Curr Genet (1988) 14:487–492

Current Genetics

Sequence analysis of the junction of the large single copy region and the large inverted repeat in the petunia chloroplast genome

Jane Aldrich, Barry W. Cherney*, Charlotte Williams, and Ellis Merlin*****

BP America, Research and Development, 4440 Warrensville Center Road, Cleveland, OH 44128, USA

Summary. We have determined the nucleotide sequence at the junction of the large single copy (LSC) region and the right and left members of the large inverted repeat, IRA and IRB, respectively, of the petunia chloroplast (cp) genome. As in *Nicotiana debneyi* and spinach (Zurawski et al. 1984), coding sequences of *rps*19 of petunia overlap the junction of IRB and LSC. Immediately into the LSC region upstream of IRA in the petunia cp genome are two small insertions relative to *N. debneyi* that occur at sites just inside IRA of *N. debneyi*. We discuss how these additions in one copy of the large inverted repeat of an *N. debneyi*-like ancestor to petunia resulted in shortening of the petunia large inverted repeat by 8 bases and in the resultant slight movement of *rps*19 farther into LSC. On a larger scale, the large inverted repeat in the tobacco, *N. debneyi* and petunia lineage relative to a spinach-like ancestor may have sustained several contractions due to deletions between short direct repeats found within IRA and the IRA/LSC junction. We also show how the large inverted repeat of *N. debneyi* instead may have been expanded relative to a tobacco-like ancestor by insertion into the large inverted repeat of bases between short inverted repeat sequences in LSC and the LSC/IRB junction.

Key words: Petunia large single copy – Inverted repeat junction nucleotide sequences – *trn*H – *rps*19 – Inverted repeat expansion/contraction

Present addresses:
* Department of Biochemistry, Georgetown University, Washington, DC 20057, USA
** Gene Trak Systems, 45 New York Avenue, Framingham, MA 01701, USA
*** CIBA-GEIGY, P.O. Box 12257, Research Triangle Park, NC 27709, USA

Offprint requests to: J. Aldrich

Introduction

The most common feature of the cp genome of angiosperms is the presence of a large inverted repeat with an average size of 20–30 kilobase pairs (kb) (Palmer and Stein 1986) and a range of 10 kb in coriander (Palmer 1985a) to 76 kb in geranium (Palmer et al. 1987a). In contrast, the large inverted repeat in more primitive land plants ranges from 9.4–17 kb, with a size of 9.4 kb for the moss *Physcomitrella patens* (Calie and Hughes 1987), 10 kb for the fern *Osmunda cinnamomea* (Palmer and Stein 1986), 11 kb for the liverwort *Marcantia polymorpha* (Ohyama et al. 1986), and 17 kb for the gymnosperm, *Ginkgo biloba* (Palmer and Stein 1986).

An exception to the expansion of the size of the large inverted repeat in the angiosperm cp genome compared to the early land plants is found in the 8 genera of 4 tribes of legumes that lack a large inverted repeat (Palmer 1985b). Cross hybridization studies using as probes cloned restriction endonuclease fragments from the inverted repeat-containing mung bean cp genome have shown that one of the two orientations of IRA has been deleted in representatives of 5 genera studied: alfalfa, wisteria, subclover, broad bean and pea (Palmer et al. 1987b). These observations suggest that the mutation resulting in this loss occurred once in an ancestral legume.

The mechanism whereby the large inverted repeat expands or contracts is unknown. However, inversions and insertions/deletions in the cp genome have been

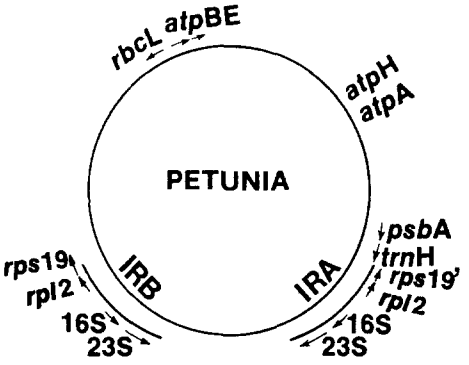

Fig. 1. Chloroplast genome map of petunia. Locations of 16S and 23S rRNA genes and location and extent of IRA and IRB are from Bovenberg et al. 1984. Loci of *rbc*L and *psb*A are from Palmer et al. 1983. Location of *trn*H is from Aldrich et al. 1988. *rps*19 and *rps*19′ and *rpl*2 loci are contributed from this report. Location of *atp*BE is from Bovenberg et al. 1984, and direction of transcription is from sequence analysis, Aldrich et al. 1985

shown to be associated with short repeated sequences (reviewed in Palmer et al. 1985a and Aldrich et al. 1988). Nucleotide sequence comparison of the junction of the LSC region with IRA and IRB from plants in a related grouping might leave a record of the nature of the expansion, contraction, or loss of the inverted repeat. To this end, we have analyzed the nucleotide sequence at the LSC/IRA and IRB border from members of the Solanaceae including petunia (this report), *N. debneyi* (Zurawski et al. 1984) and tobacco (Sugita et al. 1984) and show that short repeated sequences may be involved in the movement of the border.

Materials and methods

DNA sequencing. The 1.4 kb PstI DNA restriction endonuclease fragment of the petunia cp genome containing *trn*H (Aldrich et al. 1988) and 3′ to it, IRA, (Palmer et al. 1983) was partially sequenced according to the method of Dale et al. (1985) as described in Aldrich et al. (1988). A cloned 1.5 kb PstI DNA restriction endonuclease fragment which spans the junction of the left member of IRB and LSC (Palmer et al. 1983) was partially sequenced using synthetic oligonucleotide primers in the supercoil sequencing protocol of Chen and Seeburg 1985. The oligonucleotides were constructed at Case Western Reserve University, Cleveland, Ohio with the kind help of Dr. Pieter DeHaseth.

DNA sequence similarity determinations. DNA sequence similarity was assessed using the SEQH program (Kanahisa and Goad 1982) modified by E. Merlin and J. Cleary (unpublished) and run on a Digital Equipment Corp. VAX 11/780 computer.

Results

Figure 1 shows a schematic of the cp genome of petunia and the relative position of known genes. Both *psb*A and *trn*H are immediately adjacent to IRA in the LSC region. The determination of the location of *rps*19, *rps*19′ and *rpl*2 is described herein from sequence analysis. Our strategy for locating the actual junctions of IRA and IRB with LSC on the petunia cp genome was to first compare the petunia nucleotide sequence 3′ to *trn*H to those of spinach and *N. debneyi* (Zurawski et al. 1984), where the junction of IRA with LSC was shown to be within a few bases of *trn*H. Then, oligonucleotides that were identical, but of opposite polarity, to petunia sequences putatively assigned to IRA, based on sequence identity to bases within IRA and close to the IRA/LSC junction of spinach and *N. debneyi* (Zurawski et al. 1984) were constructed. These were used as primers in double-stranded DNA sequencing (Chen and Seeburg 1985) of the pBR322-cloned petunia 1.5 kb PstI DNA restriction endonuclease fragment that contains the junction of IRB-LSC (Palmer et al. 1983). Sequence comparisons of the bases in IRB with IRA would be expected to show 100% identity up to the junction with LSC, after which they would diverge as shown for tobacco, (Sugita et al. 1984) and spinach (Zurawski et al. 1984).

The nucleotide sequence 3′ to *trn*H in petunia is compared to the same region from spinach and *N. debneyi* (Zurawski et al. 1984) in Fig. 2A. The first base in IRA of all three plants is starred. The location of the IRA and IRB junction with the LSC region of petunia was deduced as follows. Both *N. debneyi* and petunia share nearly the same deletion of the bases located between the boxed, shaded imperfect direct repeats of spinach (Fig. 2A). Immediately following the deleted sequences in *N. debneyi* relative to spinach, close sequence similarity is apparent and starts with the first base within IRA of *N. debneyi*. By comparison to *N. debneyi* and spinach, three small insertions (Fig. 2A and B) occur in the petunia nucleotide sequence 3′ to *trn*H, after which there is close nucleotide sequence similarity to the truncated *rps*19′ of *N. debneyi* and spinach in IRA (Zurawski et al. 1984). Thus, *rps*19 must traverse the junction of IRB/LSC in petunia as it does in *N. Debneyi* and spinach.

In order to sequence across IRB/LSC of the petunia cp genome contained in the 1.5 kb PstI DNA restriction endonuclease fragment, two oligonucleotides of opposite polarity to bases that span the reversed complement of the initiation codon of *rps*19′ in IRA shown in Fig. 2 were constructed. Oligonucleotide 5′AGTGAACGTGTCACAAGCT3′ was used to prime synthesis in the 3′ direction into IRB. Its complement, 3′TCACTTGCACAGTGTTCGA5′ primed synthesis in

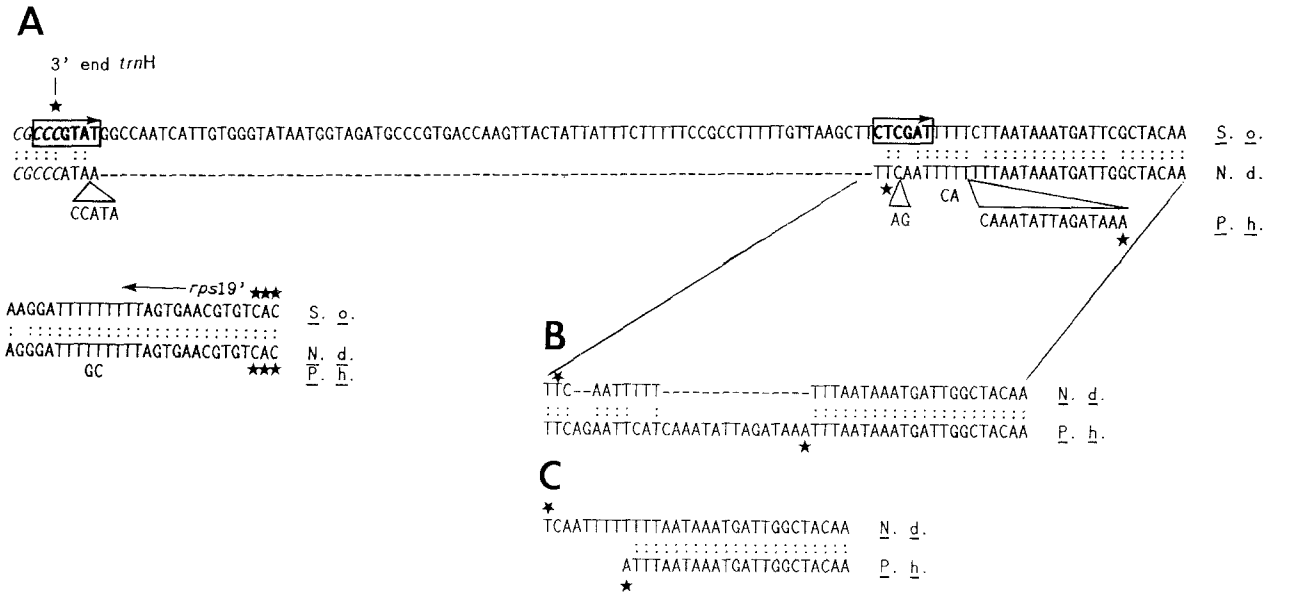

Fig. 2A–C. Nucleotide sequence comparison in the region 3′ to *trn*H of petunia, *N. debneyi* and spinach. **A** Comparison of nucleotide sequences from the 3′ end of *trn*H (*in italics*) of spinach, *N. debneyi* and petunia through IRA up to the reverse complement of the start codon of *rps*19′. Plant abbreviations: *S. o.* (*S. oleracea*, spinach) *top line*, *N. d.*, (*N. debneyi*) *second line*; *P. h.* (*Petunia hybrida*, petunia) *bottom line*. Petunia sequences are identical to *N. debneyi* except where substitutions occur or where insertions are noted by *triangles*. Sequences are aligned by the introduction of *dashes* in one sequence relative to the other and identical bases are indicated by *colons*. *Boxed*, shaded imperfect direct repeat sequences with a sequence similarity of 71.4% in spinach surround bases missing in *N. debneyi* and petunia. The first base of IRA in the three sequences is indicated by a *star*. Nucleotide sequences and location of IRA for *N. debneyi* and spinach are from Zurawski et al. 1984. IRA/LSC border determination of petunia is described in Fig. 4. The reverse complement of the start codon of *rps*19′ is indicated by *three stars*. **B** Sequence organization at the boundary of IRA and the LSC region of *N. debneyi* (*N. d.*, *top line*) and petunia (*P. h.*, *bottom line*). **C** Length of IRA of *N. debneyi* compared to petunia: first base within IRA indicated by a star

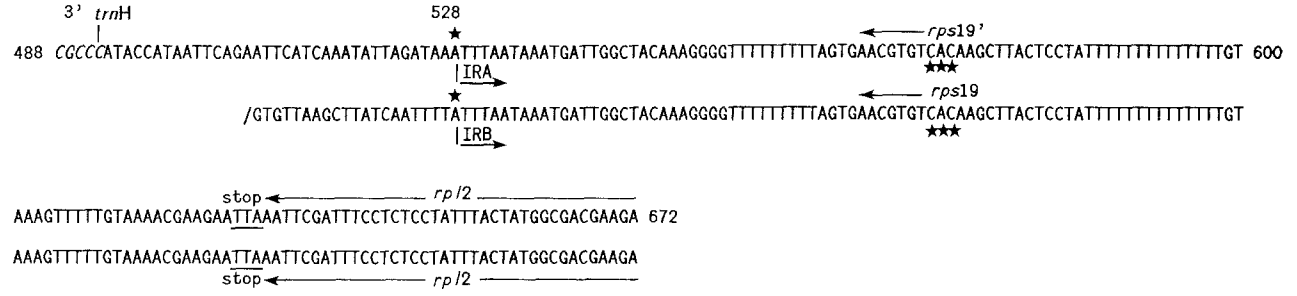

Fig. 3. Comparison of petunia nucleotide sequence from the 3′ end of *trn*H (*in italics*) into IRA (*top line*) with the sequence in the strand of the opposite polarity from within IRB up to /. Bases in the top line are numbered from the stop codon of *psb*A 5′ to *trn*H. Bases in IRB (*second line*) are aligned with their identical bases in IRA up to the first base in IRA and IRB (marked by a *star* in both lines; base position 528, top line). The locations of *rps*19, *rps*19′ and *rpl*2 were discerned by nucleotide sequence comparison with spinach and *N. debneyi* (Zurawski et al. 1984).

the 3′ direction out of IRB into the LSC region. The sequences so obtained are compared to those 3′ of *trn*H as shown in Fig. 3. Complete sequence identity is apparent up to the junctions of IRA and IRB with the LSC region where the first base within IRA and IRB is marked with an asterisk after which there is complete sequence divergence. Once the IRB/LSC junction was determined, a third oligonucleotide identical to the nucleotide sequence outside of IRB (see Fig. 3), 5′GTGTTAAGCTTATCAATT3′, was constructed to obtain nucleotide sequence information for the nucleotides immediately adjacent to the first two primers as these nucleotides were too close to the primers to be detected on sequencing gels. Detection of *rpl*2 was by sequence identity to the same gene from spinach and *N. debneyi* (Zurawski et al. 1984).

Because of the dynamism noted in the insertions/deletions at the border of IRA/LSC in the petunia/*N.*

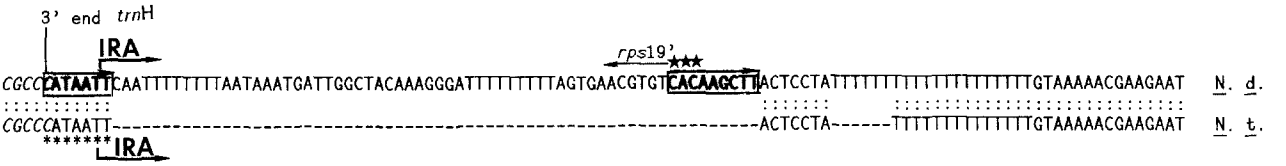

Fig. 4. Nucleotide sequence comparison of the coding strand beginning with the 3′ end of *trn*H (*in italics*) of *N. debneyi* (*N. d.*) and tobacco (*N. t.*). Nucleotide sequence similarity was maximized by introducing hyphens in one sequence relative to the other to indicate missing bases. Bases in direct repeat orientation are shown in *shaded boxes* in *N. debneyi*. Only the left member of the direct repeat present in *N. debneyi* is present in tobacco (*indicated by asterisks*). The reverse complement of the start codon of *rps*19′ is indicated by *stars*. The location of the junction of IRA with LSC of tobacco is from Sugita et al. 1984, and that of *N. debneyi* is by deduction from spinach and tobacco *rps*19 and *rps*19′ amino acid sequence similarity rather than by actual nucleotide sequence analysis of IRA/LSC and IRB/LSC (Zurawski et al. 1984). Nucleotide sequence in the 5′ region upstream of *rps*19′ coding sequences of *N. debneyi* is that of *rps*19 (Zurawski et al. 1984) as sequence information for the homologous region of *rps*19′ was not available

debneyi and spinach comparisons shown in Fig. 2, it was of interest to compare the IRA/LSC border of tobacco with *N. debneyi* (see Fig. 4). Nucleotides in *N. debneyi* that are missing in tobacco are flanked by imperfect direct repeats. The left member of the repeat, CATAATT contains bases encompassing the 3′ end of *trn*H, and the right member contains bases spanning the reverse complement of the start codon of *rps*19′, CACAAGCTT (Fig. 4). Recombination between these (now) imperfect direct repeats (sequence similarity 66.7%) could have generated the shortened tobacco sequence. Alternatively, as shown in Fig. 5, base pairing of short inverted repeated sequences at the IRB/LSC border and 3′ to it in the LSC region in a tobacco-like ancestor could have formed a stem-loop structure that was partially copy-corrected into IRA leading to the expanded inverted repeat of *N. debneyi*.

Discussion

Examination of the nucleotide sequences immediately 3′ to *trn*H of the cp genomes examined shows that this region, which encompasses the junction of LSC/IRA, is a hotspot for insertion/deletion mutations. Several putative deletion events that shortened the inverted repeat may have occurred between what are now recnized as imperfect direct repeats, as seen in the spinach, *N. debneyi* and petunia comparisons (Fig. 2A), and in the tobacco and *N. debneyi* comparison (Fig. 4). A result of the deletions of bases between the repeats, one member within IRA and the other at the IRA/LSC junction, is the shortening of IRA and the concommittant shortening of IRB and contraction of the entire large inverted repeat. If coding sequences, for example, *rps*19, span the junction of IRB/LSC, shortening of the inverted repeat has the consequence of movement of some of the bases in the coding sequence from within the inverted repeat to the LSC. The presence of a direct repeat whose members share 71.4% sequence similarity in the spinach sequence that surround the bases missing in *N. debneyi* and petunia (Fig. 2A) suggests that the latter two plants may have been derived from an ancestral spinach-like plant via deletion of sequences between these (now) imperfect direct repeats. Tobacco also has the same deletion plus the one described in Fig. 4 relative to *N. debneyi*. Consequently, relative to spinach, more of the coding sequences of *rps*19 of tobacco, *N. debneyi* and petunia are in LSC.

It is equally plausible that the directionality of movements is toward a "spreading" (Palmer 1985a) of the inverted repeat with bases from the LSC moving into IRB. The mechanism of spreading has been envisioned (Palmer 1985a) to involve short inverted repeat sequences within LSC and the border of one of the large inverted repeats, for example, IRB, at the same time IRA and IRB are paired. This leads to the formation of a stem-loop structure at the border of the large paired repeat that is replicated into IRB and "copy-corrected" during replication into IRA. Such a stem-loop structure at IRB/LSC of tobacco is shown in Fig. 5. Copy correction of a portion of this stem-loop structure could have enlarged the large inverted repeat in *N. debneyi* relative to a tobacco-like ancestral plant. A similar structure in the spinach IRB relative to tobacco, *N. debneyi* or petunia was not detected.

Another mechanism for shortening of the inverted repeat is the addition of bases just within the boundary of IRA/LSC as shown in petunia relative to *N. debneyi* and spinach (Fig. 2A). We propose that these are small additions in petunia relative to *N. debneyi* rather than deletions in *N. debneyi* relative to petunia because a) no *rps*19 sequences of other plants have the bases included in the insertions in petunia relative to *N. debneyi*, b) *rps*19′ nucleotide sequences of spinach and *N. debneyi* are colinear up to the junction of IRA/LSC of *N. debneyi* (Fig. 2A), c) *N. debneyi* and soybean (Spielmann and Stutz 1983) *rps*19′ nucleotide sequences (comparison not shown) are also colinear up to the junction of IRA/LSC of *N. debneyi*, and d) *N. debneyi* and spinach share

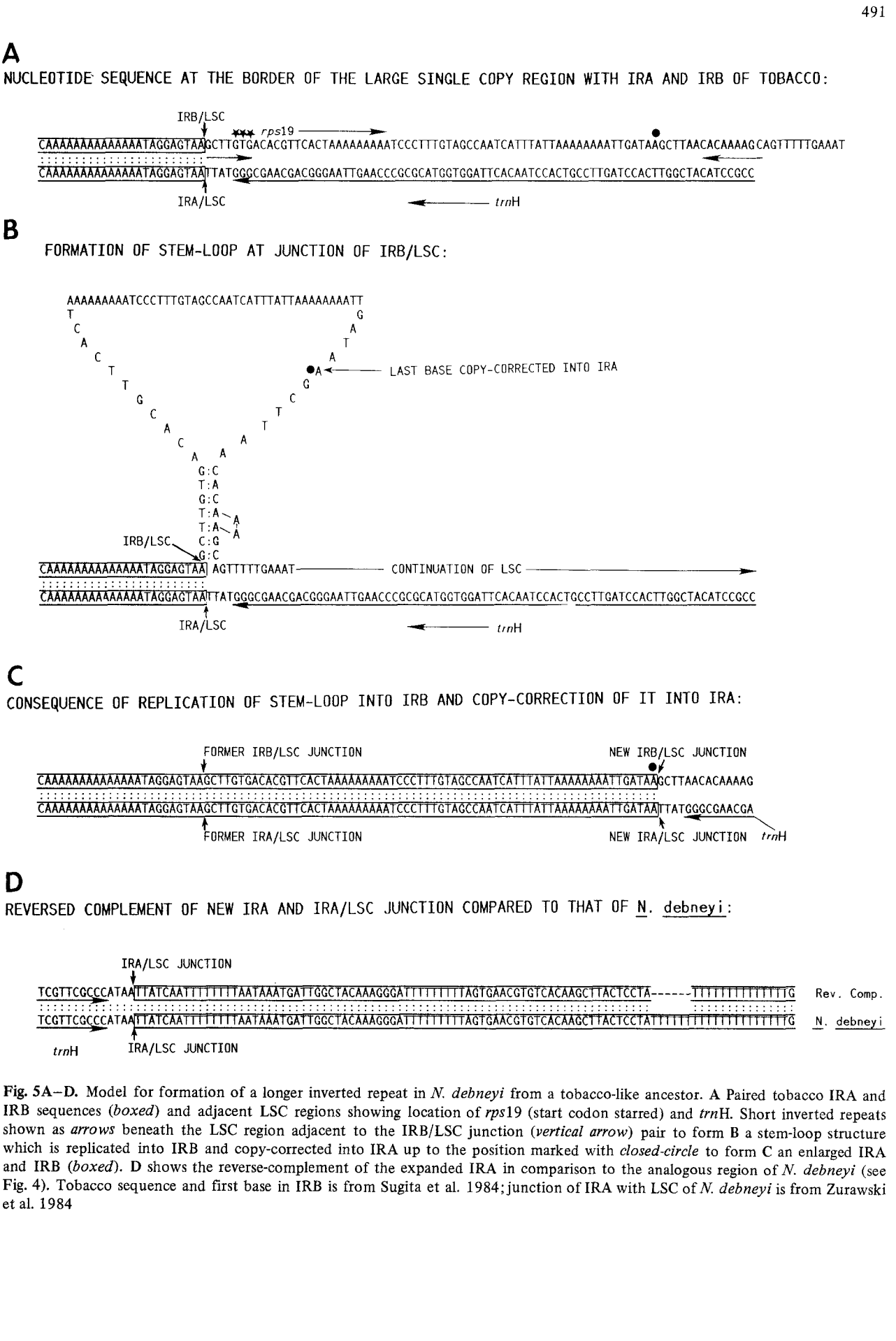

Fig. 5A–D. Model for formation of a longer inverted repeat in *N. debneyi* from a tobacco-like ancestor. A Paired tobacco IRA and IRB sequences (*boxed*) and adjacent LSC regions showing location of *rps*19 (start codon starred) and *trn*H. Short inverted repeats shown as *arrows* beneath the LSC region adjacent to the IRB/LSC junction (*vertical arrow*) pair to form B a stem-loop structure which is replicated into IRB and copy-corrected into IRA up to the position marked with *closed-circle* to form C an enlarged IRA and IRB (*boxed*). D shows the reverse-complement of the expanded IRA in comparison to the analogous region of *N. debneyi* (see Fig. 4). Tobacco sequence and first base in IRB is from Sugita et al. 1984; junction of IRA with LSC of *N. debneyi* is from Zurawski et al. 1984

90.2% sequence identity in the 61-base intergenic region between *trn*H and the reversed complement of the start codon of *rps*19′, with the deletion in *N. debneyi* relative to spinach counted as one event (Fig. 2A). These additions could have been copy-corrected into IRB to extend the size of the inverted repeat, but instead, both IRA and IRB were shortened by 8 bases compared to *N. debneyi* (see Fig. 2C). The likely reason for shortening vs expansion of the inverted repeat in this case is that the coding sequence of *rps*19 would have been disrupted by copy-correction of these additions, an event that probably would have been lethal. Shortening of the petunia inverted repeat by 8 bases results in a slight repositioning of *rps*19 into LSC from IRB of petunia relative to *N. debneyi*. It is interesting to note that a tandem duplication, CCATA, includes the last two bases of *trn*H of petunia relative to *N. debneyi* (Fig. 2A). Imperfect tandem duplications (not shown) may be also detected in the other additions in petunia compared to *N. debneyi*. Whether these duplications are the source of the additional bases cannot be discerned from the sequence record.

The directionality of the insertion/deletion mutations that have changed the position of the junction of the large inverted repeat with the large single copy region cannot yet be discerned in most cases. However, small repeated sequences, whether in a direct or inverted repeated orientation may play a pivotal role in the expansion/contraction of the large inverted repeat as they appear to in other regions of the chloroplast genome. It is interesting to note in this regard that induced nonphotosynthetic mutants of *Chlamydomonas reinhardii* have been shown to contain large inverted repeat expansion and contraction mutations (Palmer et al. 1985) whose endpoints map close to members of a family of small repeated elements.

Acknowledgements. We thank D. Brink, C. Cullis, S. Horowitz, J. D. Palmer, and L. Swofford for helpful criticisms of the manuscript.

References

Aldrich J, Cherney B, Merlin E, Christopherson L, Williams C (1985) In: Galau GA (ed) First International Congress of Plant Molecular Biology, Savannah, Ga, p 126

Aldrich J, Cherney B, Merlin E, Christopherson L (1988) Curr Genet 14:137–146

Bovenberg WA, Howe CJ, Kool AJ, Nijkamp HJJ (1984) Curr Genet 8:283–290

Calie PJ, Hughes KW (1987) Mol Gen Genet 208:335–341

Chen EY, Seeburg PH (1985) DNA 4:165–170

Dale RMK, McClure BA, Houchins JP (1985) Plasmid 13:31–40

Kanahisha M, Goad W (1982) Nucleic Acids Res 10:247–264

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H (1986) Nature 322:572–574

Palmer JD (1985a) Annu Rev Genet 19:325–354

Palmer JD (1985b) In: MacIntyre RJ (ed) Monographs in evolutionary biology: molecular evolutionary genetics. Plenum, New York, pp 131–240

Palmer JD, Stein DB (1986) Curr Genet 10:823–833

Palmer JD, Shields CR, Cohen DB, Orton TJ (1983) Theor Appl Genet 65:181–189

Palmer JD, Boynton JE, Gillham NW, Harris EH (1985) In: Steinback KE, Bonitz S, Arntzen CJ, Bogorad L (eds) Molecular biology of the photosynthetic apparatus. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 269–278

Palmer JD, Nugent JM, Herbon LA (1987a) Proc Natl Acad Sci USA 84:769–773

Palmer JD, Osorio B, Aldrich J, Thompson WF (1987b) Curr Genet 11:275–286

Spielmann A, Stutz E (1983) Nucleic Acids Res 11:7157–7167

Sugita M, Kato A, Shimada H, Sugiura M (1984) Mol Gen Genet 194:200–205

Zurawski G, Bottomley W, Whitfeld PR (1984) Nucleic Acids Res 12:6547–6558

Communicated by R. W. Lee

Received April 4, 1988 / July 27, 1988