Mol Gen Genet (1989) 217:185–194

MGG

The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals

Junzou Hiratsuka [1], Hiroaki Shimada [1], Robert Whittier [1], Takashi Ishibashi [1], Masahiro Sakamoto [1], Masao Mori [1], Chihiro Kondo [1], Yasuko Honji [1], Chong-Rong Sun*, Bing-Yuan Meng [1], Yu-Qing Li [1], Akira Kanno [2], Yoko Nishizawa [2], Atsushi Hirai [2], Kazuo Shinozaki [1], and Masahiro Sugiura [1]

[1] Center for Gene Research, Nagoya University, Chikusa, Nagoya 464-01, Japan
[2] Faculty of Agriculture, Nagoya University, Chikusa, Nagoya 464-01, Japan

Summary. The entire chloroplast genome of the monocot rice (*Oryza sativa*) has been sequenced and comprises 134525 bp. Predicted genes have been identified along with open reading frames (ORFs) conserved between rice and the previously sequenced chloroplast genomes, a dicot, tobacco (*Nicotiana tabacum*), and a liverwort (*Marchantia polymorpha*). The same complement of 30 tRNA and 4 rRNA genes has been conserved between rice and tobacco. Most ORFs extensively conserved between *N. tabacum* and *M. polymorpha* are also conserved intact in rice. However, several such ORFs are entirely absent in rice, or present only in severely truncated form. Structural changes are also apparent in the genome relative to tobacco. The inverted repeats, characteristic of chloroplast genome structure, have expanded outward to include several genes present only once per genome in tobacco and liverwort and the large single copy region has undergone a series of inversions which predate the divergence of the cereals. A chimeric tRNA pseudogene overlaps an apparent endpoint of the largest inversion, and a model invoking illegitimate recombination between tRNA genes is proposed which accounts simultaneously for the origin of this pseudogene, the large inversion and the creation of repeated sequences near the inversion endpoints.

Key words: Conserved open reading frames – Monocots – Chloroplast DNA – Sequence duplication – Multimer formation

Introduction

Chloroplasts are intracellular organelles present in plants, which contain the entire enzymatic machinery for photosynthesis. Similarly to mitochondria, chloroplasts contain their own genome distinct from the nucleus. Among land plants, this genome is generally comprised of a single circular DNA molecule, 120–160 kbp in length, divided structurally into a large single copy (LSC) and small single copy region (SSC) separated from each other by inverted repeats (IRs), which are present in two copies per genome (Palmer 1985).

To date, the entire chloroplast genomic sequence has been determined in two other plants, tobacco (*Nicotiana tabacum*) and a liverwort (*Marchantia polymorpha*) (Shinozaki et al. 1986; Ohyama et al. 1986; 1988). These studies and others indicate that chloroplasts probably code for all of their own tRNA and rRNA molecules and some, but not all, of the proteins required by their genetic apparatus or for photosynthesis (Posno et al. 1984). The remainder of these proteins must be encoded in the nucleus and imported from the cytoplasm (Schmidt and Mishkind 1986).

Although many of the open reading frames (ORFs) found within chloroplast genomes have been identified by comparing their predicted translation products with known sequences of chloroplast or prokaryotic proteins, many ORFs could not be identified. One method to help identify which of the remaining ORFs probably code for genuine chloroplast proteins is to utilize evolutionary filtering (Zurawski and Clegg 1987; Wolfe and Sharp 1988). Sequences conserved over large evolutionary distance are inferred to be evolving under constraint, presumably because they perform some necessary function. The fossil record indicates that monocots and dicots diverged 100–140 million years ago, and it has been estimated that flowering plants last shared a common ancestor with liverworts 350–400 million years ago (Stewart 1983). The published sequences of liverwort and tobacco chloroplast genomes were recently compared and conserved ORFs were identified (Wolfe and Sharp 1988). However, sequence data from monocot chloroplast genomes have been only fragmentary. To determine the entire coding potential, a complete sequence determination is required.

Studies of a fern, a gymnosperm and several angiosperms have established a consensus chloroplast gene order among vascular land plants identical to that found in tobacco (Palmer and Stein 1986). The data available for the cereal grasses wheat and maize show that their chloroplast genomes have diverged from the consensus gene order through a series of overlapping inversions within the LSC (Quigley and Weil 1985; Howe 1985).

* *Present address:* Department of Biochemistry, Fudan University, Shanghai, China

Offprint requests to: M. Sugiura

Abbreviations: PSII, photosystem II; PSI, photosystem I; RuBisCo, ribulose 1,5-bisphosphate carboxylase; IR_A and IR_B denote the inverted repeat regions distal and proximal to *ndh*F respectively

In view of the overriding world-wide agricultural importance of monocots, in particular the cereals, and their evolutionary distance from dicots, we have determined the complete sequence of the rice chloroplast genome and compared it to other complete chloroplast sequences in order to identify conserved and non-conserved genes. In addition, we have found that rice chloroplasts share the genome rearrangements observed in other grasses. Close inspection has suggested a model for the first overlapping inversion which accounts for several otherwise puzzling aspects of its border sequences. Features of this model may have general relevance to mitochondrial and chloroplast genome evolution.

Materials and methods

DNA sequence determination was accomplished by the dideoxy chain termination method (Sanger et al. 1977) utilizing cloned Klenow fragment (Takara) or modified T7 DNA polymerase (USBC). A previously described clone bank of overlapping fragments from *Oryza sativa*, cv. Nihonbare was used (Hirai et al. 1985), supplemented by additional clones from the same cultivar as necessary. Computer-assisted analysis was carried out using UWGCG and IDEAS software on a MicroVax II computer and GENETYX software on a NEC PC-98XL personal computer. The protein sequence comparison program used treated introduced gaps as mismatches, so that gap penalties were proportional to gap size. Percentage amino acid residue identity was then calculated for the portion of the proteins judged homologous on the basis of amino acid residue identity and conservative substitution.

Results and discussion

Conserved genes

The entire rice chloroplast DNA sequence was determined by the dideoxy chain termination method using a clone bank of overlapping fragments (Hirai et al. 1985). By virtue of recombination between the two IRs, chloroplast genomes exist as equimolar mixtures of two isomers differing from each other in the relative orientation of their two single copy regions (Stein et al. 1986). To simplify comparison, Fig. 1 and Table 1 depict the isomeric configuration corresponding to the one previously cloned and presented for tobacco (Sugiura et al. 1986; Shinozaki et al. 1986). The sequences determined for rice and tobacco were aligned and rice genes were identified by homology with their tobacco counterparts. Identified genes, ORFs and other notable sequence features are listed in Table 1 in order of their positions within the genome, starting from the IR_A-LSC junction and proceeding counterclockwise around the chromosome as depicted in Fig. 1.

Thirty tRNA genes were identified, corresponding in chromosomal location and anticodon to those previously identified in tobacco (Wakasugi et al. 1986). Computer-aided searches were performed in order to identify any additional tRNA genes, but no other likely functional tRNA

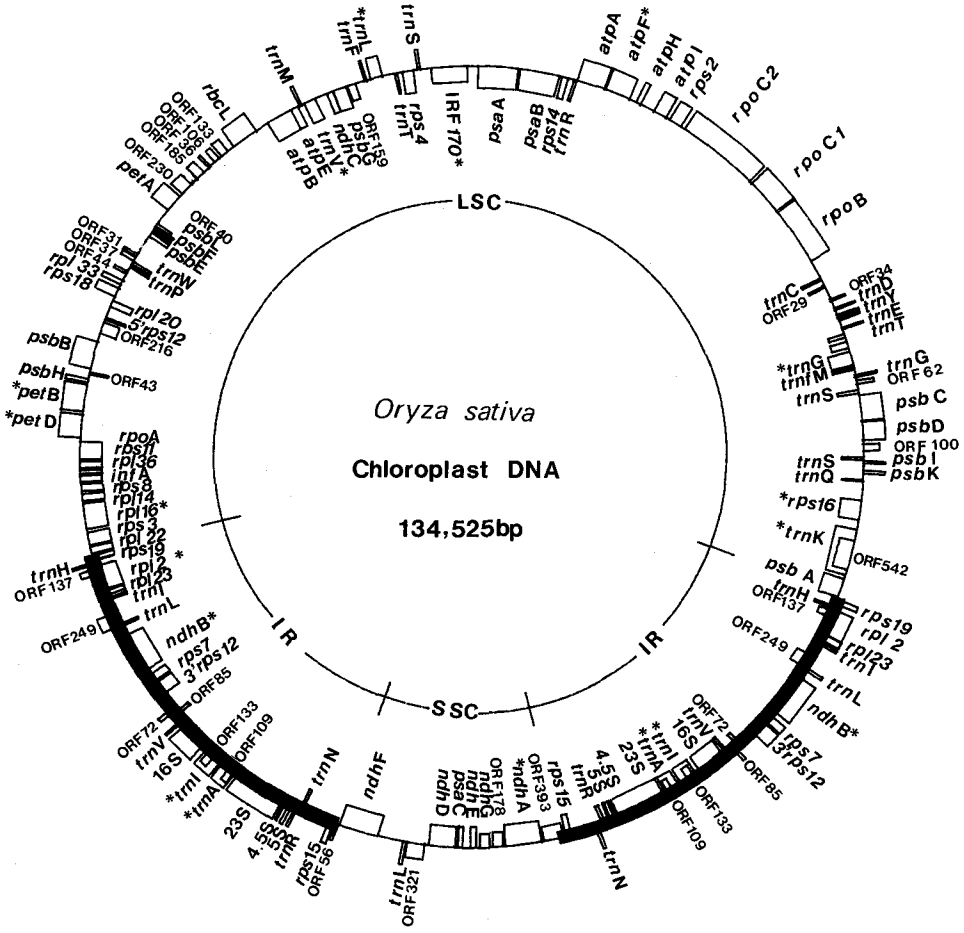

Fig. 1. Genetic circle map of the *Oryza sativa* chloroplast genome drawn to scale. Genes shown on the outside of the circle are encoded on the A strand and transcribed counter-clockwise. Genes on the inside are encoded on the B strand and transcribed clockwise. Asterisks denote split genes. LSC, large single copy region; IR, inverted repeat; SSC, small single copy region

Table 1. Genes and open reading frames (ORFs) of the rice chloroplast genome

Gene	Comment	Strand	Coding region	
			Start	End
*psb*A	PSII 32kDa protein	B	1143	82
*trn*K	tRNA-Lys (UUU) 3′exon	B	1397	1373
	5′exon	B	3931	3902
ORF542	ORF within *trn*K intron	B	3296	1668
*rps*16	Ribosomal protein S16 3′exon	B	4704	4487
	5′exon	B	5553	5514
*trn*Q	tRNA-Gln (UUG)	B	6687	6615
*psb*K	PSII K protein	A	7033	7218
*psb*I	PSII I protein	A	7608	7718
*trn*S	tRNA-Ser (GCU)	B	7916	7829
ORF100		A	8349	8651
*psb*D	PSII D2 protein	A	8900	9961
*psb*C	PSII 44 kDa protein	A	9909	11326
*trn*S	tRNA-Ser (UGA)	B	11590	11503
ORF62		A	11937	12125
*trn*G	tRNA-Gly (GCC)	A	12331	12401
ψ*trn*G	Pseudogene for tRNA-Gly (GCC)	B	12601	12528
*trn*fM	tRNA-fMet (CAU)	B	12911	12839
*trn*G	tRNA-Gly (UCC) 3′exon	B	13050	13003
	5′exon	B	13752	13729
	Homology to 3′rps12 intron	A	14277	14367
ORF91		B	14346	14077
*trn*T	tRNA-Thr (GGU)	A	15060	15131
ψ*trn*T	Pseudogene for tRNA-Thr (GGU)	A	15128	15200
*trn*E	tRNA-Glu (UUC)	A	15650	15722
*trn*Y	tRNA-Tyr (GUA)	A	15784	15867
*trn*D	tRNA-Asp (GUC)	A	16231	16304
ORF34		A	16685	16789
	Inverted repeat		16880	16927
ORF29		B	17645	17556
*trn*C	tRNA-Cys (GCA)	B	18129	18059
*rpo*B	RNA polymerase beta subunit	A	19214	22441
*rpo*C1	RNA polymerase beta′ subunit-1	A	22479	24527
*rpo*C2	RNA polymerase beta′ subunit-2	A	24727	29268
*rps*2	Ribosomal protein S2	A	29540	30250
*atp*I	ATPase a subunit	A	30501	31244
*atp*H	ATPase III subunit	A	32039	32284
*atp*F	ATPase I subunit 5′exon	A	32741	32885
	3′exon	A	33714	34111
*atp*A	ATPase alpha subunit	A	34210	35733
*trn*R	tRNA-Arg (UCU)	B	35937	35866
ψ*trn*fM/G	Chimeric pseudogene fusing *trn*fM (CAU) to *trn*G (UCC)	B	36147	36073
*rps*14	Ribosomal protein S14	B	36620	36309
*psa*B	PSI P700 apoprotein A2	B	38972	36768
*psa*A	PSI P700 apoprotein A1	B	41250	38998
IRF170	Intron-containing reading frame 3rd exon	B	42009	41851
	2nd exon	B	42968	42739
	1st exon	B	43837	43714
*trn*S	tRNA-Ser (GGA)	A	44438	44524
*rps*4	Ribosomal protein S4	B	45415	44810
*trn*T	tRNA-Thr (UGU)	B	45787	45715
*trn*L	tRNA-Leu (UAA) 5′exon	A	46558	46592
	3′exon	A	47133	47182
*trn*F	tRNA-Phe (GAA)	A	47425	47497
ORF159		B	48471	47988
*psb*G	PSII G protein	B	49309	48569
*ndh*C	NADH dehydrogenase ND3	B	49662	49300
*trn*V	tRNA-Val (UAC) 3′exon	B	50401	50367
	5′exon	B	51037	50999
*trn*M	tRNA-Met (CAU)	A	51219	51291
*atp*E	ATPase epsilon subunit	B	51817	51404
*atp*B	ATPase beta subunit	B	53310	51814
*rbc*L	RuBisCO large subunit	A	54095	55528
ORF42	Homology to rpl23	A	55806	55934
ORF133		A	55958	56359

Table 1 (continued)

Gene	Comment	Strand	Coding region	
			Start	End
ORF106		A	56553	56873
	Inverted repeat		57010	57075
ORF36		A	57222	57332
ORF185		A	57702	58259
ORF85		A	58290	58547
ORF230		A	58677	59369
*pet*A	Cytochrome f	A	59601	60563
ORF40		B	61687	61565
*psb*L	PSII L protein	B	61930	61814
*psb*F	PSII cytochrome b559	B	62072	61953
*psb*E	PSII cytochrome b559	B	62334	62083
	Homology to 3′rps12 intron	A	62982	63027
ORF31		A	63531	63626
ORF37		A	63799	63912
*trn*W	tRNA-Trp (CCA)	B	64102	64029
*trn*P	tRNA-Pro (UGG)	B	64303	64229
ORF44		A	64622	64756
*rpl*33	Ribosomal protein L33	A	65198	65398
*rps*18	Ribosomal protein S18	A	65641	66132
*rpl*20	Ribosomal protein L20	B	66714	66355
5′*rps*12	Ribosomal protein S12 exon-1	B	67503	67390
ORF216		B	68288	67638
*psb*B	PSII P680 apoprotein	A	68799	70325
*psb*H	PSII 10 kDa phosphoprotein	A	70881	71102
ORF35		A	70490	70597
ORF43		B	70777	70646
*pet*B	Cytochrome B6 5′exon	A	71232	71237
	3′exon	A	72049	72690
*pet*D	Cytochrome b/f complex subunit–5′exon	A	72883	72890
	3′exon	A	73635	74110
*rpo*A	RNA polymerase alpha subunit	B	75343	74330
*rps*11	Ribosomal protein S11	B	75838	75407
*rpl*36	Ribosomal protein L36	B	76126	76013
*inf*A	Initiation factor 1	B	76561	76238
*rps*8	Ribosomal protein S8	B	77108	76698
*rpl*14	Ribosomal protein L14	B	77619	77248
*rpl*16	Ribosomal protein L16 3′exon	B	78130	77729
	5′exon	B	79198	79190
*rps*3	Ribosomal protein S3	B	80063	79344
*rpl*22	Ribosomal protein L22	B	80568	80119
Junction of the long single copy region (LSC) with inverted repeat region B (IR_B)			80592	80593
*rps*19	Ribosomal protein S19	B	80918	80637
ORF82		B	81163	80915
*trn*H	tRNA-His (GUG)	A	81050	81124
ORF137		A	81286	81699
*rpl*2	Ribosomal protein L2	B	82664	81180
*rpl*23	Ribosomal protein L23	B	82964	82683
*trn*I	tRNA-Ile (CAU)	B	83212	83139
ORF28		A	83534	83620
ORF64		A	83685	83879
ORF249		A	83997	84746
*trn*L	tRNA-Leu (CAA)	B	84791	84711
*ndh*B	NADH dehydrogenase ND2 3′exon	B	86150	85395
	5′exon	B	87639	86863
*rps*7	Ribosomal protein S7	B	88414	87944
3′*rps*12	Ribosomal protein S12 exon-3	B	88501	88473
	exon-2	B	89273	89042
ORF72		B	90442	90227
ORF85		B	90658	90501
*trn*V	tRNA-Val (GAC)	A	90996	91067
16S rDNA	16S rRNA	A	91299	92789
*trn*I	tRNA-Ile (GAU) 5′exon	A	93100	93136
	3′exon	A	94084	94118
ORF133	ORF within trnI intron	A	93241	93642

Table 1 (continued)

Gene	Comment	Strand	Coding region	
			Start	End
*trn*A	tRNA-Ala (UGC) 5′exon	A	94183	94220
	3′exon	A	95033	95067
ORF109	ORF within trnA intron	A	94683	95012
23S rDNA	23S rRNA	A	95213	98096
4.5S rDNA	4.5S rRNA	A	98192	98286
5S rDNA	5S rRNA	A	98514	98634
*trn*R	tRNA-Arg (ACG)	A	98891	98964
ORF23		A	99016	99087
*trn*N	tRNA-Asn (GUU)	B	99287	99216
ORF63		A	100206	100397
*rps*15	Ribosomal protein S15	A	100818	101090
ORF56		A	101229	101399
Junction of IR_B with the short single copy region (SSC)			101391	101392
*ndh*F	NADH dehydrogenase ND5	B	103637	101433
ORF63		A	104352	104543
*trn*L	tRNA-Leu (UAG)	A	105074	105153
ORF321		A	105236	106201
*ndh*D	NADH dehydrogenase ND4	B	107900	106398
*psa*C	PSI C protein	B	108265	108020
*ndh*E	NADH dehydrogenase ND4L	B	109017	108712
*ndh*G	NADH dehydrogenase ND6	B	109757	109227
ORF178		B	110536	110000
*ndh*A	NADH dehydrogenase ND1 3′exon	B	111169	110631
	5′exon	B	112706	112157
Junction of the SSC with inverted repeat region A (IR_A)			113726	113727
ORF393		B	113889	112708
*rps*15	Ribosomal protein S15	B	114300	114028
ORF63		B	114912	114721
*trn*N	tRNA-Asn (GUU)	A	115831	115902
ORF23		B	116102	116031
*trn*R	tRNA-Arg (ACG)	B	116227	116154
5S rDNA	5S rRNA	B	116604	116484
4.5S rDNA	4.5S rRNA	B	116926	116832
23S rDNA	23S rRNA	B	119905	117022
*trn*A	tRNA-Ala (UGC) 3′exon	B	120085	120051
	5′exon	B	120935	120898
ORF109	ORF within *trn*A intron	B	120435	120106
*trn*I	tRNA-Ile (GAU) 3′exon	B	121034	121000
	5′exon	B	122018	121982
ORF133	ORF within *trn*I intron	B	121877	121476
16S rDNA	16S rRNA	B	123819	122329
*trn*V	tRNA-Val (GAC)	B	124122	124051
ORF85		A	124360	124617
ORF72		A	124676	124891
3′*rps*12	Ribosomal protein S12 exon-2	A	125845	126076
	exon-3	A	126617	126645
*rps*7	Ribosomal protein S7	A	126704	127174
*ndh*B	NADH dehydrogenase ND2 5′exon	A	127479	128255
	3′exon	A	128968	129723
*trn*L	tRNA-Leu (CAA)	A	130327	130407
ORF249		B	131121	130372
ORF64		B	131433	131239
ORF28		B	131584	131498
*trn*I	tRNA-Ile (CAU)	A	131906	131979
*rpl*23	Ribosomal protein L23	A	132154	132435
*rpl*2	Ribosomal protein L2	A	132454	133938
ORF137		B	133832	133418
*trn*H	tRNA-His (GUG)	B	134068	133991
ORF82		A	133955	134203
*rps*19	Ribosomal protein S19	A	134200	134481
Junction of IR_A with the LSC			134525	1

Total number of nucleotides = 134525

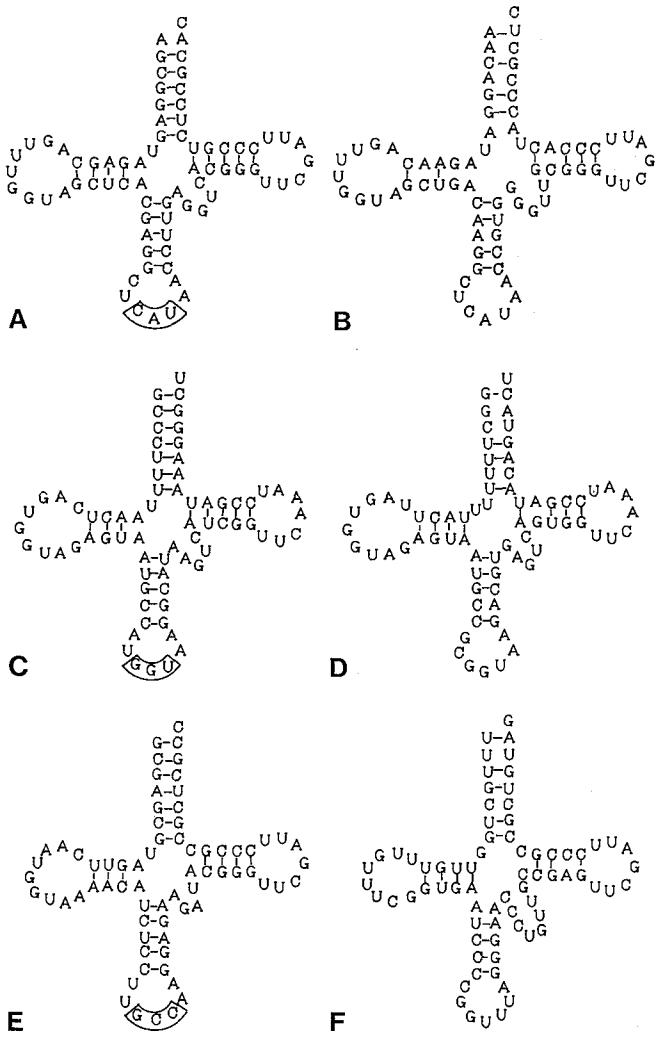

Fig. 2A–F. Rice tRNA cloverleaf structures and hypothetical cloverleaf structures for derived pseudogenes. **A** *trn*fM; **B** *ψtrn*fM/G (UCC); **C** *trn*T; **D** *ψtrn*T; **E** *trn*G; **F** *ψtrn*G

genes were detected. However, three apparent pseudogenes were found (Fig. 2). One is a chimeric pseudogene which will be discussed in greater detail below in regard to genomic rearrangements. The other two pseudogenes are located very close to the presumed functional genes they resemble: *ψtrn*T is present as a tandem direct repeat, overlapping at its 5′ end the 3′ end of *trn*T (GGU); *ψtrn*G is present as an inverted repeat downstream of *trn*G (GCC) so that 126 bp separate their 3′ termini, as drawn. The inability of these sequences to form strongly base-paired stems suggests that they are non-functional, regardless of whether they are expressed. Rice and tobacco chloroplasts, therefore, probably code for the same complement of functional tRNA molecules.

Ribosomal RNA 16S, 23S, 4.5S and 5S genes were also conserved, as were putative protein coding genes previously identified in tobacco. In addition, the predicted translation products of rice ORFs were compared to the predicted products of ORFs occurring at similar positions in tobacco and liverwort (Table 2). Most ORFs shared between tobacco and liverwort were present in rice also (compare Wolfe and Sharp 1988). As expected, predicted protein products shared greater identity with their counterparts in

Table 2. Homology between predicted ORF translation products

ORF	Position in rice		Percentage identity with rice	Number of compared amino acid residues
O.s./N.t. /M.p.[a]	Strand	Start	*N.t.*	*M.p.*
542/509A/370	B	3296	59/512	32/374
62/ 62 / 62	A	11937	87/ 62	81/ 62
34/ 34b/ 34	A	16687	100/ 34	88/ 33
29/ 29 / 29	B	17647	100/ 29	86/ 29
170/168 /167	B	43837	95/168	84/167
159/158 /169	B	48471	85/157	72/156
106/512 /316	A	56553	50/ 74	45/ 74
36/ 36b /36b	A	57222	89/ 36	71/ 31
185/184 /184	A	57702	80/185	60/183
230/229 /434	A	58 677	62/229	46/231
40/ 40 / 40	B	61687	90/ 40	85/ 40
31/ 31 / 31	A	63531	90/ 31	77/ 31
37/ 37 / 37	A	63799	100/ 37	86/ 37
44/ 44 /42b	A	64622	89/ 44	76/ 42
216/196 /203	B	68288	69/196	61/196
35/34a/ 35	A	70490	100/ 33	89/ 35
43/ 43 / 43	B	70777	98/ 43	84/ 43
26/228 /464	B	99400	84/ 25	68/ 25
321/313 /320	A	105236	70/317	54/317
176/176 /*ndh*6	B	109757	76/167	55/177
178/167 /*frx*B	B	110536	81/167	77/158
393/393 /392	B	113889	89/393	83/391

[a] *O.s., Oryza sativa*; *N.t., Nicotiana tabacum*; *M.p., Marchantia polymorpha*

tobacco than in liverwort, and homology between rice and liverwort ORFs was similar in extent to that between tobacco and liverwort ORFs. Exceptions to this trend include ORF106, discussed more fully below, and ORF216. The tobacco and liverwort ORFs corresponding to this latter ORF share 79% predicted amino acid identity with each other, notably more than either shares with rice ORF216. The loss of introns from this ORF (discussed below) coincides with an apparent loss of some evolutionary constraint upon the coding sequences.

Several ORFs conserved between tobacco and liverwort are entirely absent from rice, or present only in severely truncated form. Downstream of *rbc*L, ORF512 of tobacco and ORF316 of liverwort, whose putative translation products share 68% amino acid identity with each other over 279 residues, are represented in rice only by an ORF of 106 codons, whose predicted translation product would share no greater than 50% identity over 74 residues with either ORF. In addition, upstream of *trn*L (UAG) within the SSC, ORF55 of tobacco and ORF69 of liverwort, whose predicted protein products share 70% amino acid identity over 47 residues, have no counterpart in the rice chloroplast. Elsewhere in a region corresponding to the SSC of tobacco, but contained within the IRs of rice, a deletion of about 4.8 kbp has occurred. This deletion spans a region corresponding to most of tobacco ORFs 1244 and 228, as well as all of tobacco ORF273. Though this last missing ORF bears limited similarity to *Escherichia coli ssb*, it is not found in liverwort either. Tobacco ORF1244 exhibits sporadic homology with liverwort ORF1068, but the conserved portions are entirely removed from rice by this deletion. The most highly conserved portion of ORF228, in

```
O.s.                                fMNRYWFDTNNGSCFSMLRIQMYPQFI'
                                    :::::::::::: ::.:::.::::.
N.t. RKPLYKEKNELIKLKFFLWPNYRLEDLACMNRYWFDTNNGSRFSILRIHMYPQLKIN'
     .: . :.:  . :.: :.:..::.:::::::::.::.: :::::::.::..::: :
M.p.QKKFSKTK--IKKIKRFIWASYRFEDLACMNRFWFNTINGSRFSMLRFRMYPSLLT'
```

Fig. 3. Comparison of the predicted rice (*O.s.*) ORF26 amino acid sequence with the -COOH terminal sequences predicted for tobacco (*N.t.*) ORF228 and liverwort (*M.p.*) ORF464. Symbols: :, amino acid identity; ·, indicates conservative substitution; ′, indicates termination codon

contrast, is retained in rice and contains an initiation codon (ATG) so that an ORF of 26 codons is present in rice which shares 84% identity over 25 residues with the COOH terminal region of the predicted tobacco ORF228 product (Fig. 3). Since several very short proteins are encoded within chloroplasts (e.g. Murata et al. 1988), it is not unreasonable to speculate that this might also code for a genuine protein product. Of nine nucleotide substitutions observed between rice and tobacco in this region, seven fall within the third codon position. Although statistically a small sample, this pattern is fairly typical of conserved coding sequences (Zurawski and Clegg 1987). Alternatively, the high degree of conservation between chloroplast sequences in this area may be attributable to some promoter function for the transcript initiating upstream of *rps*15.

Gene structure

Of 20 distinct introns previously demonstrated or tentatively identified in the tobacco chloroplast genome, 17 are also present in rice and 3 are absent, even though the surrounding putative coding sequences are conserved. The genes missing introns are *rpo*C1 (1 intron) and ORF216 (2 introns). In all 3 cases the deleted sequences coincide precisely with predicted (*rpo*C1) or established (ORF216; Kohchi et al. 1988) intron borders so that the mechanism of intron loss may have involved a processed RNA intermediate. Since these introns are present in both liverwort and tobacco, we conclude that they have been specifically lost in the rice lineage. Whether the loss of these introns extends to other monocot species is not yet known. The divided structure of *rps*12, observed in other chloroplast genomes, is also maintained in rice. *Trans*-splicing between the 5′ and 3′ transcripts is presumably required to form a functional mRNA, as has been shown in tobacco (Zaita et al. 1987).

*rpl*2 is predicted to initiate with an ATG start codon in tobacco and spinach (Tanaka et al. 1986; Zurawski et al. 1984). The corresponding codon in rice, as well as maize, is ACG, and no in-frame ATG codon is present upstream or downstream within the first two-thirds of the putative coding sequence (data not shown; McLaughlin and Larrinua 1988). However, the sequence is otherwise highly conserved and contains no frameshifts within the coding region, suggesting that the gene is expressed. ACG was recently documented to function as an initiation codon for the Sendai virus in vivo (Gupta and Patwardhan 1988), and earlier in vitro evidence had suggested that it might so function in prokaryotic organisms as well (Thach et al. 1966). Other non-ATG start codons occasionally utilized, GUG, UUG and AUU, are utilized less efficiently than AUG (Kozak 1983) and a near optimal initiation codon context may be required to compensate. Indeed, rice and maize both share a sequence – AAAGGAGA – ending 13 bp upstream of the putative ACG start codon, which appears to be a much closer fit to the ribosome binding site consensus sequence than the corresponding tobacco sequence, which lacks G residues. It seems likely, therefore, that chloroplasts can utilize ACG as an initiation codon, as has previously been proposed for *psb*L in tobacco and spinach chloroplasts (Wolfe and Sharp 1988).

Genome structure

Relative to tobacco, some structural reorganization of the rice chloroplast genome is evident; the IRs encompass additional genes, kilobase-magnitude deletions have occurred, and major rearrangements are evident within the LSC. Despite the apparent expansion of the IRs into sequences corresponding to both the LSC and SSC in tobacco, deletions more than offset the additional sequence; the length of the IRs in rice (20799 bp) is shorter than in tobacco (25339 bp). In detail, *trn*H, *rps*19, *rps*15 and ORF393 are all present entirely as single copy sequences in tobacco, but the first three are completely contained within the IRs of rice, while ORF393 spans the rice SSC-IR border. Thus, of four single copy sequence-IR borders, three exhibit flux of sequence between chloroplast genomic compartments in this rice-tobacco comparison.

The largest deletions apparent in the rice chloroplast genome occur roughly at opposite ends of the IRs. The 4.8 kb deletion mentioned above occurs between *rps*15 and *trn*N, while a series of three closely spaced deletions at the other end of the IRs account together for about 6 kb of sequence, corresponding in tobacco to ORFs 1708 and 581. Rice ORF249 occupies this same region, but displays no discernible homology with tobacco, except where it overlaps *trn*L (CAA). It may be noted that both liverwort and tobacco possess extraordinarily long ORFs in this region, whose predicted translation products bear some homology with each other. Another deletion occurs in the region between *trn*D and ORF34, removing about 700 bp from the rice sequence. Numerous smaller apparent deletions also exist. Finally, in the region downstream of *rbc*L very little similarity is discernible before ORF36, other than the moderate homology pointed out between tobacco ORF512 and rice ORF106. The rice sequence is about 1.4 kb shorter than tobacco in this region. One potential cause of this sequence divergence may be the presence of a pseudo *rpl*23 gene. The pseudogene bears an internal deletion of 135 nucleotides, a frameshift and several point mutations relative to the IR copies. While this manuscript was in preparation, the same sequence, detected first in rice mitochondria, was described by Moon et al. (1988). Among related wheat species this second *rpl*23 sequence and the surrounding region appear a hotspot for deletions (Ogihara et al. 1988), but whether this results from an increased occurrence of such events or whether they are merely better tolerated is not known.

Extensive rearrangements are evident within the rice chloroplast LSC (Fig. 4). An identical gene arrangement was described for the wheat (*Triticum aestivum*) chloroplast LSC, and a scheme proposed whereby the wheat and tobacco structures could be derived from each other by three inversions (Quigley and Weil 1985). Starting from a tobacco-like ancestral gene order, an inversion occurred encompassing about 28 kb, its seeming endpoints lying between *trn*G (UCC) and *trn*R (UCU) at one end and *rps*14

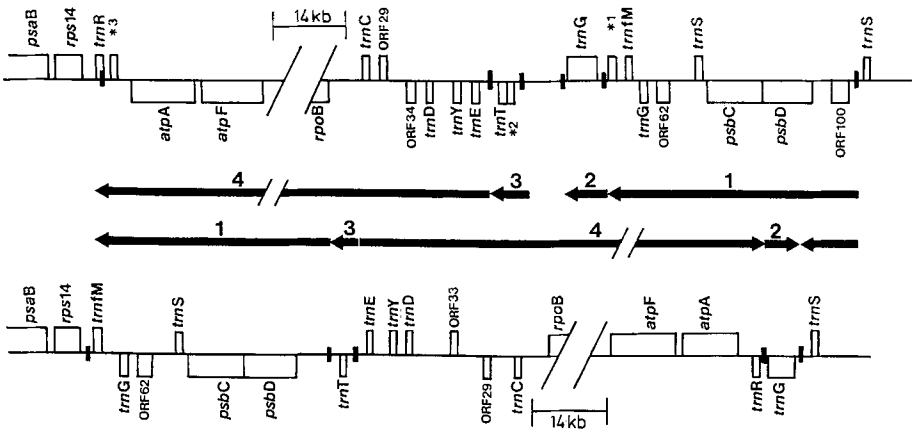

Fig. 4. Comparison of gene order and orientation within the LSCs of rice (upper gene map) and tobacco (lower gene map). *Numbered arrows* indicate the position and orientation of corresponding regions within the two genomes. *Solid vertical slashes* indicate the approximate position of rearrangement breakpoints. Sequences labeled with *numbered asterisks* in the rice gene map indicate the positions of tRNA pseudogenes: *1, ψ*trn*G; *2, ψ*trn*T; *3, ψ*trn*fM/G

and *trn*fM at the other. This led to the intermediate genome partially depicted in Fig. 6D. Two further inversions, one largely overlapping the 28 kb event, subsequently gave rise to the gene arrangement observed in rice and wheat chloroplasts. The approximate endpoints of these inversions are also indicated in Fig. 6D. Although the linear order of genes along the chromosome could be derived from tobacco by just two sequential overlapping inversions, the third event is required to account for the orientation of *trn*T (GGU). For clarity this *trn*T inversion is depicted among the latter events, although its actual order of occurrence is not clear. Studies of maize are also consistent with conservation of this gene arrangement among the cereals (Palmer and Thompson 1982), but the monocots *Spirodela oligorhiza* (duckweed) and *Oncidium excavatum* (an orchid) appear to share the tobacco gene arrangement (deHeij et al. 1983; Palmer et al. 1988). Thus, these rearrangements may be confined within the grass family.

The breakpoints delimiting the smaller inversions cannot be assigned precisely by comparison with the tobacco sequence. Between *trn*S and *psb*D, homology with tobacco is lost after nucleotide 7917 and not resumed until nucleotide 8275. The other end of this inversion event occurs between *trn*G (UCC) and *trn*T (GGU), but since the priority of this inversion with respect to the inversion of *trn*T is not evident, sequences participating in the larger inversion event might also presently lie between *trn*T and *trn*E. Sequences which cannot be assigned by homology with tobacco to either side of the breakpoints include rice nucleotides 13756–15037 and, upstream of *trn*T, nucleotides 15201–15579. Computer-assisted comparison of all these sequences revealed no inverted or direct repeats more striking than might be expected by random chance among similarly AT rich sequences.

At and near the endpoints of the largest inversion are three repeats; one lies immediately upstream of *trn*fMet, another lies just downstream, and the third occurs downstream of *rps*14, immediately upstream of a tRNA pseudogene. The same repeats have been described in wheat, and it was proposed that these repeats, inverted with respect to each other in the ancestral genome, mediated the first inversion via homologous recombination. Thus, a tobacco-like genome was converted to an intermediate one (Howe 1985). However, this model assumes the pre-existence of the repeated sequences and fails to give a clear explanation for the creation of the neighboring pseudogene. The resemblance of this pseudogene to *trn*fM has been previously

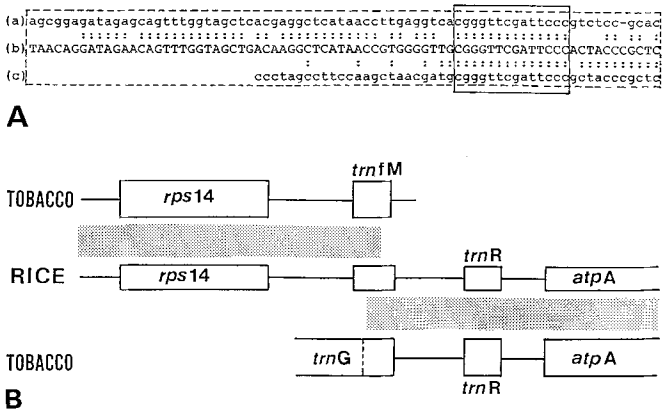

Fig. 5A and B. Sequence homology of the rice chimeric pseudogene with apparent parental *trn* genes and map of the rice region surrounding the chimeric pseudogene indicating homology with corresponding tobacco regions. **A** DNA sequence comparison of rice ψ*trn*fM/G (b) with the apparent parental tRNA genes, rice *trn*fM (a) and *trn*G(UCC) (c). **B** Homology of the ψ*trn*fM/G containing region of rice with corresponding regions of tobacco. The chimeric pseudogene itself is indicated by the *unlabeled box* in the rice gene map

noted in wheat and maize (Howe 1985; Rodermel et al. 1987). A close comparison of this tRNA pseudogene with other rice *trn* genes clearly indicates that the pseudogene is chimeric, deriving its 5′ sequence from *trn*fMet and its 3′ sequence from the second exon of *trn*G (UCC) (Fig. 5).

The chloroplast genome is highly polyploid and recent studies demonstrate that roughly one-third of its genome exists in multimeric form (X.-W. Deng, R.A. Wing and W. Gruissem, personal communication). We propose that intermolecular recombination along a 14 bp homologous region shared by the parental *trn* genes gave rise to an abnormal multimer possessing mirror-image chimeric *trn*fM/G and *trn*G/fM pseudogenes, as well as non-recombined copies of each gene (Fig. 6). Very shortly thereafter, a single deletion removed the *trn*G/fM pseudogene along with most of the duplicate genome, leaving behind only *trn*G and *trn*fM, now adjacent to each other, to create a viable genome. Thus, a single sequence of events created the chimeric pseudogene, the first inversion and the *trn*fM upstream repeat. Repeats shorter than 14 bp have been implicated in illegitimate recombination events in wheat chloroplasts (Ogihara et al. 1988) and so the proposed scheme

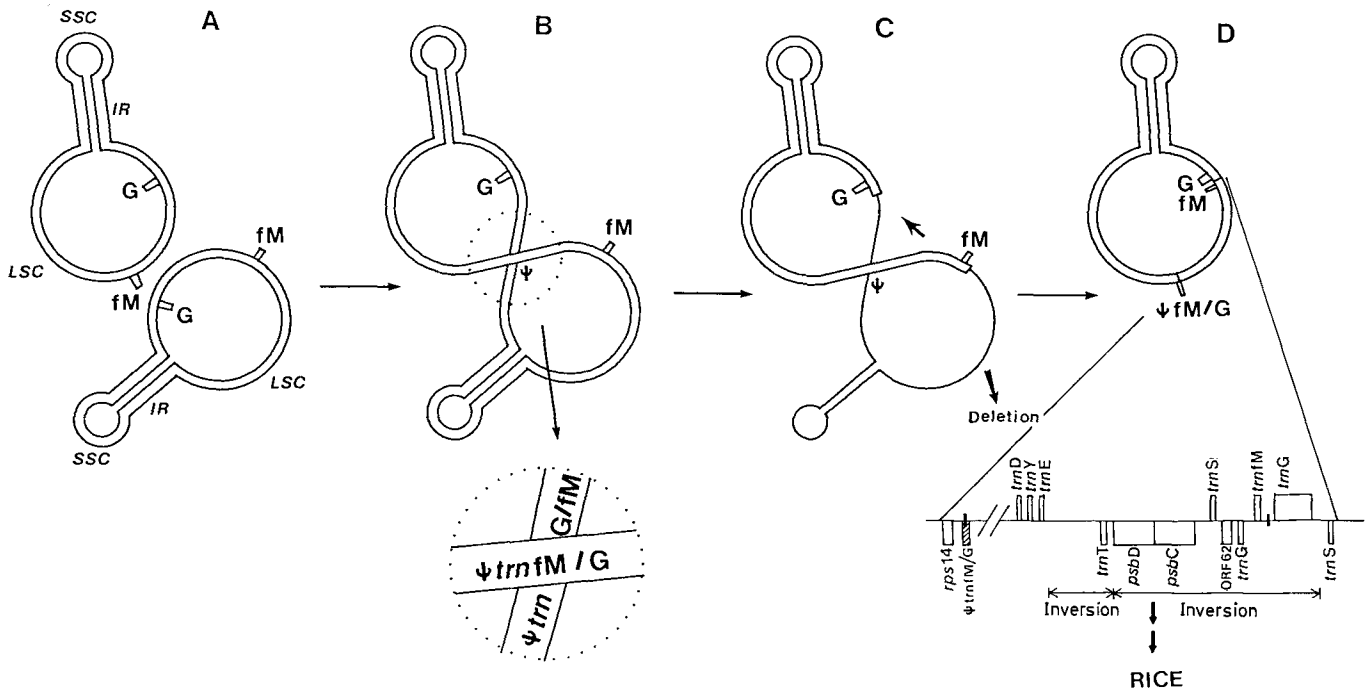

Fig. 6. Proposed model to account for chimeric pseudogene formation, origin of *trn*fM associated upstream repeat sequence, and inversion of the ancestral vascular plant chloroplast consensus gene order to yield a tobacco-rice intermediate gene order. (A) Alignment of 2 chloroplast chromosomal monomers. (B) Intermolecular recombination between *trn*fM and *trn*G(UCC) to produce an inverted dimer joined through mirror-image chimeric tRNA pseudogenes (inset). (C) Subsequent area of deletion indicated by *single line*. (D) Resulting chromosomal monomer. Inset shows corresponding tobacco-rice intermediate gene order, with the resultant inversion breakpoints indicated by *small vertical slashes*. *Arrows* below inset indicate the approximate endpoints of further inversions still required to create the gene order and orientations observed in rice and other cereals

is both consistent with the present derived genome structure and reasonable in the mechanisms it invokes.

Elements of this model may have general relevance to organelle genome evolution. It may be noted that deletion of most of a genome from a normal head to tail circular dimer would result in formation of a tandem direct repeat. Although long direct repeats are predicted to be unstable, short repeats might be viable, particularly if a selective advantage were conferred. Such an event duplicating *trn*fM and its upstream sequence may explain the creation of the third repeat sequence observed downstream of *trn*fM. If so, a second deletion must be invoked to account for the absence of the duplicated *trn*fM gene itself. The remaining duplicated sequence may indeed function beneficially in its present position. The corresponding sequence in tobacco serves as a transcript terminus (Meng et al. 1988) and so it may be active in transcript processing (Stern and Gruissem 1987).

Acknowledgements. We thank M. Tanaka, N. Hayashida, T. Wakasugi, T. Matsubayashi and K. Torazawa for valuable suggestions and discussions, and Dr. A. Hirose for continuous encouragement. This work was supported in part by a Grant-in-Aid for Special Distinguished Research from the Ministry of Education, Science and Culture, and a grant from the Ministry of Agriculture, Forestry and Fisheries. The first eight authors listed were employed by the Mitsui Plant Biotechnology Research Institute, and conducted this research at the Nagoya University Center for Gene Research under the auspices of the System of Joint Research with Industry, which is administered by the Ministry of Education, Science and Culture.

References

deHeij HT, Lustig H, Moeskops D-JM, Bovenberg WA, Bisanz C, Groot GSP (1983) Chloroplast DNAs of *Spinacia, Petunia* and *Spirodela* have a similar gene organization. Curr Genet 7:1–6

Gupta KC, Patwardhan S (1988) ACG, the initiator codon for a Sendai virus protein. J Biol Chem 263:8553–8556

Hirai A, Ishibashi T, Morikami A, Iwatsuki N, Shinozaki K, Sugiura M (1985) Rice chloroplast DNA: a physical map and the location of the genes for the large subunit of ribulose 1,5-bisphosphate carboxylase and the 32 KD photosystem II reaction center protein. Theor Appl Genet 70:117–122

Howe CJ (1985) The endpoints of an inversion in wheat chloroplast DNA are associated with short repeated sequences containing homology to *att*. Curr Genet 10:139–145

Kohchi T, Ogura Y, Umesono K, Yamada Y, Komano Y, Ozeki H, Ohyama K (1988) Ordered processing and splicing in a polycistronic transcript in liverwort chloroplasts. Curr Genet 14:147–154

Kozak M (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev 47:1–45

McLaughlin WE, Larrinua IM (1988) The sequence of the maize plastid encoded *rpl*23 locus. Nucleic Acids Res 16:8183

Meng BY, Tanaka M, Wakasugi T, Ohme M, Shinozaki K, Sugiura M (1988) Cotranscription of the genes encoding two P700 chlorophyll *a* apoproteins with the gene for ribosomal protein CS14: determination of the transcriptional initiation site by in vitro capping. Curr Genet 14:395–400

Moon E, Kao T-H, Wu R (1988) Rice mitochondrial genome contains a rearranged chloroplast gene cluster. Mol Gen Genet 213:247–253

Murata N, Miyao M, Hayashida N, Hidaka T, Sugiura M (1988) Identification of a new gene in the chloroplast genome encoding a low-molecular-mass polypeptide of photosystem II complex. FEBS Lett 235:283–288

Ogihara Y, Terachi T, Sasakuma T (1988) Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc Natl Acad Sci USA 85:8573–8577

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S-I, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572–574

Ohyama K, Fukuzawa H, Kohchi T, Sano T, Sano S, Shirai H, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S-I, Inokuchi H, Ozeki H (1988) Structure and organization of *Marchantia polymorpha* chloroplast genome. I. Cloning and gene identification. J Mol Biol 203:281–298

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19:325–354

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10:823–833

Palmer JD, Thompson WF (1982) Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell 29:537–550

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Missouri Bot Garden 75:1180–1218

Posno M, van Noort M, Debise R, Groot GSP (1984) Isolation, characterization, phosphorylation and site of synthesis of *Spinacia* chloroplast ribosomal proteins. Curr Genet 8:147–154

Quigley F, Weil JH (1985) organization and sequence of five tRNA genes and of an unidentified reading frame in the wheat chloroplast genome: evidence for gene rearrangements during the evolution of chloroplast genomes. Curr Genet 9:495–503

Rodermel S, Orlin P, Bogorad L (1987) The transcription termination region between two convergently-transcribed photoregulated operons in the maize plastid chromosome contains *rps*14, *trn*R (UCC) and a putative *trnfM* pseudogene. Nucleic Acids Res 15:5493

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schmidt GW, Mishkind ML (1986) The transport of proteins into chloroplasts. Annu Rev Biochem 55:879–912

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Stein DB, Palmer JD, Thompson WF (1986) Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus *Osmunda*. Curr Genet 10:835–841

Stern DB, Gruissem W (1987) Control of plastid gene expression: 3′ inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. Cell 51:1145–1157

Stewart WN (1983) Paleobotany and the evolution of plants. Cambridge University Press, Cambridge, UK

Sugiura M, Shinozaki K, Zaita N, Kusuda M, Kumano M (1986) Clone bank of the tobacco (*Nicotiana tabacum*) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes. Plant Sci 44:211–216

Tanaka M, Wakasugi T, Sugita M, Shinozaki K, Sugiura M (1986) Genes for the eight ribosomal proteins are clustered on the chloroplast genome of tobacco (*Nicotiana tabacum*): similarity to the S10 and *spc* operons of *Escherichia coli*. Proc Natl Acad Sci USA 83:6030–6034

Thach RE, Sundararajan TA, Dewey DF, Brown JC, Doty P (1966) Translation of synthetic messenger RNA. Cold Spring Harbor Symp Quant Biol 31:85–97

Wakasugi T, Ohme M, Shinozaki K, Sugiura M (1986) Structures of tobacco chloroplast genes for $tRNA^{Ile}$ (CAU), $tRNA^{Leu}$ (CAA), $tRNA^{Cys}$ (GCA), $tRNA^{Ser}$ (UGA) and $tRNA^{Thr}$ (GGU): a compilation of tRNA genes from tobacco chloroplasts. Plant Mol Biol 7:385–392

Wolfe KH, Sharp PM (1988) Identification of functional open reading frames in chloroplast genomes. Gene 66:215–222

Zaita N, Torazawa K, Shinozaki K, Sugiura M (1987) *Trans* splicing in vivo: joining of transcripts from the 'divided' gene for ribosomal protein S12 in the chloroplasts of tobacco. FEBS Lett 210:153–156

Zurawski G, Clegg MT (1987) Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annu Rev Plant Physiol 38:391–418

Zurawski G, Bottomley W, Whitfield PR (1984) Junctions of the large single copy region and the inverted repeats in *Spinacia oleracea* and *Nicotiana debneyi* chloroplast DNA: sequence of the genes for $tRNA^{His}$ and the ribosomal proteins S19 and L2. Nucleic Acids Res 12:6547–6558

Communicated by R.G. Herrmann

Received January 2, 1989

Note added in proof

The work of Deng et al. on chloroplast chromosomal multimers is now in press in Proc. Natl. Acad. Sci. USA 86. ORF37 has been identified in maize as the *pet*E gene, which encodes subunit 5 of the chloroplast cytochrome b_6-f complex (J. Haley and L. Bogorad, 1989, Proc. Natl. Acad. Sci. USA 86:1534–1538). ORF393 has been found to be homologous with a nuclear coded 49 kd subunit of bovine mitochondrial NADH-ubiquinone reductase (I.M. Fearnley, M.J. Runswick and J.E. Walker, 1989, EMBO J. 8:665–672). The entire sequence reported here has been communicated to the EMBO database.