes to the host nuclear genome (Timmis et al., 2004). Recent studies have revealed that transfers of segments of DNA from organellar to nuclear genomes are ongoing. Both mitochondrial and plastid DNA sequences have been integrated into the host nuclear genome (Noutsos et al., 2005; Smith

et al., 2011; Michalovova et al., 2013). Such organellar DNA-like sequences in nuclear genomes are called nuclear mitochondrial DNAs (NUMTs) and nuclear plastid DNAs (NUPTs). In previous studies, the patterns of integration in the model plants *Arabidopsis thaliana* and *Oryza sativa* have been extensively analyzed (e.g., Richly and Leister, 2004; Matsuo et al., 2005; Noutsos et al., 2005). Recently, other species also have been analyzed to reveal the patterns of integration (Smith et al., 2011; Michalovova et al., 2013; Yoshida et al., 2014). In addition, research has been conducted using tobacco plants to estimate the transfer rate (Lloyd and Timmis, 2011; Stegemann et al., 2013).

The fate of integrated DNA sequences (NUPTs and NUMTs) has been analyzed using both genomic and experimental data, and the degradation of these sequences by both small and large indels (insertions and deletions) was suggested (Richly and Leister, 2004; Matsuo et al., 2005; Sheppard and Timmis, 2009). The general patterns of genomic integration of NUPTs in 17 plant species have been analyzed (Yoshida et al., 2014). The esti-

Edited by Minoru Murata

* Corresponding author. E-mail: akiraka@cc.kyoto-su.ac.jp

DOI: <http://doi.org/10.1266/ggs.16-00039>

mated age distribution of NUPTs for those species was of two types. For some species, there was a very high proportion of young insertions of plastid genes with a rapid decrease of the number of insertions in older categories. In other species, there was a relatively even distribution of old and young insertions. However, it is not clear whether these differences are due to differences among the species. Because NUMTs originated from mitochondrial genomes, which are independent of plastid genomes, analyses of plant NUMTs provide an opportunity to compare the age distributions of NUMTs with those of NUPTs.

A distinct feature of mitochondrial genomes compared to plastid genomes is diversity in their structure and size. Mitochondrial genomes often contain repeat regions that cause intermolecular recombination and form multipartite structures. These genomes also differ in length between species. In some cases, even close relatives show large differences in genome sizes and coverage by these short repeats (Alverson et al., 2011a). Plastid genomes do not show such high levels of structural diversity among species, and so it is worth investigating whether the variable structure of mitochondrial genomes is related to the NUMT integration patterns.

In this study, we investigated the patterns of genomic integration of NUMTs for nine plant species whose mitochondrial and nuclear genome sequences are available. The results revealed two types of age distributions, as seen in NUPTs (Yoshida et al., 2014). In addition, they suggest that patterns of incorporation of NUMTs are related to the complexity of mitochondrial genomes.

MATERIALS AND METHODS

Data used We studied nine plant species for which both nuclear and mitochondrial genomic sequence data were available. Although both nuclear and mitochondrial genomic sequences of *Ricinus communis* have been reported (Chan et al., 2010; Rivarola et al., 2011), a preliminary inspection revealed that the assembly status of the nuclear genome was inadequate for analysis of NUMTs, and thus we excluded it from the current study. The species used were as follows: *Arabidopsis thaliana* (Giegé and Brennicke et al., 1999; Arabidopsis Genome Initiative, 2000), *Carica papaya* (Rice et al., unpublished NC_012116; Ming et al., 2008), *Vitis vinifera* (French-Italian Public Consortium for Grapevine Genome Characterization, 2007; Goremykin et al., 2009), *Lotus japonicus* (Sato et al., 2008; Kazakoff et al., 2012), *Glycine max* (Schmutz et al., 2010; Chang et al., 2013), *Cucumis sativus* (Huang et al., 2009; Alverson et al., 2011b), *Oryza sativa* (Notsu et al., 2002; International Rice Genome Sequencing Project, 2005), *Sorghum bicolor* (Allen et al., unpublished NC_008360; Paterson et al., 2009), and *Zea mays* (Clifton et al., 2004; Schnable et al., 2009). The

nine species include six eudicots and three monocots (all Gramineae). Accession numbers for genomic data are shown in Supplementary Table S1.

Identification of NUMTs BLAST (<http://blast.ncbi.nlm.nih.gov/>) searches against nuclear genomic sequences were conducted using each mitochondrial genome as a query sequence. For *C. papaya*, *V. vinifera* and *Z. mays*, bulk genomic sequence data were obtained from Phytozome ver. 9.1 (Goodstein et al., 2012). For *L. japonicus*, genomic data were obtained from the website of the Kazusa DNA Research Institute (www.kazusa.or.jp/lotus/index.html). The NCBI BLAST server was used for *A. thaliana*, *G. max*, *O. sativa* and *S. bicolor*, whereas the local BLAST program was used for species with bulk data. For BLAST hits, p-distance ($p = n_d/n$, where n_d is the number of nucleotide differences and n the number of aligned nucleotides) was calculated to estimate the relative integration time. A larger p-distance indicates a more ancient integration into the nuclear genome. As in the previous study of NUPTs (Yoshida et al., 2014), BLAST hits of 100 bp or longer and with relatively low p-distances (≤ 0.1) were selected. Compared with NUPTs, BLAST searches with mitochondrial genomes resulted in a larger number of partially/completely overlapping BLAST hit coordinates (start position, end position) in nuclear genomes. This is due to the larger number of repeats in mitochondrial genomes compared to plastids. To prevent overestimation of the number of NUMTs, such overlapping BLAST hits were inspected and eliminated in the following manner: first, if more than one BLAST hit had identical/nested coordinates, the longest hit with the lowest p-distance was counted as an NUMT and the others were ignored (Supplementary Fig. S1, A and B); and second, if BLAST hits partially overlapped for more than half of their length, only the hit with the longest sequence was counted as an NUMT (Supplementary Fig. S1C). If BLAST hits were adjacent to each other and with overlaps of less than half of their lengths, each hit was independently counted as an NUMT (Supplementary Fig. S1D). It should be noted that in addition to independent, recurrent integration (or simultaneous integrations) of distinct NUMTs, a single large insertion and subsequent multiple deletions can make the pattern of mosaic NUMTs. The nuclear coordinates of candidate NUMTs in each species are listed in Supplementary Table S2.

Data analyses of NUMTs The proportion of the nuclear genome that consisted of NUMTs and the cumulative length of those NUMTs were calculated and compared with nuclear genome size. Mutation directions of each nucleotide in integrated DNA sequences are different from each other (Rousseau-Gueutin et al., 2011). This makes it difficult to estimate the corrected number of

substitution events between NUMTs and mitochondrial genomes. Thus, as in Yoshida et al. (2014), p-distance was used as an indicator of relative time since integration into the nuclear genome. For each species, the age distribution of NUMTs, as indicated by their p-distance, was estimated.

The relationship between integration of NUMTs and structural diversity of mitochondrial genomes was investigated. To evaluate the structural complexity of the mitochondrial genomes of each species, the number of repeat regions in each was estimated by BLAST searches using MegaBLAST with word_size = 16. In the BLAST search, each mitochondrial genome was used as a query against its own sequence. To avoid multiple counting of repeat sequences, BLAST hits with unique coordinates in the mitochondrial genome were considered to be repeat sequences. The number of long (more than 1 kb) and short (21–1,000 bp) repeats was counted and the relationship with cumulative length of the NUMTs was examined.

For all species analyzed, there were mosaics of BLAST hits, each of which corresponds to a discrete region in the mitochondrial genome (Supplementary Fig. S1D). The existence of such mosaic BLAST hits may indicate the formation of mosaic organellar DNA fragments during or after insertion into nuclear genomes. The coordinates of each BLAST hit within the mosaic overlapped by between less than 10 bp and more than 1 kb. To compare such mosaic-structured NUMTs between species, the number of mosaic NUMTs and the length of overlap between adjacent hits were assessed based on the coordinates of BLAST hits. The hits were sorted by their coordinates (chromosome number, start position, and end position) and the end position of one NUMT (e.g., the purple-colored NUMT in Supplementary Fig. S1D) was compared with the start position of the next one (the light blue-colored

NUMT).

RESULTS AND DISCUSSION

NUMTs in plant species For each of the six study species (*A. thaliana*, *C. papaya*, *V. vinifera*, *C. sativus*, *S. bicolor* and *Z. mays*), the cumulative length of NUMTs (Table 1) was greater than that found for NUPTs (Yoshida et al., 2014). For example, in *A. thaliana*, the cumulative length of NUMTs (282.1 kb) was about 16 times greater than of NUPTs (17.7 kb). In addition to *A. thaliana*, *C. sativus* showed a large excess of NUMTs (309.2 kb) compared to NUPTs (49.0 kb). The excess of NUMTs compared with NUPTs in these species may reflect more frequent integrations from mitochondrial genomes than from plastid genomes. Several developmental processes have been proposed as pathways for integration of organellar-derived DNA: lysis of organelles during processes such as pollen development, nuclear inclusion of mitochondria, nuclear attachment of plastids, and stromules connecting plastids with nuclei (Leister, 2005). The integration of large NUMTs is more likely caused by nuclear inclusion of mitochondria, rather than the fragmental integration after lysis of organelles. Such mitochondrial inclusion events have been observed in plants (Yu and Russell, 1994) and may be one of the reasons for the greater cumulative length of NUMTs compared to NUPTs.

In our previous study, a significant enrichment of transposable elements (TEs) in the vicinity of NUPTs was observed in *A. thaliana* and *Z. mays* (Yoshida et al., 2014), suggesting the similar distribution of NUPTs and TEs in nuclear genomes. In the current study, *A. thaliana* NUMTs co-existed with TEs more frequently than expected (Supplementary Fig. S2). The average number

Table 1. Number and cumulative length of NUMTs in plant genomes for nine model species

Species	Genome size (Mb) ^a		Number	NUMTs		
	Nuclear	Mitochondrial		Length (kb)		
				All	Young NUMTs ^b	Other NUMTs ^c
<i>A. thaliana</i>	119	0.37	41	282.1	278.5	3.6
<i>C. papaya</i>	343	0.48	671	403.8	156.0	247.0
<i>V. vinifera</i>	486	0.77	1148	417.0	62.1	353.9
<i>L. japonicus</i>	301	0.38	388	117.6	47.2	70.1
<i>G. max</i>	974	0.4	493	265.9	25.0	240.2
<i>C. sativus</i>	203	1.56	397	309.2	166.6	142.6
<i>O. sativa</i>	382	0.49	1183	651.2	348.9	302.3
<i>S. bicolor</i>	697	0.47	523	174.6	54.0	120.3
<i>Z. mays</i>	2066	0.57	1456	1671.3	963.6	707.4

^aTotal length based on whole-genome sequencing.

^bCumulative length of NUMTs with p-distance < 0.01.

^cCumulative length of NUMTs with 0.01 ≤ p-distance < 0.10.

of TEs in the vicinity of NUMTs was 4.43 (S.D. = 3.82), while that in randomly extracted genomic sequences was 2.35 (S.D. = 3.44). This significant enrichment of TEs was not observed in *O. sativa* or *Z. mays*. The difference between *A. thaliana* and the other two species may be due to the genome structure of these three species: *A. thaliana* has a small genome with a relatively small number of TEs, while *O. sativa* and *Z. mays* have larger numbers of TEs in their genomes. The small genome, and, hence, the clear distinction between gene-rich and TE-rich regions within the *A. thaliana* genome, may result in the significant difference of the TE number.

Age distribution of NUMTs In our previous study of 17 plant species (Yoshida et al., 2014), NUPTs showed mainly two types of age distribution: the majority of NUPTs were relatively young for six species (p-distance <

0.01; two eudicots and four monocots), while there was an even distribution of estimated ages for the NUPTs of the remaining 11 eudicots. These same two patterns were found for NUMTs (Fig. 1). Seven of the nine study species (four eudicots and three monocots) showed a large proportion of the NUMTs falling in the youngest category of integrations, with an immediate decrease in the next-youngest category. The remaining two species, *V. vinifera* and *G. max*, did not show such a pattern. In four species, this predominance of young integrations was seen for NUMTs but not NUPTs (*A. thaliana*, *C. papaya*, *L. japonicus* and *C. sativus*). In *A. thaliana*, the excess of NUMTs in the youngest age category is due primarily to one extra-large integration: a nearly-entire mitochondrial genomic sequence is integrated into the vicinity of the centromere region of chromosome 2 (Supplementary Fig. S3) (Lin et al., 1999; Stupar et al., 2001). Long NUMTs

Fig. 1. Age distribution of NUMTs. The cumulative length of NUMTs for each p-distance interval is shown. White bars represent the proportion of these NUMTs that are long insertions (length > 10 kb).

were also observed in other species. Five species (*C. papaya*, *L. japonicus*, *C. sativus*, *O. sativa* and *Z. mays*) have more than one extra-long NUMT (longer than 10 kb) (Fig. 1, Supplementary Fig. S4). The location of extra-long NUMTs varied among nuclear chromosome regions (Supplementary Fig. S3), suggesting recurrent integrations of such long sequences. These long NUMTs were relatively recent integrations (Supplementary Fig. S4), while older NUMTs were quickly degraded (Supplementary Fig. S5). The elimination of long NUMTs (Supplementary Fig. S4) and pervasive changes caused by small indels (Supplementary Fig. S5) were similar to those observed in NUPTs, suggesting that the degradation/elimination process after integration is similar between NUPTs and NUMTs.

Relationship between nuclear genome sizes and NUMTs Two previous studies suggested that the nuclear genome size and the cumulative length of NUMTs were positively correlated (Smith et al., 2011; Michalovova et al., 2013). In contrast, Richly and Leister reported no significant correlation between both nuclear and mitochondrial genome sizes and the cumulative length of NUMTs (Richly and Leister, 2004). In the current study, nuclear genome size and cumulative length of all NUMTs had a positive relationship but it was not significant (Kendall's $\tau = 0.278$, $P = 0.359$). In previous analyses of NUPTs, we discussed the possibility that the age distribution of integrated DNAs is due to the different fitness effects of younger vs. older integrated DNAs. In that hypothesis, many recently integrated DNAs have a deleterious effect on the genome and are selected against, being either removed or reduced in size. Due to this selection, many fewer integrated DNAs survive to a relatively older age, and those DNAs that do survive are less deleterious and smaller in size. Based on that idea, we tentatively divided NUMTs into younger (p-distance < 0.01) and older NUMTs, and tested for a correlation between nuclear genome sizes and the cumulative lengths of each age class. Interestingly, while there was no relationship between nuclear genome sizes and cumulative lengths of younger NUMTs, a significant positive relationship was found between the nuclear genome sizes and cumulative lengths of older NUMTs (Table 2). To confirm that this difference is a common pattern for integrated organellar DNAs, we also tested the data for NUPTs from 17 species in our previous study (Yoshida et al., 2014) and found the same result (Table 2). One explanation is that a recent “burst” of integration events in some species may have resulted in the non-significant relationship between nuclear genome sizes and cumulative lengths of younger NUMTs/NUPTs. The other possible explanation is a difference between younger and older NUMTs in their deleterious effect on nuclear genomes, as hypothesized above. Young NUMTs may have stronger dele-

Table 2. Correlation between nuclear genome sizes and cumulative lengths of NUMTs

Test	τ	P
Nuc. genome size (9 spp.) vs. young NUMTs	-0.167	0.612
Nuc. genome size (9 spp.) vs. the other NUMTs	0.556	0.045*
Nuc. genome size (17 spp.) ^a vs. young NUPTs	0.309	0.091
Nuc. genome size (17 spp.) vs. the other NUPTs	0.426	0.017*

^a Data for NUPTs are from Yoshida et al. (2014).

* $P < 0.05$.

terious effects and their toleration by nuclear genomes may be determined by conditions such as genome architecture and the strength of the genome immune system against foreign DNAs. Genes potentially involved in such epigenetic mechanisms show functional variation within species (Shen et al., 2014) or elevated amino acid substitution rates between species (Willing et al., 2015), suggesting differences in the genome immune system between species (Springer et al., 2016). In contrast, NUMTs that survived in the genome to an older age could have always been or could have become less deleterious, and the cumulative length of those NUMTs might be determined by the size of the nuclear genome.

Relationship between complexity of mitochondrial genomes and NUMTs The structural diversity of mitochondrial genomes could affect both the number and size of NUMTs. We estimated the number of repeat sequences as an indicator of the structural complexity of the mitochondrial genomes. For all species, both long (> 1 kb) and short repeats were found (Supplementary Table S3). The largest fraction of repeat sequences was short repeats of 21 to 50 bp in length (Supplementary Fig. S6). There was an especially large number of repeat sequences in the mitochondrial genome of *C. sativus*. The number of short repeats in *C. sativus* ($n = 109,504$) was more than 30 times larger than in the species with the second-highest number of repeats, *V. vinifera* ($n = 3,065$). The number of long repeats ranged from two (*C. papaya* and *V. vinifera*) to 25 (*G. max*; Supplementary Table S3). The relationship between number of short repeats in the mitochondrial genome and cumulative length of NUMTs was not significant (Kendall's $\tau = 0.5$, $P = 0.075$; Supplementary Table S4). When the unusually high numbers from *C. sativus* were excluded, the test showed a significant correlation (Supplementary Table S4). In contrast, the number of long repeats did not show a relationship with the cumulative lengths of NUMTs. We also estimated the number of mosaic NUMTs and the overlapping sequence length for adjacent NUMTs within those mosaics (Table 3). Most overlapping regions were shorter than 10 bp for all species except *C. sativus*, for which the majority of overlaps were in the range of 21–50 bp.

Table 3. Distribution of lengths of overlapping regions for adjacent NUMTs within mosaics

Species	Number of mosaic NUMTs	Overlapping length between adjacent NUMTs					
		1–10 bp	11–20 bp	21–50 bp	51–100 bp	101–1,000 bp	> 1 kb
<i>A. thaliana</i>	9	7	0	0	0	1	1
<i>C. papaya</i>	66	61	5	0	0	0	0
<i>V. vinifera</i>	140	134	6	0	0	0	0
<i>L. japonicus</i>	42	30	5	2	0	5	0
<i>G. max</i>	33	25	2	2	1	3	0
<i>C. sativus</i>	68	14	3	35	11	5	0
<i>O. sativa</i>	224	203	15	0	0	4	2
<i>S. bicolor</i>	77	67	7	0	0	3	0
<i>Z. mays</i>	184	161	8	5	2	8	0

In previous studies, the mechanism of NUMT integration was analyzed, and nonhomologous end-joining (NHEJ) of double-stranded breaks (DSBs) was suggested as a common mechanism of integration among all eukaryotes (Ricchetti et al., 1999; Wang and Timmis, 2013; reviewed in Leister, 2005). With this mechanism, termini of nuclear DSBs and organellar DNA fragments are attached to one another due to microhomology in the sequences (Leister, 2005). Homologous sequences may influence not only de novo insertions in the nuclear genome, but also rearrangements and deletions after integration (Noutsos et al., 2005); repeats are an example of sequences that may provide microhomologies. Our results suggest a relationship between repeat sequences within mitochondrial genomes and both the cumulative length and the complexity of NUMT regions. During integration, such repeat sequences can facilitate the formation of complex structures of mitochondrial DNA by end-joining. In the rearrangement of integrated NUMTs, repeat sequences can be the cause of replication slippage. The reason why the test with *C. sativus*, which has the highest amount of short repeats, did not show clear significance is unknown. One possibility is that the extra-large mitochondrial genome size (1.56 Mb) and a high number of repeats have an antagonistic effect on the amount of integrated sequences.

Factors affecting the pattern of NUMT integration In this study, there were two types of age distribution of NUMTs across species (Fig. 1), similar to those seen for NUPTs (Yoshida et al., 2014). There was a relationship between the complexity of mitochondrial genomes and the integration pattern of NUMTs (Table 3, Supplementary Table S3). The cumulative length of NUMTs tentatively categorized by p-distance as being relatively young did not show any relationship to nuclear genome sizes, while older NUMTs showed a significant positive relationship. Although it is difficult to measure the deleterious effect of de novo NUMTs, the analysis of

epigenetic modification as an indicator of the strength of the genome immune response against foreign DNAs (Kim and Zilberman, 2014) may bring new insights into the integration pattern of NUMTs.

This work was supported by a Grant-in-Aid for Scientific Research in Innovative Areas (23113003 to A. K.) and a Research Fellowship from the Japan Society for the Promotion of Science (201508708 to T. Y.). We would like to thank Editage for English language editing and anonymous reviewers for their comments and suggestions.

REFERENCES

- Alverson, A. J., Rice, D. W., Dickinson, S., Barry, K., and Palmer, J. D. (2011b) Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *Plant Cell* **23**, 2499–2513.
- Alverson, A. J., Zhuo, S., Rice, D. W., Sloan, D. B., and Palmer, J. D. (2011a) The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. *PLoS One* **6**, e16404.
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.
- Chan, A. P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K. M., Redman, J., Chen, G., et al. (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951–956.
- Chang, S., Wang, Y., Lu, J., Gai, J., Li, J., Chu, P., Guan, R., and Zhao, T. (2013) The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. *PLoS One* **8**, e56502.
- Clifton, S. W., Minx, P., Fauron, C. M., Gibson, M., Allen, J. O., Sun, H., Thompson, M., Barbazuk, W. B., Kanuganti, S., Tayloe, C., et al. (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiol.* **136**, 3486–3503.
- French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467.
- Giegé, P., and Brennicke, A. (1999) RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. *Proc. Natl. Acad. Sci. USA.* **96**, 15324–15329.

- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186.
- Goremykin, V. V., Salamini, F., Velasco, R., and Viola, R. (2009) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Mol. Biol. Evol.* **26**, 99–110.
- Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., et al. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281.
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**, 793–800.
- Kazakoff, S. H., Imelfort, M., Edwards, D., Koehorst, J., Biswas, B., Batley, J., Scott, P. T., and Gresshoff, P. M. (2012) Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. *PLoS One* **7**, e51687.
- Kim, M. Y., and Zilberman, D. (2014) DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.* **19**, 320–326.
- Kleine, T., Maier, U. G., and Leister, D. (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu. 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.
- Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–768.
- Lloyd, A. H., and Timmis, J. N. (2011) The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Mol. Biol. Evol.* **28**, 2019–2028.
- Margulis, L. (1981) Symbiosis in Cell Evolution: Life and Its Environment on the Early Earth. W. H. Freeman, San Francisco.
- Matsuo, M., Ito, Y., Yamauchi, R., and 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.
- Mereschkowsky, C. (1905) Über natur und ursprung der chromatophoren im pflanzenreiche. *Biol. Zent. Bl.* **25**, 593–604.
- Michalovova, M., Vyskot, B., and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertion in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity* **111**, 314–320.
- Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L., et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996.
- Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., and Kadowaki, K. (2002) The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol. Genet. Genomics* **268**, 434–445.
- Noutsos, C., Richly, E., and Leister, D. (2005) Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res.* **15**, 616–628.
- Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556.
- Ricchetti, M., Fairhead, C., and Dujon, B. (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**, 96–100.
- Richly, E., and Leister, D. (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol. Biol. Evol.* **21**, 1972–1980.
- Rivarola, M., Foster, J. T., Chan, A. P., Williams, A. L., Rice, D. W., Liu, X., Melake-Berhan, A., Huot Creasy, H., Puiu, D., Rosovitz, M. J., et al. (2011) Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PLoS One* **6**, e21743.
- Rousseau-Gueutin, M., Ayliffe, M. A., and Timmis, J. N. (2011) Conservation of plastid sequences in the plant nuclear genome for millions of years facilitates endosymbiotic evolution. *Plant Physiol.* **157**, 2181–2193.
- Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., et al. (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227–239.
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183.
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115.
- Shen, X., De Jonge, J., Forsberg, S. K. G., Pattersson, M. E., Sheng, Z., Hennig, L., and Carlborg, Ö. (2014) Natural *CMT2* variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS Genet.* **10**, e1004842.
- Sheppard, A. E., and Timmis, J. N. (2009) Instability of plastid DNA in the nuclear genome. *PLoS Genet.* **5**, e1000323.
- Smith, D. R., Crosby, K., and Lee, R. W. (2011) Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biol. Evol.* **3**, 365–371.
- Springer, N. M., Lisch, D., and Li, Q. (2016) Creating order from chaos: epigenome dynamics in plants with complex genomes. *Plant Cell* **28**, 314–325.
- Stegemann, S., Hartmann, S., Ruf, S., and Bock, R. (2003) High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc. Natl. Acad. Sci. USA* **100**, 8828–8833.
- Stupar, R. M., Lilly, J. W., Town, C. D., Cheng, Z., Kaul, S., Buell, C. R., and Jiang, J. (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.
- Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135.
- Wang, D., and Timmis, J. N. (2013) Cytoplasmic organelle DNA preferentially inserts into open chromatin. *Genome Biol. Evol.* **5**, 1060–1064.
- Willing, E. M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J., Becker, C., Warthmann, N., Chica, C., Szarynska, B., et al. (2015) Genome expansion of *Arabidopsis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat. Plants* **1**, 14023.
- Yoshida, T., Furihata, H. Y., and Kawabe, A. (2014) Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. *DNA Res.* **21**, 127–140.
- Yu, H. S., and Russell, S. D. (1994) Occurrence of mitochondria in the nuclei of Tobacco sperm cells. *Plant Cell* **6**, 1477–1484.