Plant Molecular Biology (2005) 59:309–322
DOI 10.1007/s11103-005-8882-0

Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes

Christopher Saski[1], Seung-Bum Lee[2], Henry Daniell[2,*], Todd C. Wood[1], Jeffrey Tomkins[1], Hyi-Gyung Kim[1] and Robert K. Jansen[3]
[1]*Clemson University Genomics Institute, Clemson University, Biosystems Research Complex, 51 New Cherry Street, Clemson, SC, 29634, USA;* [2]*Department of Molecular Biology and Microbiology, Biomolecular Science, University of Central Florida, Building #20, 4000 Central Florida Blvd, Orlando, FL, 32816-2364, USA (*author for correspondence; e-mail daniell@mail.ucf.edu);* [3]*Patterson Laboratories 141, Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas, Austin, TX, 78712, USA*

Received 19 April 2005; accepted in revised form 16 June 2005

Key words: evolution, organization, plastid genomes, repeated sequences, Soybean

Abstract

Lack of complete chloroplast genome sequences is still one of the major limitations to extending chloroplast genetic engineering technology to useful crops. Therefore, we sequenced the soybean chloroplast genome and compared it to the other completely sequenced legumes, *Lotus* and *Medicago*. The chloroplast genome of *Glycine* is 152,218 basepairs (bp) in length, including a pair of inverted repeats of 25,574 bp of identical sequence separated by a small single copy region of 17,895 bp and a large single copy region of 83,175 bp. The genome contains 111 unique genes, and 19 of these are duplicated in the inverted repeat (IR). Comparisons of *Glycine*, *Lotus* and *Medicago* confirm the organization of legume chloroplast genomes based on previous studies. Gene content of the three legumes is nearly identical. The *rpl22* gene is missing from all three legumes, and *Medicago* is missing *rps16* and one copy of the IR. Gene order in *Glycine*, *Lotus*, and *Medicago* differs from the usual gene order for angiosperm chloroplast genomes by the presence of a single, large inversion of 51 kilobases (kb). Detailed analyses of repeated sequences indicate that many of the *Glycine* repeats that are located in the intergenic spacer regions and introns occur in the same location in the other legumes and in *Arabidopsis*, suggesting that they may play some functional role. The presence of small repeats of *psbA* and *rbcL* in legumes that have lost one copy of the IR indicate that this loss has only occurred once during the evolutionary history of legumes.

Introduction

The chloroplast is a plant organelle that contains the entire enzymatic machinery for photosynthesis. In addition to photosynthesis, several other biochemical pathways are present within chloroplasts, including biosynthesis of fatty acids, amino acids, pigments, and vitamins. The chloroplast genome of land plants generally has a highly conserved organization (Palmer, 1991; Raubeson and Jansen, 2005) with most composed of a single circular chromosome with a quadripartite structure that includes two copies of an inverted repeat (IR) that separate the large and small single copy regions (LSC and SSC). Our knowledge of the organization and evolution of chloroplast genomes has been expanding rapidly because of the large numbers of completely sequenced genomes published over the

past 10 years. Currently there are 44 completely sequenced plastid genomes (Jansen *et al*., 2005), and 27 of these are from various land plant lineages, with the best representation (20) from flowering plants. Comparative studies indicate that chloroplast genomes of land plants are highly conserved in both gene order and gene content. Several lineages of land plants have chloroplast DNAs (cpDNAs) with multiple rearrangements, including *Pinus* (Wakasugi *et al*., 1994), and the angiosperm families Campanulaceae (Cosner *et al*., 1997), Fabaceae (Palmer *et al*., 1987b, 1988; Milligan *et al*., 1989; Kato *et al*., 2000), Geraniaceae (Palmer *et al*., 1987a), and Lobeliaceae (Knox and Palmer, 1998). In most of these studies, comparisons of gene content and order have been made between distantly related taxa because only one genome sequence was available from groups with rearranged genomes. One exception is in the grasses where chloroplast genomes from four genera of crop plants (corn, wheat, sugar cane, and rice) have been sequenced (Maier *et al*., 1995; Matsuoka *et al*., 2002; Tang *et al*., 2004).

Chloroplast genetic engineering offers a number of unique advantages, including a high-level of transgene expression (DeCosa *et al*., 2001), multigene engineering in a single transformation event (DeCosa *et al*., 2001; Ruiz *et al*., 2003; Lossl *et al*., 2003), transgene containment via maternal inheritance (Daniell *et al*., 1998; Scott and Wilkenson, 1999; Daniell, 2002; Hagemann, 2004), lack of gene silencing (Lee *et al*., 2003; DeCosa *et al*., 2001; Dhingra *et al*., 2004), position effect (Daniell *et al*., 2002), pleiotropic effects (Lee *et al*., 2003; Daniell *et al*., 2001; Leelavathi *et al*., 2003) and undesirable foreign DNA (Daniell *et al*., 2004a,b). Lack of complete chloroplast genome sequences is still one of the major limitations to extend this technology to useful crops; only six published *crop* chloroplast genomes are currently available, although 200 non-crop genomes have been sequenced or are in progress. Chloroplast genome sequences are necessary for identification of spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes (Maier and Schmitz-Linneweber, 2004; Daniell *et al*., 2005). In land plants, about 40–50% of each chloroplast genome contains non-coding spacer and regulatory regions.

In this paper, we report on the complete sequence of the chloroplast genome of *Glycine max*. Soybean is considered the most important source of proteins because it is a leguminous crop. It is widely used as animal feed and for human consumption. The dry matter of soybeans contains about 20% oil and 35–40% proteins of high nutritional quality. It is also the most widely planted genetically modified crop in the world, representing in 2003 more than half of the soybean cultivated area worldwide. This includes glyphosate-tolerant cultivars, a trait that has been engineered via the nuclear genome but would offer better transgene containment if engineered via the chloroplast genome because the plastid genome of soybean is inherited maternally (Corriveau and Coleman, 1988). The primary goal of this paper is to compare the genome organization of *Glycine* with the other two completely sequenced legume chloroplast genomes (*Lotus* and *Medicago*) and with the related genome of *Arabidopsis*. In addition to examining gene content and gene order, we determine the distribution and location of repeated sequences among legumes and explore their possible role in the evolution of these genomes. Intergenic spacer and regulatory sequences will be used in future studies for chloroplast genetic engineering.

Materials and methods

DNA sources

The genome library of *Glycine max*, PI 437654, was constructed by ligating the size fractionated partial *Hind III* digests of the total cellular DNA with a pINDIGOBAC-536 vector. The average insert size of the library was 136 kb.

BAC clones containing the chloroplast genome inserts were isolated by screening the library with a soybean chloroplast probe. The first 96 positive clones from screening were pulled from the library, arrayed in a 96 well microtitre plate, copied, and archived. Selected clones were then subjected to *Hind III* fingerprinting and *Not I* digests. End-sequences were determined and localized on the chloroplast genome of *Arabidopsis thaliana* to deduce the relative positions of the clones, then one clone that covered the entire chloroplast

genome was chosen for the subsequent sequencing analysis.

DNA sequencing and data assembly

The nucleotide sequence of the BAC clone was determined by the bridging shotgun method. The purified BAC DNA was subjected to hydroshearing, end repair, and then size-fractionated by agarose gel electrophoresis. Fractions of approximately 3.0–5.0 kb were eluted and ligated into the vector pBLUESCRIPT IIKS+. The libraries were plated and arrayed into 40 96-well microtitre plates, respectively, for sequencing reactions.

Sequencing was performed using the Dye-terminator cycle sequencing kit (Perkin Elmer Applied Biosystems, USA). Sequence data from the forward and reverse priming sites of the shotgun clones were accumulated. Sequence data equivalent to eight times the size of the genome was assembled using Phred-Phrap programs (Ewing and Green, 1998).

Genome annotation

Annotation of the *Glycine* chloroplast genome was performed using DOGMA (Dual Organellar GenoMe Annotator, Wyman *et al*., 2004; http://evogen.jgi-psf.org/dogma). This program uses a FASTA-formatted input file of the complete genomic sequences and identifies putative protein-coding genes by performing BLASTX searches against a custom database of previously published chloroplast genomes. The user must select putative start and stop codons for each protein coding gene and intron and exon boundaries for intron-containing genes. Both tRNAs and rRNAs are identified by BLASTN searches against the same database of chloroplast genomes. The *Medicago* genome sequence (NC_003119) has not been annotated so we also used DOGMA to annotate this genome.

Molecular evolutionary comparisons

Gene content comparisons were performed using Multipipmaker (Schwartz *et al*. 2003). Two sets of comparisons were performed, one including four genomes (*Arabidopsis* [AP000423], and the three legumes *Glycine* [XXXXXX], *Lotus* [AP002983], and *Medicago* [AC093544]) using *Nicotiana* [Z00044] as the reference genome and a second that only included the three legumes using *Lotus* as the reference genome. Gene orders were examined by pairwise comparisons between the *Arabidopsis*, *Glycine*, *Lotus*, and *Medicago* genomes using PipMaker (Elnitski *et al*., 2002).

Repeat structure in legume chloroplast genomes was examined in two stages. First, REPuter (Kurtz *et al*., 2001) was used to identify the number and location of direct and inverted (palindromic) repeats in the three legumes and *Arabidopsis* using a minimum repeat size of 30 bp and a Hamming distance of 3 (i.e., a sequence identity of 90%). Second, the repeats identified for *Medicago* were blasted against the complete chloroplast genomes of the other two legume genomes (*Glycine* and *Lotus*) and *Arabidopsis*. Blast hits that were 20 bp and longer with a sequence identity of ≥90% were identified and extracted from these results to determine which of the repeats were shared among the four genomes examined.

Results

Size, gene content and organization of the Glycine chloroplast genome

The complete chloroplast genome size of *Glycine* is 152,218 bp (Figure 1). The genome includes of a pair of inverted repeats of 25,574 bp (IRa and IRb) of identical sequence separated by a small single copy region of 17,895 bp, and a large single copy region of 83,175 bp. The IR extends from *rps19* through a portion of *ycf1*.

The *Glycine* chloroplast genome contains 111 unique genes, and 19 of these are duplicated in the IR, giving a total of 130 genes (Figure 1). There are 30 distinct tRNAs, and seven of these are duplicated in the IR. Nineteen genes contain one or two introns, and six of these are in tRNAs. The genome consists of 60% coding regions (52% protein coding genes and 8% RNA genes) and 40% non-coding regions, including both intergenic spacers and introns. The overall GC and AT content of the *Glycine* chloroplast genome is 34% and 66%, respectively. The AT bias is higher in the non-coding regions with 70% AT vs. 62% AT in the coding regions.

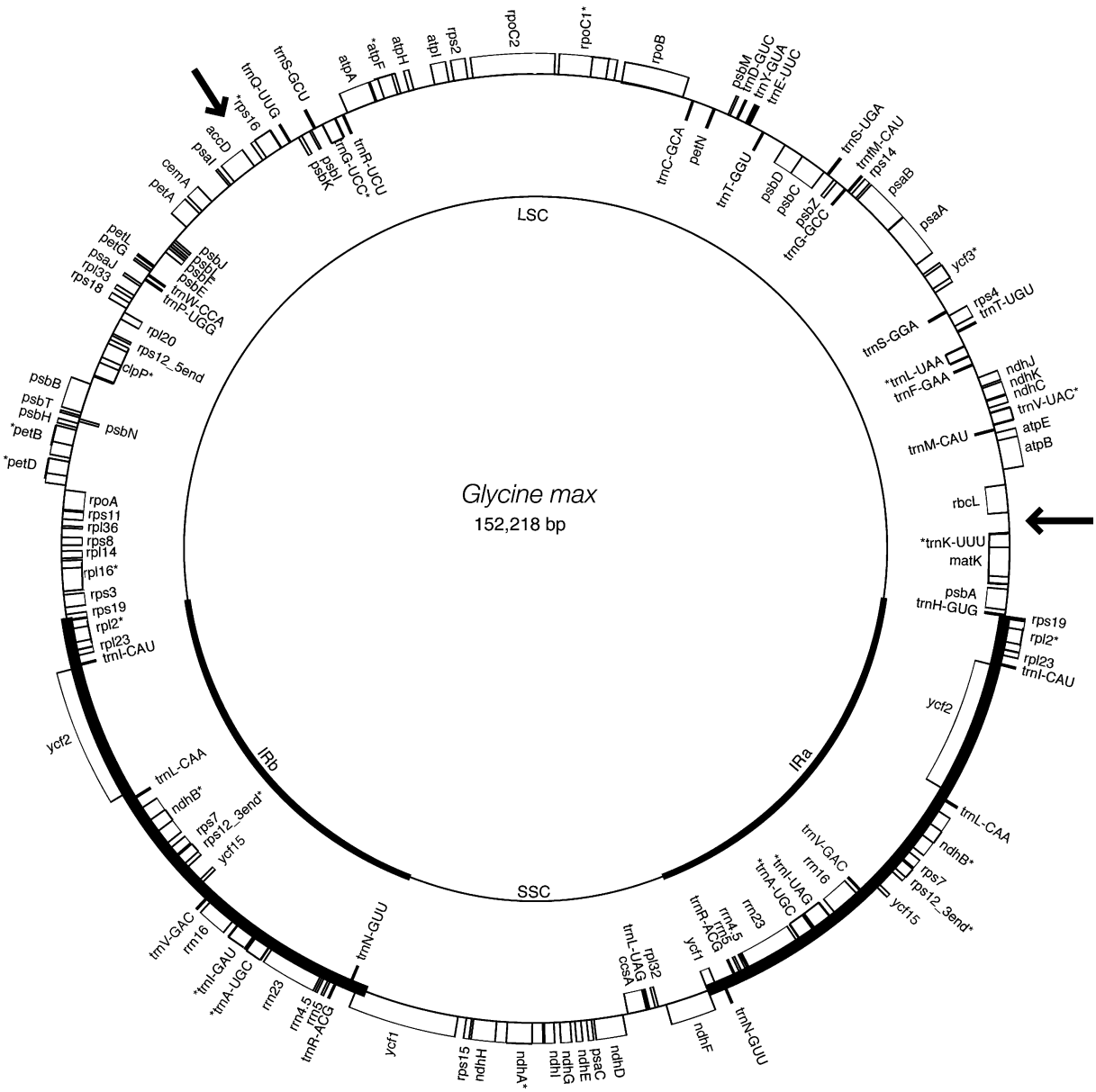

Figure 1. Gene map of *Glycine max* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb, 25,574 bp), which separate the genome into small (SSC, 17,895 bp) and large (LSC, 83,175 bp) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counter-clockwise direction. Genes containing introns are indicated by an asterisk. Arrows indicate locations of end points of th 51kb inversion.

Comparison of genome organization among legumes and Arabidopsis

Gene content

Gene content of the three sequenced legumes (*Glycine*, published here; *Lotus* [Kato *et al*., 2000; NC_002694] and *Medicago* [NC_003119] is nearly identical. *Medicago* does not have duplicate copies of the 19 genes in the IR because one copy of the IR has been lost. A comparison of gene content between the three legumes and *Arabidopsis* shows that the *rpl22* gene is missing from all 3 legumes (see arrow 1 in Figure 2A) and that *Medicago* is also missing *rps16* (see arrow 2 in Figures 2A–B).

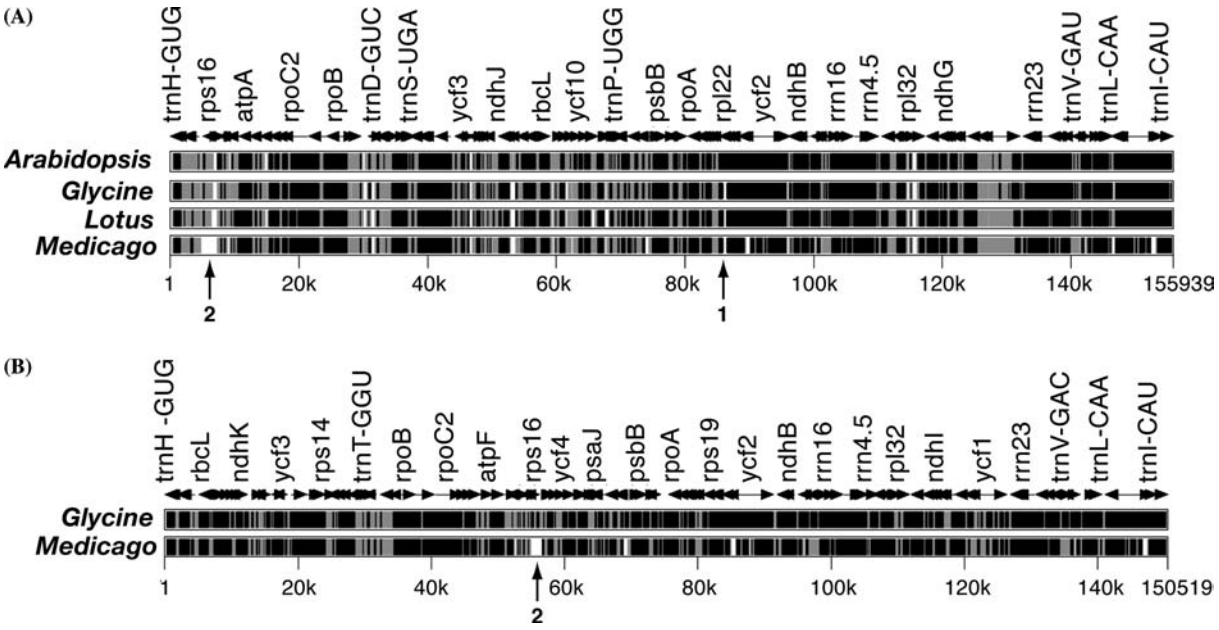

Figure 2. Multipipmaker (Schwartz *et al*. 2003) analyses of legumes and *Arabidopsis* (A, using *Nicotiana* as reference genome) and legumes (B, using *Lotus* as reference genome). Genes and their direction of transcription are indicated by horizontal arrows above each multipip diagram. Species names are listed at the left of each diagram. Levels of sequence similarity are indicated in gray (50–75%) and black (75–100%). Gene losses are indicated in white and with vertical arrows for the genes *rpl22* (1) and *rps16* (2).

Gene order

The gene order in *Glycine* differs from the usual gene order for angiosperm chloroplast genomes by the presence of a single, large inversion of approximately 51 kb that reverses the order of the genes between *rbcL* and *rps16* (see arrows in Figure 1). This same inversion is also present in *Lotus* and *Medicago* (Kato et al. 2000).

Extent of IR

The IR in Glycine is 25,574 bp long and includes 19 genes. At the IR/LSC junction the IR ends within the *rps19* gene so that 68 bp of the 5′ end of the gene is duplicated (Figure 3). The IR/SSC junction is found within *ycf1* resulting in the duplication of 478 bp of the 5′ end of this gene. Comparison of the IR region of the three completely sequenced legumes and *Arabidopsis* indicates that there is some contraction of the IR in the two legumes with an IR. At the IR/LSC boundary the IR includes 68 and 1 bp of the *rps19* gene in *Glycine* and *Lotus*, respectively. Thus, the IR in both of these legumes has contracted relative to *Arabidopsis*, which has 113 bp of the 5′ end of *rps19* duplicated. There has also been contraction of the IR in the legumes at the IR/SSC boundary relative to *Arabidopsis*. *Glycine* and *Lotus* have 478 bp and 514 bp of *ycf1* duplicated, whereas *Arabidopsis* has 1,027 bp duplicated in the IR. This contraction of the IR in these legumes accounts for the smaller size of their IR and larger size of the SSC (Figure 3).

In addition contraction of the IR boundary in legumes, IRa has been lost in *Medicago* (Figure 3). This loss has resulted in *ndhF* (usually located in the SSC) being adjacent to *trnH* (usually the first gene in the LSC at the LSC/IRa junction). Loss of one copy of the IR in some legumes provides support for monophyly of six tribes (Palmer, 1985; Wolfe, 1988; Palmer *et al*., 1987b; Lavin *et al*., 1990). Wolfe (1988) identified duplicated sequences of portions of two genes, 40 bp of *psbA* and 64 bp of *rbcL*, in the region of the IR deletion between *trnH* and *ndhF* in *Pisum sativum* and these duplications were later identified in broad bean (*Vicia faba*, Herdenberger *et al*., 1990). We found similar repeats in this region in other legumes without an IR, including two species of *Medicago* (Figure 4). The *psbA* repeat has the same length of 40 bp and it has a high sequence identity with a segment of *psbA* at coordinates 446–485 in other legumes without the IR (Figure 5A). The copies of

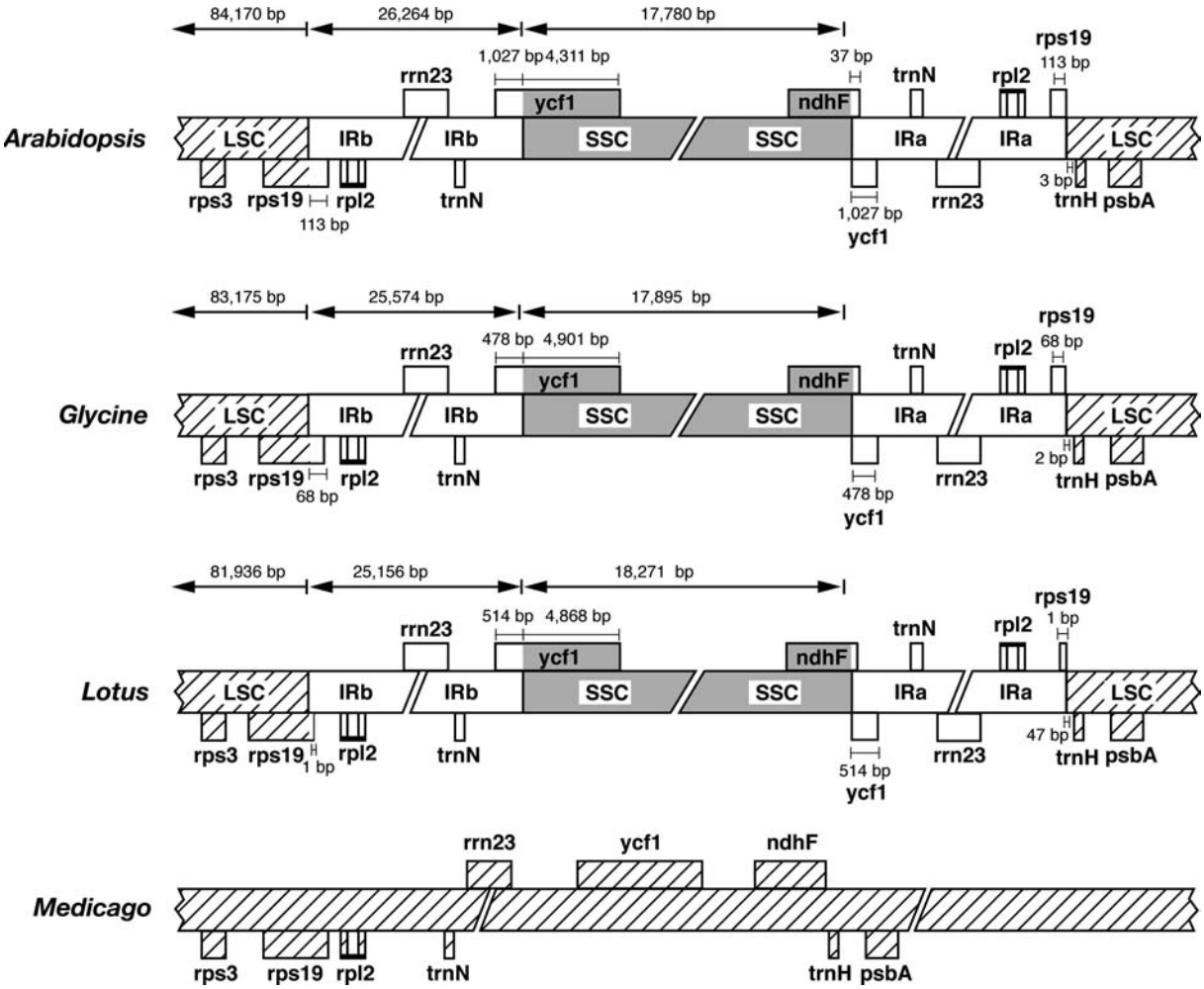

Figure 3. Comparison of boundaries of IR, SSC, and LSC among the legume and *Arabidopsis* chloroplast genomes.

the *psbA* repeat in *Pisum* and *Vicia* and in the two *Medicago* species have a 100% sequence identity with each other but the sequence identity between the *Pisum*/*Vicia* and *Medicago* repeats is 85% (Figure 4). The sequence identity of this repeat to the complete, functional copy of *psbA* is 85% for *Pisum* and *Vicia* and 95% for the two *Medicago* species (Figure 5A). The *rbcL* repeats are 39 bp long in the two *Medicago* species with a 95% sequence identity to each other (Figure 4) and 90% sequence identity to coordinates 516–554 in the complete functional copy of *rbcL* (Figure 5B). In *Vicia* and *Pisum* the *rbcL* repeat is 64 bp long with a 92% sequence identity to each other and 86–92% sequence identity to coordinates 516–579 in the complete functional copies of *Vicia* and *Pisum*, respectively (Figure 5B).

Repeat structure

Analyses using REPuter found 67 to 191 direct and inverted repeats 30 bp or longer with a sequence identity of at least 90% among the three legume chloroplast genomes examined (Figure 6). *Medicago* has the largest number of repeats with 191 and *Lotus* has the fewest with only 67. The number of repeats in the legumes is higher than the 57 repeats identified in *Arabidopsis*. The majority of the repeats (54–81%) in all four genomes are between 30–40 bp in length. The longest legume repeats are in *Lotus* and *Glycine* and are 274 and 287 bp, respectively. The largest repeat in *Glycine* is a 287 bp sequence of *ycf2* that has four identical copies, two in each IR. The two copies in each IR are separated by 1689 bp. The four copies of the 274 bp repeat in *Lotus*, which also represent a

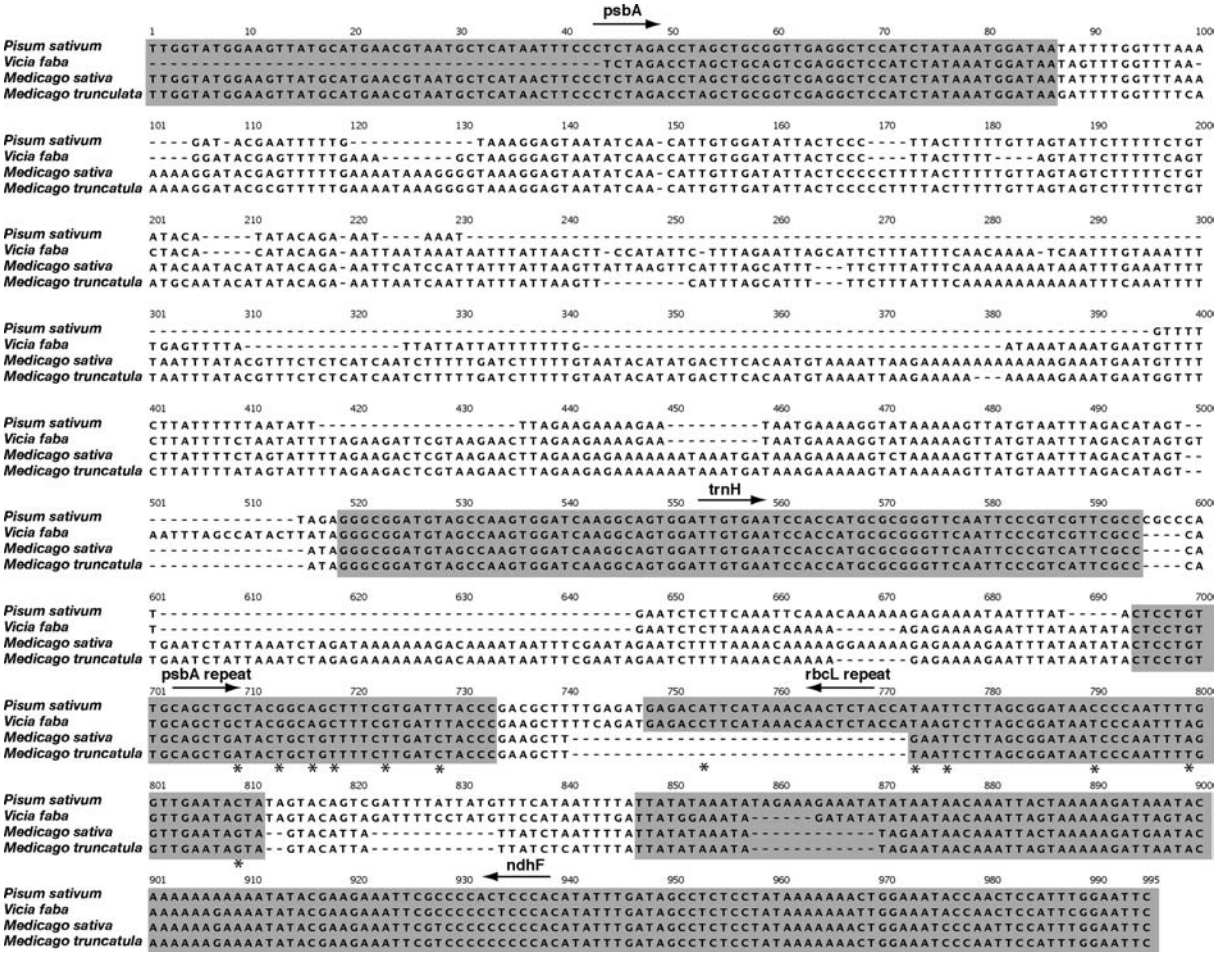

Figure 4. Sequence alignment of IR loss region between *psbA* and *ndhF* for *Medicago*, *Pisum*, and *Vicia*. Shaded regions show genes and repeat elements. Asterisks in shaded regions of repeat elements indicate positions with mismatches. Sequences for this figure were obtained from Genbank (*P. sativum* [M16899], Shapiro and Tewari, 1986; *V. faba* [X51471], Herdenberger *et al*., 1990; *M. sativa* [AY029748], D. Rosellini, unpubl.; *M. truncatula* [NC003119], Lin *et al*., unpubl.).

duplicated segment of *ycf2* in the IR, are separated by 1963 bp in each IR. The two large repeats in *Glycine* and *Lotus* are very similar with 83% sequence identity at the nucleotide level.

BlastN (Altschul *et al*. 1997) comparisons of the 191 *Medicago* repeats against the chloroplast genomes of *Arabidopsis*, *Glycine*, and *Lotus* reveal that 13 of the *Medicago* repeats show a sequence identity greater than 90% with sequences 20 bp or longer (Table 1). Five of the *Medicago* repeats are located in intergenic spacers or introns (repeats 3–7 in Table I) and the remaining eight repeats are found in four genes, *psaA*, *psaB*, *ycf1* and *ycf2*. Many of the *Medicago* repeats are also found in *Arabidopsis*. One of these is repeat 3, which represents a portion of the *psbA* gene that is found in the intergenic spacer (IGS) between *trnH* and *ndhF* and in *psbA* of *Medicago* but is only found in *psbA* of *Arabidopsis*, *Glycine*, and *Lotus* (see section on IR extent above for more details). Two repeats are restricted to legumes (repeats 10 and 13) and these are located in *ycf2*. The number of *Medicago* repeats shared with only one other genome is 1 for *Arabidopsis* (repeat 6), 2 for *Lotus* (repeats 2 and 7), and 1 for *Glycine* (repeat 8).

Discussion

The *Glycine* genome has the typical organization for land plant chloroplast genomes with two identical copies of an inverted repeat that separate

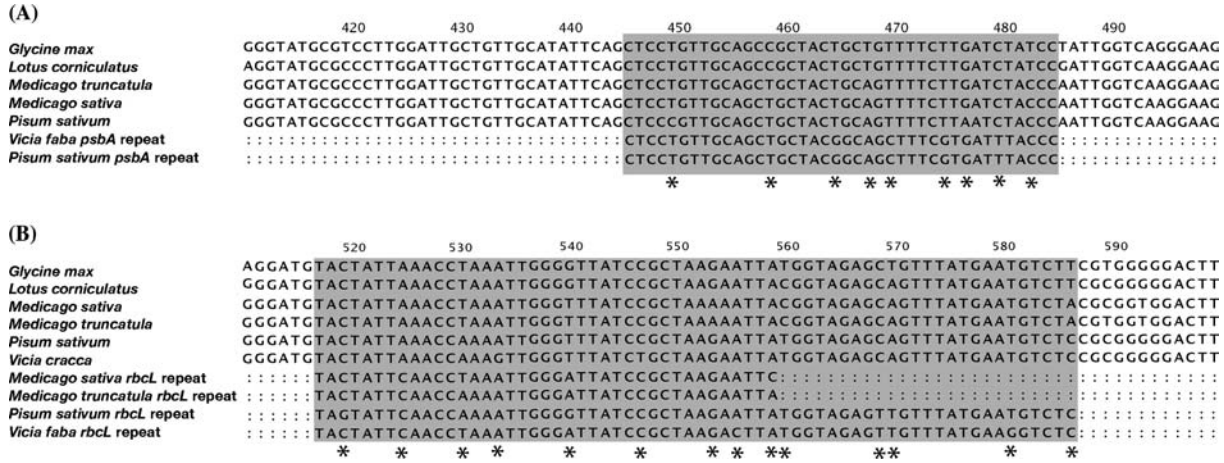

Figure 5. Sequence alignment of legume repeats for *psbA* (A) and *rbcL* (B) with functional copies of these genes. Asterisks in shaded regions indicate positions with mismatches. *psbA* sequences are from GenBank for *L. corniculatus* (AP002983), *M. truncatula* (AC093544), *M. sativa* (AY029748), *P. sativum* (M11005) and from the genome sequence of *G. max* generated in this paper (XXXXXX). *rbcL* sequences are from GenBank for *L. corniculatus* (AP002983), *M. truncatula* (AC093544), *M. sativa* (X04975), *P. sativum* (X03853) and from the genome sequence of *G. max* generated in this paper (XXXXXX). Sequences of the *psbA* and *rbcL* repeats for *P. sativum* and *V. faba* are from Shapiro and Tewari (1986, M16899) and Herdenberger *et al.* (1990, X51471), respectively.

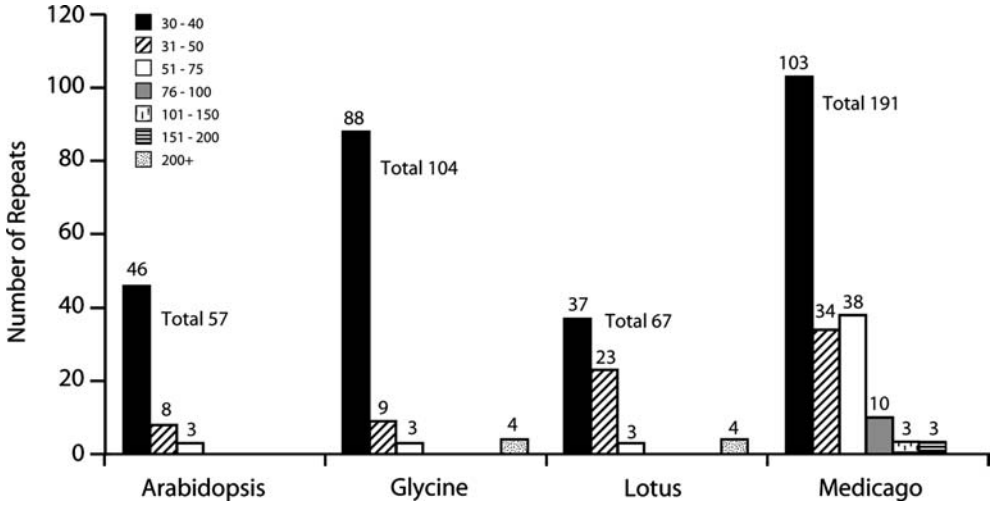

Figure 6. Histogram showing the number of repeated sequences ≥30 bp long with a sequence identity ≥90% in the three legume and *Arabidopsis* genomes using REPuter (Kurtz *et al*., 2001).

the large and small single copy regions. The size of the genome at 152,218 bp is also similar to most angiosperm chloroplast genomes that have two copies of the IR, which generally range in size from 134–164 kb (Jansen *et al*., 2005). The two IR containing legumes whose genomes have been sequenced, *Glycine* (reported here) and *Lotus* (Kato *et al*., 2000), are very similar in size with *Lotus* being 1619 bp shorter than *Glycine*. Only a small portion of this difference in length can be attributed to the expansion of the IR in *Glycine* at the IR/LSC boundary (Figure 3), a phenomenon common in flowering plants (Goulding *et al*., 1996). Therefore, most of this size variation is due to differences in sizes of intergenic spacer regions outside of the IR.

There is considerable variation in size of legume chloroplast genomes due to the loss of one copy of

Table 1. Medicago repeats in other legume chloroplast genomes and *Arabidopsis*.

Medicago repeat	*Glycine*	*Lotus*	*Arabidopsis*
1 29 bp; *ycf2*; 38,293; 39,448	4; 29 bp; 93.1%; *ycf2*; 85,182; 87,133; 148,009; 149,960	4; 29 bp; 93.1%; *ycf2*; 83,975; 85,938; 146,578; 148,481	2; 29 bp; 93.1%; *ycf2*; 87,986; 150,663
2 32 bp; *psaA* (102,048) and *psaB* (104,272)	0	1; 32 bp; 90.6%; *psaB*; 22,187	0
3 40 bp; IGS *trnH* (82)- *ndhF* and *psbA* (122,935)	1; 37 bp ; 91.9%; *psbA*; 634	1; 37 bp; 91.9%; *psbA*; 1028	1; 37 bp; 91.9%; *psbA*; 999
4 41 bp; *ndhA* intron (9,674) and *ycf3* intron (107,056)	1; 41 bp; 92.7%; *rpl16* exon 2; 80,509 1; 40 bp; 92.5%; *ndhA* intron; 117,299 1; 38 bp; 94.7%; IGS *trnS* – *ycf3*; 17,087	1; 41 bp; 92.7%; IGS *trnS* – *ycf3*; 17,209 1; 41 bp; 92.7%; *ndhA* intron; 116,769 1; 38 bp; 94.7%; IGS *rpl16* – *rps3*; 79,516 2; 38 bp; 92.1%; IGS *rps12* – *ycf15*; 96,464; 135,992	1; 38 bp; 92.%; IGS *trnS* – *ycf3*; 43,791 2; 38 bp; 94.7%; IGS *rps12* 3' end – *trnV*; 98,833; 139,816
5 42 bp; IGS *ycf15* – *rps12* 3′ end (29,070) and IGS *rps3* – *rpl16* (44,070)	1; 42 bp; 100%; *rpl16* exon 2; 80,551 1; 42 bp; 95.2%; IGS *ycf15* – *rps12* 3′ end; 137,658 1; 41 bp; 95.2%; *rps12* 3' end exon 2; 97,484 1; 39 bp; 100%; *ndhA* intron; 117,260 1; 39 bp; 94.9%; IGS *trnS* – *ycf3*; 17,049	2; 42 bp; 97.6%; IGS *ycf15* – *rps12* 3′ end; 135,953 1; 40 bp; 97.5%; *ndhA* intron; 116,809 1; 40 bp; 97.5%; IGS *rpl16* – *rps3*; 79,555 1; 39 bp; 97.4%; IGS *trnS* – *ycf3*; 17,168	2; 42 bp; 100%; IGS *trnV* – *rps12* 3′ end; 98,874; 139,777 1; 40 bp, 90%; *ndhA* intron; 120,456 1; 39 bp; 92.3%; IGS *trnS* – *ycf3*; 43,829
6 42 bp; IGS *ycf4* – *psaI* (66,222) and IGS *psaI* – *accD* (66,462)	0	0	1; 32 bp; 93.8%; IGS *accD* – *psaI* (59,241)
7 45 bp; IGS *ycf1* – *trnN* 18,846; 18,934	0	1; 20 bp; 90%, IGS *trnV* – *ndhC* (10,353)	0
8 48 bp; *ycf1*; 17,086; 17,110	1; 22 bp; 100%; *ycf1*; 109,656	0	0
9 58 bp; *psaB* (102,060) and *psaA* (104,284)	1; 52 bp; 94.2%; *psaB*; 21,977 1; 49 bp; 91.8%; *psaA*; 19,750	1; 52 bp; 90.4%; *psaB*; 22,148 1; 47 bp; 95.7%; *psaA*; 19,921	1; 58 bp; 93.1%; *psaB*; 38,720 1; 44 bp; 95.4%; *psaA*; 40,950
10 58 bp; *ycf2*; 36,489; 36,609	2; 27 bp; 92.6%; *ycf2*; 89,198; 145,944	2; 27 bp; 92.6%; *ycf2*; 88, 018; 144,438	0
11 61 bp; *ycf2*; 37,266; 37,311	2; 41 bp; 92.7%; *ycf2*; 82,228; 146,914 2; 39 bp; 92.3%; *ycf2*; 82,269; 146,873	2; 41 bp; 90.2%; *ycf2*; 87,092; 145,364 2; 41 bp; 92.7%; *ycf2*; 87,051; 145,405	2; 39 bp; 92.3%; *ycf2*; 88,164; 149,485
12 79 bp; *psaB* (102,060) and *psaA* (104,284)	1; 76 bp; 90.8%; *psaB*; 22,001	1; 47 bp; 95.7%; *psaA*; 19,921	1; 76 bp; 93.4%; *psaB*; 38,702 1; 47; 95.7%; *psaA*; 40,929
13 118 bp; *ycf2*; 36,489; 36,549	2; 27 bp; 96.3; *ycf2*; 88,018; 144,438	2; 27 bp; 96.3; *ycf2*; 89,198; 145,944	0

Only *Medicago* repeats that show a length >20 bp and a sequence identity of >90% with the other genomes are listed. Length of *Medicago* repeats (in bp) and their locations (gene names and starting coordinates) are provided in column 1. The number of copies, length (bp), percent identity, and locations (gene or region names and starting coordinates) of the repeated sequences are listed for other genomes. IGS = intergenic spacer.

the IR from members of six related tribes (Palmer, 1985; Palmer *et al*., 1987b; Lavin *et al*., 1990). A detailed examination of the IR loss region in Pea (*Pisum sativum*) and broad bean (*Vicia faba*) identified two repeated sequences of 40 and 64 bp in the region where the IR was deleted (Wolfe, 1988; Herdenberger *et al*., 1990). These repeats showed a very high sequence identity to portions of two LSC genes, *rbcL* and *psbA*. Wolfe suggested that the repeats could have been present prior to the IR loss and played a role in the deletion event. Alternatively, these repeats may have been formed as part of the IR deletion. In either case, Wolfe predicted that if other legumes that lost one copy of the IR share these repeats it would indicate that the IR deletion in legumes represents a single event. Our examination of the IR region in the three legume chloroplast genomes (Figure 4) clearly indicates that other legumes with only one copy of the IR have the *psbA* and *rbcL* repeats. Thus, this IR loss occurred only once, and it provides an excellent phylogenetic marker supporting the monophyly of six tribes of legumes. The monophyly of this group of legumes is also supported by a sequenced-based phylogeny of the plastid gene *matK* (Wojciechowski *et al*., 2004). The *psbA* repeats in *Pisum*, *Vicia* and the two *Medicago* species (Figure 4) are identical in length and have a very high sequence identity (100% for *Pisum*/*Vicia* and 85% for *Pisum*/*Medicago*). In contrast, the *rbcL* repeat (Figure 4) has diverged more in length (39 bp in *Medicago* vs. 64 bp in *Pisum* and *Vicia*) but still has a very high sequence divergence (94% for *Pisum*/*Vicia* and 95% for *Pisum*/*Medicago*). The sequenced legume genomes with both copies of the IR (*Glycine* and *Lotus*) do not have either of these repeats suggesting that the repeats originated at or shortly after the time of the deletion event.

Gene content is highly conserved in most land plant chloroplast genomes (Palmer, 1991; Raubeson and Jansen, 2005). The *Glycine* genome contains 130 genes, 19 of which represent duplicate copies in the IR. The gene content is identical to the completely sequenced *Lotus* chloroplast genome (Kato *et al*., 2000) and both of these legumes and *Medicago* lack the *rpl22* gene. The absence of *rpl22* from legume chloroplast genomes has been noted previously (Spielmann *et al*., 1988; Milligan *et al*., 1989; Gantt *et al*., 1991; Doyle *et al*., 1995). This gene represents an interesting case of gene transfer from the chloroplast to the nucleus. The nuclear encoded protein is now imported back into the chloroplast by a transit peptide (Gantt *et al*., 1991). In addition to *rpl22*, the *Medicago* genome lacks a second ribosomal protein gene, *rps16*. Sequencing studies demonstrated the loss of this gene from *Pisum sativum* (Nagano *et al*., 1991) and an extensive survey of legumes using a filter hybridization approach suggested that there have been multiple independent losses of *rps16* in legumes (Doyle *et al*., 1995). Additional losses of this gene in distantly related plant lineages (e.g., liverworts (Ohyama *et al*., 1986) and pine (Tsudzuki *et al*., 1992)) clearly indicate that this gene loss is not a very reliable phylogenetic marker.

Gene order changes in chloroplast genomes are also relatively uncommon. However, several events have been documented in legumes, including a 51 kb inversion that is shared among most papilionoid legumes (Doyle *et al*., 1996). All three of the completely sequenced legume chloroplast genomes examined here share the 51 kb inversion. The phylogenetic distribution of this inversion is congruent with chloroplast DNA-sequence phylogenies using both *trnL* intron and *matK* (Pennington *et al*., 2000; Wojciechowski *et al*., 2004).

With the exception of the IR, chloroplast genomes have very few repeated sequences (Palmer, 1991). However, a number of studies of rearranged chloroplast genomes have identified dispersed repeats (*Chlamydomonas* (Maul *et al*., 2002), *Pseudotsuga* (Hipkins *et al*., 1995), *Trachelium* (Cosner *et al*., 1997), *Trifolium* (Milligan *et al*., 1989), wheat (Bowman and Dyer, 1986; Howe, 1985), and *Oenothera* (Hupfer *et al*., 2000; Sears *et al*.1996; Vomstein and Hachtel, 1988)). The most impressive example is *Chlamydomonas* in which it was estimated that the genome comprises more than 20% dispersed repeats. All of the genomes with repeated sequences other than the IR have inversions, and this correlation has been used to suggest that repeats may have mediated these changes (Palmer, 1991). Our repeat analyses of the three legumes indicate that these genomes contain a substantial number of repeats (Figure 6). Our analyses were limited to repeats of 30 bp or longer with ≥90% sequence identity. Searches for shorter and/or more divergent repeats would likely identify many additional repeated sequences. In the legumes, the only repeats that are

found in a location where there has been a structural rearrangement are the *psbA* and *rbcL* repeats located in the IR loss region of *Medicago*. Wolfe (1988) suggested that these repeats may have played a role in the loss of the IR. However, the absence of the *psbA* and *rbcL* repeats in legumes with two copies of the IR (i.e., *Glycine* and *Lotus*) suggests that they were not involved in the IR loss.

Many of the repeats in legumes are shared with *Arabidopsis*, and they are restricted to either intergenic spacers/introns or to three genes, *psaA*, *psaB*, and *ycf2*. The *ycf2* repeat was previously identified from adzuki bean, soybean, and *Medicago* (Perry *et al*., 2002). The observation that many of the repeats in the IGS and introns are found in the same location in the other legumes and in *Arabidopsis* suggests that these conserved repeats may be much more widespread in angiosperm chloroplast genomes and that they may play some functional role.

In addition to providing insight into genome organization and evolution, availability of complete DNA sequence of chloroplast genomes should facilitate plastid genetic engineering. Thus far, transgenes have been stably integrated and expressed via the tobacco chloroplast genome to confer several useful agronomic traits, including insect resistance (McBride *et al*., 1995; Kota *et al*., 1999; DeCosa *et al*., 2001), herbicide resistance (Daniell *et al*., 1998. Iamtham and Day, 2000), disease resistance (DeGray *et al*., 2001), drought tolerance (Lee *et al*., 2003), salt tolerance (Kumar *et al*., 2004a), phytpremediation (Ruiz *et al*., 2003) and cytoplasmic male sterility (Ruiz and Daniell, 2005). The chloroplast has been used as a bioreactor to produce vaccines antigens (Daniell *et al*., 2001; Tregoning *et al*., 2003; Molina *et al*., 2004; Watson *et al*., 2004), human therapeutic proteins (Staub *et al*., 2000, Fernandez *et al*., 2003, Leelavathi and Reddy, 2003; Daniell *et al*., 2004a; Chebolu and Daniell, 2005), industrial enzymes (Leelavathi *et al*., 2003) and biomaterials (Guda *et al*., 2000; Lossl *et al*., 2003; Vitanen *et al*., 2004). Although many successful examples of plastid engineering in tobacco have set a solid foundation for various future applications, this technology has not been extended to many of the major crops. Stable plastid transformation has been recently accomplished via somatic embryogenesis using partially sequenced chloroplast genomes in soybean (Dufourmantel *et al*., 2004), carrot (Kumar *et al*., 2004a) and cotton (Kumar *et al*., 2004b; Daniell *et al*., 2005) and rice (Lee *et al*., 2005). Complete chloroplast genome sequences should provide valuable information on spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes and should help in extending this technology to other useful crops.

Acknowledgments

Investigations reported in this article were supported in part by grants from USDA 3611-21000-017-00D to Henry Daniell and from NSF DEB 0120709 to Robert K. Jansen. We thank Tim Chumley for assistance with Figure 1, Tim Chumley and Andy Alverson for comments on an earlier version of the manuscript, Gwen Gage for preparing Figures 2–6, and Cara Stockham Kenny for assistance with the analysis of repeated sequences.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389–3402.

Bowman, C.M. and Dyer, T. 1986. The location and possible evolutionary significance of small dispersed repeats in wheat ctDNA. Curr. Genet. 10: 931–941.

Chebolu, S. and Daniell, H. 2005. Chloroplast derived vaccine antigens and biopharmaceuticals: expression, folding, assembly and functionality. Curr Trends Microbiol Immunol (in press).

Corriveau, J.L. and Coleman, A.W. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. Amer. J. Bot. 75: 1443–1458.

Cosner, M.E., Jansen, R.K., Palmer, J.D. and Downie, S.R. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr. Genet. 31: 419–429.

Daniell, H. 2002. Molecular strategies for gene containment in transgenic crops. Nat. Biotechnol. 20: 581–586.

Daniell, H., Carmona-Sanchez, O. and Burns, B.B. 2004a. Chloroplast-derived vaccine antibodies, biopharmaceuticals, and edible vaccines in transgenic plants engineered via the chloroplast genome. In: S. Schillberg (Ed.), Molecular Farming, Wiley–VCH Verlag publishers, Germany, pp. 113–133.

Daniell, H., Cohill, P.R., Kumar, S. and Dufourmantel, N. 2004b. Chloroplast Genetic Engineering. In: H. Daniell and C.D. Chase (Eds.), Molecular Biology and Biotechnology of Plant Organelles, Springer Publishers, Netherlands, pp. 443–490.

Daniell, H., Khan, M. and Allison, L. 2002. Milestones in chloroplast genetic engineering: an environmentally friendly era in biotechnology. Trends Plant Sci. 7: 84–91.

Daniell, H., Kumar, S. and Duformantel, N. 2005. Breakthrough in chloroplast genetic engineering of agronomically important crops. Trends Biotechnol. 23(5): 238–245.

Daniell, H., Datta, R., Varma, S., Gray, S. and Lee, S.B. 1998. Containment of herbicide resistance through genetic engineering of the chloroplast genome. Nat. Biotechnol. 16: 345–348.

Daniell, H., Lee, S.B., Panchal, T. and Wiebe, P.O. 2001. Expression of cholera toxin B subunit gene and assembly as functional oligomers in transgenic tobacco chloroplasts. J. Mol. Biol. 311: 1001–1009.

DeCosa, B., Moar, W., Lee, S.B., Miller, M. and Daniell, H. 2001. Overexpression of the *Bt* cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. Nat. Biotechnol. 19: 71–74.

DeGray, G., Rajasekaran, K., Smith, F., Sanford, J. and Daniell, H. 2001. Expression of an antimicrobial peptide via the chloroplast genome to control phytopathogenic bacteria and fungi. Plant Physiol. 127: 852–862.

Dhingra, A., Portis, A.R. and Daniell, H. 2004. Enhanced translation of a chloroplast expressed *rbcS* gene restores SSU levels and photosynthesis in nuclear antisense *RbcS* plants. Proc. Natl. Acad. Sci. USA 101: 6315–6320.

Doyle, J.J., Doyle, J.L. and Palmer, J.D. 1995. Multiple independent losses of two genes and one intron from legume chloroplast genomes. Syst. Bot. 20: 272–294.

Doyle, J.J., Doyle, J.L., Ballenger, J.A. and Palmer, J.D. 1996. The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Mol. Phylog. Evol. 5: 429–438.

Dufourmantel, N., Pelissier, B., Garçon, F., Peltier, J.M. and Tissot, G. 2004. Generation of fertile transplastomic soybean. Plant Mol. Biol. 55(4): 479–489.

Elnitski, L., Riemer, C., Petrykowska, H., Florea, L., Schwartz, S., Miller, W. and Hardison, R. 2002. PipTools: A computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. Genomics 80: 681–690.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred II. Error probabilities. Genome Res. 8: 186–194.

Fernandez-San Millan, A., Mingo-Castel, A. and Daniell, H. 2003. Chloroplast transgenic approach to hyper-express and purify human serum albumin, a protein highly susceptible to proteolytic degradation. Plant Biotechnol. J. 1: 71–79.

Gantt, J.S., Baldauf, S.L., Calie, P.J., Weeden, N.F. and Palmer, J.D. 1991. Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J. 10: 3073–3078.

Goulding, S.E., Olmstead, R.G., Morden, C.W. and Wolfe, K.H. 1996. Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. 252: 195–206.

Guda, C., Lee, S.B. and Daniell, H. 2000. Stable expression of biodegradable protein based polymer in tobacco chloroplasts. Plant Cell Rep. 19: 257–262.

Hagemann, R. 2004. The sexual inheritance of plant organelles. In: H Daniell and C Chase (Eds.), Molecular Biology and Biotechnology of Plant Organelles, Springer Publishers, Dordrecht, The Netherlands, pp. 93–113.

Herdenberger, F., Pillay, D.T.N. and Steinmetz, A. 1990. Sequence of the *trn*H gene and the inverted repeat structure deletion of the broad bean chloroplast genome. Nucl Acids Res 18: 1297.

Hipkins, V.D., Marshall, K.A., Neale, D.B., Rottmann, W.H. and Strauss, S.H. 1995. A mutation hotspot in the chloroplast genome of a conifer (Douglas-Fir, *Pseudotsuga*) is caused by variability in the number of direct repeats derived from a partially duplicated transfer-RNA gene. Curr. Genet. 27: 572–579.

Howe, C.J. 1985. The endpoints of an inversion in wheat chloroplast DNA are associated with short repeated sequences containing homology to att-lambda. Curr. Genet. 10: 139–145.

Hupfer, H., Swaitek, M., Hornung, S., Herrmann, R.G., Maier, R.M., Chiu, W.L. and Sears, B. 2000. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome 1 of the five distinguishable *Euoenthera* plastomes. Mol. Gen. Genet. 263: 581–585.

Iamtham, S. and Day, A. 2000. Removal of antibiotic resistance genes from transgenic tobacco plastids. Nat. Biotechnol. 18: 1172–1176.

Jansen, R.K., Raubeson, L.A., Boore, J.L., dePamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., Fourcade, H.M., Kuehl, J.V., McNeal, J.R., Leebens-Mack, J. and Cui, L. 2005. Methods for obtaining and analyzing chloroplast genome sequences. Meth. Enzymol. 395: 348–384.

Kato, T., Kaneko, T., Sato, S., Nakamura, Y. and Tabata, S. 2000. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA Res. 7: 323–330.

Knox, E.B. and Palmer, J.D. 1998. Chloroplast DNA evidence on the origin and radiation of the giant lobelias in eastern Africa. Syst. Bot. 23: 109–149.

Kota, M., Daniel, H., Varma, S., Garczynski, S.F., Gould, F. and William, M.J. 1999. Overexpression of the *Bacillus thuringiensis* (*Bt*) *Cry2Aa2* protein in chloroplasts confers resistance to plants against susceptible and *Bt*-resistant insects. Proc. Natl. Acad. Sci. USA 96: 1840–1845.

Kumar, S., Dhingra, A. and Daniell, H. 2004a. Plastid expressed *betaine aldehyde dehydrogenase* gene in carrot cultured cells, roots and leaves confers enhanced salt tolerance. Plant Physiol. 136(1): 2843–2854.

Kumar, S., Dhingra, A. and Daniell, H. 2004b. Manipulation of gene expression facilitates cotton plastid transformation of cotton by somatic embryogenesis and maternal inheritance of transgenes. Plant Mol. Biol. 56(2): 203–216.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl. Acids Res. 29: 4633–4642.

Lavin, M., Doyle, J.J. and Palmer, J.D. 1990. Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. Evolution 44: 390–402.

Lee, S.B., Kwon, H.B., Kwon, S.J., Park, S.C., Jeong, M.J., Han, S.E. and Daniell, H. 2003. Accumulation of trehalose within transgenic chloroplasts confers drought tolerance. Mol. Breed. 11: 1–13.

Lee, S.M., Kang, K., Chung, H., Yoo, S.H., Xu, X.M., Lee, S.B., Cheong, J.J., Daniell, H., Kim, M., 2005. Plastid transformation in the monocotyledonous crop rice (*Oryza sativa*) and transmission of transgenes to their progeny. Mol. Breed. in press.

Leelavathi, S. and Reddy, V.S. 2003. Chloroplast expression of His-tagged GUS-fusions: a general strategy to overproduce and purify foreign proteins using transplastomic plants as bioreactors. Mol. Breed. 11: 49–58.

Leelavathi, S., Gupta, N., Maiti, S., Ghosh, A. and Reddy, V.S. 2003. Overproduction of an alkali- and thermo-stable xylanase in tobacco chloroplasts and efficient recovery of the enzyme. Mol. Breed. 11: 59–67.

Lossl, A., Eibl, C., Harloff, H.J., Jung, C. and Koop, H.U. 2003. Polyester synthesis in transplastomic tobacco (*Nicotiana tabacum* L.): significant contents of polyhydroxybutyrate are associated with growth reduction. Plant Cell Rep. 21: 891–899.

Maier, R.M., Neckermann, K., lgloi, G.L. and Kossel, H. 1995. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 251: 614–628.

Maier, R.M. and Schmitz-Linneweber, 2004. Plastid genomes. In: H. Daniell and C.D. Chase (Eds.), Molecular Biology and Biotechnology of Plant Organelles, Springer publishers, Netherlands, pp. 115–150.

Matsuoka, Y., Yamazaki, Y., Ogihara, Y. and Tsunewaki, K. 2002. Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals. Mol. Biol. Evol. 19: 2084–2091.

Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H. and Stern, D.B. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: Islands of genes in a sea of repeats. Plt. Cell 14: 1–22.

McBride, K.E., Svab, Z., Schaaf, D.J., Hogan, P.S., Stalker, D.M. and Maliga, P. 1995. Amplification of a chimeric *Bacillus* gene in chloroplasts leads to an extraordinary level of an insecticidal protein in tobacco. Bio/Technology 13: 362–365.

Milligan, B.G., Hampton, J.N. and Palmer, J.D. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol. Biol. Evol. 6: 355–368.

Molina, A., Herva-Stubbs, S., Daniell, H., Mingo-Castel, A.M. and Veramendi, J. 2004. High yield expression of a viral peptide animal vaccine in transgenic tobacco chloroplasts. Plt. Biotechnol. J. 2: 141–153.

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

Nagano, Y., Matsuno, R. and Sasaki, Y. 1991. Sequence and transcriptional analysis of the gene cluster *trnQ-zfpA-psaI*-ORF231-*petA* in pea chloroplasts. Curr. Genet. 20: 431–436.

Palmer, J.D. 1985. Evolution of chloroplast and mitochondrial DNA in plants and algae. In: RJ MacIntyre (Ed.), Monographs in Evolutionary Biology: Molecular Evolutionary Genetics, Plenum Press, New York, pp. 131–240.

Palmer, J.D. 1991. Plastid chromosomes: structure and evolution. In: RG Hermann (Ed.), The Molecular Biology of Plastids. Cell Culture and Somatic Cell Genetics of Plants vil 7A, Springer-Verlag, Vienna, pp. 5–53.

Palmer, J.D., Jansen, R.K., Michaels, H., Manhart, J. and Chase, M. 1988. Chloroplast DNA variation and plant phylogeny. Ann. Missouri. Bot. Gard. 75: 1180–1206.

Palmer, J.D., Nugent, J.M. and Herbon, L.A. 1987a. Unusual structure of Geranium chloroplast DNA – A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families. Proc. Natl. Acad. Sci. USA 84: 769–773.

Palmer, J.D., Osorio, B., Aldrich, J. and Thompson, W.F. 1987. Chloroplast DNA evolution among legumes: loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr. Genet. 11: 275–286.

Palmer, J.D., Osorio, B. and Thompson, W.F. 1988. Evolutionary significance of inversions in legume chloroplast DNAs. Curr. Genet. 14: 65–74.

Pennington, R.T., Klitgaard, B.B., Ireland, H. and Lavin, M. 2000. New insights into floral evolution of basal *Papilionoideae* from molecular phylogenies. In: PS Herendeen A Bruneau (Ed.), Advances in Legume Systematics, part 9, Kew, UK, pp. 233–248.

Perry, A.S., Brennan, S., Murphy, D.J. and Wolfe, K.H. 2002. Evolutionary re-organisation of a large operon in Adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res. 9: 157–162.

Raubeson, L.A. and Jansen, R.K. 2005. Chloroplast genomes of plants. In: R Henry (Ed.), Diversity and Evolution of Plants-Genotypic and Phenotypic Variation in Higher Plants, CABI Publishing, Wallingford, pp. 45–68.

Ruiz, O.N., Hussein, H., Terry, N. and Daniell, H. 2003. Phytoremediation of organomercurial compounds via chloroplast genetic engineering. Plt. Phys. 132: 1344–1352.

Ruiz, O.N. and Daniell, H. 2005. Engineering cytoplasmic male sterility via the chloroplast genome. Plant Physiol. 138: 1232–1246.

Scott, S.E. and Wilkenson, M.J. 1999. Low probability of chloroplast movement from oilseed rape (*Brassica napus*) into wild *Brassica rapa*. Nat. Biotechnol. 17: 390–392.

Sears, B.B., Stoike, L.L. and Chiu, W.L. 1996. Proliferation of direct repeats near the *Oenothera* chloroplast DNA origin of replication. Mol. Biol. Evol. 13: 850–863.

Shapiro, D.R. and Tewari, K.K. 1986. Nucleotide sequences of transfer RNA genes in the *Pisum sativum* chloroplast DNA. Plt. Mol. Biol. 6: 1–12.

Spielmann, A., Roux, E., von Allmen, J. and Stutz, E. 1988. The soybean chloroplast genome: completed sequence of the *rps19* gene, including flanking parts containing exon 2 of *rpl2* (upstream), but lacking *rpl22* (downstream). Nucl. Acids Res. 16: 1199.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Program, N.C.S., Green, E.D., Hardison, R.C. and Miller, W. 2003. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucl. Acids Res. 31: 3518–3524.

Staub, J.M., Garcia, B., Graves, J., Hajdukiewicz, P.T.J., Hunter, P. and Nehra, N. 2000. High-yield production of a human therapeutic protein in tobacco chloroplasts. Nat. Biotechnol. 18: 333–338.

Tang, J., Xia, H., Cao, M., Zhang, X., Zeng, W., Hu, S., Tong, W., Wang, J., Wang, J., Yu, J., Yang, H. and Zhu, L. 2004. A comparison of rice chloroplast genomes. Plt. Phys. 135: 412–420.

Tregoning, J.S., Nixon, P., Kuroda, H., Svab, Z., Clare, S., Bowe, F., Fairweather, N., Ytterberg, J., van Wijk, K.J., Dougan, G. and Maliga, P. 2003. Expression of tetanus toxin Fragment C in tobacco chloroplasts. Nucl. Acids Res. 31(4): 1174–1179.

Vitanen, P.V., Devine, A.L., Kahn, S., Deuel, D.L., Van-Dyk, D.E. and Daniell, H. 2004. Metabolic engineering of the chloroplast genome using the *E. coli ubi*C gene reveals that corismate is a readily abundant precursor for 4-hydroxybenzoic acid synthesis in plants. Plt. Phys. 136: 4048–4060.

Tsudzuki, J., Nakashima, K., Tsudzuki, T., Hiratsuka, M., Shibata, M., Wakasugi, T. and Sugiura, M. 1992. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequence of the *trn*Q, *trn*K, *psb*A, *trn*I, and *trn*H and the absence of *rps16*. Mol. Gen. Genet. 232: 206–214.

Vomstein, J. and Hachtel, W. 1988. Deletions, insertions, short inverted repeats, sequences resembling att-lambda, and frame shift mutated open reading frames are involved in chloroplast DNA differences in the genus *Oenothera* subsection *Munzia*. Mol. Gen. Genet. 213: 513–518.

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T. and Sugiura, M. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. USA 91: 9794–9798.

Watson, J., Koya, V., Leppla, S.H. and Daniell, H. 2004. Expression of *Bacillus anthracis* protective antigen in transgenic chloroplasts of tobacco, a non-food/feed crop. Vaccine 22: 4374–4384.

Wojciechowski, M.F., Lavin, M. and Sanderson, M.J. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. Amer. J. Bot. 91: 1846–1862.

Wolfe, K.H. 1988. The site of deletion of the inverted repeat is pea chloroplast DNA contains duplicated gene fragments. Curr. Genet. 13: 97–99.

Wyman, S.K., Boore, J.L. and Jansen, R.K. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255.