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*' 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[1]. The complete sequence of the mitochondrial genome in the model plant *Arabidopsis*[2] 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-biogenesis.
- 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*[3], and in the even smaller mtDNAs of green algae, such as *Prototheca wickerhamii*[4], and in some densely packed protist genomes, such as *Reclinomonas americana*[5]. 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 tRNA 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 *nad9*).
- 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 *nad10* subunit of complex I, which is encoded mitochondrially in some fungi and algae, is encoded in the nuclear genome in *Arabidopsis*[6]. 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 *orfB* gene[4,7]. 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

1360 - 1385/99/$ – see front matter © 1999 Elsevier Science Ltd. All rights reserved. PII: S1360-1385(99)01502-2

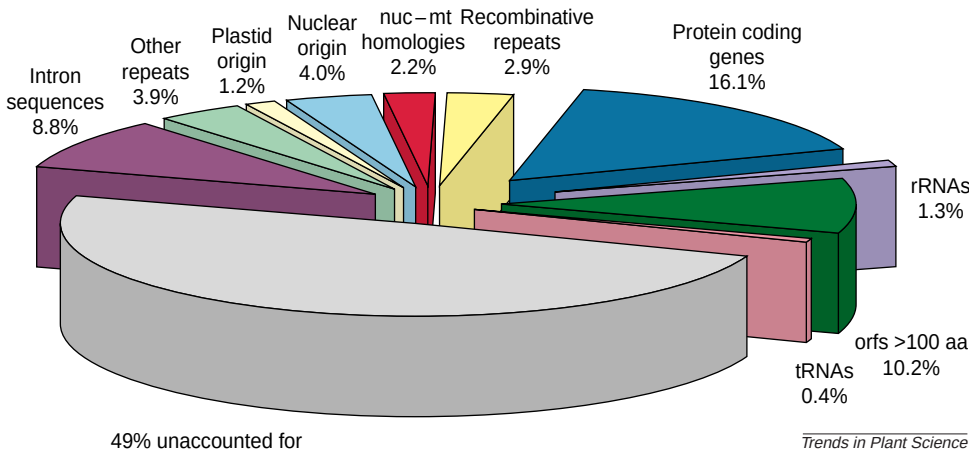

Fig. 1. Genome sequences in *Arabidopsis*. To date, half of the mitochondrial genome sequences in *Arabidopsis* are accounted for. Some of the nuclear–mitochondrial homologies are probably sequence transfers from the mitochondrial genome to the nucleus. Duplications of sequences listed in two categories of repeats might cover parts of genes or total genes. Intron sequences exclusively belonging to group II include several *trans*-splicing introns, whose total size has been averaged conservatively as 800 nucleotides. To date, unaccounted sequences show no overt similarity to any identified sequence in the databases (GenBank release 111.0; PIR-Protein release 60.0; SWISS-PROT release 37.0). Abbreviations: aa, amino acids; orfs, open-reading-frames; nuc–mt, nuclear–mitochondrial.

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[3] are not found in the *Arabidopsis* mitochondrial DNA. The genes *rps1*, *rps2*, *rps8*, *rps10*, *rps11*, *rps13*, *rps19* and *rpl6* are absent. The missing genes, *rps19* and *rpl6*, together with *rps3*, are encoded in a cistron configuration in the mitochondrial genome of other flowering plants, such as *Zea mays*, *Petunia hybrida* and *Oenothera berteriana*[8,9]. 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 RPS13 function[10]. In the bacterial ribosome, RPS13 connects RPS19 to the mRNA–rRNA, a function that might have been taken over by *rps19*.

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[11]. The *rpl2* 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 *rpl2* gene must have occurred earlier in the mitochondrial genome of ancestral land plants, because in other plant species both parts of *rpl2* are found as separate open-reading-frames (*orf*) in the mitochondrial genome[8,9].

Incomplete set of tRNA genes

In the mitochondrial genome of *Arabidopsis*, 22 tRNA genes are identified on the basis of their 'classic' cloverleaf structures. There are four duplicated genes and 18 different tRNA 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 tRNA 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 sequences transcribed in the plastid are recognized, and that outside of the respective tRNA 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 tRNA genes partially compensates for the loss of mitochondrial information. To date, only tRNA 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[12]. In addition to the functional tRNA genes, a fragment of the plastid tRNA-Ile group I intron has been integrated into the *Arabidopsis* mitochondrial genome (nucleotide position 138 280–138 360) without the respective tRNA gene sequences. The resident tRNA-Ile is of genuine mitochondrial descent.

The 18 different tRNAs 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, tRNA genes for five amino acids are lacking altogether, and these tRNAs (and possibly others) therefore have to be imported from the nucleus. Extrapolating from the situation described for the mitochondrial tRNA complements in larch (*Larix*), maize, wheat, bean and potato, where the set of endogenous and imported tRNAs has been evaluated experimentally, it is thought that several more nuclear-encoded tRNA species will be found[13,14]. In larch, more than half of the tRNAs are imported from the cytosol; in potato, 11 out of a total of 31 mitochondrial tRNA species are imported from the cytosol[13] (Table 2).

The tRNA genes for the five amino acids Ala, Arg, Leu, Thr and Val are missing in all higher plant mitochondrial genomes, and must also be imported from the cytoplasm in *Arabidopsis*. In addition the sole gene for tRNA-Phe has been lost from *Arabidopsis* mtDNA. In the mitochondrial genomes of higher plants, a group of 'plant minimal' tRNA genes specific for the six amino acids Asp, Gln, Glu, Ile, Met and Tyr are usually encoded[14]. However, in *Arabidopsis*, the imported respective plastid gene has substituted the mitochondrial tRNA-Asp of this set of amino acids.

In *Marchantia*, two tRNA species, a tRNA-Ile and a tRNA-Thr, are missing in the mitochondrial genome, and are imported products of nuclear genes[15,16]. In *Arabidopsis*, we conclude that 13 tRNAs 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 tRNA-Leu species in potato and bean mitochondria. The *Arabidopsis* sequence thus corroborates the experimental data and the conclusions drawn from the direct analysis of the tRNA population in the mitochondrial compartment in plants[13].

Table 1. Comparison of the coding information in mitochondrial genomes[a]

	Complex I								
	nad1	***nad2***	***nad3***	***nad4***	***nad4L***	***nad5***	***nad6***	***nad7***	***nad9***
Arabidopsis	324[b] (4)	307[c] (4)	118	494 (3)	100	668[b] (4)	206	393 (4)	191
Marchantia	328	488 (1)	118 (1)	495 (1)	100 (2)	669 (1)	198	ψ (2)	–
Prototheca	340	510	117	523	90	689	207	400	191
Chondrus	326	497	121	491	101	665	204	–	–
Podospora	366 (4)	556	137 (1)	519	89 (1)	593 (4)	221	–	–
Homo	316	348	114	459	98	602	175	–	–

	Complex II			Complex III	Complex IV			Complex V	
	sdh2	***sdh3***	***sdh4***	***cob***	***cox1***	***cox2***	***cox3***	***atp1***	***atp6***
Arabidopsis	–	–	94	393	527	260 (1)	265	507	385; 349
Marchantia	–	137	86	404[d] (3)	522 (9)	251 (2)	265 (2)	513 (2)	251
Prototheca	–	–	–	384	515 (3)	258	263	509	307
Chondrus	253	127	94	380	532	254	272	–	252
Podospora	–	–	–	387 (3)	540 (16)	255 (2)	269	–	264 (1)
Homo	–	–	–	377	469	227	261	–	227

			Cytochrome-c biogenesis					Protein transport	Other orfs
	atp8	***atp9***	***ccb206***	***ccb256***	***ccb453***	***ccb382***	***ccb203***	***mttB***	***orf25***
Arabidopsis	158	85	206	256	543 (1)	382	203	286[f]	192
Marchantia	172	74 (1)	–	–	–	509[e]	509[e]	244	183
Prototheca	234	74	–	–	–	–	–	234	183
Chondrus	137	76	–	–	–	–	–	262	183
Podospora	50	–	–	–	–	–	–	–	–
Homo	67	–	–	–	–	–	–	–	–

	Ribosomal RNAs			Ribosomal proteins					
	5S	***srrn***	***lrrn***	***rps1***	***rps2***	***rps3***	***rps4***	***rps7***	***rps8***
Arabidopsis	1	1	1	–	–	556 (1)	362	148	–
Marchantia	1	1	1	270	237	429	196	230	152
Prototheca	1	1	1	–	274	500	511	222	–
Chondrus	–	1	1	–	–	240	–	–	–
Podospora	–	1	1	–	–	–	–	–	–
Homo	–	1	1	–	–	–	–	–	–

	rps10	***rps11***	***rps12***	***rps13***	***rps14***	***rps19***	***rpl2***	***rpl5***	***rpl6***	***rpl16***
Arabidopsis	–	–	125	–	ψ	–	307	185	–	179
Marchantia	102	125	126	120	99 (1)	91	501 (1)	187	101	134
Prototheca	107	128	125	119	99	77	–	231	194	–
Chondrus	–	116	127	–	–	–	–	–	–	138
Podospora	–	–	–	–	–	–	–	–	–	–
Homo	–	–	–	–	–	–	–	–	–	–

[a]Mitochondrial gene contents differ between *Arabidopsis*, a liverwort (*Marchantia polymorpha*), a green alga (*Prototheca 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, MTPACG 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 Ψ. Intron numbers are in parentheses.

[b]Includes two *trans*-splicing introns. [c]Includes one *trans*-splicing intron. [d]A pseudogene is also present. [e]*Marchantia orf509* covers *ccb382* and *ccb203* of *Arabidopsis*. [f]No ATG start codon.

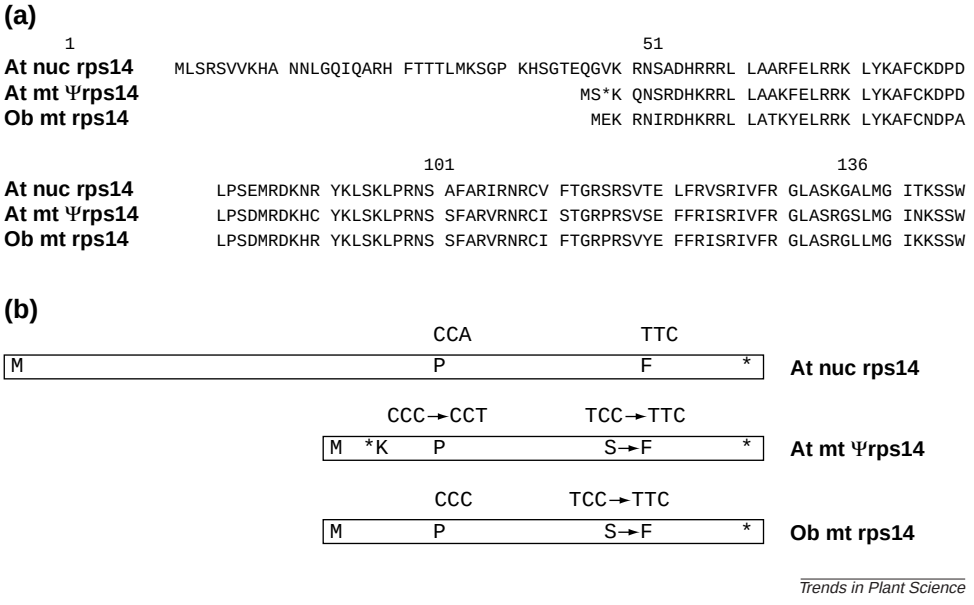

Fig. 2. The functional gene for ribosomal protein S14 has moved from the mitochondrial to the nuclear genome in *Arabidopsis*. (a) In the nuclear genome (At nuc), a complete protein sequence including an N-terminal extension is encoded (BAC T31E10; Accession no. AC004077), whereas in the mitochondrial 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 berteriana* (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.

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 described[17]. Searches of the *Arabidopsis* mitochondrial genome using varied stringencies have found no similarity to the respective maize or bacterial sequences anywhere in the genome. The respective 4.5S RNA is thus either not required for *Arabidopsis* mitochondrial function, or its gene is located in the nucleus. The import of the RNA product from the cytosol would offer the interesting prospect of another class of RNA molecules being transported into the organelle in addition to the well documented import of tRNAs.

Genes for proteins involved in cytochrome c biogenesis

In the *Arabidopsis* mitochondrial genome, four genes specify proteins 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-workers[18], 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 transporter with four subunits moves haem into position for presentation by the downstream membrane protein CCL1 (Ref. 18). Two of the ABC transporter subunits (HELB and HELC) are encoded in the mitochondrial genome of *Arabidopsis*, the genes for the other two presumably have been moved to the nucleus. The *Arabidopsis orf453*, which codes for one subunit, is split into two reading frames in *Marchantia* (*orf169* and *orf322*) without loss of function.

The membrane-anchored CCL1 protein is encoded by a single gene in *M. polymorpha* mitochondria (*orf509*), but by two genes (*orf382* and *orf202*) in *Arabidopsis* mitochondria, which code for the N- and C-terminal regions, respectively. These split frames have also been found in oilseed rape[19], but not in wheat[8,9], *Oenothera*[8,9] or *M. polymorpha*[3], suggesting that during the evolutionary history of the Brassicaceae, genomic recombination 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 oilseed rape genes.

Gene for a novel protein transport pathway

Five-years ago an open-reading-frame (*orfx*, now renamed *mttB*) was found in plant mitochondria, for which similarities and clearly homologous genes were identified in bacteria, notably in *E. coli*[20]. In this bacterium, the homologous gene is essential, because it cannot be deleted without impairing viability. Recent functional analysis of this bacterial gene revealed a novel pathway specific for membrane targeting and secretion of cofactor-containing proteins, such as iron–sulphur clusters, of which the *mttB* gene encodes one subunit[21]. A homologous pathway might be responsible for the correct localization and assembly of such FeS-containing protein complexes in the inner mitochondrial membrane, the respective polypeptides can be encoded by mitochondrial or nuclear genes.

Conserved open-reading-frames

The single intron encoded reading frame, termed *mat-r*, 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 protein[22]. Only one other protein-coding gene in the *Arabidopsis* mitochondrial genome is also conserved in other plants (including algae and several protists)[7,23] and thus probably represents a *bona fide* gene. However, the function of this gene (*orf25*) 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 *mttB* gene (nucleotide position 157 491–158 351) has no conventional ATG start codon, the first in-frame codon AAT specifies asparagine. The intron-encoded *mat-r*-reading frame begins with a GGG codon, such non-canonical start codons are seen often in maturase genes. The *ccb203* gene begins with a GTG codon instead of the normal ATG. The GTG codon has been found as a start codon in other plant mitochondrial genes, such as

Table 2. The tRNA complement encoded by the mitochondrial genomes varies considerably between different plant species[a]

Origin	*Arabidopsis thaliana*	*Marchantia polymorpha*	*Solanum tuberosum*	*Vicia faba*	*Helianthus annuus*	*Petunia hybrida*	*Triticum aestivum*	*Zea mays*
Mitochondria	12	29	25	ND	11	12	9	10
Chloroplast	6	0	5	ND	6	5	6	6
Nucleus	13	2	11	⩾ 8	ND	ND	⩾ 3	ND

[a]About a third of the mitochondrially encoded tRNA genes are actually derived from chloroplast genes integrated into the mitochondial genome. ND = not determined.

in the *rpl16* gene in *Oenothera*[24] and in the *rpl16* gene in a mutant of *Arabidopsis*[25]. The genomic ACG start codon in the *nad1* open-reading-frame is altered by RNA editing to the normal ATG codon, which is found in all other genes in the *Arabidopsis* mitochondrial DNA. Thus, in total, there are three genes with unusual start codons in the *Arabidopsis* mitochondrial genome, with no evidence that changes by RNA editing in the mRNA alters these codons, although their functionality remains to be tested experimentally.

Novel reading frames – additional genes in unique open-reading-frames?

In the *Arabidopsis* mitochondrial genome, 156 reading frames with >100 amino acids have been identified in the 367 kb. Of the open-reading-frames with >150 amino acids, eight are derived by duplications and extensions from 'classic' mitochondrial genes that are encoded intact elsewhere in the mitochondrial genome. Probably most of these open-reading-frames have no function in the organelle and are just chance arrangements[26,27]. The other reading frames are potential genes, but do not show overt similarity to any of the entries in the databases to date, including all the other analysed 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[28–30] and their consequences on the coded polypeptide sequences. This survey has been initiated, and has identified one novel gene already, the *sdh4* coding region[31,32].

Pseudogenes

In the *Arabidopsis* mitochondrial genome, two types of pseudogenes are found scattered around the genome. The first type exemplified by the *rps14* and *tRNA-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 mitochondria 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 *cox2* gene, and to have joined 180 nucleotides duplicated from *rps12* with 90 nucleotides amplified from the *nad6* gene[26].

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*[5] 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[33].

Only open-reading-frames that contain sequences related to retrotransposons show similarities that are typical of the respective enzymes of nucleic 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 becomes available, or the origin of other similarities between the two genomes can be ascertained.

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[35] 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 3. Several fragments of the plastid genome can be identified within the mitochondrial DNA of *Arabidopsis*[a]

Homologous to plastid sequence	Transferred fragment	Conservation (%)	Position			
			From mitochondria	To mitochondria	From chloroplast	To chloroplast
	74	90.7	38524	38598	21995	22069
Mitochondrial repeat1	49	82.0	47163	47212	79428	79475
5′ Region of rps4	61	72.3	53739	53800	29807	29871
tRNA–Ser	174	80.8	62286	62460	44765	44944
LSU RuBisCo	583	96.1	77541	78124	55490	56076
psbD	119	88.2	102070	102189	33553	33670
ndhB exon a, tRNA–Asn	952	95.8	105017	105969	108984	109932
ndhB exon a, tRNA–Asn		95.8	105017	105969	128717	129665
	51	73.1	111231	111282	75769	75820
Spacer rps7/orf58	145	93.2	129208	129353	105675	105819
Spacer rps7/orf58		93.2	129208	129353	132830	132974
tRNA–Ile intron	56	100.0	138284	138340	103290	103346
tRNA–Ile intron	56	100.0	138284	138340	135313	135369
Mitochondrial repeat1		82.0	181328	181377	79428	79475
tRNA–Met	76	87.0	194895	194971	52061	52132
	442	93.0	205724	206166	23373	23808
tRNA–Trp	276	71.8	250004	250280	66165	66438
	128	93.0	254287	254415	39544	39668
tRNA–Asp	83	94.0	275104	275187	29788	29871
Photosystem-II gene	529	79.2	334345	334874	350	889
tRNA–His	87	94.3	359663	359750	1	87
orf1708	561	94.6	364684	365245	105193	105756
orf1708		94.6	364684	365245	132903	133467

[a]Besides the six active and employed tRNA genes, fragments of protein genes can be identifed, 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 *ndhB* gene, the plastid equivalent of a respiratory chain complex I gene, which covers 952 traceable nucleotides. Altogether 4502 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.

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 tRNA genes, sequence similarity is maintained even better, with virtually identical nucleotides in the mitochondrial and plastid genes. 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 (*Cryptonectria parasitica*) and a potato parasitic fungus (*Rhizoctonia solani*)[36]. These similarities and the observation of an analogous fragment in the mitochondrial genome in bean (*Vicia faba*) suggest that these RNA viruses can target their RNA into the mitochondria 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 nucleic acid transfers by RNA intermediates[37,38] because these events definitely require a reverse transcriptase activity. Together with the verified import of tRNAs into mitochondria[14], 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 *cox2* gene in the Fabaceae, must lose any need for RNA editing and organellar intron splicing, this suggests a mature RNA intermediate[39,40].

Less frequently, DNA transfers and integration might occur, such as the insertion of the 12 kb plastid DNA in the mitochondrial genome of maize[35]. One of the mitochondrial fragments inserted into the nucleus in *Arabidopsis* appears to be derived from unedited, and thus probably genomic, DNA sequences. The *cox2* gene segment in the nuclear genome (Table 4) 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 4. Sequences of mitochondrial origin in the nuclear genome of *Arabidopsis*[a]

From *Arabidopsis* mitochondria	To *Arabidopsis* mitochondria	Length (nt)	Conservation (%)	Homologous to	BAC	GenBank Accession no.
28749	28857	47	89	*rps19*	MQL5	AB018117
42290	42392	102	96	*cox2* exon a	F21P24	AC004401
58352	58631	279	82	*rps14*	T31E10	AC004077
138123	138433	310	100	*nad7* exon e	TM018A10	AF013294
204299	204191	47	89	*rps19*	MQL5	AB018117
260156	260305	149	98	*rps12*	T14C8	AC006219
288635	289231	596	99	*nad1* exon b		L05401

[a]Similarities 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 *rps19*, for which non-functional remnants are still found in the mitochondrial organelle. Similarities between both genomes that have no known function in either are not included, this includes 8 kb of the mitochondrial genome at 80% or more nucleotide identity. About 3 kb identified in the polyubiquitin gene *ubq13* in the *Arabidopsis* nucleus alongside a fragment of mitochondrial origin[42] are also not included.

To date, with about half of the nuclear genome available, only seven examples can be identified clearly. These include the mitochondrial *nad1* fragment in one of the nuclear ubiquitin loci[42] (the polyubiquitin *ubq13* gene). Similarity to the mitochondrial *nad1* 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 *nad1* locus to 596 nucleotides. The adjacent 2.9 kb at this nuclear locus are almost identical (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 *rps19* (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 mitochondrial 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*[43].

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[42] 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 secondary acquired trait that is not connected to the amount of information 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*[7].

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[44] 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 animals. 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[45].

Acknowledgements

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

References

1 Quetier, F. and Vedel, F. (1977) Heterogeneous population of mitochondrial DNA molecules in higher plants, *Nature* 268, 365–368

2 Unseld, M. *et al.* (1997) The mitochondrial genome in *Arabidopsis* contains 57 genes in 366,924 nucleotides, *Nat. Genet.* 15, 57–61

3 Oda, K. *et al.* (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA: a primitive form of plant mitochondrial genome, *J. Mol. Biol.* 223, 1–7

4 Wolff, G. *et al.* (1994) Complete sequence of the mitochondrial DNA of the chlorophyte alga *Prototheca wickerhamii*, *J. Mol. Biol.* 237, 75–86

5 Lang, B.L. *et al.* (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature, *Nature* 378, 493–497

6 Heiser, V., Brennicke, A. and Grohmann, L. (1996) The plant mitochondrial 22 kDa (PSST) subunit of respiratory chain complex I is encoded by a nuclear gene with enhanced transcript level in flowers, *Plant Mol. Biol.* 31, 1195–1204

7 Gray, M.W. *et al.* (1998) Genome structure and gene content in protist mitochondrial DNAs, *Nucleic Acids Res.* 26, 865–878

8 Levings, C.S., III and Brown, G.G. (1989) Molecular biology of plant mitochondria, *Cell* 56, 171–179

9 Schuster, W. and Brennicke, A. (1994) The plant mitochondrial genome: structure, information content, RNA editing and gene transfer, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 45, 61–78

10 Sanchez, H. *et al.* (1996) Transfer of rps19 to the nucleus involves the gain of an RNP-binding motif which may functionally replace RPS13 in *Arabidopsis* mitochondria, *EMBO J.* 15, 2138–2149

11 Aubert, D., Bisanz-Seyer, C. and Herzog, M. (1992) Mitochondrial *rps14* is a transcribed and edited pseudogene in *Arabidopsis*, *Plant Mol. Biol.* 20, 1169–1174

12 Brennicke, A. *et al.* (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants, *FEBS Lett.* 325, 140–145

13 Kumar, R. *et al.* (1996) Striking differences in mitochondrial tRNA import between different plant species, *Mol. Gen. Genet.* 252, 404–411

14 Marechal-Drouard, L., Weil, J-H. and Dietrich, A. (1993) Transfer RNAs and transfer RNA genes in plants, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 44, 13–32

15 Akashi, K. *et al.* (1996) Occurrence of nuclear encoded tRNA-Ile in mitochondria of the liverwort *Marchantia polymorpha*, *Curr. Genet.* 30, 181–185

16 Akashi, K. *et al.* (1997) Accumulation of nuclear-encoded tRNA-Tyr(AGU) in mitochondria of the liverwort *Marchantia polymorpha*, *Biochim. Biophys. Acta* 1350, 262–266

17 Yang, A. and Mulligan, R.M. (1996) Identification of a 4.5S-like ribonucleoprotein in maize mitochondria, *Nucleic Acids Res.* 24, 3601–3606

18 Kranz, R. *et al.* (1998) Molecular mechanisms of cytochrome c biogenesis: three distinct systems, *Mol. Microbiol.* 29, 383–396

19 Handa, H., Bonnard, G. and Grienenberger, J-M. (1996) The rapeseed mitochondrial gene encoding a homologue of the bacterial protein Ccl1 is divided into two independently transcribed reading frames, *Mol. Gen. Genet.* 252, 292–302

20 Sünkel, S., Brennicke, A. and Knoop, V. (1994) RNA editing of a conserved reading frame in plant mitochondria increases its similarity to two overlapping reading frames in *Escherichia coli*, *Mol. Gen. Genet.* 242, 65–72

21 Weiner, J.H. *et al.* (1998) A novel and ubiquitous system for membrane targeting and secretion of cofactor-containing proteins, *Cell* 93, 93–101

22 Thomson, M.C. *et al.* (1994) RNA editing of mat-r transcripts in maize and soybean increase similarity of the encoded protein to fungal and bryophyte group II intron maturases: evidence that mat-r encodes a functional protein, *Nucleic Acids Res.* 22, 5745–5752

23 Leblanc, C. *et al.* (1997) Origin and evolution of mitochondria: what have we learnt from red algae? *Curr. Genet.* 31, 193–207

24 Bock, H., Brennicke, A. and Schuster, W. (1994) Rps3 and rpl16 genes do not overlap in *Oenothera* mitochondria: GTG as a potential translation initiation codon in plant mitochondria? *Plant Mol. Biol.* 24, 811–818

25 Sakamoto, W. *et al.* (1997) An unusual mitochondrial atp9–rpl16 cotranscript found in the maternal distorted leaf mutant of *Arabidopsis*: implication of GUG as an initiation codon in plant mitochondria, *Plant Cell Physiol.* 38, 975–979

26 Marienfeld, J.R. *et al.* (1997) Mosaic open reading frames in the *Arabidopsis* mitochondrial genome, *Biol. Chem.* 378, 859–862

27 Marienfeld, J.R., Unseld, M. and Brennicke, A. (1996) Genomic recombination of the mitochondrial atp6 gene in *Arabidopsis* at the protein processing site creates two different presequences, *DNA Res.* 3, 287–290

28 Covello, P.S. and Gray, M.W. (1989) RNA editing in plant mitochondria, *Nature* 341, 662–666

29 Gualberto, J. *et al.* (1989) RNA editing in wheat mitochondria results in the conservation of protein sequences, *Nature* 341, 660–662

30 Hiesel, R. *et al.* (1989) RNA editing in plant mitochondria, *Science* 246, 1632–1634

31 Giegè, P. *et al.* (1998) An ordered *Arabidopsis* mitochondrial cDNA library on high-density filters allows rapid systematic analysis of plant gene expression: a pilot study, *Plant J.* 15, 721–726

32 Giegè, P., Knoop, V. and Brennicke, A. (1998) A complex II subunit gene in plant mitochondria, *Curr. Genet.* 34, 313–317

33 Hedtke, B., Börner, T. and Weihe, A. (1997) Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*, *Science* 277, 809–811

34 Knoop, V. *et al.* (1996) *copia*-, *gypsy*- and LINE-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis*, *Genetics* 142, 579–585

35 Stern, D.B. and Lonsdale, D.M. (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common, *Nature* 229, 698–702

36 Marienfeld, J.R. *et al.* (1997) Viral nucleic acid sequence transfer between fungi and plants, *Trends Genet.* 13, 260–261

37 Schuster, W. and Brennicke, A. (1987) Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *Oenothera*: is genetic information transferred between organelles via RNA? *EMBO J.* 6, 2857–2863

38 Schuster, W. and Brennicke, A. (1988) Interorganellar sequence transfer: plant mitochondrial DNA is nuclear, is plastid, is mitochondrial, *Plant Science* 54, 1–10

39 Covello, P.S. and Gray, M.W. (1992) Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: evidence for RNA-mediated gene transfer, *EMBO J.* 11, 3815–3820

40 Nugent, J.M. and Palmer, J.D. (1991) RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution, *Cell* 66, 473–481

41 Meyerowitz, E.M. and Somerville, C.R. (1994) Arabidopsis, Cold Spring Harbor Laboratory Press, New York

42 Sun, C-W. and Callis, J. (1993) Recent stable insertion of mitochondrial DNA into an *Arabidopsis* polyubiquitin gene by nonhomologous recombination, *Plant Cell* 5, 97–107

43 Kobayashi, Y. *et al.* (1997) Interorganellar gene transfer in bryophytes: the functional *nad7* gene is nuclear encoded in *Marchantia polymorpha*, *Mol. Gen. Genet.* 256, 589–592

44 Cho, Y. *et al.* (1998) Explosive invasion of plant mitochondria by a group I intron, *Proc. Natl. Acad. Sci. U. S. A.* 95, 14244–14249

45 Giegè, P. and Brennicke, A. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in open reading frames, *Proc. Natl. Acad. Sci. U. S. A.* (in press)

Joachim Marienfeld is at the IbF, Schenkendorffstrabe 1, D-22085 Hamburg, Germany; Michael Unseld is at LIONbioscience AG, Im Neuenheimer Feld 515, D-69120 Heidelberg, Germany; Axel Brennicke* is at the Allgemeine Botanik, Universität 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).