DNA Research **9**, 157–162 (2002)

Evolutionary Re-organisation of a Large Operon in Adzuki Bean Chloroplast DNA caused by Inverted Repeat Movement

Antoinette S. Perry, Sinéad Brennan, David J. Murphy, Tony A. Kavanagh, and Kenneth H. Wolfe*

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland

(Received 2 July 2002; revised 26 July 2002)

Abstract

We have sequenced two sections of chloroplast DNA from adzuki bean (*Vigna angularis*), containing the junctions between the inverted repeat (IR) and large single copy (LSC) regions of the genome. The gene order at both junctions is different from that described for other members of the legume family, such as *Lotus japonicus* and soybean. These differences have been attributed to an apparent 78-kb inversion that spans nearly the entire LSC region and which is present in adzuki and its close relative, the common bean. This 78-kb rearrangement broke the large *S10* operon of ribosomal proteins into two smaller operons, one at each end of the LSC, without affecting the gene content of the genome. It disrupted the physical and transcriptional relationship between the six-gene *rpl23–rpl14* cluster and the four-gene *rps8–rpoA* cluster that is conserved in most land plants. Analysis of the endpoints of the rearrangement indicates that it probably occurred by means of a two-step process of expansion and contraction of the IR and not by a 78-kb inversion.

Key words: plastome; operon evolution

1. Introduction

The chloroplast genome of most land plants is highly conserved with respect to its size and structure, ranging from 120 to 217 kb.[1] Most of this variation in size can be accounted for by changes in the size of the large inverted repeat (IR), which is a feature of nearly all land plant chloroplast genomes. The IR is made up of two completely identical segments, IR_A and IR_B, which are typically 10–25 kb but can range from 6 to 76 kb in length.[2,3] The two segments of the repeat are separated by large and small single copy regions (LSC and SSC, respectively).

A major constraint on rearrangements in chloroplast (cp) DNA appears to be its organisation into large clusters of polycistronically transcribed genes.[4–6] An example of such is the *S10* operon of ribosomal proteins, homologous to the *S10*, *spc* and *alpha* operons of *Escherichia coli*.[7,8] In legume species this operon contains 10 genes (*rpl23–rpl2–rps19–rps3–rpl16–rpl14–rps8–rpl36–rps11–rpoA*)[9] coding for nine ribosomal proteins (*rpl* and *rps* genes) and the alpha subunit of chloroplast RNA polymerase (*rpoA*). In most other plants, the *S10* operon also includes the genes *rpl22* and *infA*, but both of these have been transferred from cpDNA to nuclear DNA in legumes.[10,11] An additional gene (*trnI*, coding for tRNA-Ile) can be included at the 5′ end of the operon transcript in some species, depending on which promoter is utilized.[6,12] *S10* is the largest operon in the chloroplast genome. It initiates within the IR and is transcribed across the junction (J_{LB}) between IR_B and the LSC as a single polycistronic transcript.[6,12] Therefore, this operon is partly duplicated in most plants (e.g., *Lotus*; Fig. 1) because it initiates at identical positions within the two segments of the IR. At the second IR/LSC junction (J_{LA}), the operon is truncated and transcription is terminated. A second operon, *psbA–trnH*, is transcribed in the opposite direction towards IR_A.[12]

Despite this evolutionary constraint, the chloroplast genomes of legumes are known to have undergone more rearrangements than other angiosperms, including complete loss of the IR in one group of legumes.[4] Palmer et al.[4] found that two species with the IR, mung bean (*Vigna radiata*) and common bean (*Phaseolus vulgaris*), contain an inversion of a 78-kb segment relative to all other land plants. This large rearrangement encompasses nearly the entire LSC region and disrupts the *S10* operon. Within the Phaseoleae tribe, this inversion is the only genomic difference that distinguishes *Vigna* and *Phaseolus* from soybean (*Glycine max*).

Although it is commonly described as an inversion,[13,14] Palmer et al.[4] proposed two ways in which the 78-kb rearrangement could have occurred. The first, which they

Communicated by Masahiro Sugiura

* To whom correspondence should be addressed. Tel. +353-1-6081253, Fax. +353-1-6798558, E-mail: khwolfe@tcd.ie

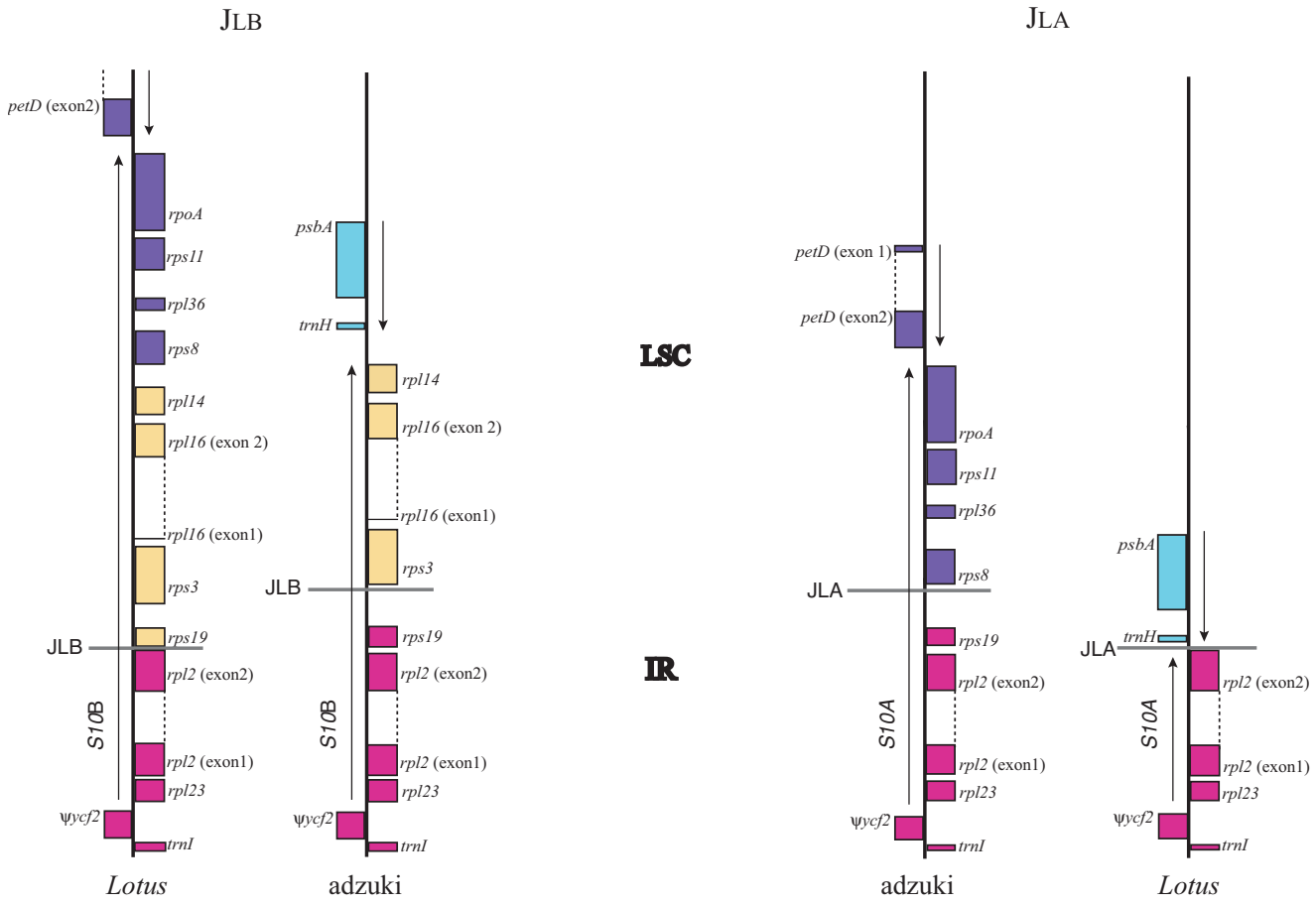

Figure 1. Organisation of the *S10* operon in *Lotus* and adzuki illustrating the effect of the 78-kb rearrangement on the genes in the operon and at the IR/LSC junctions.

considered the more likely for reasons of both parsimony and simplicity, is a straightforward inversion that occurred after the Phaseolinae subtribe (containing mung, common, and adzuki beans) diverged from the Glycininae lineage (including soybean).[4,14] The second model is based on the close proximity of the 78-kb segment termini to the IR/LSC junctions. They proposed a two-step process involving expansion of the IR, resulting in duplication of approximately 4 kb within the *S10* operon. This step is followed by contraction of the IR to the same starting point but with a deletion of the original copy of the gene cluster. The result is a relocation of a block of genes from one end of the LSC to the other (see Fig. 4 in Palmer et al.[4]). This "expansion-contraction" model produces a gene map identical to the inversion model.

The 78-kb inversion endpoints have been located close to the J_{LA} and J_{LB} junctions by restriction mapping,[4] but have not been analysed at the DNA sequence level. To understand the effect of the rearrangement upon the *S10* operon, we sequenced and analysed the two IR/LSC junctions and their surrounding regions in adzuki bean (*Vigna angularis*), which has a restriction map very similar to that of mung bean.[4,15] In doing so, we have addressed two questions. Firstly, where do the precise boundaries of the 78-kb rearrangement lie? This information has enabled us to investigate the nature of this event and its effects on the organisation of the chloroplast genome. Secondly, what effects has this rearrangement had on the *S10* operon? We show that it disrupts the transcriptional linkage between member genes of the *S10* operon, breaking it into two smaller operons, but retaining intact copies of all ten genes.

2. Materials and Methods

2.1. Plasmid DNAs

Plasmid clones harbouring adzuki bean (*Vigna angularis* cv. Erimo-shozu) chloroplast *Pst* I fragments of 17.2 kb and 16.2 kb[15] were provided by Dr. Tetsuo Mikami. The clones were transformed into *E. coli* host strain DH5α and plasmid DNAs were purified using a QIAGEN miniprep kit. Sequencing the ends of these plasmids showed that they extended from a common *Pst* I site in *rps7* in the IR, to *Pst* I sites in *rpl14* (17.2-kb clone) and *rps11* (16.2-kb clone). The parts of these plas-

mids containing *S10* operon genes were sequenced completely on both strands by Agowa (Berlin, Germany) using a primer walking strategy.

2.2. PCR amplification

PCR was used to obtain the sequence of the 3′ ends of both operons. Adzuki bean seeds of Chinese origin were purchased from Munster Wholefoods (Farranfore, Co. Kerry, Ireland) and grown in the laboratory. Plant tissue was ground to a pulp using a pestle and mortar and then homogenised in extraction buffer (2% SDS, 1 M Tris HCl, 50 mM EDTA, 50 mM NaCl, 0.75% β-mercaptoethanol). This was followed by two phenol:chloroform extractions and treatment with RNAseA (10 mg/ml). The DNA sample was then cleaned through a CL-6B sepharose column. Extracted DNAs were used directly for PCR amplification without further purification.

A 0.9-kb PCR product was obtained by amplification using primers from adzuki *rpl14* (5′-ATTCGAATCCTAGGAGCTAG-3′) and a multiple legume species alignment of the 3′ end of *psbA* (5′-ATCCGTAGTTGATAGTCAAGGTCG-3′). Similarly a 1.6-kb product was obtained using primers from adzuki *rps11* (5′-TGTGACTATTACAGATGTACGAGGTC-3′) and a multiple alignment of the 3′ end of *petD* (5′-TGGTTAGGTATTGGAGCAACATTACC-3′). PCR products were cleaned through a CL-6B sepharose column and precipitated with ethanol. Pellets were resuspended in 10 μl of distilled H_2O and sequenced by Agowa.

2.3. Sequence analysis

Gene content in the sequenced clones and PCR products was determined by using each DNA sequence as a query in a BLASTX similarity search against the non-redundant protein database (nr) at the NCBI website. tRNA genes were found using BLASTN. The results were filtered to show results only from the fully sequenced and annotated *Lotus japonicus* genome.[9] The endpoints of the rearrangement were compared by BLASTN. The adzuki bean sequences have been deposited in GenBank (accession numbers AF536225 and AF536226).

3. Results and Discussion

3.1. Gene organization at the ends of the 78-kb rearrangement in adzuki bean

We determined the sequence around the endpoints of the 78-kb rearrangement in adzuki bean cpDNA and compared it to the unrearranged genome of *Lotus japonicus*.[9] The endpoints of the rearrangement are very close to the junctions between the IR and the LSC in both species (Fig. 1).

In *Lotus*, a complete (10-gene) copy of the *S10* operon is transcribed out of the IR and across J_{LB}. We will refer to this copy as *S10B*. The second copy of the operon, designated *S10A*, is truncated and contains only the first two genes (*rpl23* and *rpl2*). The *S10A* operon terminates at the J_{LA} junction, where it meets the *psbA–trnH* operon which is transcribed in the opposite direction. The organization of the *S10B* and *S10A* operons in *Lotus* is similar to that in other angiosperms, except for the missing *rpl22* and *infA* genes. There are also small variations among species as to how much of *rps19* is located inside the IR. In soybean, 68 bp of *rps19* are within the IR,[16,17] so that its *S10A* operon includes a truncated gene designated *rps19′*. In *Lotus*, only 1 bp of *rps19* is located inside the IR, so that in effect *rps19* is only present in the *S10B* operon.

In adzuki, we find that the IR is larger than in *Lotus* and includes a complete copy of *rps19* as well as *rpl23* and *rpl2*. The adzuki *S10B* operon contains the first six of the ten genes in the *Lotus S10B* operon: *rpl23–rpl2–rps19–rps3–rpl16–rpl14*, after which *trnH* and *psbA* are found on the opposite strand. The normal linkage between *rpl14* and *rps8* in *S10B* has been broken in adzuki, and instead the last four genes (*rps8* to *rpoA*) are fused to the *S10A* operon, giving it the structure *rpl23–rpl2–rps19–rps8–rpl36–rps11–rpoA*, after which *petD* is found in the opposite orientation (Fig. 1).

The 78-kb rearrangement of cpDNA in the legume subtribe Phaseolinae was first reported by Palmer et al.,[4] who mapped the breakpoints in mung bean and common bean by Southern analysis with small probes. They estimated that one breakpoint was in the spacer between *rps8* and *infA* (which is normally located between *rps8* and *rpl36* but is now known to be absent from all legume *cpDNAs*[11]), and the other breakpoint was between *trnH* and *rpl2*. Their map is largely confirmed by the adzuki sequence data, except for the placements of *rps8* (which is in *S10A*, not *S10B*), and *rps19* (which is in the IR, not solely in *S10B*). The novel *rpl16–rpl14–trnH–psbA* arrangement in Phaseolinae was previously sequenced in common bean by Carelse et al.,[18] who mistakenly described the *rpl16* and *rpl14* genes as pseudogenes. We have re-sequenced these genes from *Phaseolus vulgaris* and find them to be intact as in adzuki (C. Jones and T. A. Kavanagh, data not shown).

3.2. ycf2 *pseudogene*

The spacer between *trnI* and *rpl23* in both adzuki and *Lotus* is unusually large and contains a duplicated copy of a 265-bp segment of the *ycf2* gene, which lies about 2 kb further along the IR. We assume that this sequence is a pseudogene, similar to other duplicated gene fragments in legume cpDNAs,[19] and have designated it as *ψycf2* (Fig. 1). In *Lotus*, *ψycf2* has 100% sequence identity to *ycf2* itself (we cannot make this comparison in adzuki because we have not sequenced its *ycf2*). We also found a similar *ψycf2* segment at precisely the same location in soybean (EST sequence, GenBank accession

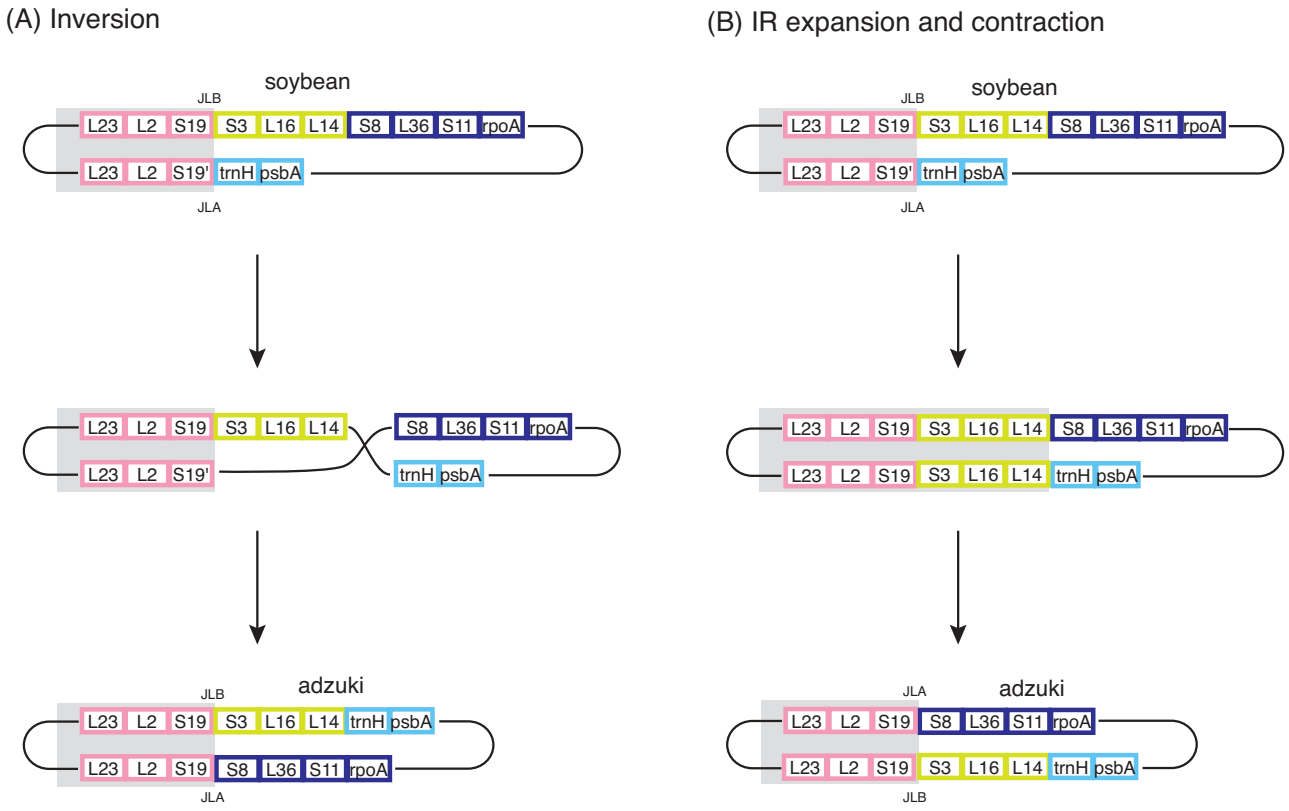

Figure 2. The two models proposed for the evolution of the present-day arrangement of the *S10* operon in Phaseolinae from a soybean-like ancestor. (A), A straightforward inversion brought about by recombination of the spacer regions between *rpl14–rps8* and *rps19′–trnH*. (B), An expansion of the IR (shaded area) into the LSC, followed by a deletion of most of the original gene sequence that became duplicated. The two adzuki structures shown at the bottom of the diagram are indistinguishable, because cpDNA exists as a mixture of isomers with the two possible relative orientations of the LSC and SSC regions.[21] Note that in B, the designations of J_{LA} and J_{LB} in adzuki are reversed as compared to soybean.

number BI785044), and in *Medicago truncatula* cpDNA (GenBank accession number NC_003119) although in *Medicago* there is extensive sequence divergence between *ψycf2* and *ycf2*.

3.3. IR/LSC junctions

Shifts in the border positions between the IR and LSC at J_{LA} and J_{LB} have been reported in several species of angiosperms, demonstrating that the IR/LSC boundaries are not static.[9,17] We find that the IR/LSC junctions have moved in adzuki bean compared with its relatives, soybean and *Lotus* (Fig. 1). In adzuki the IR/LSC junction lies 559 bp downstream from the translation stop codon of *rps19* so that J_{LB} is located in the intergenic spacer between *rps19* and *rps3*, and J_{LA} is located in the spacer between *rps19* and *rps8*.

This structure indicates that the IR (at the IR/LSC junction) has expanded by 770 bp in adzuki as compared to soybean, the most closely related species to adzuki that has been sequenced. This 770 bp encompasses the remaining 211 bp of *rps19* that is not duplicated in soybean, and 559 bp of noncoding DNA downstream of *rps19* forming most of the *rps19–rps3* and *rps19–rps8* intergenic spacers. At J_{LB} *rps3* begins 24 bp into the LSC, while at J_{LA} *rps8* starts 53 bp into the LSC.

3.4. Evolutionary Models

Knowledge of the precise gene order and content at the IR/LSC boundaries in adzuki allows us to consider the two competing models proposed for the evolution of the large rearrangement in the Phaseolinae.[4] These models are a 78-kb inversion and an IR expansion/contraction, as summarized in Fig. 2, which shows how an adzuki-like genome could have been formed from a soybean-like ancestor. For discussion of these models, we have assumed a soybean-like ancestor (with a *rps19′* sequence) rather than a *Lotus*-like ancestor (lacking *rps19′*), both for clarity and because soybean is a closer relative of adzuki.[20]

Under the inversion model (Fig. 2A), breakpoints occurred in the ancestor's *rpl14–rps8* and *rps19′–trnH* spacers. Almost the entire LSC was then inverted, forming new *rpl14–trnH* and *rps19′–rps8* spacers. The difficulty with this model is that adzuki contains a complete *rps19* gene at the new *rps19–rps8* junction. Therefore, even if

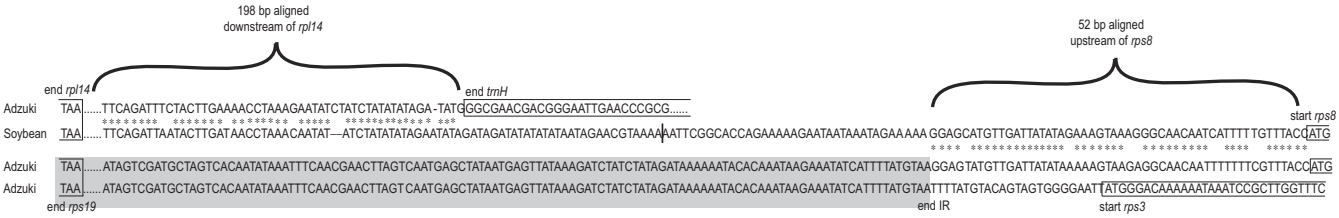

Figure 3. Sequence evidence for the expansion contraction model. Adzuki bean nucleotide sequence around the rearrangement endpoint in *S10B* (*rpl14–trnH*) is aligned with soybean *rpl14–rps8* spacer. The homology between the two (shown by asterisks) ends at *trnH* in adzuki. Homology between the *rpl14–rps8* spacer sequence in soybean and *rps19–rps8* in adzuki begins at the start of the LSC at J_{LA} in adzuki, which is the location of the second rearrangement endpoint. Soybean sequence is from ESTs (GenBank accession numbers BQ298413 and BE607846). The vertical line in soybean indicates the position of a short unsequenced gap between these EST sequences, estimated to be less than 10 bp.

the rearrangement occurred by an inversion mechanism, the IR in Phaseolinae must also have expanded by about 770 bp before the inversion happened, which is unparsimonious.

Under the IR expansion/contraction model (Fig. 2B), the IR in the ancestor expanded so that all the genes between *rpl23* and *rpl14* were within the IR. This formed a new J_{LA} between *rpl14* and *trnH*. The IR then contracted again by deleting *rps3–rpl16–rpl14* from the *S10B* operon, leaving the gene arrangement now seen in adzuki and simultaneously creating new IR/LSC junctions at both ends of the LSC. Contrary to what was envisaged by Palmer et al.,[4] the contraction of the IR did not quite trim it back to its former J_{LA} endpoint (with a short *rps19′* sequence as seen in soybean) but instead left the IR somewhat longer, with 559 bp downstream of *rps19* remaining duplicated.

Sequence evidence favouring the IR expansion/contraction model for the origin of the 78-kb rearrangement is summarized in Fig. 3. First, the spacer between *rpl14* and *trnH* in adzuki can be aligned with the soybean sequence downstream of *rpl14*, almost up to the first base of *trnH*. This is similar to the structure of junctions between *trnH* and *rps19′* seen in other species[17] and strongly suggests that the *rpl14–trnH* junction in adzuki was at one time the endpoint of an extended IR. For this structure to have been formed by an inversion, it is necessary to propose that a recombination breakpoint occurred precisely at the end of *trnH*, which seems unlikely. Second, the sequences upstream of *rps8* can be aligned between soybean and adzuki, but this alignment ends abruptly at the adzuki IR/LSC junction (J_{LA}). This is exactly the structure that would be expected if deletion of the three-gene cluster *rps3–rpl16–rpl14* from *S10B* had occurred in an ancestor of adzuki (see Fig. 2B): the deletion event causes the IR/LSC boundary to shift back to *rps19*, and all of the newly-defined LSC sequence adjacent to this boundary is derived from the old *rpl14–rps8* spacer and thus aligns with the soybean *rpl14–rps8* spacer.

We therefore reject the hypothesis of a straightforward inversion for the evolution of the 78-kb rearrangement in the Phaseolinae, and instead support the alternative model of expansion and contraction of the IR as the most likely mechanism. If this is correct, other Phaseolinae species may exist that retain the expanded IR. The only drawback to this proposal is that we were unable to identify any significant repeated sequences shared by the soybean *rps19–rps3* and *rpl14–rps8* spacers, such as might be expected to catalyse a deletion of the genes between them. However, we were also unable to find any repeats between the adzuki *rpl14–trnH* and *rps19–rps8* spacers such as might have catalysed an inversion.

4. Conclusion

Comparisons of chloroplast genome organisation not only provide us with valuable information for understanding the processes of chloroplast evolution, but also give insight into the mechanisms underlying genomic rearrangement. The expansion-contraction model proposed by Palmer et al.[4] for the evolution of the 78-kb rearrangement in Phaseolinae is supported by our sequence analysis in adzuki bean. This rearrangement disrupted the physical and transcriptional linkage between the 5′ and 3′ ends of a large operon that is conserved in all other land plants. The consequence is that the *S10A* and *S10B* operons must both be essential in Phaseolinae, whereas *S10A* has no apparent function in most angiosperms.[12] That both adzuki and common bean can tolerate such a rearrangement leads us to the conclusion that the translational outcome of the *S10* operon must be unaffected. It is notable that the rearrangement did not create any truncated ribosomal protein genes (similar to *rps19′*) which might, if translated, interfere with ribosome assembly. However, further studies are necessary to establish the transcriptional levels of the *S10A* and *S10B* operons in adzuki. In other angiosperms, the *S10A* transcript is quickly degraded[12] resulting in equal transcript abundance for all ten ribosomal protein genes, but in adzuki it seems impossible to maintain this stoichiometry.

Acknowledgements: We thank Dr. Tetsuo Mikami for plasmids and Claire Jones for assistance. This study was supported by Science Foundation Ireland.

References

1. Palmer, J. D., Nugent, J. M., and Herbon, L. A. 1987, Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families, *Proc. Natl. Acad. Sci. USA*, **84**, 769–773.
2. Palmer, J. D. 1985, Comparative organization of chloroplast genomes, *Annu. Rev. Genet.*, **19**, 325–354.
3. Shinozaki, K., Ohme, M., Tanaka, M. et al. 1986, The complete nucleotide sequence of tobacco chloroplast genome: its gene organisation and expression, *EMBO J.*, **5**, 2043–2049.
4. Palmer, J. D., Osorio, B., and Thompson, W. F. 1988, Evolutionary significance of inversions in legume chloroplast DNAs, *Curr. Genet.*, **14**, 65–74.
5. Woodbury, N. W., Roberts, L. L., Palmer, J. D., and Thompson, W. F. 1988, A transcription map of the pea chloroplast genome, *Curr. Genet.*, **14**, 75–89.
6. Kanno, A. and Hirai, A. 1993, A transcription map of the chloroplast genome from rice (*Oryza sativa*)., *Curr. Genet.*, **23**, 166–174.
7. Tanaka, M., Wakasugi, T., Sugita, M., Shinozaki, K., and 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.
8. Sugiura, M. 1992, The chloroplast genome, *Plant Mol. Biol.*, **19**, 149–168.
9. 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.
10. 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.
11. Millen, R. S., Olmstead, R. G., Adams, K. L. et al. 2001, Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus, *Plant Cell*, **13**, 645–658.
12. Tonkyn, J. C. and Gruissem, W. 1993, Differential expression of the partially duplicated chloroplast *S10* ribosomal protein operon, *Mol. Gen. Genet.*, **241**, 141–152.
13. Bruneau, A. J., Doyle, J. J., and Palmer, J. D. 1990, A chloroplast DNA structural mutation as a subtribal character in the Phaseoleae (Leguminosae)., *Syst. Bot.*, **15**, 378–386.
14. Bruneau, A. J., Doyle, J. J., and Doyle, J. L. 1994, In: Crisp, M. and Doyle, J. J. (eds) Advances in Legume Systematics, Phylogeny, Royal Botanic Gardens, *Kew*, **7**, 309–330.
15. Kato, S., Yamaguchi, H., Shimamoto, Y., and Mikami, T. 2000, The chloroplast genomes of azuki bean and its close relatives: a deletion mutation found in weed azuki bean, *Hereditas*, **132**, 43–48.
16. Spielmann, A. and Stutz, E. 1983, Nucleotide sequence of soybean chloroplast DNA regions which contain the *psbA* and *trnH* genes and cover the ends of the large single copy region and one end of the inverted repeat, *Nucleic Acids Res.*, **11**, 7157–7167.
17. 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.
18. Carelse, O., Chetsanga, C. J., and Mubumbila, M. V. 1994, Identification of a *trnH* (GUG) gene and of several pseudogenes downstream of *psbA* gene in the common bean chloroplast genome, *Plant Sci.*, **98**, 47–52.
19. Wolfe, K. H. 1988, The site of deletion of the inverted repeat in pea chloroplast DNA contains duplicated gene fragments, *Curr. Genet.*, **13**, 97–99.
20. Doyle, J. J., Doyle, J. L., Ballenger, J. A., Dickson, E. E., Kajita, T., and Ohashi, H. 1997, A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: Taxonomic correlations and insights into the evolution of nodulation, *Amer. J. Bot.*, **84**, 541–554.
21. Palmer, J. D. 1983, Chloroplast DNA exists in two orientations, *Nature*, **301**, 92–93.