LBNL-59348

Complete chloroplast genome of *Trachelium caeruleum*: extensive rearrangements are associated with repeats and tRNAs

Rosemarie C. Haberle[1], Matthew L. Fourcade[2], Jeffrey L. Boore[2], and Robert K. Jansen[1]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas, Austin TX 78712; [2]DOE Joint Genome Institute, Walnut Creek CA 94598

Abstract

Chloroplast genome structure, gene order and content are highly conserved in land plants. We sequenced the complete chloroplast genome sequence of *Trachelium caeruleum* (Campanulaceae) a member of an angiosperm family known for highly rearranged chloroplast genomes. The total genome size is 162,321 bp with an IR of 27,273 bp, LSC of 100,113 bp and SSC of 7,661 bp. The genome encodes 115 unique genes, with 19 duplicated in the IR, a tRNA (*trnI-CAU*) duplicated once in the LSC and a protein coding gene (*psbJ*) duplicated twice, for a total of 137 genes. Four genes (*ycf15*, *rpl23*, *infA* and *accD*) are truncated and likely nonfunctional; three others (*clpP*, *ycf1* and *ycf2*) are so highly diverged that they may now be pseudogenes. The most conspicuous feature of the *Trachelium* genome is the presence of eighteen internally unrearranged blocks of genes that have been inverted or relocated within the genome, relative to the typical gene order of most angiosperm chloroplast genomes. Recombination between repeats or tRNAs has been suggested as two means of chloroplast genome rearrangements. We compared the relative number of repeats in *Trachelium* to eight other angiosperm chloroplast genomes, and evaluated the location of repeats and tRNAs in relation to rearrangements. *Trachelium* has the highest number and largest repeats, which are concentrated near inversion endpoints or other rearrangements. tRNAs occur at many but not all inversion endpoints. There is likely no single mechanism responsible for the remarkable number of alterations in this genome, but both repeats and tRNAs are clearly associated with these rearrangements.

Land plant chloroplast genomes are highly conserved in structure, gene order and content. The chloroplast genomes of ferns, the gymnosperm *Ginkgo*, and most angiosperms are nearly collinear, reflecting the gene order in lineages that diverged from lycopsids and the ancestral chloroplast gene order over 350 million years ago (Raubeson and Jansen, 1992). Although earlier mapping studies identified a number of taxa in which several rearrangements have occurred (reviewed in Raubeson and Jansen, 2005), an extraordinary number of chloroplast genome alterations are concentrated in several families in the angiosperm order Asterales (sensu APGII, Bremer *et al.,* 2003). Gene mapping studies of representatives of the Campanulaceae (Cosner, 1993; Cosner *et al.,*1997, 2004) and Lobeliaceae (Knox *et al.,* 1993; Knox and Palmer, 1999) identified large inversions, contraction and expansion of the inverted repeat regions, and several insertions and deletions in the cpDNAs of these closely related taxa. Detailed restriction site and gene mapping of the chloroplast genome of *Trachelium caeruleum* (Campanulaceae) identified seven to ten large inversions, families of repeats associated with rearrangements, possible transpositions, and even the disruption of operons (Cosner *et al.,* 1997). Seventeen other members of the Campanulaceae were mapped and exhibit many additional rearrangements (Cosner *et al.*, 2004). What happened in this lineage that made it susceptible to so many chloroplast genome rearrangements? How do normally very conserved chloroplast genomes change? The cause of rearrangements in this group is unclear based on the limited resolution available with mapping techniques. Several mechanisms have been proposed to explain how rearrangements occur: recombination between repeats, transposition, or temporary instability due to loss of the inverted repeat (Raubeson and Jansen, 2005). Sequencing whole chloroplast genomes within the Campanulaceae offers a unique opportunity to examine both the extent and mechanisms of rearrangements within a phylogenetic framework.

We report here the first complete chloroplast genome sequence of a member of the Campanulaceae, *Trachelium caeruleum*. This work will serve as a benchmark for subsequent, comparative sequencing and analysis of other members of this family and close relatives, with the goal of further understanding chloroplast genome evolution. We confirmed features previously identified through mapping, and discovered many additional structural changes, including several partial to entire gene duplications, deterioration of at least four normally conserved chloroplast genes into gene fragments, and the nature and position of numerous repeat elements at or near inversion endpoints.

The focus of this paper is on analyses of sequences at or near these rearrangements in *Trachelium caeruleum*. Inversions are believed to occur due to the presence of repeat elements subject to homologous recombination (Palmer, 1991; Knox *et al.,* 1993). Repeats may facilitate inversions or other genome rearrangements (Achaz *et al.,* 2003), and higher incidences of repeats have been correlated with greater numbers of rearrangements (Rocha, 2003). Alternatively, repeats may proliferate within a genome as a result of DNA strand repair mechanisms following a rearrangement event such as an inversion. Gene mapping studies previously identified five families of dispersed repeats in *Trachelium* at or near inversion endpoints (Cosner *et al.,* 1997). Here we examine the sequences of these repeats and identify, map and characterize numerous additional repeats within the genome. We compare the number and size of repeats in typical unrearranged angiosperm chloroplast genomes to what we find in the highly rearranged chloroplast genome of *Trachelium*. The *Trachelium* chloroplast genome has the highest number and the largest repeats of diverse origin of any sequenced angiosperm chloroplast genome. These repeats are generally clustered at or near rearrangements and they are of diverse origins: partial or entire chloroplast gene duplications, noncoding chloroplast sequences or novel DNA with no clear sequence identity to any existing chloroplast DNA sequences. The *Trachelium* chloroplast genome represents the most highly rearranged sequenced genome of land plants and its bizarre organization is clearly associated with the high incidence of dispersed repetitive DNA.

Materials and Methods

Sample acquisition, cpDNA isolation and DNA sequencing. *Trachelium caeruleum* plants were purchased from a local nursery and grown in the UT-Austin research greenhouses. Plants were placed in the dark for 24 hours prior to harvesting leaves, and a voucher specimen (RCHaberle XXX) is deposited at TEX. A chloroplast-DNA enriched sample was isolated from living material using the sucrose gradient method (Palmer 1986). The DNA was sheared into approximately 3 kb pieces using a Hydroshear device (Genemachines, San Carlos, CA, USA) and then shotgun cloned using a pUC18 plasmid vector to create libraries for sequencing. Colonies were picked, amplified using RCA and sequenced using BigDye™ terminators (Applied Biosystems). Detailed sequencing protocols are available at http://www.jgi.doe.gov/sequencing/protocols/protsproduction.html. PHRED /PHRAP (Ewing and Green, 1998) were used to assemble the 600-800 bp reads, and visualized using CONSED (Gordon *et al.*, 1998) and SEQUENCHER (Gene Codes Corp 2003). The draft sequence included areas of low coverage as well as gaps between contigs. We developed primers that amplified chloroplast enriched DNA from the original isolation to enhance coverage in low areas and to fill in gaps with a minimum of two reads with a PHRED/PHRAP quality score (q value) greater than 20 (Jansen *et al.*, 2005). Using SEQUENCHER, we manually reconstructed part of the second copy of the inverted repeat (IR) as automated PHRED/PHRAP assembly cannot distinguish between reads that belong in one or the other copy. This allowed us to produce a complete circular genome with both copies of the IR for annotation and analysis.

Genome annotation and analysis. The *Trachelium* chloroplast genome sequence was annotated by submitting a FASTA-formatted input file to DOGMA (Dual Organellar GenoMe Annotator, Wyman *et al.*, 2004; http://evogen.jgi-psf.org/dogma). This program locates and identifies protein-coding genes based on BLASTX searches of a custom database of 15 previously published chloroplast genomes. Genes for tRNAs and rRNAs are located by BLASTN searches of the same database. To visualize gene content and divergence between *Trachelium* and other angiosperms, the *Trachelium* sequence was run against the published sequences of eight other angiosperm chloroplast genomes (*Nicotiana* (Z00044), *Amborella* (AJ506156), *Calycanthus* (AJ428413), *Nymphaea* (AJ627251), *Arabidopsis* (AP000423), *Spinacia* (AJ400848), *Lotus* (AP002983), and *Zea* (X86563)) using MULTIPIPMAKER (Schwartz *et al.* 2003).

To compare the gene order of *Trachelium* to typical, unrearranged angiosperm chloroplast genomes, we numbered the genes in the *Nicotiana* chloroplast genome from 1-116, starting from *trnH* (1) at coordinate 6 and proceeding through the large single copy, one copy of the inverted repeat and the small single copy to *ycf1* (116) at coordinate131, 594. This included protein coding genes, tRNAs, rRNAs and ycfs (hypothetical reading frames) but not orfs (open reading frames) specific to *Nicotiana*.

We used this numbering system to number genes in *Trachelium* to map the gene order relative to *Nicotiana*.

Size and locations of direct and inverted repeats in the *Trachelium* chloroplast genome were determined by running REPUTER (Kutrz *et al.*, 2001) at a repeat length ≥ 30 bp with Hamming distance of 3. Repeats were mapped onto the *Trachelium* chloroplast genome, and those located at or near inversion endpoints and other sites of rearrangement were characterized by BlastN searches in GenBank. We ran the same REPUTER analyses against the identical eight angiosperm chloroplast genomes as were used for MULTIPIPMAKER to assess the number of repeats in chloroplast genomes with few if any rearrangements. BlastN searches of intergenic regions between blocks of inverted gene sequences were performed against GenBank.

Results

Organization of the *Trachelium* chloroplast genome. The complete chloroplast genome sequence of *Trachelium caeruleum* is 162,321 bp, with an IR of 27,273 bp separating a large single copy (LSC) region of 100,113 bp and a small single copy (SSC) region of 7,661 bp (**Fig. 1**). The G + C content is 38.3%; within coding regions it is 40.59%, in noncoding regions it is 35.6%. Coding regions comprise 59.67% of the genome.

Gene content. *Trachelium* has 115 unique genes, 19 of which are duplicated in the IR, plus one (*trnI-CAU*) duplicated one time in the LSC and another (*psbJ*) duplicated twice, giving a total of 137 genes. *Trachelium* has 75 protein coding genes of known function, five ycfs (hypothetical chloroplast reading frames), four rRNAs and 30 tRNAs; unlike most land plant chloroplast genomes it has a number of partial or entire gene duplications, and several truncated genes (discussed below). Seventeen genes contain introns; the intron in *rps16* is absent in *Trachelium*. Whole genome alignment of the *Trachelium* chloroplast genome against eight published angiosperm chloroplast genomes shows high conservation of some coding regions as well as marked divergence in others (**Fig 2a).** The genes *ycf15, rpl23, infA*, and *accD* are abbreviated and likely nonfunctional. The truncated *ycf15* in *Trachelium* aligns with the first 191 bases of the *Oenothera γycf15*. In *Trachelium*, the last 50 bp of the 3’ end of *rpl23* is all that remains.

This occurs at an inversion endpoint in the LSC between gene blocks 1-14 and 86-69. A 34 bp repeat of this *rpl23* gene fragment occurs at another inversion endpoint in the LSC (between gene blocks 38-36 and 19-15). The gene *infA* is reduced to a fragment consisting of 191 bp of the middle of the gene, but lacking the 5' and 3' ends. We found a 290 bp fragment of *accD* in the *Trachelium* genome, embedded in the highly diverged *ycf1* gene in the IR in the vicinity of a number of other rearrangements.

Three other genes, *clpP*, *ycf1* and *ycf2*, have diverged greatly from other angiosperms and it is not known if they are functional. MULTIPIPMAKER alignment shows that these seven reduced or altered genes still align with intact copies of these genes in other angiosperm chloroplast genomes (**Fig. 2b**).

Gene order. *Nicotiana* exhibits the typical gene order of angiosperm chloroplast genomes, and we compared the gene order of *Nicotiana* to *Trachelium*. We found 18 conserved blocks of genes in *Trachelium* rearranged relative to the *Nicotiana* chloroplast genome (**Fig. 1)**. These blocks ranged in size from 4 to 17 kb with otherwise unrearranged blocks of genes relocated and reoriented in comparison to chloroplast genomes of most angiosperms**.** The gene order in *Trachelium* is further altered by the insertion of entire genes or fragments of genes from other parts of the genome between some of the otherwise conserved blocks of genes.

Location of tRNAs in relation to rearrangements. In *Trachelium*, there are four locations in the LSC in which tRNAs occur at the ends of conserved gene blocks: *trnT-UGU*, *trnM-CAU*, *trnC-GCA*, and one copy of *trnI-CAU*, and two that occur in the IR and are duplicated (*trnL-CAA* and *trnN-GUU*) (**Fig.1, arrows).** In two other cases, a tRNA has been relocated to lie in between gene blocks. The second copy of *trnI-CAU* (gene number 87) occurs in the LSC between conserved blocks 39-46 and 35-20; *trnV-GAC* (gene number 94) is moved from its normal IR location to the LSC between gene blocks 86-69 and 66-55.

Repeats in the *Trachelium* chloroplast genome

Number and sizes of repeats. Repeat analysis in the chloroplast genomes of *Trachelium* and eight other angiosperms shows that all genomes have multiple repeats, many of which are mono- or dinucleotide strings and likely to be microsatellites. Although repetitive DNA typically occurs in angiosperm chloroplast genomes, *Trachelium* has the highest number of repeats and the largest repeats among all genomes compared, including *Zea*, which has three rearrangements (Palmer and Thompson, 1982) (**Fig. 3)**. REPUTER has limitations that affect the results because it over reports the number of repeats by counting every paired repeat (so counting repeats with multiple copies additional times) and recounting nested repeats. However, since all nine chloroplast genomes were analyzed in the same way, the overall result is that *Trachelium* has many more and larger repeats, and suggests a strong correlation between the number of repeats and the extensive rearrangements in the genome.

In *Trachelium*, many repeat elements were found at some but not all of the inversion endpoints and at or near other rearrangements, such as gene duplications and disruption of operons. The length, orientation and coordinates for these repeats were pinpointed (REPUTER output, **supplementary material)** and mapped**.** A total of 767 direct and inverted repeats ≥ 30 bp with a Hamming distance of 3 was identified in *Trachelium* with 483 direct and 284 inverted repeats, and mapped onto the *Trachelium* genome (**Fig. 4, middle circle)**. Of these, 303 occurred either as parts of genes or in intergenic spacers in conserved blocks of genes. 464 repeats occur either between inverted gene blocks or near other rearrangements, clearly showing an association between repeats and rearrangements. These were characterized by BlastN searches against GenBank. Repeats at or near inversion endpoints are of diverse origins: derived from protein coding regions within the chloroplast genome (i.e. partial to entire gene duplications, discussed below), a tRNA (*trnI-CAU*), noncoding cpDNA, and novel DNA not previously identified as being chloroplast in origin (**Table 1**).

Inversion endpoints as hotspots for rearrangements and repeats. Multiple rearrangements and repeats of diverse origin are concentrated between blocks of inverted genes in the *Trachelium* chloroplast genome. For example, between conserved blocks of genes 86-69 and 66-55 in the LSC region, 2.7 kb of sequences normally found in multiple other sites in an unrearranged chloroplast genome are found in the space where genes 68 (*clpP*) and 67 (*5'rps12*) would typically be found between *psbB* (gene 69) and *rpl20* (gene 66) (**Fig. 4, shaded area A**). *trnV-GAC* (gene 94) which is normally found within the IR, has been moved from there into this endpoint in the LSC. Additionally, a duplicate, functional copy of *psbJ* (gene 55) is also inserted into this area (**Fig. 4, r6**). Finally, repeats of non-coding cpDNA sequences from different areas of the genome are located within this one particular hotspot between *psbB* and *rpl20*, flanking *trnV* and *psbJ* (**Fig.4**, **r4, r5, r7**).

Another inversion endpoint with a complex content occurs between blocks of genes 66-55 and 39-46 (**Fig. 4, shaded area B**). The copy of *psbJ* (**Fig. 4**, **r6**) at this endpoint is the original copy, in its operon with *psbF, psbE, and psbL.* A 105 bp repeat of noncoding chloroplast DNA sequence (**Fig. 4**, **r7**) is shared with the copy of *psbJ* between 86-69 and 66-55 but not with the third copy of *psbJ* located approximately 25 kb counterclockwise of the original copy of *psbJ*, within block 35-20. All three copies of *psbJ* are direct repeats on the plus strand. Finally, a small repeat of part of the *clpP* exon 1 is located here (**Fig. 4**, **r8**). The entire, presumably functional copy of *clpP* is located in the IR (see below).

The most complex rearrangements in the *Trachelium* chloroplast genome occur within the IR in association with multiple repeats (**Fig. 4, shaded area C**). A 4.6 kb portion of sequence normally found in the LSC as well as several smaller duplicated sequences from the LSC and the IR are inserted into one heterogeneous area between gene blocks 95-102 and 116-110 (**Fig. 5**). The first two genes of the *clpP* operon, c*lpP* and the *5'rps12* (genes 68 and 67, respectively), were moved here in their entirety, and a 1013 bp repeat of sequence normally found adjacent to the start of this operon was duplicated and moved as well (**r3).** This includes an identical copy of the first 300 bp of *psbB* (gene 69) plus an intergenic spacer between the functional copy of *psbB* and *trnV-GAC* at the 86-69/66-55 inversion endpoint in the LSC. Exon 1 of *clpP* contains a large insertion. Besides the disruption of the *clpP-5'rps12-rpl20* operon by this rearrangement, *clpP* and *5'rps12* are separated by additional insertions of part of the 3rd exon of *ycf3* (**r10**) and a 457 bp repeat of non-coding sequence (**r11)** from the vicinity of the functional copy of *ycf3* within the LSC. Immediately adjacent to this area is a very divergent copy of *ycf1*, into which a 290 bp vestige of the *accD* is inserted. An identical 499 bp repeat of the 5' end of the *23S rrn* gene (**r12**) is found between *ycf1* and *rps15.*

DISCUSSION

Genome Organization. The complete *Trachelium* chloroplast genome sequence is far more complex than originally described based on restriction site and gene mapping (Cosner *et al.*, 1997). Although many other genomes have been identified as having multiple rearrangements, the *Trachelium* genome shows an extraordinary number of genome rearrangements, including partial to entire gene duplications, several gene reductions and the loss of an intron, numerous large inversions and a concentration of repeats and tRNAs at or near inversion endpoints.

Gene duplications have been infrequently reported in chloroplast genomes. The *psbA* duplication in some ferns (Stein *et al.*, 1992), and numerous duplications of normally single copy genes in *Pelargonium* (Palmer *et al.*, 1987) have been attributed to expansion of the IR. Partial duplications of tRNAs have been reported in taxa known for rearranged chloroplast genomes, in grasses (Hiratsuka *et al.*, 1989; Tsai and Strauss, 1989; Hipkins *et al.*, 1995) and legumes (Mubumbila *et al.*, 1984). Wolfe (1988) suggested that the partial duplications of *rbcL* and *psbA* in *Pisium* are associated with loss of one copy of the IR; this was recently supported by the findings of Saski *et al*., (2005) as having occurred simultaneous to the loss of the IR in an entire clade of legumes. The duplication of *psaM* and several tRNAs in black pine (Wakasugi *et al.*, 1994) may be due to the inherent instability caused by severe reduction of the IR. Duplications of tRNAs have been recently reported in otherwise unrearranged chloroplast genomes, for example *Arabidopsis* and related taxa in the Brassicaceae (Koch *et al.*, 2005). The presence of three copies of *psbJ* in the LSC of *Trachelium* is puzzling. One of the *psbJ* duplications in *Trachelium* may have been caused by expansion and contraction of the IR and subsequent inversions, as this copy occurs at an inversion endpoint. The second duplication occurs within an otherwise unrearranged block of genes. This suggests some other mechanism may be responsible, perhaps a duplicative transposition, which has been suggested in the generation of dispersed repeats in conifers (Tsai and Strauss, 1989) and subclover (Milligan *et al.*, 1989). There is no direct evidence of transposable elements within the *Trachelium* genome, although they may have been present transiently. Both of these duplications must have occurred relatively recently as they have 100% sequence identity to the original copy, and are being maintained in the genome, presumably as functional copies. The duplication of *trnI-CAU* is also associated with an inversion because the extra copy occurs between two conserved gene blocks. A 20 kb inversion in rice and generation of a tRNA pseudogene was attributed to recombination between two tRNA genes (Hiratsuka *et al*., 1989). Another hypothesis for the *trnI-CAU* duplication in *Trachelium* entails generation of a tandem repeat of *trnI-CAU* that was subsequently moved during the course of inversions. Whether the duplication is responsible for the inversion due to nonhomolgous recombination between one of the adjacent tRNAs and the original copy of *trnI-CAU*, or the result of an error in the repair of a double strand break cannot be determined from these data. There is only a single base pair difference between the copies.

Another striking feature of the *Trachelium* genome in relation to other angiosperm chloroplast genomes is the partial loss of four genes: *ycf15*, *rpl23*, *infA* and *accD*. These four genes have been lost or altered in other chloroplast genomes but not all four in the same genome. The conserved reading frame *ycf15* has been shown to be variable among angiosperm chloroplast genomes with conserved 5' and 3' ends and an intervening 250 bp in some taxa that renders it a pseudogene (Schmitz-Linneweber *et al.*, 2001). *rpl23* is a pseudogene in spinach (Thomas *et al.*, 1988, Schmitz-Linneweber *et al.*, 2001) and a pseudogene copy persists in grasses in the LSC, with intact copies in the IR (Morton and Clegg, 1993). In *Trachelium*, *rpl23* is neatly severed after 50 bp of well-conserved sequence; the truncation may have occurred as a result of an inversion and/or in a process involving recombination between repeats, as there is a partial repeat of this fragment elsewhere in the LSC. Millen *et al.* (2001) found 24 independent losses or reduction of *infA* in a survey of 308 angiosperms, including *Campanula*, *Trachelium* and *Platycodon* (Campanulaceae) and two members of their sister family, Lobeliaceae, but the gene is present in other members of the Asterales. This indicates that the loss or reduction of *infA* occurred in the recent common ancestor of the Campanulaceae/Lobeliaceae clade. Our sequence of *infA* in *Trachelium* confirms the earlier evidence found in southern hybridization data that the gene is reduced to a fragment. Earlier mapping studies of *Trachelium* and other members of the Campanulaceae and Lobeliaceae reported that *accD* is absent in both families, and this is a synapamorphy supporting their sister relationship (Downie and Palmer, 1992; Cosner *et* *al*., 1997; Knox and Palmer, 1999). *accD* is also lost in the Poaceae and close relatives (Hiratasuka *et al.*,1989; Downie and Palmer, 1992; Maier *et al.,* 1995; Katayama and Ogihara, 1996; Ogihara *et al.,* 2002) from a “hotspot” between *rbcL* and *psaI* into which an *rpl23* pseudogene has been inserted. The fragment of *accD* that we found in the IR of the *Trachelium* genome in the vicinity of a number of other rearrangements suggests that there may be something intrinsic to this gene that makes it vulnerable to rapid change. The loss of functional copies of *ycf15, rpl23, infA* and *accD* and the extreme divergence of *ycf1, ycf2, and clpP* may be the result of multiple inversion events, in which repetitive motifs within the genes have made them more susceptible as targets of nonhomologous recombination, or a replication error like slipped strand mispairing. *infA* has been transferred to the nucleus in at least four unrelated lineages (Millen *et al.,* 2001). It is conceivable that the other fragmented genes have been transferred to the nucleus as well, as a result of the general instability of this genome.

Large inversions and the evolutionary influence of repeats. Even in unrearranged chloroplast genomes, small inversions occur regularly in intergenic areas, caused by short (11-24 bp) inverted repeats forming hairpins that can easily flip-flop (Kelchner and Wendel, 1996; Kelchner, 2000; Kim and Lee, 2005). Larger (> 200 bp) inversions are found in some angiosperm chloroplast genomes, but generally not more than a few within a genome. A number of possible mechanisms have been proposed for these events. Inversions may occur in a specific location due to the presence of short repeat elements subject to homologous recombination (Palmer, 1991; Knox *et al*.,1993). In the grasses, with three large inversions, repeats flank the borders of a 28-kb inversion and may have facilitated the inversion, or nonhomologous recombination between tRNA genes may have caused the rearrangement (Hiratsuka *et al*., 1989; Sugiura 1989). A 54 kb inversion in *Oenothera elata* has a series of small inverted repeats at each end (Hupfer *et al.,* 2000). In the Ranunculaceae, Hoot and Palmer (1994) found up to six inversions in certain taxa, ranging in size from 5.6 kb to 53.6 kb. They proposed that certain inversions might have positioned repeat sequences in a way that would cause subsequent inversions. They also noted the similarity of inversion endpoints in *Anenome* to those reported by Knox *et al.,* (1993) in other rearranged chloroplast genomes of lineages quite distant to the Ranunculaceae, including the Lobeliaceae, sister to the Campanulaceae.

Comparison of the size and number of repeats in *Trachelium* and eight other angiosperm chloroplast genomes shows that there is a background of repeats even in unrearranged chloroplast genomes. Polymorphic, simple sequence repeats (SSR) < 15 bp have been identified in many chloroplast sequences and in all completely sequenced land plant chloroplast genomes (Provan *et al.,* 2001). Short dispersed repeats have been associated with inversion endpoints and occur in a number of taxa (in *Pelargonium*, Palmer *et al.,* 1987, wheat, Howe, 1985, Quigley and Weil, 1985, Bowman and Dyer, 1986; Bowman *et al.,* 1988; Ogihara *et al.,* 1988; rice, Shimada and Sugiura, 1989; sub clover, Milligan *et al.,* 1989; Douglas fir, Tsai and Strauss, 1989). A recent comparison of four chlorophyte algal chloroplast genomes showed a strong correlation between the number of repeats in the chloroplast genome and the degree of rearrangement (Pombert *et al*., 2005). The most highly rearranged green algal chloroplast genome is *Chlamydomonas reinhardtii* (Maul *et al*., 2002), which also has the greatest number of repeats in this lineage (Pombert *et al*., 2005).

In *Trachelium*, the most conspicuous alterations in the chloroplast genome are its large (> 4 kb), multiple inversions and relocation of blocks of genes. *Trachelium* also has the most and largest repeats in comparison with eight other angiosperm chloroplast genomes. It has a concentration of repeats of diverse origin at or near these inversion endpoints and other rearrangements, such as the cluster of rearrangements within the IR. With only a few exceptions, the repeats were direct repeats, not the inverted repeats one would expect to be associated with inversions (Palmer, 1991). It is possible that short inverted repeats were responsible for some inversions in the *Trachelium* genome, but were reoriented to direct repeats as a result of inversion, or have diverged or been eliminated over time. Our parameters for repeat searches were quite stringent at ≥ 30 bp and a Hamming distance of 3. Many repeats that REPUTER reports smaller than 30 bp are likely not significant biologically in terms of rearrangements because they represent strings of mono- or dinucleotides and are frequently polymorphic (Marshall *et al*., 2001). Less stringent searches yield many more repeats: with a 20 bp window and Hamming distance of 4, we found over 30,000 repeats (data not shown).

Chloroplast genome inversions have also been attributed to nonhomologous recombination between different tRNAs (Knox *et al*., 1993; Hoot and Palmer, 1994). There are tRNAs present at or near twelve out of seventeen inversion endpoints in the highly rearranged chloroplast genome of the charophyte *Chaetospheridium globosum* (Turmel *et al.,* 2002). In *Trachelium,* tRNAs may be implicated in some inversions because there are tRNAs at the ends of ten of eighteen rearranged blocks of genes.

The most rearranged region in the *Trachelium* genome is in the IR, where within a 12.5 kb area there are two partial gene duplications of genes present in their entirety in the LSC (*psbB* and *ycf3*), a partial duplication of *23Srrn*, and the relocation of *clpP* and *5'rps12* from the LSC, with a remnant of *accD* nearby within the highly divergent gene *ycf1*. Although a series of inversions might explain the relocation of these coding sequences into the IR, selection is believed to constrain against inversions between the LSC and the IR, and within operons and genes (Palmer, 1991). Loss of one copy of the IR in an ancestor to *Trachelium* would remove the constraint against LSC/IR inversions, however gene mapping data of other Campanulaceae shows this to be unlikely (Cosner, 1993; Cosner *et al.,* 1997). Several of these other taxa lack the *clpP-5'rps12* occurrence in the IR, and are basal to *Trachelium*, but share the same IR/SSC boundary unique to the Campanulaceae/Lobeliaceae. This suggests that the transfer of these genes into the IR occurred after establishment of this boundary. Transposition may be the most parsimonious explanation for how these genes and gene fragments became concentrated into this one area. The sole known example of a transposon in a chloroplast genome occurs in the highly rearranged cpDNA of the green alga *Chlamydomonas reinhardtii*, with two copies of a disabled transposable element, "Wendy" (Fan and Mosig, 1995). This is not found in any other chloroplast genome including any close relatives of *C.reinhardtii*. This element invaded *C.reinhardtii* after its divergence from other species and is responsible for its unique genome relative to its close relatives. This suggests that it is possible for a mobile element to invade a chloroplast genome and generate structural rearrangements, although no transposon-like element has been found in the *Trachelium* chloroplast genome.

These explanations for complex genome rearrangements are predicated on the idea of a circular genome, which replicates in a manner that maintains the integrity of the genome (Kolodner and Tewari, 1975). Recent fluorescent microscopy studies show that chloroplast genomes may exist at least part of the time as multigenomic, branched structures or as linear strands (Oldenburg and Bendich, 2004a; 2004b). The presence of even transient single strands or dimers would increase the possibility of inter- and intramolecular recombination, but how chloroplast genomes persist as stable and conservatively evolving units is not clear, if this scenario is accurate.

Models of evolution. Several programs are currently available that model pairwise genome rearrangements to generate scenarios representing the minimum number of steps for a given ancestral gene order to evolve to the subject genome. Unfortunately, these programs are inadequate to deal with the complexities of a genome as highly rearranged as *Trachelium*'s. GRIMM (Tesler, 2002) is limited to comparing genomes with identical gene content and permutations only by inversions, with no allowance for transposition. DERANGE2 (Blanchette, 1996) allows differential weighing for inversions, transpositions and inverted transpositions, but also requires identical gene content.

Neither allows weighing inversions or other rearrangements between regions in the genome, for example, between the single copy and IR regions. The *Trachelium* gene order data has to be so severely distorted to run either of these programs that neither can provide a biologically realistic model of how this genome evolved.

Based on gene mapping, Cosner *et al.* (1997) proposed three different evolutionary models for how the *Trachelium* chloroplast genome may have originated from an ancestral unrearranged chloroplast genome. Each model proposed events that are unusual in chloroplast genomes. They favored a model based on seven inversions, contraction and expansion of the IR, and a transposition to move *clpP* and *5'rps12* from the LSC into the IR. The two other models avoided using transposition as a mechanism by allowing only inversions and extreme and temporary contraction or loss of the IR with subsequent regrowth.   Not included in any of the models were partial to entire gene duplications or losses, or cpDNA insertions and deletions, many of which were not identified through mapping.   Although our data do not refute the earlier preferred model of Cosner *et al.*, (1997), the additional information provided by whole genome sequencing indicates that the rearrangement of the *Trachelium* genome was far more circuitous than originally proposed.   Some of these previously undetected events such as the duplication of one copy of *psbJ* and the truncation and partial duplication of *rpl23* occur between blocks of inverted genes. Did these duplications arise concurrently with the inversion? Did they occur prior to the inversion and contribute to it? The *Trachelium* sequence shows gene duplications, partial gene losses, a proliferation of repeats of different origins concentrated near rearrangements, and multiple cases of possible transposition that complicate evolutionary models based on inversions and other events described earlier.

The evidence for the mechanisms responsible for structural rearrangements in the *Trachelium* chloroplast genome may have been lost over evolutionary time. The genome has such a high incidence of rearrangements and an accumulation of repeats that something must have happened within this lineage that has made it susceptible to instability. Earlier mapping studies found at least 42 inversions in 18 Campanulaceae chloroplast genomes, at least 8 examples of possible transpositions, and multiple different IR expansions/contractions (Cosner, 1993; Cosner *et al.*, 2004). Preliminary analysis of the draft sequences of seven other Campanulaceae chloroplast genomes (R.Haberle and R.Jansen, unpubl.) suggests that there are many other rearrangements in these genomes and that many of these are also associated with repeats.

Chloroplasts are eubacterial in origin, and a homologue of one of the bacterial recombination proteins, *RecA*, has been reported in the chloroplasts of *Pisum* (Cerruti *et al.*, 1993), *Arabidopsis* (Cerutti *et al.*, 1995), and *Chlamydomonas reinhardtii* (Nakazato *et al*., 2003). There is likely a nuclear-encoded *Rec*-like repair system that contributes to the overall high fidelity of chloroplast genomes. Could an ancestral chloroplast genome within the Campanulaceae lineage have developed a faulty repair mechanism?

Completion of the *Trachelium* chloroplast genome sequence raises many questions that can be best addressed through comparative analysis with complete genome sequences of other, closely related taxa. Are the rates of structural or nucleotide substitutions in this group any faster than groups with unrearranged chloroplast genomes? Do they have an accelerated rate of transfer of genes to the nucleus? Do they share partial loss or deterioration of the same genes, and are there stages of these changes apparent among these relatives? Are the repeats in other members of the Campanulaceae associated with repeats and the same rearrangements as in *Trachelium*? Using tools of comparative chloroplast genomics may reveal clues to the underlying causes of structural evolution in this unusual group that would be obscured over time in more distantly related taxa.

Acknowledgements

Funding for this project was provided by a grant from NSF (DEB 0120709) to RKJ and JLB. The authors thank Stacia Wyman for computational assistance, and Tim Chumley and Gwen Gage for technical assistance in generating figures 1 and 4. We also thank Andrew Alverson and Katie Hansen for their helpful comments and suggestions on an earlier version of this manuscript. This paper represents a portion of a Ph.D. thesis of RCH in Botany at the University of Texas at Austin.

References

Achaz, G., E. Coissac, et al. (2003). "Associations between inverted repeats and the structural evolution of bacterial genomes." Genetics 164(4): 1279-1289.

Blanchette, M., T. Kunisawa and D. Sankoff (1996). "Parametric genome rearrangement." Gene 172: 11-17.

Bowman, C. M. and T. A. Dyer (1986). "The location and possible evolutionary significance of small dispersed repeats in wheat ctdna." Curr. Genet. 10(12): 931-941.

Bowman, C. M., R. F. Barker, et al. (1988). "In wheat ctdna, segments of ribosomal-protein genes are dispersed repeats, probably conserved by nonreciprocal recombination." Curr. Genet. 14(2): 127-136.

Bremer, B., K. Bremer, et al. (2003). "An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II." Bot. J. Linn. Soc. 141(4): 399-436.

Cerutti, H., A. M. Johnson, et al. (1995). "Inhibition of chloroplast DNA recombination and repair by dominant-negative mutants of Escherichia-coli recA." Mol.Cell. Biol. 15(6): 3003-3011.

Cerutti H., H. Z. I., A.T. Jagendorf (1993). "Treatment of pea (Pisum sativum L.) protoplasts with DNA-damaging agents induces a 39-kilodalton chloroplast protein immunologically related to Esherichia coli RecA." Plant Physiol. 102(1): 155-163.

Cosner, M. E. (1993). Phylogenetic and molecular evolutionary studies of chloroplast DNA variation in the Campanulaceae. Ph.D. thesis. Department of Plant Biology. Columbus, Ohio, The Ohio State University.

Cosner, M. E., R.K. Jansen, J.D. Palmer, S.R. Downie (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.

Cosner, M. E., L. A. Raubeson, et al. (2004). "Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes." BMC Evol Biol 4.

Downie, S. R. and J. D. Palmer. (1992). Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Plant molecular systematics. P. Soltis, D. Soltis, and J. J. Doyle. New York, Chapman and Hall: 14-35.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. error probabilities." Genome Res 8(3): 186-194.

Fan, W. H., M. A. Woelfle, et al. (1995). "2 Copies of a DNA element, Wendy, in the chloroplast chromosome of Chlamydomonas-reinhardtii between rearranged gene clusters." Plant Mol Biol 29(1): 63-80.

Gordon, D., C. Abajian, et al. (1998). "Consed: A graphical tool for sequence finishing." Genome Res 8(3): 195-202.

Hipkins, V. D., K. A. Marshall, et al. (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(6): 572-579.

Hiratsuka, J., H. Shimada, et al. (1989). "The complete sequence of the rice (*Oryza sativa*) chloroplast genome - intermolecular recombination between distinct transfer-RNA genes accounts for a major plastid DNA inversion during the evolution of the cereals." Mol Gen Genetics 217(2-3): 185-194.

Hoot, S. B. and J. D. Palmer (1994). "Structural rearrangements, including parallel inversions, within the chloroplast genome of *Anemone* and related genera." J Mol. Evol. 38(3): 274-281.

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(2): 139-145.

Howe, C. J., R. F. Barker, et al. (1988). "Common features of 3 inversions in wheat chloroplast DNA." Curr.Genet. 13(4): 343-349.

Hupfer, H., M. Swiatek, et al. (2000). "Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable Euoenothera plastomes." Mol. Gen. Genet. 263(4): 581-585.

Jansen, R. K., L. A. Raubeson, et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. Molecular Evolution: Producing the Biochemical Data, Part B. Methods Enzymol. 395: 348-384.

Katayama, H. and Y. Ogihara (1996). "Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA." Curr. Genet. 29(6): 572-581.

Kelchner, S. A. (2000). "The evolution of non-coding chloroplast DNA and its application in plant systematics." Ann. Missouri Bot. Gard. 87(4): 482-498.

Kelchner, S. A. and J. F. Wendel (1996). "Hairpins create minute inversions in non-coding regions of chloroplast DNA." Curr. Genet. 30(3): 259-262.

Kim, K. -J., K. –S. Choi, and R. K. Jansen. (2005). Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae)." Mol Biol Evol 22(9): 1783-1792.

Kim, K. -J. and H. L. Lee (2005). "Widespread occurrence of small inversions in the chloroplast genomes of land plants." Molecules and Cells 19(1): 104-113.

Knox, E. B., S. R. Downie, et al. (1993). "Chloroplast genome rearrangements and the evolution of giant lobelias from herbaceous ancestors." Mol. Biol. Evol. 10(2): 414-430.

Knox, E. B. and J. D. Palmer (1999). "The chloroplast genome arrangement of *Lobelia thuliniana* (Lobeliaceae): expansion of the inverted repeat in an ancestor of the Campanulales." Pl. Syst. Evol. 214(1-4): 49-64.

Koch, M. A., C. Dobes, et al. (2005). "Evolution of the trnF(GAA) gene in *Arabidopsis* relatives and the Brassicaceae family: monophyletic origin and subsequent diversification of a plastid pseudogene." Mol. Biol. Evol. 22(4): 1032-1043.

Kolodner, R. D. and K. K. Tewari (1975). "Chloroplast DNA from higher-plants replicates by both cairns and rolling circle mechanism." Nature 256(5520): 708-711.

Kurtz, S., J. V. Choudhuri, et al. (2001). "REPuter: the manifold applications of repeat analysis on a genomic scale." Nucleic Acids Res. 29(22): 4633-4642.

Maier, R. M., K. Neckermann, et al. (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(5): 614-628.

Marshall, H. D., C. Newton, et al. (2001). "Sequence-repeat polymorphisms exhibit the signature of recombination in lodgepole pine chloroplast DNA." Mol Biol Evol 18(11): 2136-2138.

Maul, J. E., J.W. Lilly, L. Cui, C.W. dePamphilis, W. Miller, E.H. Harris, and D.B. Stern. (2002). "The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats." Plant Cell 14: 2659-2679.

Millen, R. S., R. G. Olmstead, et al. (2001). "Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus." Plant Cell 13(3): 645-658.

Milligan, B. G., J. N. Hampton, et al. (1989). "Dispersed repeats and structural reorganization in subclover chloroplast DNA." Mol Biol Evol. 6:355-368.

Morton, B. R. and M. T. Clegg (1993). "A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbclL* in the grass family (Poaceae)." Curr. Genet. 24(4): 357-365.

Mubumbila, M., E. J. Crouse, et al. (1984). "Transfer-RNAs and transfer-RNA genes of *Vicia fabia* chloroplasts." Curr. Genet. 8(5): 379-385.

Nakazato, E., H. Fukuzawa, et al. (2003). "Identification and expression analysis of cDNA encoding a chloroplast recombination protein REC1, the chloroplast *RecA* homologue in *Chlamydomonas reinhardtii*." Biosci. Biotech. Biochem.   67(12): 2608-2613.

Ogihara, Y., T. Terachi, and T.Sasakuma. (1988). "Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species." PNAS 85(22): 8573-8577.

Formatted: Spanish (Colombia)

Ogihara, Y., K. Isono, T. Kohima, A. Endo, M. Hanaoka, T. Shiina, T. Terachi, S. Utsugi, M. Murata, N. Mori, *et al.* (2002). "Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA." Mol. Gen.Genomics 266(5): 740-746.

Oldenburg, D. J. and A. J. Bendich (2004a). "Changes in the structure of DNA molecules and the amount of DNA per plastid during chloroplast development in maize." J.Mol.Biol. 344(5): 1311-1330.

Oldenburg, D. J. and A. J. Bendich (2004b). "Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms." J. Mol. Biol. 335(4): 953-970.

Palmer, J. D., and W.F. Thompson (1982). "Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost." Cell 29: 537-550.

Palmer, J. D. and D. B. Stein (1982). "Chloroplast DNA from the fern *Osmunda cinnamomea* : physical organization, gene localization and comparison to angiosperm chloroplast DNA." Curr.Genet. 5(3): 165-170.

Palmer, J. D., J. M. Nugent, et al. (1987). "Unusual structure of geranium chloroplast DNA - a triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families." PNAS 84(3): 769-773.

Palmer, J. D. (1991). Plastid chromosomes: structure and evolution. Molecular biology of plastids. L. a. I. K. V. Bogorad. San Diego, Academic Press. 7A: 5-53.

Palmer, J. D. (1986). "Isolation and structural analysis of chloroplast DNA." Meth. Enzym. 395: 167-186.

Pombert, J.-F., C. Otis, C.Lemieux, and M.Turmel (2005). "The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages." Mol. Biol. Evol. 2005(9): 1903-1918.

Provan, J., W. Powell, et al. (2001). "Chloroplast microsatellites: new tools for studies in plant ecology and evolution." Trends Ecol.Evol. 16(3): 142-147.

Quigley, F. and J. H. Weil (1985). "Organization and sequence of 5 transfer-RNA genes and of an unidentified reading frame in the wheat chloroplast genome - evidence for gene rearrangements during the evolution of chloroplast genomes." Curr. Genet. 9(6): 495-503.

Raubeson, L. A. and R. K. Jansen (1992). "Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants." Science 255 (5052): 1697-1699.

Raubeson, L. A. and R. K. J. (2005). Chloroplast genomes of plants. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. R. J. Henry. Cambridge, MA, CABI Publishing: 45-68.

Rocha, E. P. C. (2003). "DNA repeats lead to the accelerated loss of gene order in bacteria." Trends Genet. 19(11): 600-603.

Saski, C., S-B.Lee, H. Daniell, T.C. Wood, J.Tomkins, H-G Kim, and R.K. Jansen (2005). "Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes." Pl.Mol. Biol.59:309-322.

Schmitz-Linneweber, C., R.M. Maier, J-P. Alcaraz, A. Cottet, R.G. Herrmann, R.Mache (2001). "The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization." Pl. Mol. Biol. 45: 307-315.

Schwartz, S., L. Elnitski, et al. (2003). "MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences." Nucleic Acids Res. 31(13): 3518-3524.

Shimada, H. and M. Sugiura (1989). "Pseudogenes and short repeated sequences in the rice chloroplast genome." Curr. Genet. 16(4): 293-301.

Stein, D. B., D. S. Conant, et al. (1992). "Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns." PNAS 89(5): 1856-1860.

Sugiura, M. (1989). "The chloroplast chromosomes in land plants." Ann. Rev. Cell Biol. 5: 51-70.

Tesler, G. (2002). "GRIMM: genome rearrangements web server." Bioinformatics 18(3): 492-493.

Thomas, F., O. Massenet, et al. (1988). "Expression of the *rp123*, *rp12* and *rps19* genes in spinach chloroplasts." Nucleic Acids Res. 16(6): 2461-2472.

Tsai, C. H. and S. H. Strauss (1989). "Dispersed repetitive sequences in the chloroplast genome of Douglas-fir." Curr. Genet. 16(3): 211-218.

Turmel, M., C. OTis, and C. Lemieux (2002). "The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants." PNAS 99: 11275-11280.

Wakasugi, T., J. Tsudzuki, et al. (1994). "Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*." PNAS 91(21): 9794-9798.

Wakasugi, T., M. Sugita, T. Tsudzuki and M. Sugiura (1998). "Updated gene map of tobacco chloroplast DNA." Plant Mol.Biol. Rep. 16: 231-241.

Wolfe, K. H. (1988). "The site of deletion of the inverted repeat in pea chloroplast DNA contains duplicated gene fragments." Curr.Genet. 13(1): 97-99.

Wyman, S. K., R. K. Jansen, et al. (2004). "Automatic annotation of organellar genomes with DOGMA." Bioinformatics 20(17): 3252-3255.

Table 1. Characterization of selected repeats in *Trachelium* identified by REPUTER ≥ 30 bp with a Hamming distance 3

Repeat	Size (bp)	Percent Sequence Identity	Number of Copies	Blast Results
R1	85	74	2	Plant nuclear
R2	35	100	2	Part of *rpl23* cp gene
R3	1014	100	3	Part of *psbB* cp gene
R4	34	94	3	Non cp
R5	432	99.5	2	Non cp
R6	499	100	3	*PsbJ/orf99* cp genes
R7	105	98	2	Non cp
R8	37	89	2	Part of *clpP* exon 1 cp gene
R9	174	99	2	*TrnI-CAU* cp gene
R10	121	97	3	Part of *ycf3* exon 3 cp gene
R11	457	99.5	3	Cp intergenic spacer
R12	487	100	4	Part of *23s rrn* cp gene

Fig.1 Complete chloroplast genome of *Trachelium caeruleum*. Heavy lines on the middle circle indicate the extent of the inverted repeat (IR), separating the large single copy (LSC) region from the small single copy region (SSC). The outer circle of numbered arrows indicates conserved blocks of genes relative to an unrearranged genome such as *Nicotiana*, Genes are numbered consecutively 1-116 in *Nicotiana*, but in *Trachelium* have been rearranged in location and/or orientation. Genes shown on the inside of the middle circle are transcribed clockwise, while genes on the outside are transcribed counterclockwise. Asterisks by gene names indicate truncation. Small arrows perpendicular to gene block arrows show locations of tRNAs in relation to rearrangement

Fig.2a. Whole genome alignment of the *Trachelium* chloroplast genome with eight other angiosperm chloroplast genomes using *Nicotiana* as the reference genome, generated by MULTIPIPMAKER Schwartz (*et al*. 2003). The top line indicates the genes in *Nicotiana*. Sequence identity of both genes and intergenic spacers to other genomes is shown with black (75 - 100 %), grey (50-75%), and white (< 50%). Arrows show divergent or altered genes in *Trachelium.*

Fig 2b. Closeup of MULTIPIPMAKER alignment of selected regions. *Nicotiana* serves as the reference genome, and the heavy black bars represent its chloroplast genes. A. *atpB*, highly conserved across the nine angiosperm chloroplast genomes, B (*rpl23*), C(*ycf15*),D (*infA*) and E (*accD*) genes are reduced to gene fragments in *Trachelium*; F (*ycf2*), G (*clpP*) and H (*ycf1*) are highly divergent genes.

Fig. 3. Direct and inverted repeat size and frequency in *Trachelium* and eight other angiosperm chloroplast genomes identified with REPUTER (Kurtz et al. 2001) at a repeat length ≥ 30bp with a Hamming distance of 3. Vertical bars represent repeats clustered in classes of 30-39, 40-49, 50-75, 76-200, and 201-3500. *Trachelium* has far more, and far larger repeats than the other angiosperm chloroplast genomes, including *Zea* which has 3 large inversions and highly divergent gene content compared to *Nicotiana*.

Fig. 4. The *Trachelium* chloroplast genome map showing location of repeats identified by REPUTER, ≥ 30bp with a Hamming distance of 3 in relation to blocks of rearranged genes. The outer circle of numbered arrows identifies conserved gene blocks and their orientation in relation to *Nicotiana*. The circle of hashmarks indicates the location of 767 direct and inverted repeats reported by REPUTER. Numbered arrows **r1-12** identify selected large repeats of diverse origin, characterized in **Table 1**. Shaded areas **A**, **B**, **C** show rearrangement hotspots. **A** indicates an inversion endpoint where repeats of different origin (gene duplications, noncoding cpDNA, and non chloroplast DNA) are clustered between two blocks of unrearranged genes. **B** indicates another accumulation of diverse repeats at an inversion endpoint. **C** is the area in the IR where repeats and possible duplicative transpositions are amassed.

Fig.5 **C** shaded area in detail, in IRb. Asterisks indicate gene fragments. Genes numbered 69, 36, 50 and 98 reflect partial duplications of genes; the entire, functional copies of 69 (*psbB*) and 36 (*ycf3*) are located in the LSC. *accD* (gene 50) is normally found between *rbcL* and *psaI* in the LSC, and was previously reported as lost in the

Lobeliaceae and Campanulaceae. The *23S* duplication (gene 98) occurs 13.3 kb away from the intact copy within IRb.

Figure 1

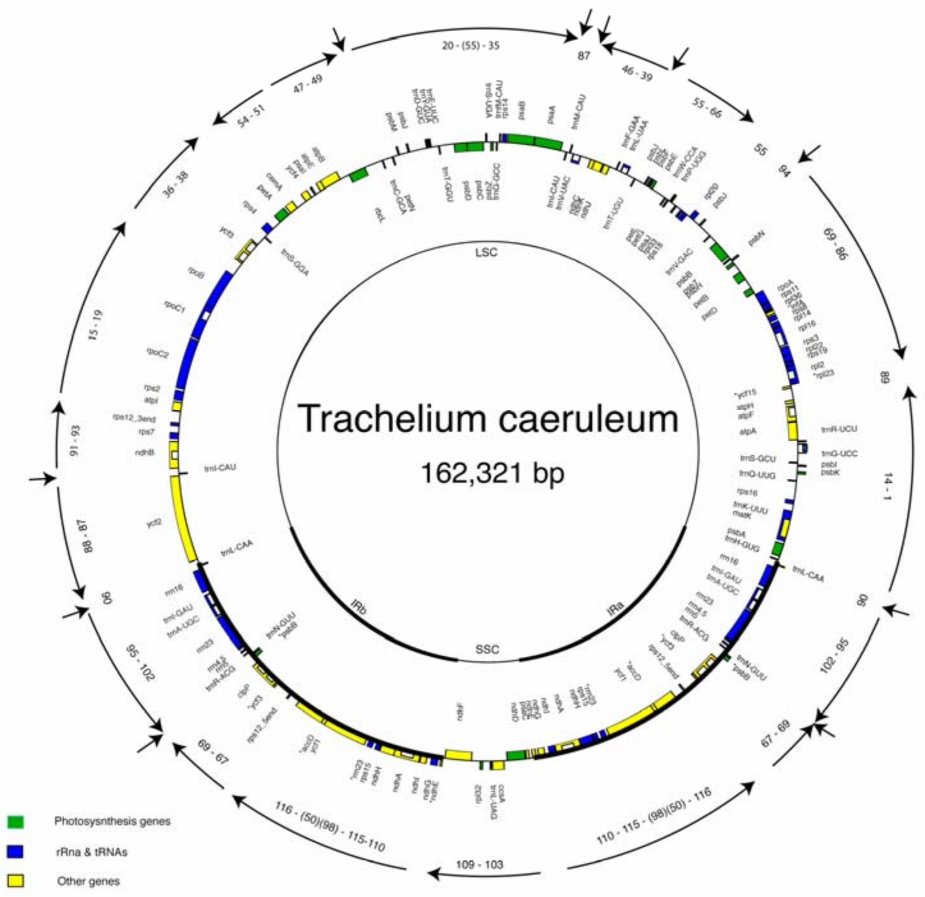

Trachelium caeruleum
162,321 bp
LSC
IRb
IRa
SSC
20 - (55) - 35
87
46 - 39
55 - 66
55
94
69 - 86
89
14 - 1
90
102 - 95
67 - 69
110 - 115 - (98)(50) - 116
109 - 103
116 - (50)(98) - 115-110
69 - 67
95 - 102
90
88 - 87
91 - 93
15 - 19
36 - 38
54 - 51
47 - 49
psbM
psbJ
trnD-GUC
trnY-GUA
trnE-UUC
trnS-UGA
trnfM-CAU
rps14
psaB
psaA
trnM-CAU
trnF-GAA
trnL-UAA
petE
trnW-CCA
trnP-UGG
rpl20
psbJ
psbN
rpoA
rps11
rpl36
rpl14
rpl16
rps3
rpl22
rps19
rpl2
*rpl23
*ycf15
atpH
atpF
atpA
trnR-UCU
trnG-UCC
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
psbA
trnH-GUG
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
trnL-CAA
clpP
*ycf3
rps12_5end
*accD
ycf1
trnN-GUU
*psbB
*rrn23
rps15
ndhH
ndhA
ndhI
ndhG
ndhE
psaC
ndhD
ndhF
rpl32
trnL-UAG
ccsA
trnT-UGU
trnV-GAC
trnI-CAU
trnV-UAC
ndhJ
psbB
petB
petD
trnT-GGU
psbD
psbC
trnG-GCC
trnC-GCA
petN
rbcL
rps4
petA
cemA
atpB
ycf3
trnS-GGA
rpoB
rpoC1
rpoC2
rps2
atpI
rps12_3end
rps7
ndhB
trnI-CAU
ycf2
trnL-CAA
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
trnN-GUU
*psbB
clpP
*ycf3
rps12_5end
*accD
ycf1
Photosysnthesis genes
rRna & tRNAs
Other genes

Figure2

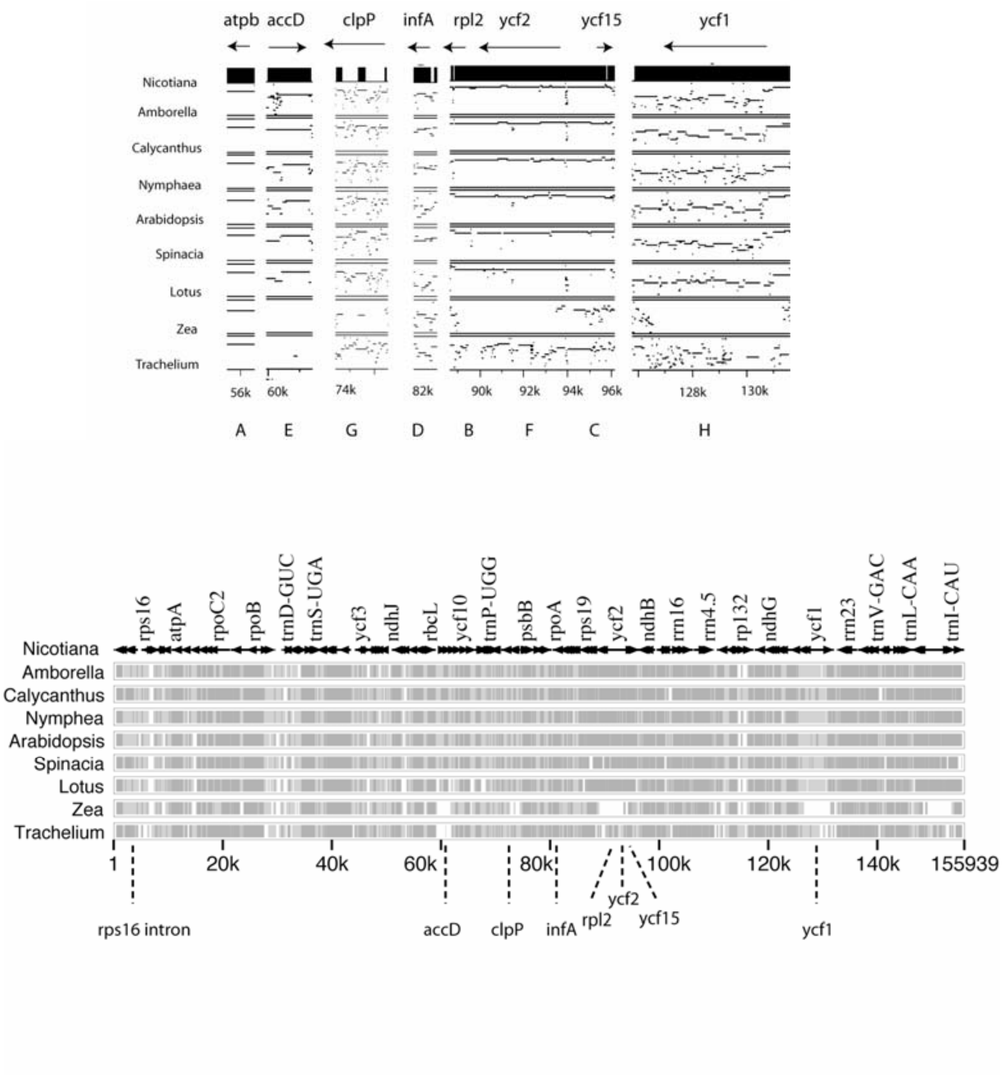

atpb
accD
clpP
infA
rpl2
ycf2
ycf15
ycf1
Nicotiana
Amborella
Calycanthus
Nymphaea
Arabidopsis
Spinacia
Lotus
Zea
Trachelium
56k
60k
74k
82k
90k
92k
94k
96k
128k
130k
A
E
G
D
B
F
C
H
rps16
atpA
rpoC2
rpoB
trnD-GUC
trnS-UGA
ycf3
ndhJ
rbcL
ycf10
trnP-UGG
psbB
rpoA
rps19
ycf2
ndhB
rrn16
rrn4.5
rp132
ndhG
ycf1
rrn23
trnV-GAC
trnL-CAA
trnI-CAU
Nicotiana
Amborella
Calycanthus
Nymphea
Arabidopsis
Spinacia
Lotus
Zea
Trachelium
1
20k
40k
60k
80k
100k
120k
140k
155939
rps16 intron
accD
clpP
infA
rpl2
ycf2
ycf15
ycf1

Figure 3

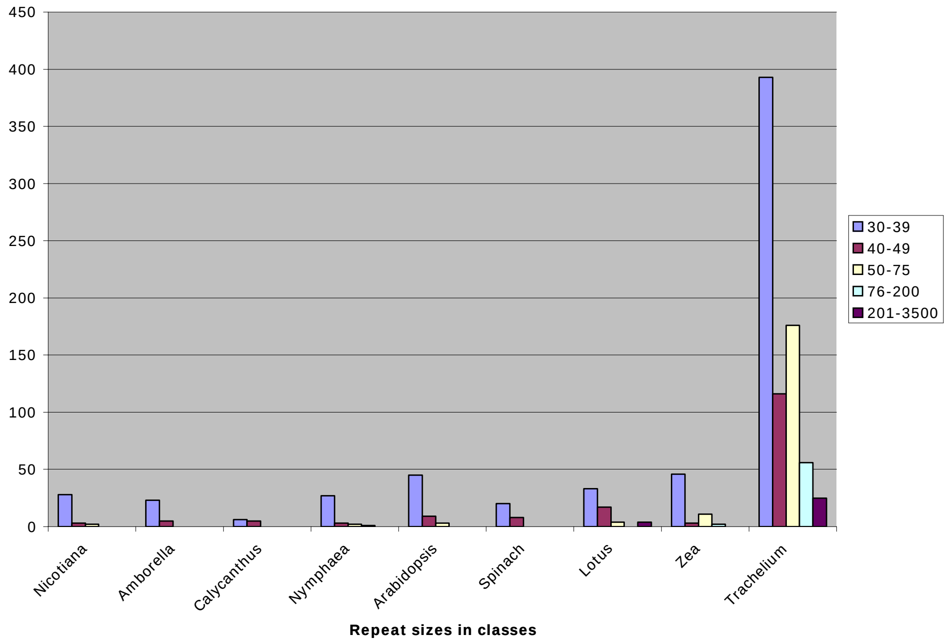
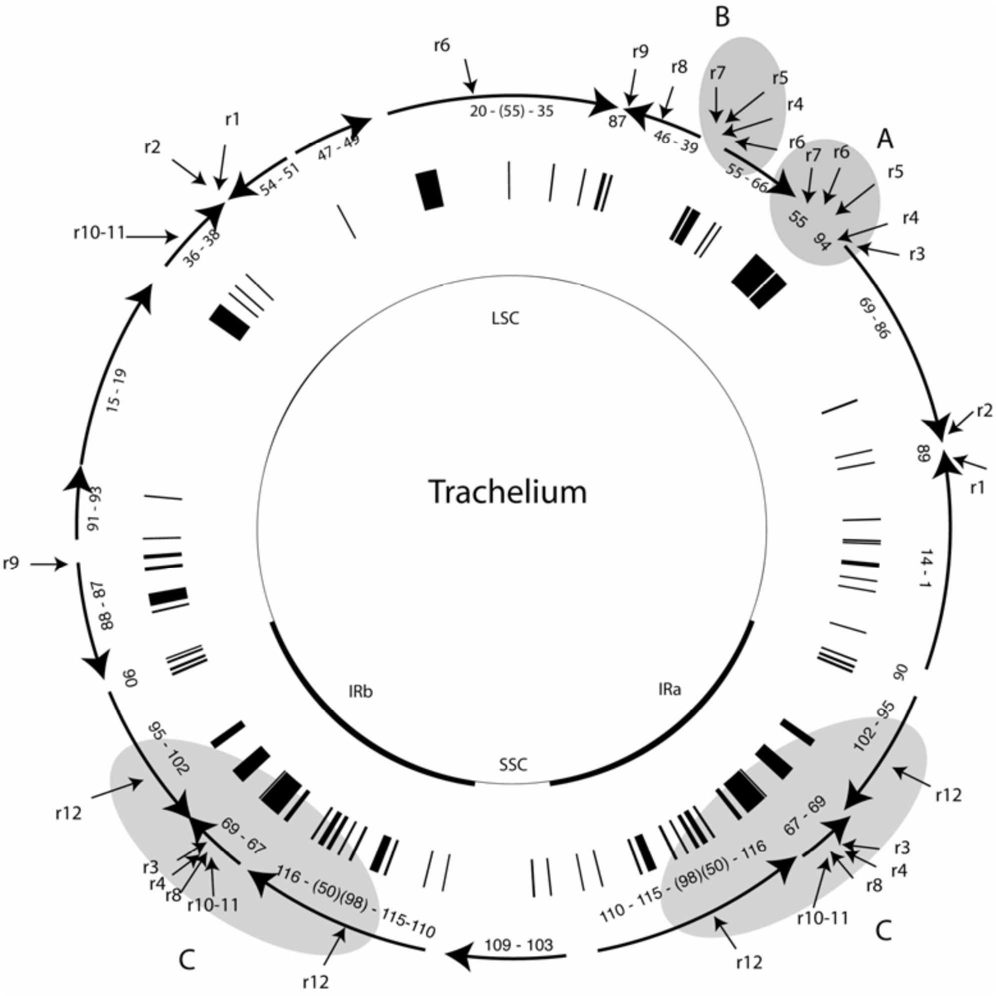
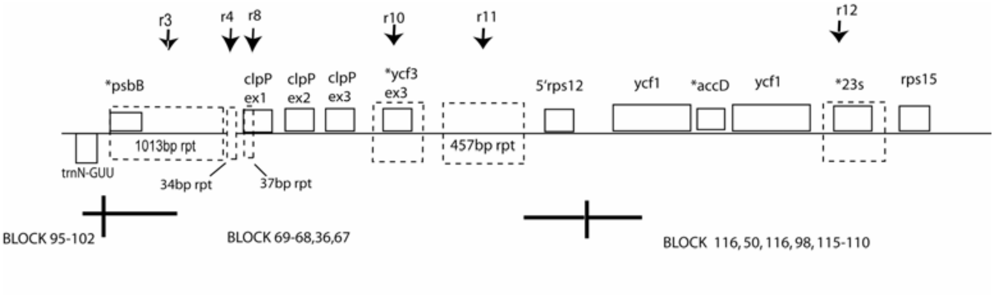

450
400
350
300
250
200
150
100
50
0
Nicotiana
Amborella
Calycanthus
Nymphaea
Arabidopsis
Spinach
Lotus
Zea
Trachelium
Repeat sizes in classes
30-39
40-49
50-75
76-200
201-3500

Figure 4



Trachelium
LSC
IRa
IRb
SSC
A
B
C
C
20 - (55) - 35
87
46 - 39
55 - 66
55
94
69 - 86
89
14 - 1
90
102 - 95
67 - 69
110 - 115 - (98)(50) - 116
109 - 103
116 - (50)(98) - 115-110
69 - 67
95 - 102
90
88 - 87
91 - 93
15 - 19
36 - 38
54 - 51
47 - 49
r1
r2
r3
r4
r5
r6
r7
r8
r9
r10-11
r12

Figure 5


r3
r4
r8
r10
r11
r12
*psbB
clpP ex1
clpP ex2
clpP ex3
*ycf3 ex3
5'rps12
ycf1
*accD
ycf1
*23s
rps15
1013bp rpt
457bp rpt
trnN-GUU
34bp rpt
37bp rpt
BLOCK 95-102
BLOCK 69-68,36,67
BLOCK 116, 50, 116, 98, 115-110