The mitochondrial genome of *Arabidopsis* is composed of both native and immigrant information

Joachim Marienfeld, Michael Unseld and Axel Brennicke

Plants contain large mitochondrial genomes, which are several times as complex as those in animals, fungi or algae. However, genome size is not correlated with information content. The mitochondrial genome (mtDNA) of *Arabidopsis* specifies only 58 genes in 367 kb, whereas the 184 kb mtDNA in the liverwort *Marchantia polymorpha* codes for 66 genes, and the 58 kb genome in the green alga *Prototheca wickerhamii* encodes 63 genes. In *Arabidopsis*’s mtDNA, genes for subunits of complex II, for several ribosomal proteins and for 16 tRNAs are missing, some of which have been transferred recently to the nuclear genome. Numerous integrated fragments originate from alien genomes, including 16 sequence stretches of plastid origin, 41 fragments of nuclear (retro)transposons and two fragments of fungal viruses. These immigrant sequences suggest that the large size of plant mitochondrial genomes is caused by secondary expansion as a result of integration and propagation, and is thus a derived trait established during the evolution of land plants.

To date, the largest mitochondrial genomes have been found in higher plants. The enormous difference in mitochondrial genome size between animals and plants has been puzzling ever since the first genome size estimates some 20-years ago. The complete sequence of the mitochondrial genome in the model plant *Arabidopsis* allows an in-depth comparison of its coding capacity with other completely analysed mitochondrial genomes of the liverwort *Marchantia polymorpha*, several algae, fungi and animals. These genome comparisons should clarify the question of a potential correlation between the resident information and the genome sizes. The genome sizes range from small animal genomes with 15 or 16 kb, to the intermediate protist, fungal and algal mtDNAs with 20–100 kb, up to the largest genomes with 200–2400 kb in plants. In fungi, a major part of the genome expansion can be explained by the presence of introns, but the more than tenfold increase in flowering plants has remained largely unexplained.

Several additional genes in plant mitochondria, which are not found in animal or fungal mitochondrial genomes, suggest that there is increased information content in plants. These include genes coding for:

- Ribosomal proteins
- Additional subunits of respiratory chain complexes
- Genes involved in cytochrome-c-biosynthesis
- Reading frames conserved between different plant species, but the function of which is unclear.

Most of these genes are also found in the smaller mtDNA of the liverwort *M. polymorpha*, and in the even smaller mtDNAs of green algae, such as *Prototheca wickerhamii*, and in some densely packed protost genomes, such as *Reclinomonas americana*. Although the mitochondrial genome in *M. polymorpha* is 100 kb larger than the mtDNA in *P. wickerhamii*, it encodes only four additional genes (i.e. a rRNA and three ribosomal proteins). The remainder of the additional sequence in *M. polymorpha* is composed of increased intergenic regions of unknown origin, more and larger introns and sequence duplications.

Analysis of the *Arabidopsis* mitochondrial genome has extended this mitochondrial genome comparison to include flowering plants, with another size increase of 180 kb (Ref. 2). Surprisingly, instead of more genes there is less bona fide information coded in the *Arabidopsis* mitochondrial genome than in the mtDNA of *M. polymorpha*, which is half the size, and in algal genomes, which are seven-times smaller. In this review, we analyse the complete sequence of the *Arabidopsis* mitochondrial genome with respect to our understanding of the coding potential and function of the mitochondrial genome in plants, and we focus on the origin of the additional, mostly non-coding sequences, in the genome (Fig. 1).

**Mitochondrial genes**

**Genes for respiratory chain functions**

The *Arabidopsis* mitochondrial genome encodes all the ‘classic’ subunits of the different complexes in the respiratory chain (i.e. those which are usually encoded in the mitochondrial genomes of all eukaryotes: Table 1). These genes specify subunits of:

- Complex I, the NADH-dehydrogenase (nine polypeptides, genes *nad1–7, nad4L* and *nad5*).
- Complex II, the succinate dehydrogenase (1 subunit, *sdh4*).
- Complex III, the cytochrome-c reductase (1 polypeptide, *cytB*).
- Complex IV, the cytochrome-c oxidase (three subunits, *cox1–3*).

The nad1/10 subunit of complex I, which is encoded mitochondrially in some fungi and algae, is encoded in the nuclear genome in *Arabidopsis*. Genes for three subunits of complex II are found in the mitochondrial genomes of some red and brown algae, two units in *M. polymorpha*, but only one in *Arabidopsis*. Four subunits of the F1/F0 ATPase are encoded in the *Arabidopsis* mitochondrial genome: subunits 1, 6, 8 and 9 (*atp1, atp6, atp8* and *atp9*). The poorly conserved subunit 8, which is coded by the mitochondrial genome in animals is also specified in the mitochondria of higher plants by the *orth* gene. Although no significant primary similarity is observed between *Arabidopsis* and the mammalian reading frame, similarity has been traced through the algal and protist sequences.

**Ribosomal protein genes**

In plant and algal mitochondrial as well as plastid genomes, genes for ribosomal proteins can generally be defined by their similarity to the respective bacterial genes. The structural similarities between the organellar and prokaryotic ribosomal protein genes...
are indicative of the common origin postulated by the endosymbiotic theory. Half of the 16 genes that code for ribosomal proteins in the *M. polymorpha* mitochondrial genome are not found in the *Arabidopsis* mitochondrial DNA. The genes *rps1*, *rps2*, *rps8*, *rps16*, *rpl11*, *rpl12*, *rpl19* and *rpl20* are absent. The missing genes, *rps19* and *rpl6*, together with *rps12, rps13, rps15* and *rpl16* are encoded in a cistron configuration in the mitochondrial genome of other flowering plants, such as *Zea mays*, *Petunia hybrida* and *Oenothera biennis*.

Only the central *rps3* gene of this co-transcribed gene cluster has been retained in *Arabidopsis* mitochondrial DNA. In *Arabidopsis*, the gene for *rps19* has been translocated to the nuclear genome, and the gene for *rps13* has been lost completely. Instead, the *rps19* gene, which is now nuclear, has acquired an RNA-binding domain at its N-terminus, which is postulated to substitute for the *rps19* gene function. In the bacterial ribosome, RPS13 connects the catalytic domain at its N-terminus, which is postulated to substitute for the *rps19* gene function.

The functional *rps14* gene sequence has likewise moved to the nuclear genome (Fig. 2), whereas in the mitochondrial genome only an incomplete fragment is identified, a remnant of the previously active gene. The *rps2* gene is partially missing in the *Arabidopsis* mitochondrial genome, where a shorter reading frame codes for only 307 amino acids. The missing part probably has been transferred to the nuclear genome after the disruption of the gene. The interruption of the *rps2* gene must have occurred earlier in the mitochondrial genome of ancestral land plants, because in other plant species both parts of *rps2* are found as separate open-reading frames (orf) in the mitochondrial genome.

Incomplete set of *rRNA* genes

In the mitochondrial genome of *Arabidopsis*, 22 *rRNA* genes are identified on the basis of their ‘classic’ cloverleaf structures. There are four duplicated genes and 18 different *rRNA* genes. Classified by their comparative similarities to those in *M. polymorpha* and non-plant mitochondria, 12 of these are ‘native’ resident genes (i.e. they originated from the original endosymbiont). The remaining six *rRNA* genes in the *Arabidopsis* genome show greater similarity to the respective plastid genes and are presumably derived from transferred DNA fragments of the plastid genome (Table 2).

Considering that only *rRNA* genes transcribed in the plastid are recognized, and that outside of the respective *rRNA* sequence itself no similarities to genomic plastid sequences are seen, these transfers might have occurred as genomic DNA transfers or might have an RNA-based origin in the plastid compartment. Integration of plastid *rRNA* genes partially compensates for the loss of mitochondrial information. To date, only *rRNA* genes have been successfully duplicated and transferred from the plastid to the mitochondrion, although numerous plastid DNA fragments are found in the mitochondrial genomes of *Arabidopsis* and other plants.

In addition to the functional *rRNA* genes, a fragment of the plastid *rRNA* II gene group I intron has been integrated into the *Arabidopsis* mitochondrial genome (nucleotide position 138 280–138 360) without the respective *rRNA* sequence. The resident *rRNA* II is of genuine mitochondrial descent.

The 18 different *rRNAs* specified by the *Arabidopsis* mitochondrial genome are not sufficient to decipher the entire set of codons found in the protein-coding genes. In *Arabidopsis* mitochondria, *rRNA* genes for five amino acids are lacking altogether, and these *rRNAs* (and possibly others) therefore have to be imported from the nucleus. Extrapolating from the situation described for the mitochondrial *rRNA* complements in larch (*Larix*), eucalypt, maize, bean and potato, where the set of endogenous and imported *rRNAs* has been evaluated experimentally, it is thought that several more nuclear-encoded *rRNA* species will be found. In larch, more than half of the *rRNAs* are imported from the plastid, whereas none of the mitochondrial *rRNA* species are imported from the cytosol (Table 2).

The *rRNA* genes for the five amino acids Asp, Gln, Glu, Ile, Met and Tyr are usually encoded. However, in *Arabidopsis*, the imported respective plastid gene has substituted the mitochondrial *rRNA* set for all of these amino acids. In *Marchantia*, two *rRNA* species, a *rRNA* L4a and a *rRNA* D7a, are missing in the mitochondrial genome, and are imported products of nuclear genes. In *Arabidopsis*, we conclude that 13 *rRNAs* are imported into the mitochondrial compartment, almost half of the required set (Table 2). This number might be even higher, depending on overlapping specificities and redundancies, such as observed for the nuclear-encoded *rRNA* L3 species in potato and bean mitochondria. The *Arabidopsis* sequence thus corroborates the experimental data and the conclusions drawn from the direct analysis of the *rRNA* population in the mitochondrial compartment in plants.
Table 1. Comparison of the coding information in mitochondrial genomes

<table>
<thead>
<tr>
<th></th>
<th>Complex I</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nad1</td>
<td>nad2</td>
<td>nad3</td>
<td>nad4</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>324 (4)</td>
<td>307 (4)</td>
<td>118</td>
<td>494 (3)</td>
</tr>
<tr>
<td>Marchantia</td>
<td>328</td>
<td>488 (1)</td>
<td>118 (1)</td>
<td>495 (1)</td>
</tr>
<tr>
<td>Prototricha</td>
<td>340</td>
<td>510</td>
<td>117</td>
<td>523</td>
</tr>
<tr>
<td>Chondrus</td>
<td>326</td>
<td>497</td>
<td>121</td>
<td>491</td>
</tr>
<tr>
<td>Podospora</td>
<td>366 (4)</td>
<td>556</td>
<td>157 (1)</td>
<td>519</td>
</tr>
<tr>
<td>Homo</td>
<td>316</td>
<td>348</td>
<td>114</td>
<td>459</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Complex II</th>
<th></th>
<th></th>
<th>Complex III</th>
<th></th>
<th>Other orfs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sdh2</td>
<td>sdh3</td>
<td>cob</td>
<td></td>
<td>atp1</td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>–</td>
<td>–</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marchantia</td>
<td>–</td>
<td>137</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototricha</td>
<td>–</td>
<td>–</td>
<td>364</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chondrus</td>
<td>253</td>
<td>127</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Podospora</td>
<td>–</td>
<td>–</td>
<td>307</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Complex IV</th>
<th></th>
<th></th>
<th>Complex V</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cox1</td>
<td>cox2</td>
<td>cox3</td>
<td>atp4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marchantia</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototricha</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chondrus</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Podospora</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ribosomal RNAs</th>
<th>Ribosomal proteins</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>ssrRNA</td>
<td>epp1</td>
<td>epp2</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Marchantia</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Prototricha</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Chondrus</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Podospora</td>
<td>–</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Homo</td>
<td>–</td>
<td>1</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>epp10</th>
<th>epp11</th>
<th>epp12</th>
<th>epp13</th>
<th>epp14</th>
<th>epp19</th>
<th>epp2</th>
<th>epp5</th>
<th>epp6</th>
<th>epp16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>–</td>
<td>–</td>
<td>125</td>
<td></td>
<td>–</td>
<td>6</td>
<td>–</td>
<td>307</td>
<td>185</td>
</tr>
<tr>
<td>Marchantia</td>
<td>102</td>
<td>125</td>
<td>126</td>
<td>120</td>
<td>99 (1)</td>
<td>91</td>
<td>501 (1)</td>
<td>187</td>
<td>101</td>
</tr>
<tr>
<td>Prototricha</td>
<td>107</td>
<td>128</td>
<td>125</td>
<td>119</td>
<td>99</td>
<td>77</td>
<td>231</td>
<td>194</td>
<td>–</td>
</tr>
<tr>
<td>Chondrus</td>
<td>–</td>
<td>116</td>
<td>127</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>138</td>
</tr>
<tr>
<td>Podospora</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Homo</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*Mitochondrial gene contents differ between Arabidopsis, a liverwort (Marchantia polymorpha), a green alga (Prototricha wickerhamii), a red alga (Chondrus crispus), the fungus Podospora anserina and Homo sapiens. Only Arabidopsis open-reading-frames (orfs), for which homologs can be identified in one or more of the other species are shown. Protein coding genes are subdivided into groups coding for subunits of the different respiratory chain complexes I (nad), II (sdh), III (cytb) and IV (cox) of the inner mitochondrial membrane. Genes for subunits of the large (rpl) and small (rps) subcomplexes of the mitochondrial ribosomes, and for proteins involved in the biogenesis of functional cytochrome c, vary in their mitochondrial presence between different organisms. Database loci are MTACG, MPOMTCG, PWU02970, MTCCGNME, MTTPACG and HUMMTC, respectively. The presence of genes encoding functional RNAs in a given genome is indicated by a 1. The deduced number of amino acids for each orf is given. A pseudogene is indicated by C. Intron numbers are in parentheses.

7 Includes two trans-splicing introns. 8 Includes one trans-splicing intron. 9 A pseudogene is also present. 10 Marchantia orf509 covers ccb382 and ccb203 of Arabidopsis. 11 No ATG start codon.
in Arabidopsis, which codes for one subunit, is split into two reading frames presumably have been moved to the nucleus. The mitochondrial genome of ABC transporter subunits (HELB and HELC) are encoded in the by the downstream membrane protein CCL1 (Ref. 18). Two of the porter with four subunits moves haem into position for presentation animal mitochondria have evolved a different mode of assembly. been retained only in plant and protist mitochondria; fungal and the alpha-proteobacteria. This complex biogenesis pathway has documented import of tRNAs. ecules being transported into the organelle in addition to the well would offer the interesting prospect of another class of RNA mol-

The membrane-anchored CCL1 protein is encoded by a single gene in M. polymorpha mitochondria, which code for the N- and C- terminal regions, respectively. These split frames have also been found in oilseed rape30, but not in wheat31, Oenothera32 or M. polymorpha33, suggesting that during the evolutionary history of the Brassicaceae, genomic recombina
tion disrupted this frame without deleterious consequences to protein function. The two Arabidopsis genes are separated by 24 kb and the C-terminal part is encoded upstream of the N-terminal fragment, confirming that the two genes are transcribed independently and translated in a similar way to the respective organelle encoded genes.

Gene for a novel protein transport pathway Five years ago an open-reading frame (orf509, now renamed mtB) was found in plant mitochondria, for which similarities and clearly homologous genes were No gene for a 4.5S RNA in the Arabidopsis mitochondrial genome In maize mitochondria, a candidate gene for a 4.5S RNA with similarity to the respective bacterial genes involved in protein secretion has been described34. Searches of the Arabidopsis mitochondrial genome using varied stringencies have found no simi-

**Fig. 2.** The functional gene for ribosomal protein S14 has moved from the mitochondrial to the nuclear genome in Arabidopsis. (a) In the nuclear genome (At nuc), a complete protein sequence including an N-terminal extension is encoded (BAC T3J1E5, Accession no. AC004977), whereas in the mitochondrial genome (At mt), only a rudimentary gene is found, which is interrupted by a translational stop (asterisk) and a frameshift two nucleotides downstream of the second RNA-editing site (not indicated). In the mitochondrial genome of the flowering plant Oenothera biennis (Ob mt), an intact frame is encoded. (b) RNA editing in the mitochondrial sequences changes one nucleotide in Oenothera and two in Arabidopsis. In the Arabidopsis nuclear gene, no editing is required at these positions. A translational stop is indicated by an asterisk.

**Genes for proteins involved in cytochrome c biogenesis** In the Arabidopsis mitochondrial genome, four genes specify pro-
teins that are homologous to bacterial polypeptides involved in cytochrome-c-biogenesis (Table 1). Their presence indicates that plant mitochondria use pathway I of the classification devised by Robert Kranz and co-workers35, which is inherited from the original endosymbiont, and is thus also similar to cytochrome assembly in the alpha-proteobacteria. This complex biogenesis pathway has been retained only in plant and protist mitochondria; fungal and animal mitochondria have evolved a different mode of assembly.

In the bacterial and plant mitochondrial pathway, an ABC trans-
porter with four subunits moves haem into position for presentation of FeS-containing protein complexes in the inner mitochondrial membrane, the respective polypeptides are encoded by mitochondrial or nuclear genes.

**Conserved open-reading frames** The single intron encoded reading frame, termed mat-S, found in most flowering plant mtDNAs, is encoded within the most distal intron of the nad1 gene in Arabidopsis, and probably codes for an intron maturase-like protein36. Only one other protein-coding gene in the Arabidopsis mitochondrial genome is also conserved in other plants (including algae and several protists)23 and thus probably represents a bona fide gene. However, the function of this gene (orf253) is unknown, but it should be deducible soon using evolutionary similarities to one or other of the homologs in other organisms.

**Translational start codons** In Arabidopsis, the start gene (nucleotide position 157 491–158 351) has no conventional ATG start codon, the first in-frame codon AAA specifies asparagine. The intron-encoded mat-S-reading frame begins with a GGG codon, which is not usually a start codon. Other similar cases have been seen often in maturase genes. The ccb203 gene begins with a GTG codon instead of the normal ATG. The GTG codon has been GTG codon instead of the normal ATG. The GTG codon has been
Changes by RNA editing in the mRNA alters these codons, although which is found in all other genes in the reading-frame is altered by RNA editing to the normal ATG codon, in the of 5 bacterial-like polymerase genes are absent, suggesting cofactors are all imported from the cytoplasm. The DNA- and RNA-polymerases, as well as transcription or replication nuclear encoded function for these open-reading-frames. Further parameters to translational analysis is required to define possible expression and analyzed mitochondrial genomes. Detailed transcriptional and entries in the databases to date, including all the other frames are potential genes, but do not show overt similarity to any of the entries in the databases to date, including all the other analyzed mitochondrial genomes. Detailed transcriptional and translational analysis is required to define possible expression and function for these open-reading-frames. Further parameters to positively identify these open-reading-frames as potential genes requires functional investigation beyond traditional transcriptional analysis, including a search for RNA-editing events and their consequences on the coded polypeptide sequences. This survey has been initiated, and has identified one novel gene already, the sah4 coding region.

Pseudogenes

In the Arabidopsis mitochondrial genome, two types of pseudogenes are found scattered around the genome. The first type exemplified by the rpl14 and irnA-Phe pseudogenes, are degenerated copies of genes, which have no intact copy elsewhere in the mitochondrial genome. Presumably these genes became obsolete upon successful functional gene transfer to, or substitute from, the nucleus, which now provides the mitochondrion with the respective function.

The second type is composed of partial and subsequently degenerated copies of genes, of which intact copies are present elsewhere in the mitochondrial DNA. Such pseudogenes have been found, for example, to duplicate ~300 nucleotides from the cco2 gene, and to have joined 180 nucleotides duplicated from epl12 with 90 nucleotides amplified from the nad6 gene.

All functions for replication and transcription are nuclear encoded, DNA- and RNA-polymerases, as well as transcription or replication cofactors are all imported from the cytoplasm. The Reclinomonas americana bacterial-like polymerase genes are absent, suggesting that transcription fully relies on nuclear-encoded proteins, notably the phage-type RNA polymerase with a mitochondrial import sequence, which has been identified in the Arabidopsis nuclear genome.

Only open-reading-frames that contain sequences related to retrotransposons show similarities that are typical of the respective enzymes of nuclear acid metabolism, such as reverse transcriptase and RNase H (Ref. 34). However, these similarities are imported from the nuclear genome as part of the mobile transposon sequence and are probably not functional. Considering the size of the potential proteins encoded by the largest of these reverse transcriptase–like open-reading-frames, which have ~380 amino acids, there is a distinct chance that such a protein might be functional. The respective enzymatic activity needs to be identified to corroborate gene expression, but it is thought that such proteins would have little activity because indiscriminate reverse transcriptase activity in the mitochondrial compartment would be lethal to organelle and cell functions.

Nuclear sequences in the mitochondrial genome

A major contribution to the size expansion of the mitochondrial genomes in plants can be attributed to integrated and propagated sequences originating in the nuclear genome. So far the sequence similarities have exclusively identified fragments of elements that are mobile in the nuclear genome, mostly retrotransposons. However, fragments that are similar to transposable elements of bacteria and animals have also been found; interspecific transfer might be responsible for these mobile sequences. However, the small size of some of these sequence stretches might bias the ranking and thus suggest a false relationship. More closely related elements might be found in the nuclear genome of Arabidopsis, which should enable the origin of some of these mitochondrial similarities to be traced more conclusively. Altogether ~15 kb (i.e. almost 5% of the genome) are derived from such transposable elements of the nuclear genome. To date, no other sequences of nuclear origin can be identified until more sequence information about the nuclear genome. To date, no other sequences of nuclear origin can be identified until more sequence information about the nuclear genome. To date, no other sequences of nuclear origin can be identified until more sequence information about the nuclear genome.

Plastid sequences in the mitochondrial genome

Much less DNA from the plastid genome is recognizably part of the mitochondrial genome compared with the nuclear genome. About 1.23% (i.e. 4.502 nucleotides) of the mitochondrial DNA has been imported from the plastid and integrated and propagated in the mitochondrial DNA (Table 3). The unique 12 kb plastid DNA fragment integrated in the mitochondrial DNA in maize is not found in Arabidopsis, supporting the suggestion that most individual transfer events are comparatively recent evolutionary events. A recent transfer, in preference to an ancient transfer (and subsequent loss) of the plastid sequences in Arabidopsis, is supported by the high degree of similarity between even non-functional

---

**Table 2. The RNA complement encoded by the mitochondrial genomes varies considerably between different plant species**

<table>
<thead>
<tr>
<th>Origin</th>
<th>Arabidopsis thaliana</th>
<th>Marchantia polymorpha</th>
<th>Solanum tuberosum</th>
<th>Vicia faba</th>
<th>Helianthus annuus</th>
<th>Petunia hybrida</th>
<th>Triticum aestivum</th>
<th>Zea mays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondria</td>
<td>12</td>
<td>29</td>
<td>25</td>
<td>ND</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>ND</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Nuclear</td>
<td>13</td>
<td>2</td>
<td>11</td>
<td>ND</td>
<td>8</td>
<td>ND</td>
<td>3</td>
<td>ND</td>
</tr>
</tbody>
</table>

*About a third of the mitochondrially encoded tRNA genes are actually derived from chloroplast genes integrated into the mitochondrial genome. ND = not determined.*
regions in the mitochondrial genome, such as the psbD or ndhB exon a fragments (Table 3). Where constraint by function is applied to the integrated sequence in the mitochondrion, such as to some of the RNA genes, sequence similarity is maintained even better, with virtually identical nucleotides in the mitochondrial and plastid genome. Other non-functional sequences are generally expected to deteriorate in evolutionary time, eventually beyond recognition.

Viral sequences in the mitochondrial genome

Two open-reading frames in the Arabidopsis mitochondrial genome show significant similarity to RNA-dependent RNA polymerases that are typical of RNA viruses. The greatest similarity between these open-reading frames can be seen in RNA viruses found in plant pathogenic fungi: the chestnut blight fungus (Cryophyllum parasitica) and a potato parasitic fungus (Potato virus solanum). These similarities and the observation of an analogous fragment in the mitochondrial genome in bean (Phaseolus lotus) suggest that these RNA viruses can target their RNA into the mitochondrion of fungi as well as of plants. In plants, the RNA virus, or fragments of the virus, would have to be reverse transcribed into DNA to become integrated into the genome, where they would be detectable until similarity has drifted enough to have become too low for identification.

Sequence transfers via DNA and RNA

Sequences derived from RNA viruses and integrated into the DNA genome of plant mitochondria suggest that at some point the RNA has to be reverse transcribed into mitochondrial DNA. Whether this is achieved by the expression of one or more of the reverse transcriptase sequences introduced with the retrotransposon sequences from the nuclear genome needs to be investigated thoroughly so that an enzymatic activity can be ascribed to one or more of these reading frames in the mitochondrial genome.

The identification of integrated sequences derived from RNA-viruses further supports the feasibility of nuclear acid transfers by RNA intermediates because these events definitely require a reverse transcriptase activity. Together with the verified import of RNA into mitochondria, and the dominant presence of retrotransposon sequences from the nuclear compartment and the overwhelming majority of transcribed sequences among the transfers from the plastid genome, these observations make a strong case for intercompartmental transfers via RNA intermediates being the dominant route. Successful gene transfers from the mitochondrial genome to the nucleus, such as those described for the cos2 gene in the Fabaceae, must lose any need for RNA editing and organellar intron splicing, this suggests a mature RNA intermediate.

Less frequently, DNA transfers and integration might occur, such as the insertion of the 12 kb plastid DNA in the mitochondrial genome of maize. One of the mitochondrial fragments inserted into the nucleus in Arabidopsis appears to be derived from unedited, and thus probably genomic, DNA sequences. The cos2 gene segment in the nuclear genome contains three RNA-editing sites, all of which are unedited.

Mitochondrial sequences in the nuclear genome of Arabidopsis

The mitochondrial genome sequence of Arabidopsis has provided a detailed description of the mitochondrial sequences transferred to and integrated into the nuclear genome of this plant (Table 4). The full extent of the mitochondrial fragments in the nuclear genome will become available with the sequence analysis of the nuclear DNA (Ref. 41), which is in progress.

Table 3. Several fragments of the plastid genome can be identified within the mitochondrial DNA of Arabidopsis

<table>
<thead>
<tr>
<th>Homologous to plastid sequence</th>
<th>Transferred fragment</th>
<th>Conservation (%)</th>
<th>From mitochondria</th>
<th>To mitochondria</th>
<th>From chloroplast</th>
<th>To chloroplast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondrial repeat1</td>
<td>74</td>
<td>80.7</td>
<td>36524</td>
<td>36598</td>
<td>21999</td>
<td>22869</td>
</tr>
<tr>
<td>5' Region of psaA</td>
<td>61</td>
<td>72.3</td>
<td>57393</td>
<td>53800</td>
<td>29807</td>
<td>29871</td>
</tr>
<tr>
<td>LSU RibulCo</td>
<td>583</td>
<td>96.1</td>
<td>77541</td>
<td>78124</td>
<td>55490</td>
<td>56076</td>
</tr>
<tr>
<td>psbD</td>
<td>159</td>
<td>88.2</td>
<td>102070</td>
<td>102189</td>
<td>33553</td>
<td>33470</td>
</tr>
<tr>
<td>ndhB exon a, LSU–Ann</td>
<td>952</td>
<td>95.8</td>
<td>105177</td>
<td>105969</td>
<td>108984</td>
<td>109932</td>
</tr>
<tr>
<td>ndhB exon a, LSU–Ann</td>
<td>952</td>
<td>95.8</td>
<td>105969</td>
<td>128737</td>
<td>129665</td>
<td></td>
</tr>
<tr>
<td>Spacer rps7/orf58</td>
<td>145</td>
<td>93.2</td>
<td>129208</td>
<td>129353</td>
<td>105193</td>
<td>105756</td>
</tr>
<tr>
<td>Spacer rps7/orf58</td>
<td>145</td>
<td>93.2</td>
<td>129353</td>
<td>132030</td>
<td>132974</td>
<td></td>
</tr>
<tr>
<td>LSU ribonuclease</td>
<td>56</td>
<td>100.0</td>
<td>138284</td>
<td>138340</td>
<td>135313</td>
<td>135369</td>
</tr>
<tr>
<td>Mitochondrial repeat1</td>
<td>56</td>
<td>100.0</td>
<td>138284</td>
<td>138340</td>
<td>103290</td>
<td>103346</td>
</tr>
<tr>
<td>LSU–Met</td>
<td>76</td>
<td>87.0</td>
<td>194895</td>
<td>194971</td>
<td>52061</td>
<td>52132</td>
</tr>
<tr>
<td>LSU–Top</td>
<td>442</td>
<td>93.0</td>
<td>205724</td>
<td>206166</td>
<td>23373</td>
<td>23808</td>
</tr>
<tr>
<td>rRNA–Asp</td>
<td>128</td>
<td>93.0</td>
<td>254287</td>
<td>254415</td>
<td>39544</td>
<td>39685</td>
</tr>
<tr>
<td>Photosystem-II gene</td>
<td>529</td>
<td>79.2</td>
<td>334345</td>
<td>334874</td>
<td>350</td>
<td>889</td>
</tr>
<tr>
<td>rRNA–His</td>
<td>87</td>
<td>94.3</td>
<td>359863</td>
<td>359780</td>
<td>1</td>
<td>87</td>
</tr>
<tr>
<td>orf1708</td>
<td>561</td>
<td>94.6</td>
<td>364684</td>
<td>365245</td>
<td>105193</td>
<td>105756</td>
</tr>
<tr>
<td>orf1708</td>
<td>561</td>
<td>94.6</td>
<td>364684</td>
<td>365245</td>
<td>132903</td>
<td>133467</td>
</tr>
</tbody>
</table>

*Besides the six active and employed tRNA genes, fragments of protein genes can be identified, such as the large subunit of the Rubisco (LSU Rubisco), the PSBD polypeptide and a subunit of photosystem II. The largest fragment of plastid origin is exon a of the rps7 gene, the plastid equivalent of a respiratory chain complex I gene, which covers 932 nucleotides. Altogether 4302 nucleotides are of unambiguous plastid origin at a cut-off value of 70% nucleotide identity compared to the Arabidopsis chloroplast sequence (http://genome-www.stanford.edu/Arabidopsis/). Abbreviations: orf, open-reading frame; PSBD, photosystem II subunit D.
the total size of the mitochondrial fragment of the
have also been duplicated for some distance and transferred, bringing
flanking this exon, including the partial group II intron sequences,
which are identical in the nucleus and in mitochondria. The regions
coding region is limited to the 82 nucleotides that make up exon b,
which are identical in the nucleus and in mitochondria. The regions
flanking this exon, including the partial group II intron sequences,
have also been duplicated for some distance and transferred, bringing
the total size of the mitochondrial fragment of the $exon^{a}$ locus to 596
nucleotides. The adjacent 2.9 kb at this nuclear locus are almost iden-
tical (99.9% identity) to that at a distant non-coding region of the
mitochondrial genome. This similarity is probably derived from
the mitochondrial genome by the inherent recombinogenic activity,
although this direction has not been unambiguously identified.

Among the seven clear sequence transfers to the nuclear genome, two effective gene translocations are included, the $ypt19$ (Ref. 10) and the $rps14$ gene transfers (Fig. 2). The chances of
such transfers being successful if starting from a random event are
small, because in a single step they must:

- Include the complete coding sequence.
- Eliminate any need for RNA editing and organellar intron excision.
- Add a protein import sequence and the appropriate gene
  expression signals$^{12}$. Considering the odds against this, many errors in gene transfers to
the nucleus would be expected. With only five unsuccessful mito-
ochondrial fragment transfers we are left with insufficient traces of
these transfers. These transfers must thus either be rapidly deleted
from the nuclear genome or must somehow select for the functional
completion of the numerous successful gene transfers documented
in many plants including M. polymorpha$^{11}$.

Additional similarities between mitochondrial and nuclear genomes

The overall number of mitochondrial sequences integrated into the
nuclear genome might be increased by the future assignment of
additional similarities between the two genomes (Fig. 1). About 8 kb
in 35 fragments ranging between 70 and 2885 nucleotides have
>90% identical nucleotides, the largest being a fragment inserted
together with the mitochondrial nad1 fragment in one of the nuclear
ubiquitin loci$^{13}$ as mentioned previously. None of the other
sequences can be assigned to any gene or function in either of the
genomes, which is necessary for identifying the transfer direction.

Further similarities include large portions of several BAC clones
with sequences that are almost identical to the mitochondrial
genome (BAC Accession nos: AC006225, AC007143, AC007729).

These might result from cloning events or represent genuine large
insertions in the nuclear genome, this will be decided soon by the
complete sequence of the nuclear genome.

**Evolutionary blow-up of the mitochondrial genome in plants**

The large size of plant mitochondrial genomes appears to be a sec-
ondary acquired trait that is not connected to the amount of infor-
mation encoded. During evolution from common ancestors to
land plants and algae, information was transferred to the nuclear
genome in both lineages, but considerably less information was
transferred from the mitochondrial genomes in algae than in
higher plants. This conclusion is supported by the observation that
in all instances higher plants contain less information in the form
of genes in their mitochondrial DNA than the much smaller
mtDNAs in the algae, excepting the secondary shrinking of the
mtDNA in Chlamydomonas$^{14}$. The secondary size expansion of plant mitochondrial
genomes thus appears to coincide to some extent with the move
from water to land habitats and has continued throughout the
evolution of the flowering plants. Although the increase in
genome size can be attributed to the by now classic features of
additional introns, such as the late invasion of an aggressive
group I intron$^{15}$ and larger intragenic regions, the substantial
contribution of integrated plastid and nuclear sequences is
unique to land plants. Whatever the nature of the pressure on
genome compactness might be, it has been reduced in plant
mitochondria, but maintained in many algae, protists and ani-
mals. Accordingly, plant mitochondrial genomes tolerate and
propagate excess sequences including imported plastid, nuclear
and viral DNA fragments.

An additional contributing factor that might explain the origin of
the additional sequences in plant mitochondria could be the
concomitant appearance of genome expansion and RNA editing. If
RNA-editing specificity is indeed mediated by double-stranded
RNAs obtained through sense and partial antisense RNA sequences,
a considerable amount of genomic sequence would be required to
specify the 441 nucleotides edited in the open-reading-frames of
**Arabidopsis** mitochondria.

**Acknowledgements** We thank present and past members of the laboratory for technical,
theoretical and moral support and stimulating discussions, particu-
larly Petra Brandt for her contribution to the sequence analysis.
Support by the Bundesministerium für Bildung, Wissenschaft,
Forschung und Technologie is gratefully acknowledged.

<table>
<thead>
<tr>
<th>From <strong>Arabidopsis</strong></th>
<th>To <strong>Arabidopsis</strong></th>
<th>Length (nt)</th>
<th>Conservation (%)</th>
<th>Homologous to</th>
<th>BAC GeneBank Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>22749</td>
<td>22837</td>
<td>93</td>
<td>100</td>
<td>22749</td>
<td>MQL5</td>
</tr>
<tr>
<td>42290</td>
<td>42392</td>
<td>1010</td>
<td>93</td>
<td>42290</td>
<td>F2IP24</td>
</tr>
<tr>
<td>58352</td>
<td>58631</td>
<td>2790</td>
<td>82</td>
<td>58352</td>
<td>T31E10</td>
</tr>
<tr>
<td>139123</td>
<td>138633</td>
<td>4990</td>
<td>100</td>
<td>139123</td>
<td>A013294</td>
</tr>
<tr>
<td>204299</td>
<td>204191</td>
<td>792</td>
<td>99</td>
<td>204299</td>
<td>A018117</td>
</tr>
<tr>
<td>260056</td>
<td>260305</td>
<td>1449</td>
<td>99</td>
<td>260056</td>
<td>T14C8</td>
</tr>
<tr>
<td>288635</td>
<td>289231</td>
<td>596</td>
<td>99</td>
<td>288635</td>
<td>A006219</td>
</tr>
</tbody>
</table>

*Similariestes between mitochondrial and nuclear genomes are only listed when they are part of or cover part of a clearly mitochondrial gene, and thus are identified as having moved into the nucleus: these cover 1436 nucleotides of known genes. Included are two successful gene transfers for $rps14$ and $ypt19$, for which non-functional remnants are still found in the mitochondrial organelle. Similariestes between both genomes that have no known function in either are not included, this includes 8 kb of the mitochondrial genome at 90% or more nucleotide identity. About 3 kb identified in the polyubiquitin gene transfers for $ubq13$ in the **Arabidopsis** mitochondria nucleus. A fragment of mitochondrial origin is also not included.*

\[\text{December} 1999, \text{Vol.} 4, \text{No.} 12\]
References


10. Sanchez, H. et al. (1996) Transfers of mrp8 to the nucleus involves the gain of an RSP-binding motif which may functionally replace RPS13 in Arabidopsis mitochondria. EMBO J. 15, 2128–2149


42. Kobayashi, Y. et al. (1997) Interorganellar sequence transfer: plant mitochondrial DNA in nuclear, in plant mitochondria, Cell 84, 7–10


Joachim Marienfeld is at the IfB, Schenkendorfstrasse 1, D-22085 Hamburg, Germany; Michael Unseld is at LIQBiocience AG, Im Neuenheimer Feld 515, D-69120 Heidelberg, Germany; Axel Brennicke* is at the Algenemeine Botanik, Universitat Ulm, Albert-Einstein-Allee, D-89069 Ulm, Germany.

*Author for correspondence (tel. +49 731 502 2610; fax +49 731 502 2626; e-mail:axel.brennicke@biologie.uni-ulm.de).