LBNL-56163

Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences

Robert K. Jansen[1], Linda A. Raubeson[2], Jeffrey L. Boore[3], Claude W. dePamphilis[4], Timothy W. Chumley[1], Rosemarie C. Haberle[1], Stacia K.Wyman[5], Andrew J. Alverson[1], Rhiannon Peery[2], Sallie J. Herman[2], H. Matthew Fourcade[3], Jennifer V. Kuehl[3], Joel R. McNeal[4], James Leebens-Mack[4], and Liying Cui[4]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas at Austin, Austin TX 78712

[2]Department of Biological Sciences, Central Washington University, Ellensburg, WA 98926

[3]Evolutionary Genomics Department, DOE Joint Genome Institute, Walnut Creek, CA 94598

[4]Department of Biology, Huck Institutes of Life Sciences, and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA, 16802

[5]Department of Computer Sciences, University of Texas at Austin, Austin TX 78712

Abstract

During the past decade there has been a rapid increase in our understanding of plastid genome organization and evolution due to the availability of many new completely sequenced genomes. Currently there are 43 complete genomes published and ongoing projects are likely to increase this sampling to nearly 200 genomes during the next five years. Several groups of researchers including ours have been developing new techniques for gathering and analyzing entire plastid genome sequences and details of these developments are summarized in this chapter. The most important recent developments that enhance our ability to generate whole chloroplast genome sequences involve the generation of pure fractions of chloroplast genomes by whole genome amplification using rolling circular amplification, cloning genomes into Fosmid or BAC vectors, and the development of an organellar annotation program (DOGMA). In addition to providing details of these methods, we provide an overview of methods for analyzing complete plastid genome sequences for repeats and gene content, as well as approaches for using gene order and sequence data for phylogeny reconstruction. This explosive increase in the number of sequenced plastid genomes and improved computational tools will provide many insights into the evolution of these genomes and much new data for assessing relationships at deep nodes in plants and other photosynthetic organisms.

I. Introduction

1. Historical Overview of Chloroplast Genomics

The study of chloroplast genomes dates back to the 1950s when plant biologists first discovered that chloroplasts contain their own DNA (see Sugiura, 2003 for a review). Early work used electron microscopy, cloning, comparative restriction site mapping, and gene mapping to characterize genome structure–gene order and organization (Palmer, 1991; Sugiura, 1992). Such comparisons yielded numerous phylogenetic studies based on restriction site polymorphisms and gene order changes (reviewed in Downie and Palmer, 1992; Jansen *et al*., 1998; Olmstead and Palmer, 1994). The publication of complete plastid sequences for *Nicotiana* (Shinozaki *et al*., 1986) and *Marchantia* (Ohyama *et al*., 1986) provided the first opportunity for nucleotide level, whole genome comparisons (Morton, 1994; Wolfe *et al*., 1987). Currently the list of completely sequenced plastid genomes has increased to 43 and now includes a wide diversity of taxonomic groups. The number of sequenced chloroplast genomes is growing rapidly: 17 of these 43 genomes (Table 1) have appeared in the last two years. In spite of the availability of so many complete genome sequences, our understanding of chloroplast genome evolution is still limited because this remains a very small sampling of plastid-containing species and because previous sequencing efforts were not designed to address phylogenetic or molecular evolutionary issues. A number of groups (e.g., algae and various lineages of land plants, including bryophytes, ferns and fern allies, gymnosperms, and certain angiosperm groups, especially monocots other than the cereal grasses) remain poorly sampled. However, several groups of scientists are now focusing their sequencing efforts at filling these gaps and the number of completely sequenced chloroplast genomes will continue to increase dramatically in the next few years (for details of three such projects see http://megasun.bch.umontreal.

ca/ogmp/projects/sumprog.html, http://www.jgi.doe.gov/programs /comparative/second_levels/ chloroplasts/jansen_project_home /chlorosite.html, and http://ucjeps.berkeley.edu/TreeofLife/).

2. Brief Overview of Chloroplast Genome Structure and Evolution

Plastid genomes vary in size from 35-217 kilobases (kb) but the vast majority from photosynthetic organisms are between 115 and 165 kb (Table 1). The 43 completely sequenced genomes (Table 1) encode from 63 (*Toxiplasma*) to 209 (*Porphyra*) genes with most containing 110–130 genes. The majority of these genes code for proteins, mostly involved in photosynthesis or gene expression, with the remainder being transfer RNA or ribosomal RNA genes. Although the number of genes may be similar between even distantly related lineages, the exact gene complement may be quite different. Although gene content is largely consistent within land plants, Martin et al. (2002) found only 44 protein-coding genes to be common among 15 chloroplast genomes representing all major lineages of photosynthetic organisms. A few genes have evidently been gained during plastid genome evolution, but the vast majority of gene content changes represent gene losses, some of which have been lost independently in different lineages (Martin *et al*., 2002; Maul *et al*., 2002). In all plastid genomes, most genes are part of polycistronic transcription units, suggestive of bacterial operons (Fig. 1, Mullet *et al*., 1992; Palmer, 1991). Plastid operons often have multiple promoters that enable a subset of genes to be transcribed within the operon (e.g. Kuroda and Maliga, 2002; Miyagi *et al*., 1998). Both group-I and group-II types of self-splicing introns are found in cpDNAs; the majority are group-II (Palmer, 1991). A unique intron type, known as a “twintron” that contains an intron within an intron, is found in *Euglena* (Copertino and Hallick, 1991) and possibly other organisms (Maier *et*

al., 1995). Although intron content is quite variable among algal genomes, it is highly conserved among land plant cpDNAs.

Most land plant (and some algal) genomes have a quadripartite organization (Fig. 1), comprised of two copies of a large inverted repeat (IR) and two sections of unique DNA, which are referred to as the large and small single copy regions (LSC and SSC, respectively). The gene content and organization of the chloroplast genome change by several mechanisms. Transposition has been suggested as a mechanism of genomic change in chloroplasts (e.g., in *Trachelium* in the Campanulaceae (Cosner *et al*., 1997) and in *Trifolium* in the Fabaceae (Milligan *et al*., 1989) but few definitive examples have been documented. Only one clear case of transpositional gain has been documented in *Chlamydomonas* (Fan *et al*., 1995), where a transposable element that is no longer active has been characterized. The frequency of the other types of rearrangements, including gene and intron gains and losses, expansion and contraction of the inverted repeat, and inversions, varies from group to group. Most genomes have very few gene order changes, at least in comparison to close relatives. However, several lineages have cpDNAs that are highly rearranged. The most notable examples are in the algae (e.g., *Chlamydomonas* (Maul *et al*., 2002), conifers (e.g., *Pinus*; (Wakasugi *et al*., 1994)), and several angiosperm lineages (e.g., Campanulaceae (Cosner *et al*., 1997); Fabaceae (Milligan *et al*., 1989), Geraniaceae (Palmer *et al*., 1987), and Lobeliaceae (Knox and Palmer, 1998)). Two recent reviews summarize the types of genomic rearrangements in cpDNAs of algae (Simpson and Stern, 2002) and land plants (Raubeson and Jansen, 2005). Gene order changes in plastid genomes have proven useful for resolving phylogenetic relationships within a number of different plant groups (Raubeson and Jansen, 2005).

3. Overview of this Chapter

This chapter will focus on the methods used to gather and analyze plastid genomic sequences. This will include methods for (1) isolating chloroplasts and purified cpDNA, (2) amplifying, cloning, and sequencing cpDNA, (3) assembling drafts and finishing genomes, (4) annotating chloroplast genomes, and (5) analyzing genome sequence and structure. Most of the steps are equally applicable to the plastid genomes of nonphotosynthetic plants, except for the initial isolation steps, which typically involve generation of a large insert genomic library. In our treatment of genomic analysis, we will focus on evolutionary issues, and even then we will not be able to be comprehensive. In addition to reviewing methods that others have used, this chapter will provide some more detailed protocols used by our group in an ongoing project for which we are sequencing 60 plastid genomes from seed plants.

II. Whole Chloroplast Genome Sequencing

Until quite recently, chloroplast genomes have been sequenced by cloning cpDNA into plasmid vectors, selecting cpDNA-containing clones, and then sequencing the clones using both plasmid and chloroplast-specific primers. This process is very labor-intensive and involves isolation of highly purified cpDNA, which can be quite difficult for many taxa. More recently, faster and more cost effective approaches have been developed. Currently, there are four basic approaches to sequencing entire chloroplast genomes: (1) isolation of pure cpDNA, followed by random shearing, shotgun cloning, and sequencing; (2) amplification using long PCR of large segments of the genome, followed by cloning, and then sequencing of the products using chloroplast-specific primers; (3) amplification of the entire genome using rolling circular amplification (RCA) followed by shearing of the RCA product and shotgun cloning and

sequencing of the fragments; and (4) construction of BAC (Bacterial Artificial Chromosome) or Fosmid libraries from total DNA preparations, preferably ones that are enriched for cpDNA, followed by shearing, cloning, and sequencing. We first will outline our general genomic sequencing methods and then go on in additional sections to describe the unique parts of each of the four above-mentioned approaches, with an emphasis on those used by our group.

Draft sequences of chloroplast genomes from our group are being produced at the DOE Joint Genome Institute (JGI) in Walnut Creek, CA, USA. This facility is a very high throughput operation that relies on robotics for many of the steps in the process. Details of JGI protocols can be found at http://www.jgi.doe.gov/Internal/protocols/prots_production.html but a general description is given here. Our approach is to shear the DNA, select approximately 3 kb fragments, and clone these fragments into plasmid vectors. *E. coli* are then transformed with the recombinant plasmids and spread onto large plates from which colonies are robotically picked and placed into 384-well plates containing the appropriate growth medium. Picking of colonies from the library is random, so the percentage of wells in the plates that contain cpDNA clones will be proportional to the percentage of cpDNA (as opposed to nuclear or mitochondrial "contaminant") in the DNA sample used to create the library. The inserts are sequenced from the 384-well plates using forward and reverse plasmid primers yielding about 500-800 bp of sequence from each end of the insert. Sequencing proceeds until the depth of coverage, from many overlapping sequence reads, enables the assembly of the reads into one contiguous genomic sequence. In this approach most steps are performed robotically minimizing human effort compared to earlier methods. The tradeoff is that, unlike directed approaches such as chromosome walking with custom primers, the genome must be sequenced to a depth of 6 -10X coverage to ensure accurate characterization of the entire genome.

1. Isolation of Chloroplast DNA

If pure cpDNA can be obtained in sufficient quantity, it can serve as the template for the sequencing approach just described. Many methods have been developed for isolating purified cpDNA from plants (Palmer, 1986). Most of these methods involve three basic steps: separation of plastids from other organelles, lysis of the chloroplasts, and purification of DNA. The most commonly applied methods use sucrose or percoll gradients (Palmer, 1986), DNAse I treatment (Kolodner and Tewari, 1979), or high salt buffers (Bookjans *et al*., 1984) to isolate purified cpDNA (or more realistically, a total DNA preparation enriched for cpDNA). The use of sucrose gradients is most generally applicable at least in land plants and a detailed protocol is provided in Table 2. Basically, sucrose step gradients are used to obtain chloroplasts that are then lysed and the DNA is recovered from the lysate. We include several modifications of the basic method that have been used by our group to improve the quality and quantity of cpDNA. Consistent problems are encountered with two aspects of cpDNA isolations, using this method or any other; (1) collecting a sufficient quantity of chloroplasts while eliminating nuclear contamination and (2) lysing the chloroplasts and releasing the membrane-bound cpDNA. Nuclear DNA tends to adhere to the outer chloroplast membrane, leading to the first challenge. Regarding the second challenge, chloroplasts can be surprisingly difficult to lyse. If harsh enough detergents are used to lyse the chloroplasts abruptly then the DNA is degraded. Since the DNA is bound to the thylakoid membranes the membranes must be solubilized to release the DNA, but if the chloroplast is lysed too gently the DNA remains bound to the membrane and is lost. Our modifications to the basic procedure help in reducing these problems but do not totally overcome them.

Two other approaches to cpDNA isolation are the DNAse I (Kolodner and Tewari, 1979), which is used as a modification of the sucrose gradient technique, and the high salt (Bookjans et al., 1984) methods (see http://www.jgi.doe.gov/programs/comparative/second_levels /chloroplasts/jansen_project_home/cpDNA_protocols.html for protocols). In the DNAse I method the chloroplast pellet in step 7 (Table 1) is treated with DNAse I to destroy nuclear DNA. This treatment also will destroy any cpDNA that is not protected within intact plastids. Thus, although the purity of cpDNA is very high, the yield is much lower and much more leaf material is needed to obtain sufficient cpDNA. In our experience, this method yields very pure cpDNA when it works but it has only worked for two species of the many that we have attempted (*Lactuca sativa* [Fig. 2] and *Ginkgo biloba*). Even in those cases, sufficient quantities of cpDNA for shearing and shotgun cloning were not always recovered. The second alternative method employs a high NaCl (1.25 M) concentration in the isolation and wash buffers and it does not involve any step-gradient centrifugation. The high salt concentration is supposed to significantly reduce nuclear contamination. According to Bookjans et al. (1984), the undissociated chromatin or nuclear DNA tends to stick to chloroplast membranes because of electrostatic interactions. The high salt concentration diminishes these electrostatic interactions yielding a DNA prep that is enriched in cpDNA. We have had only limited success with this approach; one isolation by this method yielded cpDNA of sufficient purity and quantity to proceed to genomic sequencing (*Ranunculus macranthus*, Fig. 2). However, the use of high salt wash buffers in combination with the sucrose gradient technique has proved to be quite valuable for decreasing nuclear DNA contamination in chloroplast preps.

The methods just described can also be used (stopping prior to lysis) to collect chloroplasts for use in whole genome amplifications (described below). Other workers are

experimenting with the use of a Fluorescence Activated Cell Sorter (FACS) to separate chloroplasts from mitochondria and nuclei (D. Mandoli, personal communication). This method may be particularly valuable when there is limited tissue available. Once purified chloroplasts have been obtained from the FACS, they can be further processed using one of the methods below. Another advantage of the FACS approach is that it may also provide purified fractions of both mitochondria and nuclei in addition to chloroplasts.

2. Whole Genome Amplification

If purified chloroplasts can be obtained, they can serve as a template from which to produce abundant cpDNA via Rolling Circular Amplification (RCA), a powerful new approach for performing whole genome amplification. This process involves an isothermal amplification using bacteriophage Phi29 polymerase, which is capable of performing strand displacement DNA synthesis for more than 70 kb without disassociating from the template (Dean *et al.*, 2002). This feature, combined with the stability of this polymerase and its low error rate, make this enzyme a powerful tool for template preparation. RCA involves the use of random hexamer primers that are exonuclease resistant, necessary because the DNA polymerase has a 3'-5' exonuclease proofreading activity. Most applications of RCA have been directed toward performing human genome amplification and a kit for this purpose (Repli-G™) is available from Molecular Staging Inc. Our group has been using this kit routinely for amplifying entire chloroplast genomes and we have modified the Repli-G™ protocol to improve cpDNA amplification (see Table 3 for protocol). We have had considerable success with the RCA approach for a wide diversity of seed plants. Figure 3 shows restriction digests of RCA products for two taxa that had sufficient quality and quantity of cpDNA to proceed with genome

sequencing. One possible further modification of this protocol would be to develop genome-specific primers for chloroplast or mitochondrial genomes, which would enable the amplification of the chloroplast and mitochondrial genomes from total DNA isolations. Although the low temperature of the RCA reaction limits the specificity of annealing for these primers, experiments are in progress, focusing on buffer modifications that show promise for increasing the specificity of the amplification.

3. Long PCR and Sequencing

A third approach for obtaining DNA template from which to generate whole chloroplast genome sequences involves PCR-amplifying of large fragments of the genome using conserved chloroplast primers. This approach has been employed recently to sequence three basal angiosperm genomes (Goremykin *et al*., 2003a, b, 2004). Goremykin et al. developed conserved primers by aligning sequences from seven seed plant genomes (*Arabidopsis*, *Nicotiana*, *Oenothera*, *Oryza*, *Pinus*, *Spinacia*, and *Zea*). These primers then were used to amplify long fragments ranging in size from 4 to 20 kb and covering the entire chloroplast genome. The long PCR products were then sheared into smaller pieces, shotgun cloned, and sequenced. Although this approach worked well for Goremykin's group, it does have several disadvantages: (1) The primer combinations may not work for seed plant genomes that have experienced gene order changes or substantial sequence divergence at priming sites; (2) The method relies on PCR, which can sometimes be problematic for some DNAs or segments of the genome; (3) It would be difficult to extend this approach to algae or spore-bearing plants as little or no published chloroplast genome sequence information is available to direct primer design in

these groups; and (4) Numerous PCR and cloning reactions are required, consuming more time than some of the other available methods.

4. Cloning Chloroplast Genomes for Sequencing

Finally, a more labor-intensive but highly useful approach for obtaining sequencing template involves cloning the genome into either BAC or Fosmid vectors. This approach is superior to plasmid cloning as the insert is much larger – 40 to 150 kb. The larger insert size reduces the amount of screening involved and allows the clones to be sequenced via the JGI method described above. A number of BAC and Fosmid cloning kits are available commercially; we have been using the Epicentre CopyControl™ Kit (Cat. #CCF0S110). Our group has used the Fosmid cloning approach to sequence plastid genomes from parasitic plants as well as normal photosynthetic plants. The details of our Fosmid protocol can be found in McNeal et al., (in prep.) but here we provide a general outline for this procedure. DNA is isolated using a modified CTAB method (Doyle and Doyle, 1987) with 1% PEG 8000 in the extraction buffer. The DNA must then be end-repaired for cloning into vectors that require blunt-ended, 5' phosphorylated ends. Pulse Field Gel Electrophoresis (PFGE) is used to separate fragments in the 40 - 50 kb range for Fosmid cloning and in the 100 - 150 kb range for BAC cloning. DNA of the correct size is excised and recovered from the gel, and its concentration is measured, preferably by flourimetry, to ensure the proper ratio of template to vector for efficient ligation. Clones are plated and then transferred to 384-well plates for easy referencing and gridding onto nylon filters. We use robotics to pick, transfer, and grid clones quickly and efficiently. Plants with larger nuclear genome sizes have a proportionally higher ratio of nuclear to plastid clones and, thus, require a greater number of clones to be arrayed for screening to

ensure enough plastid clones will be found to cover the entire plastid genome. When the DNA used for Fosmid or BAC cloning is enriched for cpDNA, fewer clones need to be screened. Macroarrays are screened using hybridization probes generated by PCR-amplification of genes scattered throughout the plastid genome. Once positively hybridizing plastid clones have been identified, a minimal set of Fosmid clones are selected that cover the entire plastid genome (usually 2-5). End sequencing and PCR assays of each clone aid in the selection of minimally overlapping clones, which together cover the genome completely. One caveat of this method is that the macroarray hybridizations may also detect recent mitochondrial or nuclear plastid gene transfers. However, single or low-copy nuclear transfers are much less likely to be found than true plastid genome fragments, which occur in many more copies per cell. End-sequencing and PCR assays of each clone should eliminate all but the largest and most recent mitochondrial transfers from passing as plastid clones. For BAC libraries, only one or two clones are needed to get complete coverage of the genome, depending on genome size. The clones are then sheared, shotgun cloned, and sequenced as described for other methods. One 384-well plate is sequenced for each Fosmid clone (with both plasmid primers to yield 768 reads) or 2-3 plates for each BAC clone in order to obtain 6-10x coverage of the insert. Additional sequencing may be required to close gaps or verify regions with low coverage.

Our group has used the Fosmid cloning method to successfully create libraries for a number of photosynthetic (*Ipomoea, Lindenbergia,* and *Yucca*) and non-photosynthetic parasitic and mycotrophic plants (*Corynaea*, *Cuscuta*, *Cytinus*, *Monotropa*, *Orobanche*, and *Prosopanche*). Researchers preparing BAC libraries typically screen for "contaminant" clones containing chloroplast genome fragments. In collaboration with Pietro Piffanelli (CIRAD-AMIS, Montpellier, France) we have obtained plastid genome sequences from cpDNA-

containing BACs identified from his *Musa* and *Elaeis* libraries. While Fosmid or BAC library construction is certainly more technically demanding and time-consuming than cpDNA isolation or RCA amplification of plastid genomes, the libraries will have a broader utility and we have found, generally, less finishing of draft genome sequences is required when the shotgun sequencing libraries are made from well-chosen Fosmid or BAC templates.

III. Assembling, Finishing, and Annotating Genomes

1. Assembling Draft Genome Sequences

When preparing a draft genomic sequence from cpDNA or RCA product, we first generate one 384-well plate of sequences using both forward and reverse primers (768 reads). Vector and quality trimming of the resulting sequences is performed using Phred (Ewing and Green, 1998). Using BLASTN (Altschul *et al*., 1997), trimmed reads are then used to query a nucleotide sequence database of previously sequenced chloroplast genomes. If the BLASTN search indicates that 60% or more of the reads are chloroplast sequences, we then proceed to sequence four more plates for a total of 3,840 reads, although additional plates are sometimes required. If less than 60% of the library is cpDNA we do not proceed with additional sequencing but instead work to obtain purer cpDNA preps from which to construct a new library. When sequencing from Fosmid/BAC clones, we prepare a separate library for each clone. One plate per Fosmid clone library or two plates per BAC clone library usually provide sufficient coverage.

Individual reads generated from the plates are assembled into contiguous sequences ("contigs") using Phrap (Ewing and Green, 1998) and the resulting contigs are analyzed in Consed, a powerful software package used for sequence finishing

(http://www.phrap.org/consed/consed.html, Gordon *et al*., 1998). Consed has numerous useful features (Fig. 4), including an overview of the assembly, numerous editing options, a method for tearing contigs into pieces and performing mini-reassembly, an option for designing finishing primers, and options for adding new reads. The Assembly View option (see Fig. 4 for an example) provides a wealth of information to evaluate the draft genome sequence, including the depth of coverage, the possible arrangement of the contigs, and cross matches of sequences between contigs. For chloroplast genomes that are not highly rearranged, one generally does not encounter many problems with the assembly, but highly rearranged genomes often require considerable work interactively reassembling the sequences due to the high frequency of repeated sequences. This will require examination of each of the contigs to identify possible mis-assemblies and the removal and/or relocation of misplaced reads. Examination of the assembly will also reveal regions of the draft where there are few high quality reads and more sequencing is needed. Effective integrated use of Phred, Phrap and Consed, takes considerable time to master. Phred and Phrap, however, are necessary for sequence assembly and Consed is extremely valuable for assessment of draft assemblies and identifying regions where directed sequencing is necessary to finish the genome sequence. The most finishing will likely be required when purified cpDNA or RCA product is sheared, shotgun cloned and sequenced "randomly" whereas the least finishing is required when Fosmid or BAC clones are used as the template.

2. Finishing Genomic Sequences

Finishing draft chloroplast genomic sequences involves four basic steps: (a) make a preliminary identification of genes occurring in each contig using the chloroplast genome

annotation program DOGMA (Dual Organellar GenoMe Annotator, Wyman *et al*., in press); (b) examine depth of coverage within each contig to identify regions of low sequence coverage; (c) design primers that flank gaps and regions of low coverage and perform PCR and sequencing to fill in necessary regions; and (d) determine the extent of the inverted repeat (IR) and if necessary confirm using PCR and sequencing across the inverted IR/single copy (SC) junctions. These four finishing steps will be described in more detail in the following sections. All of these steps may not be necessary. For example, drafts generated by sequencing BAC or Fosmid clones often do not require finishing if the screening and selection of clones was done correctly. However, when purified cpDNA or RCA product is used some finishing will be necessary. The amount of finishing will depend on the purity of the cpDNA or RCA product. High purity cpDNA could yield the entire chloroplast genome in one contig with no areas of low coverage, although in our experience this rarely happens unless the purity of the cpDNA or RCA product is exceptionally high or more than five plates of sequences are done. Even in these cases, it is still necessary to confirm the boundaries of the IR because both copies of the IR will assemble together in Consed.

(a) Identify genes in contigs with DOGMA – DOGMA is a web-based program developed by our group that makes this step in the finishing process very easy (see section III.3 below for more details about the program, Wyman *et al*., in press). DOGMA identifies which genes are likely to occur in each contig. Knowledge of the gene content assists in determining the arrangement of the contigs so that primer pairs that span gaps can be developed. For genomes where previous gene mapping data is available one simply compares the gene content of the contigs to the gene map to arrange the contigs. When no gene map is available, comparison of gene orders in the contigs to already sequenced chloroplast genomes often can

provide valuable information for deciding how the contigs likely are arranged. It is often possible to use the already sequenced genomes to estimate the location and sizes of gaps and to develop more universal primers to amplify through the gaps.

(b) Examine depth of coverage in contigs – Generally our methods generate contigs that have 6X-10X coverage, but this certainly will depend on which genome sequencing method we have employed and the quality of the sequencing template. Our group has decided that, for each nucleotide, a minimum of two reads each with a Phred/Phrap quality score (q value) exceeding 20 is necessary for satisfactory genome coverage. In general, coverage is much higher except in those regions where we fill in gaps. However, if areas with sparse coverage occur within contigs, primers are designed and additional sequence data are gathered.

(c) Design primers to fill in gaps by PCR and sequencing – Once all gaps and areas of low coverage have been identified primers are designed that flank these regions. We generally design 18-20 bp primers in coding regions that are adjacent to the gaps or regions of low coverage. We attempt to make the primers as universal as possible by comparing the primer sequences with previously sequenced chloroplast genomes so that the primers could be used in the finishing of other chloroplast genomes. In some cases, we need to design primers that are not in coding regions. This is more difficult because primers in non-genic regions may have multiple priming sites. We usually can avoid this problem by searching the genome for the primer sequence using the Consed Autofinish feature (Gordon *et al.*, 2001). For larger gaps additional primers must be made to sequence through the gap. In many cases, the size of the gap is unknown so it may be a matter of trial and error to determine what extension time to use in the PCR reaction.

(d) Confirm extent of inverted repeats – Chloroplast genomes that are sequenced using long PCR or BAC/Fosmid clones may include each copy of the IR in separate contigs in which case defining the extent of the IR is straightforward. However, in many cases all the individual IR sequencing reads generated by shotgun cloning of purified cpDNA or RCA product will be assembled together making it difficult to determine the precise IR/SC boundaries. Several tricks can be used to get a general idea of these boundaries, especially if parts of the IR are present in different contigs. In general, the assembly view in Consed shows a higher depth of reads in the IR (Fig. 4). Also, another very useful feature of Consed shows subclone pairing. This provides information about the positions of forward and reverse reads from the same clone. If ends of the same clone match in distant regions or in different contigs this may be due to a sequence being part of one of the IR/SC junctions. In most cases, these methods for identifying possible IR boundaries are not definitive, and it is necessary to design primer pairs (two for each of the four IR/SC boundaries) that span the IR/SC junctions. Amplification and sequencing of these regions is needed to confirm the boundaries. Once this has been confirmed, the IR sequence must be copied, inverted and inserted into the appropriate location to complete the chloroplast genome sequence.

3. Annotation using DOGMA

Annotation of chloroplast genomes traditionally has been a very tedious and error-prone task. The annotations currently in Genbank are not consistent in terms of gene names, and they are not usually updated when the identities and functions of hypothetical chloroplast reading frames (ycfs) or open reading frames (orfs) are clarified. In the past, most chloroplast genome sequences were annotated by performing BLASTN and BLASTX searches on Genbank. Many

of these problems were alleviated upon completion of DOGMA, a web-based program designed by our group to assist in the annotation of chloroplast and animal mitochondrial genomes (see http://phylocluster. biosci.utexas.edu/dogma/, Wyman *et al*., in press)**.** This program takes a FASTA-formatted input file of the complete (or partial) genomic sequences and identifies putative protein-coding genes by performing BLASTX searches against a custom database of 15 published chloroplast genomes of green plants (Fig. 5A). Errors in the Genbank entries have been corrected in the database and names of genes and their products have been standardized following Martin et al. (2002). Sequence identity is highly conserved for both tRNAs and rRNAs in chloroplast genomes, so these genes are identified by BLASTN searches against a database of the same 15 chloroplast genomes. DOGMA also uses a custom program to infer the stem loop structure of tRNAs and draw candidate secondary structure diagrams.

DOGMA has many other features to aid in annotation of chloroplast genomes (Fig. 5B). One DOGMA panel displays all of the putative genes color-coded by gene type. Selection of a gene in this lower panel generates an upper panel that shows the five or more most similar sequences from the database compared to the sequence under analysis along with potential start and stop codons (Fig. 5B). The user then must select the most likely start and stop codons to identify each putative gene. For genes with introns, DOGMA will identify putative exon boundaries by BLAST; the user must verify these boundaries and use DOGMA to connect the exons. Another window appears that records the annotation information that can be used to generate a Sequin file for submitting the annotation to Genbank. Selection of the gene name in the top panel also generates a window with the actual BLAST results. In the lower panel of the annotation window there are additional buttons (Fig. 5B). The 'extract sequences' button enables the user to extract certain sets of sequences from the annotation, including protein coding genes

(either nucleotide or amino acid sequences), intergenic regions, introns, tRNAs, or rRNAs (Fig. 5B). This feature is particularly useful for extracting sequences to add to a data matrix for phylogenetic analyses. The text summary button generates a tabular form of the annotation with coordinates for the genes and other information about each gene. More details about the features of this program can be found by downloading the cp tutorial at http://phylocluster.biosci.utexas.edu /dogma/.

In the future DOGMA will be modified in several ways: (1) the chloroplast database will be expanded to include more sequences especially from underrepresented groups such as algae; (2) mitochondrial genomes from plants, fungi and protists will be added; (3) an option will be included to allow individual users to develop their own custom database; (4) an ORF finder will be added to search for putative new genes; and (5) methods will be developed to deal with RNA editing of start and stop codons, a phenomenon that is common to plant mitochondrial genomes and chloroplast genomes of some plants (Bock, 2000).

IV. Analysis of Genome Sequences

The analysis of whole genome sequences is an immense scientific field for which numerous databases and computational tools exist, some relevant to the study of chloroplast genomes (Table 4). Some of these are simply a listing of available genome sequences with accession numbers to access the sequences on Genbank, whereas others provide additional information about chloroplast gene names, details of the characteristics of the genomes, databases of corrected annotations, gene orders, universal primer sequences, and searchable databases. All of these are valuable resources for anyone who is working on comparative chloroplast genomics. In the sections below we will discuss chloroplast genome analysis in

terms of phylogenetic comparisons of gene content and gene order, detection of repeats, and use of coding sequences for phylogenetic studies.

1. Whole genome comparisons and repeat analysis

A number of computational tools exist for whole genome comparisons, though most of these were not designed specifically for chloroplast genomes. We have used several of these tools to compare gene content, examine genome-wide sequence similarity, look for repeated sequences, and identify putative regulatory motifs with the primary goal of improving our understanding of genome evolution. Our primary goal in using these programs has been to improve our understanding of both the patterns and mechanisms of chloroplast genome evolution. Below we briefly review a few of these tools and how we have applied them to comparisons of chloroplast genomes. Table 4 includes information about accessing these programs.

MultiPipmaker (Table 4, Schwartz *et al.*, 2003) allows the user to compare multiple chloroplast genomes. The program generates alignments of whole genomes in comparison to a reference genome. The output from MultiPipmaker includes a stacked set of percent identity plots (Fig. 6) referred to as a "MultiPip," that illustrate sequence similarity among the genomes in coding and non-coding regions. This output is helpful in identifying potential genes and regulatory elements. Visual inspection of the Multipip also is useful for identifying putative gene losses or gene duplications, for identifying unannotated genes or conserved nongenic regions, and for assessing overall sequence similarity among genomes (see Maul *et al.*, 2002 for a chloroplast genome comparison). PipMaker (Elnitski *et al.*, 2002), the pairwise version of this

tool, also has been used to identify repeated sequences by aligning a genome against itself (see Pombert *et al*., 2004 for an example of this application using a plant mitochondrial genome).

With the exception of the large inverted repeat that is present in most taxa, chloroplast genomes generally are considered to have very few repeated sequences (Palmer, 1991). However, repeated sequences have been identified in a number of genomes, including *Chlamydomonas* (Maul *et al*., 2002), *Pseudotsuga* (Hipkins *et al*., 1995), *Trachelium* (Cosner *et al*., 1997), *Trifolium* (Milligan *et al*., 1989), wheat (Bowman and Dyer, 1986; Howe, 1985), and *Oenothera* (Hupfer, 2000; Sears *et al*., 1996; Vomstein and Hachtel, 1988). The most striking example to date is the *Chlamydomonas* chloroplast genome of which more than 20% is composed of short dispersed repeats. In most of these cases, repeats appeared to be associated with rearranged blocks of genes. Thus, characterization of repeat structure in chloroplast genomes could provide insights into mechanisms of gene order changes.

Several programs have been developed that are designed to identify repeats and group them into classes. The two programs that we have found most useful are REPuter and RepeatFinder (Table 4). REPuter (Kurtz *et al*., 2001; Kurtz and Schleiermacher, 1999) includes a search algorithm that finds various types of repeats, including direct and inverted repeats (Fig. 7). The user specifies the desired repeat type, minimum repeat length, and the percent identity (Hamming Distance) and the program locates all repeats that meet these criteria. The program also provides a graphic visualization of the location of the repeats in the genome (Fig. 7A). REPuter can be accessed and run directly using a web browser (http://www.genomes.de/), though this platform does not allow the user to modify the default options. We recommend that users download the standalone version, which is available for UNIX platforms at no cost. RepeatFinder (Volfvovsky *et al*., 2001) is a software tool for clustering repeats into classes. It

takes as input repeats that have been identified by another program such as REPuter. This program must be downloaded and setup on a UNIX platform. Both REPuter and RepeatFinder have been used together to examine repeat structure in plant mitochondrial genomes (Bartoszewski *et al*., 2004; Pombert *et al*., 2004).

2. Gene content and order for phylogeny reconstruction

Chloroplast genomes in many groups are highly conserved in gene content, though there are significant differences in these features in comparisons between algal and land plant genomes (Raubeson and Jansen, 2005; Simpson and Stern, 2002). Martin *et al*. (2002) estimated that only 44 of the 274 plastid-encoded genes are retained in all plastid genomes, and approximately half (117) of the ones that are missing have been lost or transferred to the nucleus. Among green plants there is considerable conservation of both gene content and gene order. For example, the gene organization of the earliest diverged green alga sequenced so far, *Mesostigma*, is very similar in structure to land plant cpDNAs with 81% of its genes being found in the same clusters as in land plants (Lemieux *et al*., 2000). More recent comparisons with the green alga *Chlamydomonas* also revealed a high incidence of gene loss among algal chloroplast genomes but a much higher level of similarity among green plants (Maul *et al*., 2002; Simpson and Stern, 2002). The large number of gene losses among plastid genomes, often occurring in parallel in different lineages (Martin *et al*., 2002; Maul *et al*., 2002), suggests that the use of gene content for phylogeny reconstruction may be of limited value and, in most cases, the utility of these types of characters may be restricted to selected groups.

Gene order of the chloroplast genome is generally highly conserved, especially among land plants. Previous studies have demonstrated the phylogenetic utility of gene rearrangements

for resolving relationships at deep nodes, though in most cases only one or a few characters were available. Some notable examples include a 30 kb inversion that identified the lycopsids as the basal lineage of vascular plants (Raubeson and Jansen, 1992), three inversions that supported monophyly of the Poaceae and indicated its relationship to Joinvilleaceae and Restionaceae (Doyle *et al*., 1992), and a 22 kb inversion that identified the basal clade in the Asteraceae (Jansen and Palmer, 1987). These types of changes make powerful phylogenetic markers, and subsequent phylogenetic studies using DNA sequence data corroborated these relationships first identified by gene order changes. The best example of the utility of gene order data for phylogeny reconstruction is in the angiosperm family Campanulaceae (Cosner *et al*., 1994; Cosner *et al*., 2000; Cosner *et al*., in press). Gene mapping studies of 18 genera in this family identified numerous changes in gene order, which were caused by inversion, expansion and contraction of the IR, and possibly transposition. The situation in the Campanulaceae is so complicated that it is not possible to define clearly the evolutionary events responsible for these rearrangements. However, phylogenetic analyses of the gene order data have generated a well-resolved phylogeny for 18 taxa (Fig. 8), and the dataset exhibits lower levels of homoplasy than phylogenies inferred from *rbcL* or ITS sequences for the same taxa (Cosner *et al*., in press).

A number of groups have been developing computational methods for using gene order data for phylogeny reconstruction (Table 4, Bourque and Pevzner, 2002; Cosner *et al*., 2000; Cosner *et al*., in press; Larget *et al*., 2002; Moret *et al*., 2001; Wang *et al*., 2002). The approaches are designed to analyze highly rearranged genomes using several different phylogenetic approaches, including distance, parsimony and Bayesian methods. Most of these algorithms are designed for genomes that have a single chromosome with equal gene content, though some more recent studies have begun to implement methods for multiple chromosomes

(Bourque and Pevzner, 2002) and unequal gene content (Tang and Moret, 2003). The utility of most of these algorithms has been tested using simulation studies, however, the Campanulaceae chloroplast genomes have been used as benchmark empirical data-set for assessing speed and accuracy of these methods (Bourque and Pevzner, 2002; Moret *et al*., 2001). A more detailed review of algorithms for phylogenetic analysis of gene order data can be found in the chapter in this volume by B. Moret and T. Warnow. The availability of many new completely sequenced chloroplast genomes in the future should provide a much expanded empirical data-set for the development of new algorithms that use gene order data for phylogeny reconstruction.

3. Phylogenetic and molecular evolutionary analysis of genomic sequences

Completely sequenced chloroplast genomes provide a rich source of nucleotide and amino acid sequence data that can be used to address phylogenetic and molecular evolutionary questions. Several recent studies have attempted to use entire suites of sequences (e.g., all shared protein-coding genes) from completely sequenced genomes to resolve a number of phylogenetic issues, including relationships among grasses (Matsuoka *et al*., 2002), identification of the basal lineage of flowering plants (Goremykin *et al*., 2003a, b, 2004; Leebens-Mack *et al*., in prep) and land plants (Kugita, 2003), and relationships among land plants and green algae (Lemieux *et al*., 2000; Turmel *et al*., 1999). Phylogenies based on all or at least many of the shared genes among completely sequenced chloroplast genomes also have been used to address questions about the origins of plastids and about patterns of gene loss or transfer (Chu *et al*., 2004; Martin *et al*., 2002; Maul *et al*., 2002). The latter studies have supported several phylogenetic conclusions: (1) there has been a single primary endosymbiotic origin of plastids;

(2) extensive gene loss and/or transfer to the nucleus has occurred; and (3) multiple, independent secondary endosymbiotic events have occurred.

Use of many or all of the genes from the chloroplast genome provides many more characters for phylogeny reconstruction in comparison with previous studies that have relied on only a few genes to address the same questions. However, one current problem with the whole genome approach is that taxon sampling is quite limited and can result in misleading estimates of relationship. A recent example of this problem is the study by Goremykin et al. (Goremykin *et al*., ; 2003b; Goremykin *et al*.) that suggests that *Amborella* may not be the basal angiosperm, a result that contradicts many recent phylogenies based on sequences of a few genes (Barkman *et al*., 2000; Graham and Olmstead, 2000; Mathews and Donoghue, 1999; Parkinson *et al*., 1999; Qiu *et al*., 1999; Soltis *et al*., 1999). Phylogenetic analyses of expanded taxon sets have demonstrated that inadequate taxon sampling caused Goremykin's anomalous result (Leebens-Mack *et al*., in prep; Soltis and Soltis, 2004). In the future, increased availability of more completely sequenced chloroplast genomes will facilitate phylogenetic inference. Much denser taxon sampling is necessary before many of the advantages of whole genome sequencing can be fully realized and investigators must seriously consider the effects of long branch attractions (Felsenstein, 1978). Three other problematic issues have been identified and must be considered especially in broad phylogenetic comparisons. Compositional bias among plastids from divergent lineages can generate incorrect tree topologies (Lockhart *et al*., 1999); alignment of coding regions can be very difficult, especially when addressing phylogenetic issues at deep nodes (Chu *et al*., 2004); and tree topologies are very sensitive to the model of evolution being used (Leebens-Mack *et al*., in prep; Martin *et al*., 1998). A number of studies have attempted to address these issues by developing more realistic models of amino acid substitutions for

chloroplast-encoded genes (Adachi *et al*., 2000; Morton and So, 2000), by examining lineage and locus specific rate heterogeneity among chloroplast genomes (Muse and Gaut, 1997), and by developing alternative methods for using sequences from whole chloroplast genomes (Chu *et al*., 2004; Lockhart *et al*., 1999; Rivas *et al*., 2002).

V. Summary and Future Directions

It is currently a very exciting time for the field of comparative chloroplast genomics. The first chloroplast genome sequences were published 18 years ago and now there are 43 genomes available, almost two-thirds of which have been completed during the past four years (Table 1). In this chapter, we have described many recent developments, which, by improving methods for gathering and analyzing chloroplast genome sequences, are providing the necessary framework for greatly expanding the number of sequenced genomes in the near future. The most significant advancements include RCA for amplification of entire genomes and DOGMA software (Wyman *et al*., in press) for annotation. Several ongoing projects on seed plants, land plants, and algae are likely to result in the availability of nearly 200 completely sequenced genomes during the next five years (see http://megasun.bch.umontreal. ca/ogmp/projects/sumprog.html, http://www.jgi.doe.gov/programs /comparative/second_levels/ chloroplasts/jansen_project_home /chlorosite.html, and http://ucjeps.berkeley.edu/TreeofLife/ for more detailed information about ongoing projects). This increased taxon sampling to include more representatives of all of the major lineages of plants, ultimately will provide unprecedented opportunities for addressing phylogenetic questions at deep nodes. These data also will provide important new insights into both patterns and mechanisms of chloroplast genome evolution. Another outcome of these efforts will be the development of new algorithms, new models of

chloroplast sequence and genome evolution, and improved computational tools for using both gene order and sequence data for phylogeny reconstruction. Finally, the chloroplast genomic data and the computational methods will be of great value to plant molecular biologists interested in the functional attributes of chloroplast genes and their interaction with other plant organelles.

Acknowledgments

Our research on chloroplast genome sequencing and gene order phylogeny is supported by NSF grants (Biocomplexity grant DEB 0120709 to RKJ, LAR, JB, and CWD; IIS 0113654 to RKJ, T. Warnow, and B. Moret; and RUI/DEB 0075700 to LAR) . SKW acknowledges support from a NSF IGERT fellowship (DGE 0114387) and RCH acknowledges Research Internship support from the Graduate School at UT-Austin. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC03-76SF00098. We thank Ms. Sheila Plock for technical assistance.

Literature Cited

Adachi, J., P. J. Waddell, W. Martin and M. Hasegawa (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**, 348-358.

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

Barkman, T. J., G. Chenery, J. R. McNeal, J. Lyons-Weiler, W. J. Ellisens, G. Moore, A. D. Wolfe and C. W. dePamphilis (2000). Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 13166-13171.

Bartoszewski, G., N. Katzir and M. J. Havey (2004). Organization of repetitive DNAs and the genomic regions carrying ribosomal RNA, *cob*, and *atp9* genes in the cucurbit mitochondrial genomes. *Theoretical and Applied Genetics* **108**, 982-992.

Bock, R. (2000). Sense from nonsense: How the genetic information of chloroplasts is altered by RNA editing. *Biochimie* **82**, 549-557.

Bookjans, G., B. M. Stummann and K. W. Henningsen (1984). Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic-strength. *Analytical Biochemistry* **141**, 244-247.

Bourque, G. and P. A. Pevzner (2002). Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research* **12**, 26-36.

Bowman, C. M. and T. Dyer (1986). The location and possible evolutionary significance of small dispersed repeats in wheat ctDNA. *Current Genetics* **10**, 931-941.

Chu, K. H., J. Qi, Z. G. Yu and V. Anh (2004). Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Molecular Biology and Evolution* **21**, 200-206.

Copertino, D. W. and R. B. Hallick (1991). Group-II Twintron - an intron within an intron in a chloroplast cytochrome-B-559 gene. *Embo Journal* **10**, 433-442.

Cosner, M. E., R. K. Jansen and T. G. Lammers (1994). Phylogenetic relationships in the Campanulales based on *rbcL* sequences. *Plant Systematics and Evolution* **190**, 79-95.

Cosner, M. E., R. K. Jansen, B. M. E. Moret, L. A. Raubeson, L.-S. Wang, T. Warnow and W. S. (2000). A new fast heuristic for computing the breakpoint phylogeny and experimental analyses

of real and synthetic data. 8th International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA.

Cosner, M. E., R. K. Jansen, J. D. Palmer and 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. *Current Genetics* **31**, 419-429.

Cosner, M. E., L. A. Raubeson and R. K. Jansen (in press). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biology*.

Dean, F. B., S. Hosono, L. H. Fang, X. H. Wu, A. F. Faruqi, P. Bray-Ward, Z. Y. Sun, Q. L. Zong, Y. F. Du, J. Du, M. Driscoll, W. M. Song, S. F. Kingsmore, M. Egholm and R. S. Lasken (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5261-5266.

Downie, S. R. and J. D. Palmer (1992). Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. *In* "Molecular Systematics of Plants", (P. S. Soltis, D. E. Soltis and J. J. Doyle, Eds.), pp. 14-35, Chapman and Hall, New York.

Doyle, J. J., J. I. Davis, R. J. Soreng, D. Garvin and M. J. Anderson (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proceedings of the National Academy of Sciences of the United States of America* **89**, 7722-7726.

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

Ewing, B. and P. Green (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-194.

Fan, W. H., M. A. Woelfle and G. Mosig (1995). 2 copies of a DNA element, Wendy, in the chloroplast chromosome of *Chlamydomonas reinhardtii* between rearranged gene clusters. *Plant Molecular Biology* **29**, 63-80.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**, 401-410.

Gordon, D., C. Abajian and P. Green (1998). Consed: A graphical tool for sequence finishing. *Genome Research* **8**, 195-202.

Gordon, D., C. Desmarais and P. Green (2001). Automated finishing with Autofinish. *Genome Research* **11**, 614-625.

Goremykin, V., K. I. Hirsch-Ernst, S. Wolfl and F. H. Hellwig (2003a). Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution* **20**, 1499-1505.

Goremykin, V., K. I. Hirsch-Ernst, S. Wolfl and F. H. Hellwig (2003b). The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* - structural and phylogenetic analyses. *Plant Systematics and Evolution* **242**, 119-135.

Goremykin, V., K. I. Hirsch-Ernst, S. Wolfl and F. H. Hellwig (2004). The chloroplast genome of *Nymphaea alba* : Whole-genome analyses and the problem of identifying the most basal angiosperm. *Molecular Biology and Evolution* **21**, 1445-1454.

Graham, S. W. and R. G. Olmstead (2000). Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* **87**, 1712-1730.

Hipkins, V. D., K. A. Marshall, D. B. Neale, W. H. Rottmann and S. H. Strauss (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. *Current Genetics* **27**, 572-579.

Howe, C. J. (1985). The endpoints of an inversion in wheat chloroplast DNA are associated with short repeated sequences containing homology to att-lambda. *Current Genetics* **10**, 139-145.

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

Jansen, R. K. and J. D. Palmer (1987). A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America* **84**, 5818-5822.

Jansen, R. K., J. L. Wee and D. Millie (1998). Comparative utility of restriction site and DNA sequence data for phylogenetic studies in plants. *In* "Molecular Systematics of Plants II: DNA Sequencing", (D. E. Soltis, P. S. Soltis and J. J. Doyle, Eds.), pp. 87-100, Chapman and Hall, New York.

Knox, E. B. and J. D. Palmer (1998). Chloroplast DNA evidence on the origin and radiation of the giant lobelias in eastern Africa. *Systematic Botany* **23**, 109-149.

Kolodner, R. and K. K. Tewari (1979). Inverted repeats in chloroplast DNA from higher plants. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 41-45.

Kugita, M., Kaneko, A., Yamamoto Y., Takeya, Y., Matsumoto, T. and Yoshinaga, K. (2003). The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. *Nucleic Acids Research* **31**, 716-721.

Kuroda, H. and P. Maliga (2002). Overexpression of the *clpP* 5 '-untranslated region in a chimeric context causes a mutant phenotype, suggesting competition for a *clpP*-specific RNA maturation factor in tobacco chloroplasts. *Plant Physiology* **129**, 1600-1606.

Kurtz, S., J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye and R. Giegerich (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**, 4633-4642.

Kurtz, S. and C. Schleiermacher (1999). REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**, 426-427.

Larget, B., D. L. Simon and J. B. Kadane (2002). Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 681-693.

Leebens-Mack, J. H., C. dePamphilis, W., L. A. Raubeson, R. K. Jansen, L. Cui, J. L. Boore, M. Fourcade, J. V. Kuehl and T. W. Chumley (in prep). The utility of whole chloroplast genome sequencing for reconstructing and dating deep nodes in the angiosperm phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*.

Lemieux, C., C. Otis and M. Turmel (2000). Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* **403**, 649-652.

Lockhart, P. J., C. J. Howe, A. C. Barbrook, A. W. D. Larkum and D. Penny (1999). Spectral analysis, systematic bias, and the evolution of chloroplasts. *Molecular Biology and Evolution* **16**, 573-576.

Maier, U. G., S. A. Rensing, G. L. Igloi and M. Maerz (1995). Twintrons are not unique to the *Euglena* chloroplast genome - Structure and evolution of a plastome *cpn60* Gene from a cryptomonad. *Molecular & General Genetics* **246**, 128-131.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa and D. Penny (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12246-12251.

Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa and K. V. Kowallik (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162-165.

Mathews, S. and M. J. Donoghue (1999). The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947-950.

Matsuoka, Y., Y. Yamazaki, Y. Ogihara and K. Tsunewaki (2002). Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals. *Molecular Biology and Evolution* **19**, 2084-2091.

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

Milligan, B. G., J. N. Hampton and J. D. Palmer (1989). Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Molecular Biology and Evolution* **6**, 355-368.

Miyagi, T., S. Kapoor, M. Sugita and M. Sugiura (1998). Transcript analysis of the tobacco plastid operon *rps2/atpI/H/F/A* reveals the existence of a non-consensus type II (NCII) promoter upstream of the *atpI* coding sequence. *Molecular and General Genetics* **257**, 299-307.

Moret, B. M. E., L. S. Wang, T. Warnow and S. K. Wyman (2001). New approaches for reconstructing phylogenies from gene order data. 9th International Conference on Intelligent Systems in Molecular Biology, Copenhage, Denmark.

Morton, B. R. (1994). Codon use and the rate of divergence of land plant chloroplast genes. *Molecular Biology and Evolution* **11**, 231-238.

Morton, B. R. and B. G. So (2000). Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. *Journal of Molecular Evolution* **50**, 184-193.

Mullet, J. E., D. Christopher, J. Rapp, B. Y. Meng, M. Y. Kim, J. M. Kim and D. Laflamme (1992). Dynamic regulation of plastid gene expression during chloroplast biogenesis. *Photosynthesis Research* **34**, 94-94.

Muse, S. V. and B. S. Gaut (1997). Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* **146**, 393-399.

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

Olmstead, R. G. and J. D. Palmer (1994). Chloroplast DNA systematics - A review of methods and data analysis. *American Journal of Botany* **81**, 1205-1224.

Palmer, J. D. (1986). Isolation and structural analysis of chloroplast DNA. *Methods in Enzymology* **118**, 167-186.

Palmer, J. D. (1991). Plastid chromosomes: structure and evolution. *In* "The molecular biology of plastids. Cell culture and somatic cell genetics of plants." (R. G. Hermann, Ed.), 7A, pp. 5-53, Springer-Verlag, Vienna.

Palmer, J. D., J. M. Nugent and L. A. Herbon (1987). Unusual structure of Geranium chloroplast DNA - A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 769-773.

Parkinson, C. L., K. L. Adams and J. D. Palmer (1999). Multigene analyses identify the three earliest lineages of extant flowering plants. *Current Biology* **9**, 1485-1488.

Pombert, J. F., C. Otis, C. Lemieux and M. Turmel (2004). The complete mitochondrial DNA sequence of the green alga *Pseudendoclonium akinetum* (Ulvoplayceae) highlights distinctive evolutionary trends in the chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. *Molecular Biology and Evolution* **21**, 922-935.

Qiu, Y. L., J. H. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. D. Chen, V. Savolainen and M. W. Chase (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404-407.

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

Raubeson, L. A. and R. K. Jansen (2005). Chloroplast genomes of plants. *In* "Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants", (R. J. Henry, Ed.), pp. CAB International.

Rivas, R. D., J. J. Lozano and A. R. Ortiz (2002). Comparative analysis of chloroplast genomes: Functional annotation, genome-based phylogeny, and deduced evolutionary patters. *Genome Research* **12**, 567-583.

Sandbrink, J. M., P. Vellekoop, R. Vanham and J. Vanbrederode (1989). A method for evolutionary studies on RFLP of chloroplast DNA, applicable to a range of plant species. *Biochemical Systematics and Ecology* **17**, 45-49.

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

Sears, B. B., L. L. Stoike and W. L. Chiu (1996). Proliferation of direct repeats near the *Oenothera* chloroplast DNA origin of replication. *Molecular Biology and Evolution* **13**, 850-863.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchishinozaki, C. Ohto, K. Torazawa, B. Y. Meng, M. Sugita, H. Deno, T. Kamogashira, K. Yamada, J. Kusuda, F. Takaiwa, A. Kato, N. Tohdoh, H. Shimada and M. Sugiura (1986). The complete nucleotide sequence of the tobacco chloroplast genome - Its gene organization and expression. *Embo Journal* **5**, 2043-2049.

Simpson, C. L. and D. B. Stern (2002). The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. *Plant Physiology* **129**, 957-966.

Soltis, D. E. and P. S. Soltis (2004). *Amborella* not a "basal angiosperm"? Not so fast. *American Journal of Botany* **91**, 997-1001.

Soltis, P. S., D. E. Soltis and M. W. Chase (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**, 402-404.

Sugiura, M. (1992). The chloroplast genome. *Plant Molecular Biology* **19**, 149-168.

Sugiura, M. (2003). History of chloroplast genomics. *Photosynthesis Research* **76**, 371-377.

Tang, J. J. and B. M. E. Moret (2003). Phylogenetic reconstruction from gene-rearrangement data with unequal gene content. *In* "Algorithms and Data Structures, Proceedings", 2748, pp. 37-46.

Triboush, S. O., N. G. Danilenko and O. G. Davydenko (1998). A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. *Plant Molecular Biology Reporter* **16**, 183-189.

Turmel, M., C. Otis and C. Lemieux (1999). The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 10248-10253.

Volfvovsky, N., B. J. Hass and S. L. Salzberg (2001). A clustering method for repeat analysis in DNA sequences. *Genome Biology* **2**, 1-11.

Vomstein, J. and W. Hachtel (1988). Deletions insertions, short inverted repeats, sequences resembling att-lambda, and frame shift mutated open reading frames are involved in chloroplast DNA differences in the genus *Oenothera* subsection *Munzia*. *Molecular & General Genetics* **213**, 513-518.

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

Wang, L.-S., R. K. Jansen, B. M. E. Moret, L. A. Raubeson and T. Warnow (2002). Fast phylogenetic methods for analysis of genome rearrangement data: An empirical study. Proceedings of the 7th Pacific Symposium Biocomputing PSB 2002, Lihue, Hawaii.

Wolfe, K. H., W. H. Li and P. M. Sharp (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 9054-9058.

Wyman, S. K., R. K. Jansen and J. L. Boore (in press). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*.

Table 1. Alphabetical list of 43 complete plastid genome sequences as of August 16, 2004 (see http://megasun.bch.umontreal.ca/ogmp/projects/other/cp_list.html, http://www.ncbi.nlm.nih.gov:80/genomes/static/euk_o.html, and http://www.rs.noda.tus.ac.jp/~kunisawa/order/front.html for access to these genomic sequences). All listed genomes are chloroplasts except as noted.

Species	NCBI Classification	Accession Number	Year Completed	Genome Size (bp)
Adiantum capillus-veneris	Embryophyta	AY178864	2003	150,568
Amborella trichopoda	Embryophyta	AJ506156	2003	162,686
Anthoceros formosae	Embryophyta	AB086179	2003	161,162
Arabidopsis thaliana	Embryophyta	AP000423	1999	154,478
Atropa belladonna	Embryophyta	AJ316582	2003	156,687
Calycanthus fertilis var. ferax	Embryophyta	AJ428413	2003	153,337
Chaetosphaeridium globosum	Streptophyta	AF494278	2002	131,183
Chlamydomonas reinhardtii	Chlorophyta	BK000554	2004	203,828
Chlorella vulgaris	Chlorophyta	AB001684	1997	150,613
Cyanidioschyzon merolae	Rhodophyta	AB002583/ AY286123	2003	149,987 149,705
Cyanidium caldarium	Rhodophyta	AF022186	1999	164,921
Cyanophora paradoxa	Glaucocystophyceae	U30821	1995	135,599
Eimeria tenella[1]	Alveolata	AY217738	2003	34,750
Epifagus virginiana[2]	Embryophyta	M81884	1993	70,028
Euglena gracilis	Euglenozoa	X70810	1993	143,171
Euglena longa	Euglenozoa	AJ294725	2001	73,345
Gracilaria tenuistipitata	Rhodophyta	AY673996	2004	183,883
Guillardia theta	Cryptophyta	AF041468	1998	121,524
Lotus corniculatus	Embryophyta	AP002983	2001	150,519
Marchantia polymorpha	Embryophyta	X04465	1986	121,024
Medicago truncatula	Embryophyta	AC093544	2001	124,033
Mesostigma viride	Chlorophyta	AF166114	2000	118,360
Nephroselmis olivacea	Chlorophyta	AF137379	1999	200,799
Nicotiana tabacum	Embryophyta	Z00044	1986	155,939
Nymphaea alba	Embryophyta	AJ627251	2004	159,930
Odontella sinensis	Stramenopiles	Z67753	1996	119,704
Oenothera elata	Embryophyta	AJ271079	2000	163,935
Oryza nivara	Embryophyta	AP006728	2004	134,494

Oryza sativa	Embryophyta	X15901/ AY522329/ AY522331	1989/ 2004/ 2004	134,525/ 134,496/ 134,551
Physcomitrella patens	Embryophyta	AP005672	2003	122,890
Pinus koraiensis	Embryophyta	AY228468	2003	116,866
Pinus thunbergii	Embryophyta	D17510	1996	119,707
Porphyra purpurea	Rhodophyta	U38804	1996	191,028
Psilotum nudum	Embryophyta	AP004638	2002	138,829
Saccharum hybrid	Embryophyta	AE009947	2004	141,182
Saccharum officinarum	Embryophyta	AP006714	2004	141,182
Spinacia oleracea	Embryophyta	AJ400848	2000	150,725
Toxoplasma gondii[1]	Alveolata	U87145	1999	34,996
Triticum aestivum	Embryophyta	AB042240	2001	134,545
Zea mays	Embryophyta	X86563	1995	140,384

[1]plastid genome remnant, nonphotosynthetic protist
[2]plastid genome, nonphotosynthetic flowering plant

Table 2. Isolation of chloroplasts or cpDNA by Sucrose Step Gradient Centrifugation (see Palmer, 1986; Sandbrink *et al*., 1989).

1. Prior to extraction, place plants in the dark for 1-2 days to reduce chloroplast starch levels. Approximately 100 gm or more of leaf tissue is required to get sufficient quantities of cpDNA. If the chloroplast isolation is being prepared for RCA at least 20 gm of leaf tissue is generally necessary. The quality of the plant tissue is probably the most important criterion for a successful isolation. Leaves that are fresher and younger are far superior to older, senescing leaves.

2. Wash healthy greens leaves in tap water if visibly dirty and cut into small pieces (ca. 2-10 cm^2 in surface area).

3. Place 25-50 gm of cut leaves in 400 ml of ice-cold isolation buffer. Steps 3-5 are done in a cold room at 4˚ C or on ice. We have found that the isolation buffer in Sandbrink et al. (1989) often yields a much purer chloroplast pellet (see recipes at end of protocol). This buffer contains higher concentrations of salts and 2-mercapto ethanol.

4. Homogenize in a pre-chilled blender for 5 five-second bursts at high speed.

5. Filter through four layers of cheesecloth and squeeze remaining liquid through the cloth. Then filter through one layer of miracloth (Calbiochem, catalog number 475855) without squeezing.

6. Divide filtrate into multiple centrifuge bottles and centrifuge at 1000g for 15 min at 4˚ C. Pour off supernatant.

7. Resuspend pellet from 10-50 gm of starting material in 7 ml of ice-cold wash buffer using a soft paintbrush and by vigorous swirling.

8. Gently load the resuspended pellet onto a step gradient consisting of 18 ml of 52% sucrose, over-layered with 7 ml of 30% sucrose. The overlay should be added with sufficient mixing to create a diffuse interface. It is best to pour the sucrose gradients 1-2 days prior to the extraction and allow them to sit at 4˚ C to allow for mixing of the interface. To enhance the purity of your cpDNA isolation, it is best to use more sucrose gradients, each with material from a smaller amount of tissue, so that the nuclei can better penetrate the chloroplast band. At least six sucrose gradients are recommended for 100-200 gm of starting material. When preparing chloroplasts (rather than cpDNA) we will use three gradients for just 20 gm of tissue. We also have experimented with modifying the percentage of sucrose in the step gradients. We have found that the optimal percentage varies from one taxon to the next. For example, 52/30% gradients work well for most angiosperms, *Ginkgo* and conifers but we found that a 44-48 % sucrose in the bottom layer yielded DNA with a much higher proportion of cpDNA for cycads.

9. Centrifuge the step gradients at 25,000 RPM for 30-60 min at 4° C in a SW-27 (Beckman) or AH-627 (Sorvall) swinging bucket rotor.

10. Remove the chloroplast band from the 30 - 52% interface using a wide bore pipet, dilute with 3-10 volumes wash buffer, and centrifuge at 1,500g for 15 min at 4° C. We have found that the use of the Sandbrink wash buffer often improves the purity of the cpDNA. Multiple cycles of washing, pelleting, and resuspending of the chloroplasts often much purer cpDNA.

11. Resuspend the chloroplast pellet in wash buffer to a final volume of 2 ml. Depending on the size of the final pellet it may be necessary to resuspend the pellet in a larger volume and then divide resuspended pellet into separate tubes with no more than 3 ml per tube. If you are planning to use the chloroplasts for RCA this is the point at which you proceed to the RCA protocol in Table 3.

12. Add one-tenth volume of a 10 mg/ml solution of self-digested (2 hr at 37° C) Pronase (Calbiochem, catalog number 537088) and incubate for 2 min at room temperature.

13. Gently add one-fifth volume of 1X lysis buffer and mix in by slowly inverting the tube several times over a period of 10-15 min at room temperature. We experimented with higher concentrations of lysis buffer (a 5X lysis buffer versus the normal 1X buffer) and with doing the lysis at higher temperatures for longer periods of time (37 ˚C for 15-60 min). In general, we found that the 5X lysis buffer incubated at 37 ˚C gave much higher yields of cpDNA. We also tried several alternative lysis buffers that used Hexadecyltrimethylammonium Bromide (CTAB, Milligan *et al.*, 1989) or sodium dodecyl sulphate (SDS, Triboush *et al.*, 1998), but in general we did not have much success with these buffers.

14. Centrifuge for 10 min at room temperature in a clinical centrifuge to remove residual starch and cell wall debris from the chloroplast lysate. Transfer lysate to a new tube. This step is optional.

15. Add 1.0 g of technical grade cesium chloride (CsCl) per 1 ml of lysate and add ethidium bromide (EtBr) to a final concentration of 200 mg/ml. Fill remaining volume of ultracentrifuge tubes with a premixed solution of 1 g CsCl per 1 ml of TE buffer.

16. Centrifuge the small CsCl/EtBr gradients (5 ml) in a vertical rotor for 5-8 hr at 65,000 RPM at 20 ˚C.

17. Remove the band from gradient and if necessary reband in a second gradient or move on to step 18. High molecular weight chloroplast DNA will be very viscous and easily removed "en masse" from near the center of the gradient.

18. Remove ethidium bromide by at least three extractions with isopropanol saturated with NaCl and H_20 and dialyze against at least three changes of 2 liters of dialysis buffer over a period of 1-2 days.

19. Check purity of cpDNA by doing restriction digests and agarose gel electrophoresis.

20. Store the chloroplast DNA at 4 ˚C for short-term and at –20 ˚C for long-term use. Digests of cpDNA produce well-defined bands whereas nuclear DNA produces so many bands that it appears as a smear on the gel.

Standard isolation buffer
0.35 M sorbitol
50 mM tris-HCl, pH 8.0
5 mM EDTA
0.1% BSA (w/v, Sigma A-4503)
1.5 mM 2-mercapto ethanol

Sandbrink isolation buffer
1.25 M NaCl
50 mM tris-HCl, pH 8.0
5 mM EDTA
1% BSA (w/v, Sigma A-4503)
10 mM 2-mercapto ethanol
5% poly pyrrolidone (PVP-40)

Standard wash buffer
0.35M sorbitol
50 mM Tris-HCl, pH 8.0
25 mM EDTA

Sandbrink wash buffer
10 mM Tris-HCl, pH 8.0
5 mM EDTA
10 mM 2-mercapto ethanol
100 ug/ml proteinase K

52% Sucrose Solution
52% Sucrose (w/v)
50mM Tris, pH8.0
25 mM EDTA

30% Sucrose Solution
30% Sucrose (w/v)
50mM Tris, pH8.0
25 mM EDTA

1X Lysis Buffer
5% sodium sarcosinate (w/v)
50 mM Tris pH 8.0
25 mM EDTA

5X Lysis Buffer
20% sodium sarcosinate (w/v)
50 mM Tris pH 8.0
25 mM EDTA

Dialysis Buffer
10 mM Tris, pH 8.0
10mM NaCl
0.1 mM EDTA

Table 3. Whole chloroplast genome amplification using RCA

A. Setting up the RCA Reaction

1. Thaw RCA kit (Repli-g, Molecular Staging Inc.) reaction components (1X PBS, 4X Mix, Solution B, polymerase) on ice. Prepare the alkaline lysis solution (Solution A) if necessary.
2. Activate Solution A by adding DTT (must be made fresh before using): for each reaction 31.5 μl of solution A and 3.5 μl 1M of DTT is needed. This can be done while waiting for lysis in the next step or while components are thawing in the previous step.
3. Add 3 μl of the 5X lysis buffer to 15 μl of isolated chloroplasts (from step 11 in Table 2) and incubate for 15 min at 37 ºC. We have attempted to quantify the amount of chloroplasts in this step but it turns out that this is futile. The success of subsequent steps is more dependent on the quality and purity of the chloroplasts rather than on the number of chloroplasts that are added to the lysis reaction. We have found that the amount of the chloroplast prep added needs to be optimized for each taxon.
4. Add 50 μl of 1X PBS to the lysate.
5. Add 35 μl of the resulting solution to 35 μl of activated solution A and incubate on ice for 10 min.
6. While alkaline lysis is proceeding, prepare the reaction cocktail (35 μl H_20 + 12.5 μl 4X Mix + 0.5 μl polymerase) and aliquot it to the reaction tubes. This is based on 2 μl of lysate being added to each reaction – adjust volume of water accordingly if using more or less of the lysate.
7. Stop the alkaline lysis by adding 35 μl of neutralization solution to the lysate.
8. Take 2 μl of lysate and add to each reaction.
9. Incubate at 30 ºC for 16 hr; terminate with 3 min at 65 ºC. Generally the solution looks cloudy if the reaction has worked. Store in refrigerator or freezer until proceeding to B.

B. Checking for RCA product

1. Run 2 μl of product on mini-gel to determine if the RCA was successful.
2. If there is product on the mini-gel proceed with restriction digests.
3. Do restriction enzyme digests of 2 μl of RCA product using *BstB1* and *EcoR1* following the manufacturers recommendations in 20 μl reactions. Some enzymes do not digest RCA product very well. We have tested a number of enzymes and found that *BstB1* and *EcoR1* work best.

   2 μl RCA product
   2 μl of appropriate enzyme buffer
   sufficient H_20 to end up with a total volume of 20 μl
   10 - 20 units of enzyme
4. Load entire digest into 1% agarose gel and run dye marker to 10 cm
5. Stain, visualize, and photograph gel to assess the quality of the RCA product (see Fig. 2 for an example).

   Stocks: 5 M KOH (28g KOH pellets + H_20 to 100 ml; exothermic!)
   0.5 M EDTA (18.6g EDTA + 80 ml H_20, pH to 8.0; raise volume to 100ml)
   Lysis Solution (Solution A): 0.4 ml 5M KOH + 0.1 ml 0.5M EDTA + 4.5 ml H_20
   5X Lysis Buffer: 20% sarcosyl, 50 mM Tris pH 8, 25 mM EDTA.

Table 4. Chloroplast genome online databases and software.

Database/Software	Url	Features
Organelle Genome Megasequencing Program - OGMP	http://megasun.bch.umontreal.ca/ogmp/ projects/other/cp_list.html	Lists all sequenced chloroplast genomes with NCBI classification, accession numbers, and links to GenBank
ExPASy	http://us.expasy.org/txt/plastid.txt	Lists names of chloroplast and cyanelle proteins with abbreviations; also gives list of completely sequenced plastid genomes
NCBI - Organelle Genomes	http://www.ncbi.nlm.nih.gov:80/ genomes/static/euk_o.html	Lists all completely sequenced organelle genomes with accession numbers, genome size, and date of submission
Dual Organellar GenoMe Annotator (DOGMA)	http://phylocluster.biosci.utexas.edu/dogma/	A program for annotation of chloroplast and animal mitochondrial genomes
Plastid Gene Order Database	http://www.rs.noda.tus.ac.jp/~kunisawa /order/front.html	Provides corrected annotations for chloroplast genomes with tools to view gene orders and extracting sequences
Genomemine	http://www.genomics.ceh.ac.uk/cgi-bin/ gmine/ gminemenu.cgi?action=listorganelles &sort=genome	Provides list of all sequenced genomes with details of accession number, size, numbers of orfs, percent coding, and base frequency
DOE Joint Genome Institute (JGI) Organelle Genomics	http://www.jgi.doe.gov/programs/comparative /top_level/organelles.html	Provides access to several ongoing projects in organelle genomics and access to various tools for annotating and analyzing chloroplast and mitochondrial genomes
A Data Base of PCR Primers for the Study of the Chloroplast	http://fbva.forvie.ac.at/200/1859.html	Contains information about universal primers for chloroplast genomes

Genome in Plants		
BPAnalysis	http://www.cs.washington.edu/homes/blanchem/software.html	A program that computes minimal breakpoint trees from gene order data
Derange2	http://www.cs.washington.edu/homes/blanchem/software.html	A program that computes an approximation of minimal edit distances between pairs of gene orders
Genome Rearrangements In Man and Mouse (GRIMM)	http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM/index.html	Rearrangement algorithms for genomes, which computes the minimum possible number of rearrangement steps, and determines a possible evolutionary scenario using this number of steps.
Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms (GRAPPA)	http://www.cs.unm.edu/~moret/GRAPPA/	A program for constructing phylogenies using gene order data
Multiple Genome Rearrangements (MGR)	http://www-cse.ucsd.edu/groups/bioinformatics/MGR/index.html	A tool for constructing phylogenies based on gene order data
PipMaker and MultiPipmaker	http://pipmaker.bx.psu.edu/pipmaker/	Used to align 2 (PipMaker) or multiple (MultiPipmaker) genomes and provide dotmatrix and percent identity plot (PIP) diagrams of whole genomes
REPuter	http://www.genomes.de/	A program for identifying repeated sequences in genomes and provides an excellent visualization of the location and sequence of various types of repeats
FootPrinter	http://bio.cs.washington.edu/software.html	A program to identify putative regulatory elements in DNA sequences that requires a

		phylogeny
RepeatFinder	http://www.tigr.org/software/	Organizes repeats into classes
RepeatMasker	http://www.repeatmasker.org/	A program that screens DNA sequences for interspersed repeats

Figure 1. Gene map of tobacco chloroplast genome (from Raubeson and Jansen, 2005). The inner circle shows the four major regions of the genome – the two copies of the inverted repeat (IR_A and IR_B) and the large and small single copy regions (LSC and SSC). The outer circle represents the tobacco genome with the transcribed regions shown as boxes proportional to gene size. Genes inside the circle are transcribed in a clockwise direction, and genes outside of the circle are transcribed counterclockwise. The IR extent is shown by the increased width of the circle representing the tobacco genome. Genes with introns are marked with asterisks (*). Arrows between the gene boxes and gene names show those operons known to occur in tobacco cpDNA. Genes coding for products that function in protein synthesis are dark gray; genes coding for products that function in photosynthesis are stippled; genes coding for products with various other functions are lighter gray.

Figure 2. Gel photo showing chloroplast DNA isolations for *Lactuca* (Asteraceae) using DNAseI method and *Ranunuculus* using the NaCl method (see section II.1). Lanes 1 and 2 and 4 and 5 were digested with KpnI and HaeII, respectively; lane 3 is a lambda DNA digest used as a size marker.

Figure 3. Gel photo showing results of whole chloroplast genome amplification using RCA of isolated chloroplasts of *Ginkgo* and *Podocarpus*. Lane 2 shows uncut RCA product, and lanes 3 - 5 show 2 ul of RCA product cut with restriction enzymes. Lanes 1 and 6 are two different size markers. Quality of RCA product can be assessed by performing digests and running gels such as those shown here. Nuclear contamination would appear as a smear while the cpDNA forms discrete bands. The relative proportion of smear to bands is assessed visually

from the gel photo. Upon sequencing, this *Podocarpus* RCA product was found to be over 80% cpDNA and the *Ginkgo* product over 60% cpDNA.

Figure 4. Screen shot of two Consed (Gordon *et al*., 1998) windows. The left panel shows the main window with the list of contigs, individual reads, and several other features. The right panel shows the assembly view of four contigs of *Nuphar*, illustrating contig order, read depth, and inconsistent forward-reverse subclone pairs.

Figure 5. Two web browser windows from DOGMA (Wyman *et al*., in press). **A**. The main window for submitting FASTA formatted input files of complete genome sequences or contigs of portions of the genome. A number of optional settings are available for the genetic code for BLASTx, percent identity for protein coding genes and RNAs, e value, and the number of BLAST hits to return. **B**. A view of the annotation window with three panels: lower panel has several option buttons for extracting sequences, deleting/adding genes, and generating a Sequin formatted file or text file; middle panel shows tentative gene identifications, clicking on a gene will display that gene, its BLAST hits, and putative start and stop codons in the upper panel; upper panel shows the BLAST hits for the *psbA* gene and some putative stop codons. The sequin information window is also shown here. This is the window used to commit to the start and stop codon and it generates an entry compatible with Sequin.

Figure 6. MultiPipmaker (Schwartz *et al*., 2003) output of various published chloroplast genome sequences. The reference genome *Nicotiana* (Z00044) was analyzed against eight other genome sequences, including *Amborella* (AJ506156), *Arabidopis* (AP000423), *Calycanthus*

(AJ428413), *Lotus* (AP002983), *Nymphaea* (AJ627251), *Oenothera* (AJ271079), *Spinacia* (AJ400848), and *Triticum* (AB042240). **A.** MultiPip view showing sequencing identity (50-100%) among genomes with identity increasing with darker shades. Positions of genes and selected gene names are shown at top, names of taxa are on left. **B.** Selected region of the MultiPip showing sequence identity between 50-100%. Arrows on top of map indicate position of selected genes and numbers above gene indicate the exons for genes with introns. Note that this diagram shows that *accD* is absent from *Triticum*.

Figure 7. Reputer (Kurtz and Schleiermacher, 1999) output views of an analysis of the *Medicago* chloroplast genome (AC093544). The search examined forward and inverted repeats > 20 bp in length with 90% sequence identity. **A**. The visualization window is shown for forward repeats > 30 bp in length. **B**. A portion of the display of repeats found with the size of repeat, the coordinates in the genome, the hamming distance, e value, and the DNA sequence of the repeat given.

Figure 8. Campanulaceae phylogeny based on a maximum parsimony analysis of gene order changes (modified from Cosner *et al*., in press). Number and type of each genomic change are indicated as e = endpoint of IR, IV = inversion, IS = insertion >5 kb, T = transposition, and D = deletion/divergence. Only three endpoint characters are homoplasious, changing twice on the tree. Brackets indicate the major clades of Campanulaceae.

Figure 1

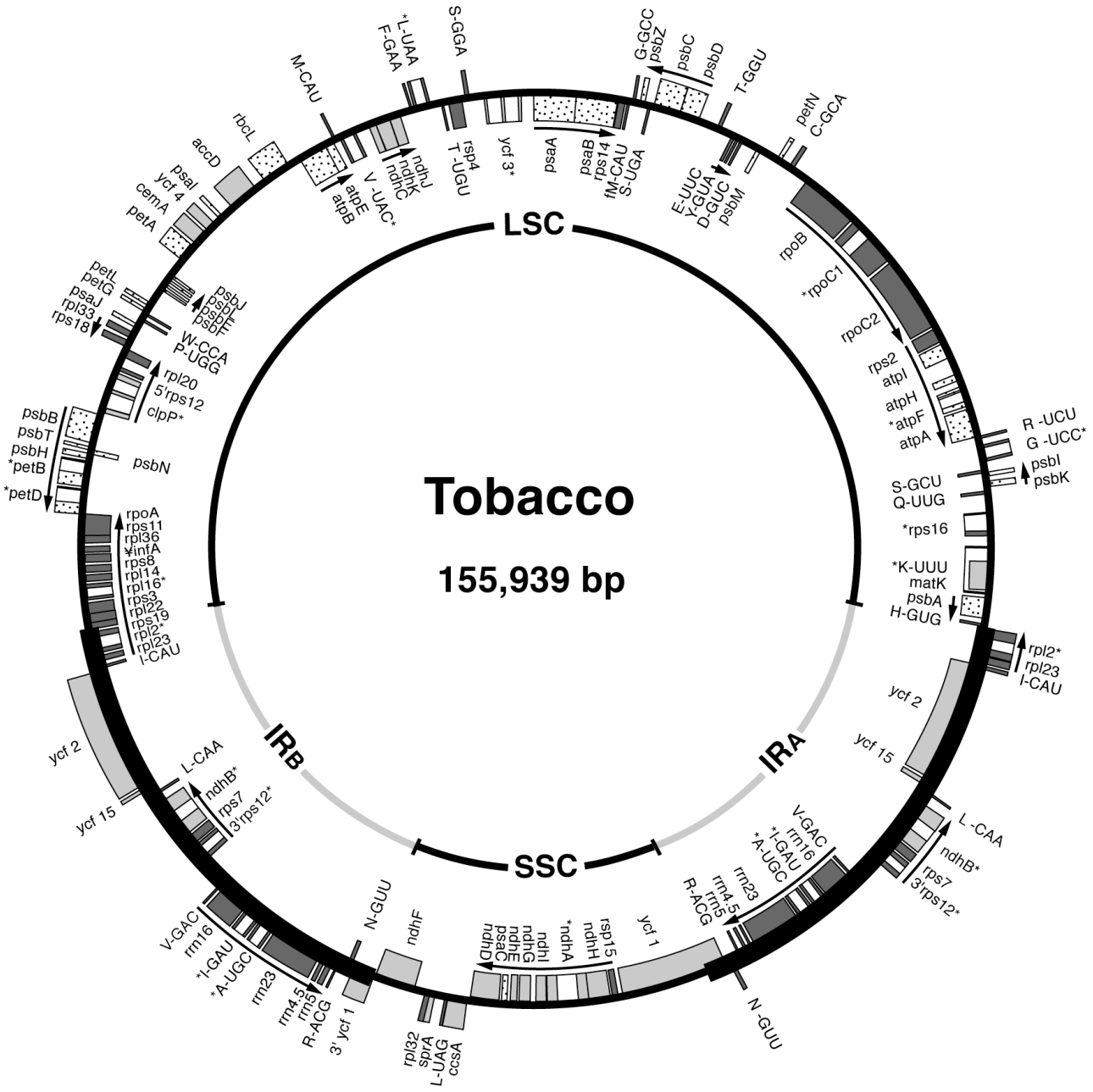

Tobacco
155,939 bp
LSC
IRA
SSC
IRB
S-GGA
*L-UAA
F-GAA
M-CAU
rbcL
accD
psaI
ycf 4
cemA
petA
petL
petG
psaJ
rpl33
rps18
psbJ
psbL
psbF
psbE
W-CCA
P-UGG
rpl20
5'rps12
clpP*
psbB
psbT
psbH
*petB
*petD
psbN
rpoA
rps11
rpl36
¥infA
rps8
rpl14
rpl16*
rps3
rpl22
rps19
rpl2*
rpl23
I-CAU
ycf 2
ycf 15
L-CAA
ndhB*
rps7
3'rps12*
V-GAC
rrn16
*I-GAU
*A-UGC
rrn23
rrn4.5
rrn5
R-ACG
3' ycf 1
N-GUU
ndhF
rpl32
sprA
L-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
*ndhA
ndhH
rsp15
ycf 1
N -GUU
ndhJ
ndhK
ndhC
V -UAC*
atpE
atpB
rsp4
T-UGU
ycf 3*
psaA
psaB
rps14
fM-CAU
S-UGA
G-GCC
psbZ
psbC
psbD
T-GGU
E-UUC
Y-GUA
D-GUC
psbM
petN
C-GCA
rpoB
*rpoC1
rpoC2
rps2
atpI
atpH
*atpF
atpA
R -UCU
G -UCC*
psbI
psbK
S-GCU
Q-UUG
*rps16
*K-UUU
matK
psbA
H-GUG
rpl2*
rpl23
I-CAU
ycf 2
ycf 15
L -CAA
ndhB*
rps7
3'rps12*
V-GAC
rrn16
*I-GAU
*A-UGC
rrn23
rrn4.5
rrn5
R-ACG

Figure 2

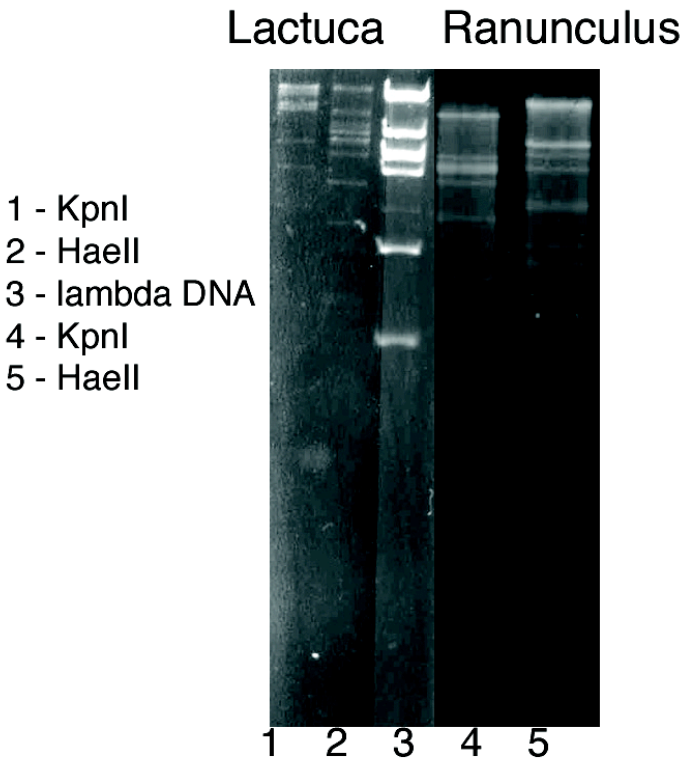

Lactuca
Ranunculus
1 - KpnI
2 - HaeII
3 - lambda DNA
4 - KpnI
5 - HaeII
1
2
3
4
5

Figure 3

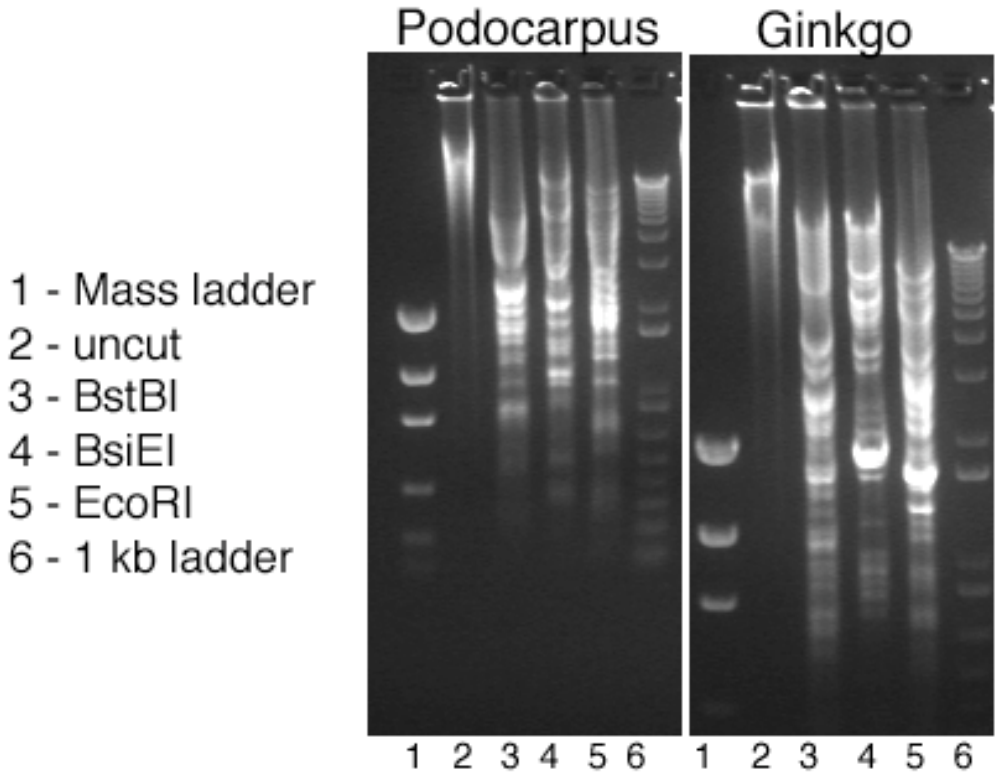

Podocarpus
Ginkgo
1 - Mass ladder
2 - uncut
3 - BstBI
4 - BsiEI
5 - EcoRI
6 - 1 kb ladder
1 2 3 4 5 6
1 2 3 4 5 6

Figure 4

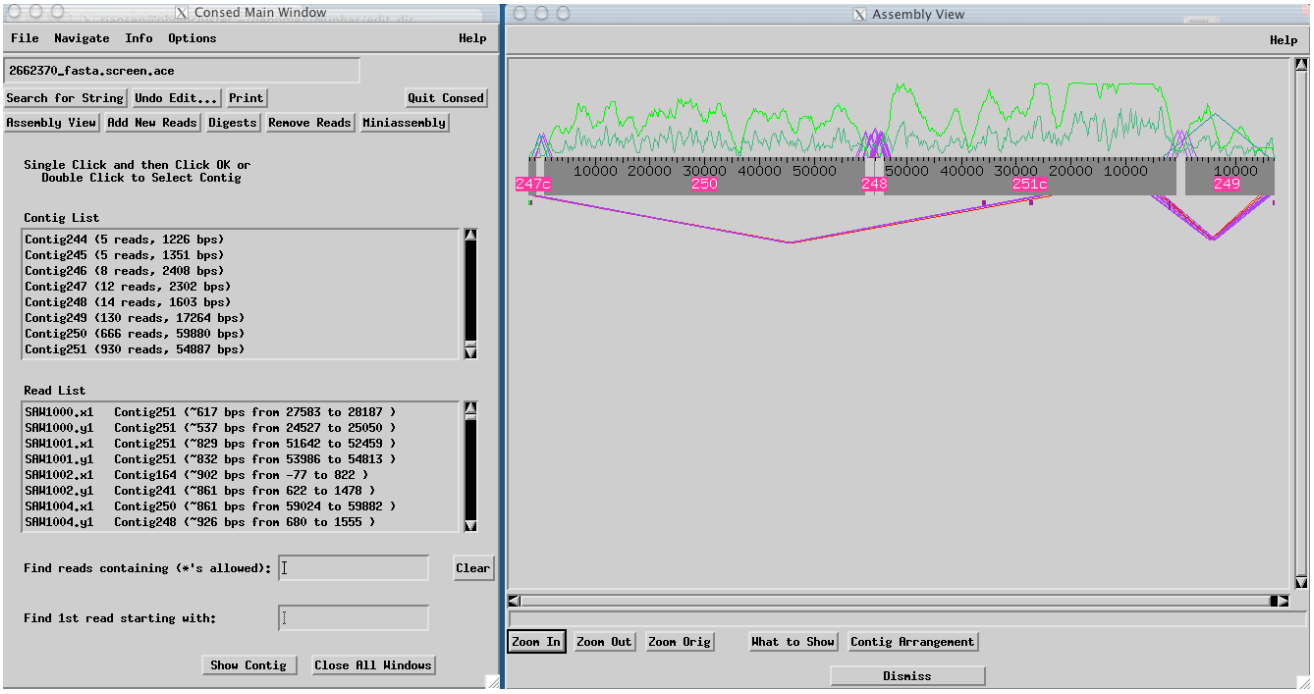
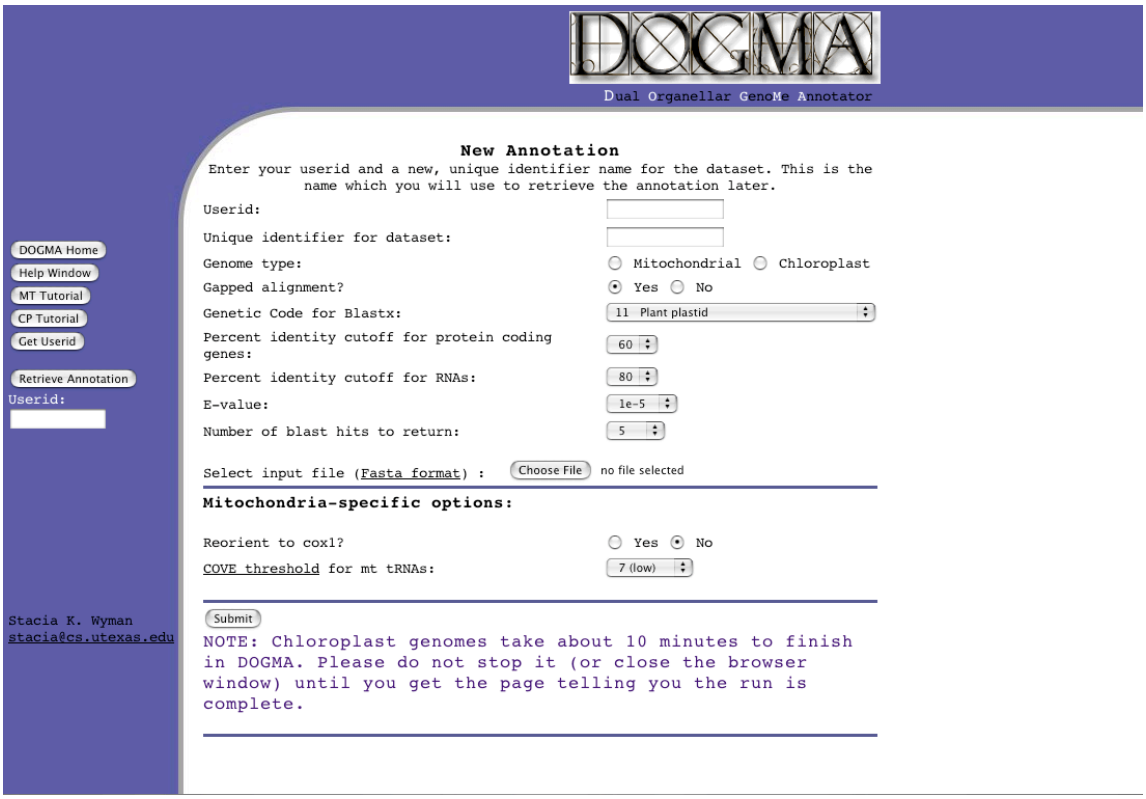

Consed Main Window
File Navigate Info Options Help
2662370_fasta.screen.ace
Search for String Undo Edit... Print Quit Consed
Assembly View Add New Reads Digests Remove Reads Miniassembly
Single Click and then Click OK or
Double Click to Select Contig
Contig List
Contig244 (5 reads, 1226 bps)
Contig245 (5 reads, 1351 bps)
Contig246 (8 reads, 2408 bps)
Contig247 (12 reads, 2302 bps)
Contig248 (14 reads, 1603 bps)
Contig249 (130 reads, 17264 bps)
Contig250 (666 reads, 59880 bps)
Contig251 (930 reads, 54887 bps)
Read List
SAW1000.x1 Contig251 (~617 bps from 27583 to 28187)
SAW1000.y1 Contig251 (~537 bps from 24527 to 25050)
SAW1001.x1 Contig251 (~829 bps from 51642 to 52459)
SAW1001.y1 Contig251 (~832 bps from 53986 to 54813)
SAW1002.x1 Contig164 (~902 bps from -77 to 822)
SAW1002.y1 Contig241 (~861 bps from 622 to 1478)
SAW1004.x1 Contig250 (~861 bps from 59024 to 59882)
SAW1004.y1 Contig248 (~926 bps from 680 to 1555)
Find reads containing (*'s allowed):
Clear
Find 1st read starting with:
Show Contig Close All Windows
Assembly View
Help
10000 20000 30000 40000 50000 50000 40000 30000 20000 10000 10000
247c 250 248 251c 249
Zoom In Zoom Out Zoom Orig What to Show Contig Arrangement
Dismiss

Figure 5

A


DOGMA
Dual Organellar GenoMe Annotator
New Annotation
Enter your userid and a new, unique identifier name for the dataset. This is the name which you will use to retrieve the annotation later.
Userid:
Unique identifier for dataset:
Genome type:
Mitochondrial
Chloroplast
Gapped alignment?
Yes
No
Genetic Code for Blastx:
11 Plant plastid
Percent identity cutoff for protein coding genes:
60
Percent identity cutoff for RNAs:
80
E-value:
1e-5
Number of blast hits to return:
5
Select input file (Fasta format) :
Choose File
no file selected
Mitochondria-specific options:
Reorient to cox1?
Yes
No
COVE threshold for mt tRNAs:
7 (low)
Submit
NOTE: Chloroplast genomes take about 10 minutes to finish in DOGMA. Please do not stop it (or close the browser window) until you get the page telling you the run is complete.
DOGMA Home
Help Window
MT Tutorial
CP Tutorial
Get Userid
Retrieve Annotation
Userid:
Stacia K. Wyman
stacia@cs.utexas.edu

B

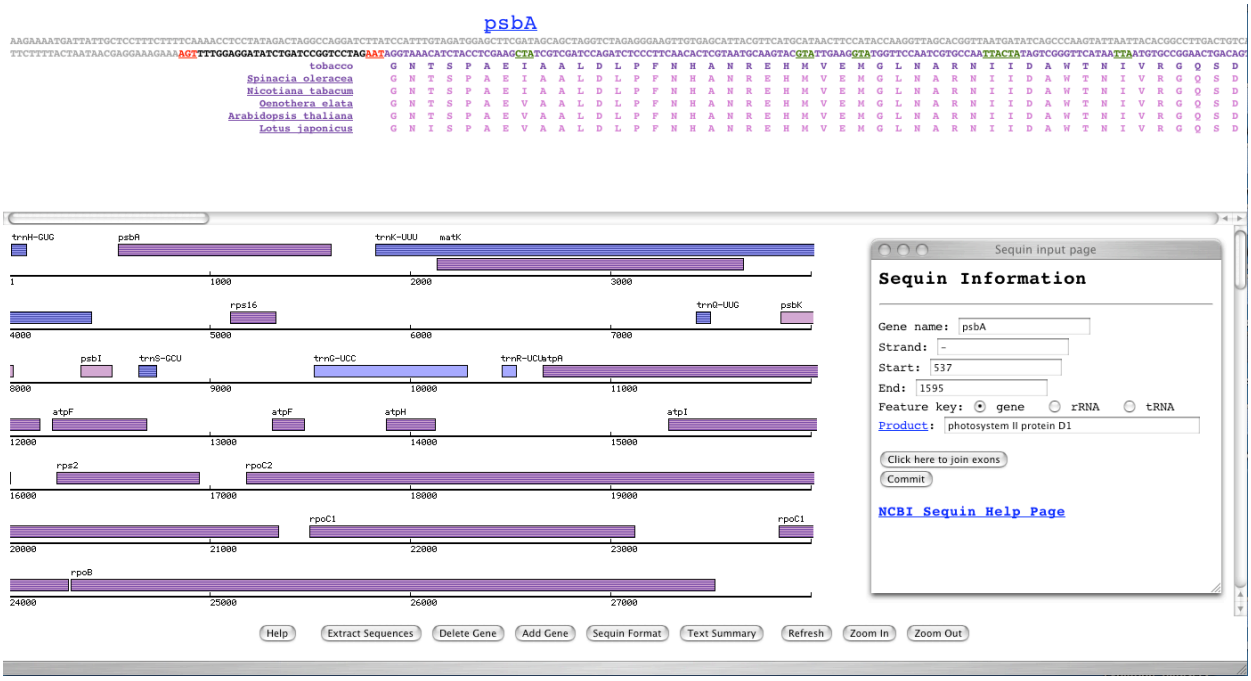

psbA
tobacco
Spinacia oleracea
Nicotiana tabacum
Oenothera elata
Arabidopsis thaliana
Lotus japonicus
trnH-GUG
psbA
trnK-UUU
matK
rps16
trnQ-UUG
psbK
psbI
trnS-GCU
trnG-UCC
trnR-UCU
atpA
atpF
atpF
atpH
atpI
rps2
rpoC2
rpoC1
rpoC1
rpoB
Sequin input page
Sequin Information
Gene name: psbA
Strand: -
Start: 537
End: 1595
Feature key: gene
rRNA
tRNA
Product: photosystem II protein D1
Click here to join exons
Commit
NCBI Sequin Help Page
Help
Extract Sequences
Delete Gene
Add Gene
Sequin Format
Text Summary
Refresh
Zoom In
Zoom Out

Figure 6

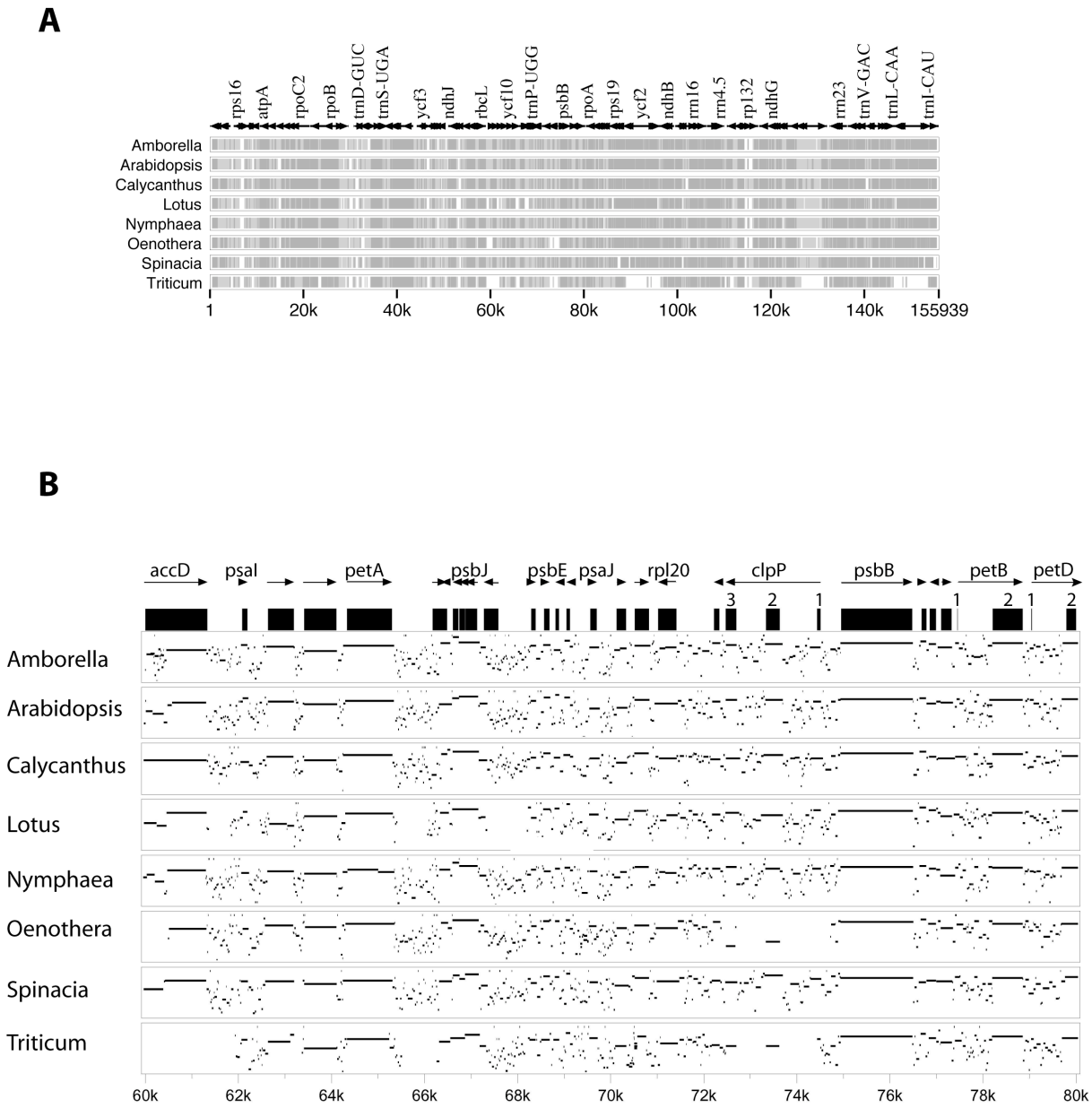
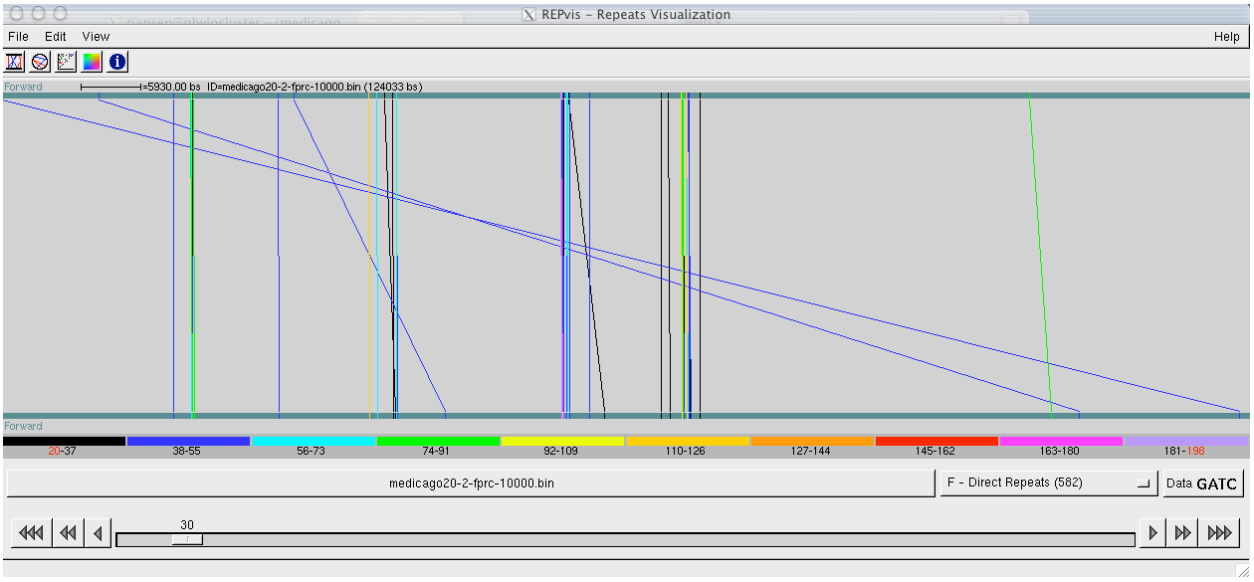

A
rps16
atpA
rpoC2
rpoB
trnD-GUC
trnS-UGA
ycf3
ndhJ
rbcL
ycf10
trnP-UGG
psbB
rpoA
rps19
ycf2
ndhB
rrn16
rrn4.5
rp132
ndhG
rrn23
trnV-GAC
trnL-CAA
trnI-CAU
Amborella
Arabidopsis
Calycanthus
Lotus
Nymphaea
Oenothera
Spinacia
Triticum
1
20k
40k
60k
80k
100k
120k
140k
155939
B
accD
psaI
petA
psbJ
psbE
psaJ
rpl20
clpP
psbB
petB
petD
3
2
1
1
2
1
2
Amborella
Arabidopsis
Calycanthus
Lotus
Nymphaea
Oenothera
Spinacia
Triticum
60k
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k

Figure 7

A.


REPvis - Repeats Visualization
File Edit View
Help
Forward
=5930.00 bs ID=medicago20-2-fprc-10000.bin (124033 bs)
Forward
20-37
38-55
56-73
74-91
92-109
110-126
127-144
145-162
163-180
181-198
medicago20-2-fprc-10000.bin
F - Direct Repeats (582)
Data GATC
30

B.

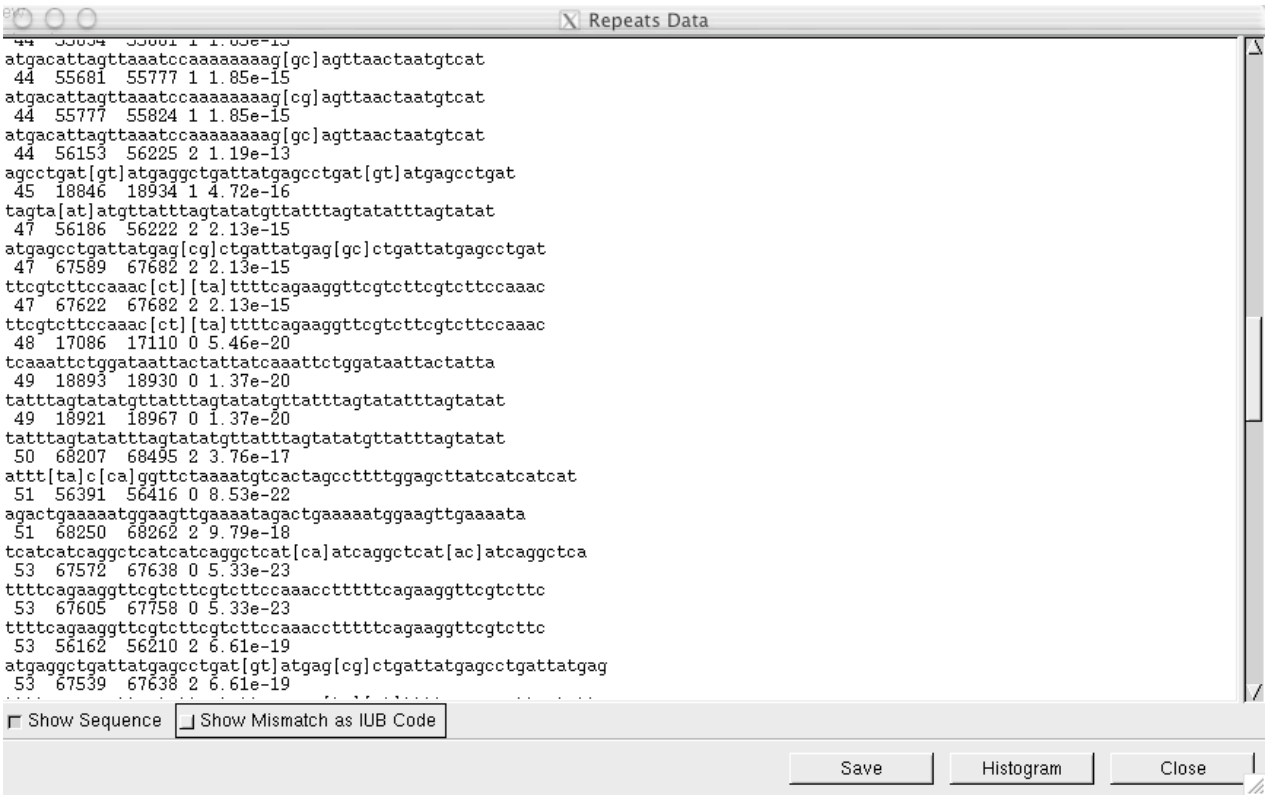
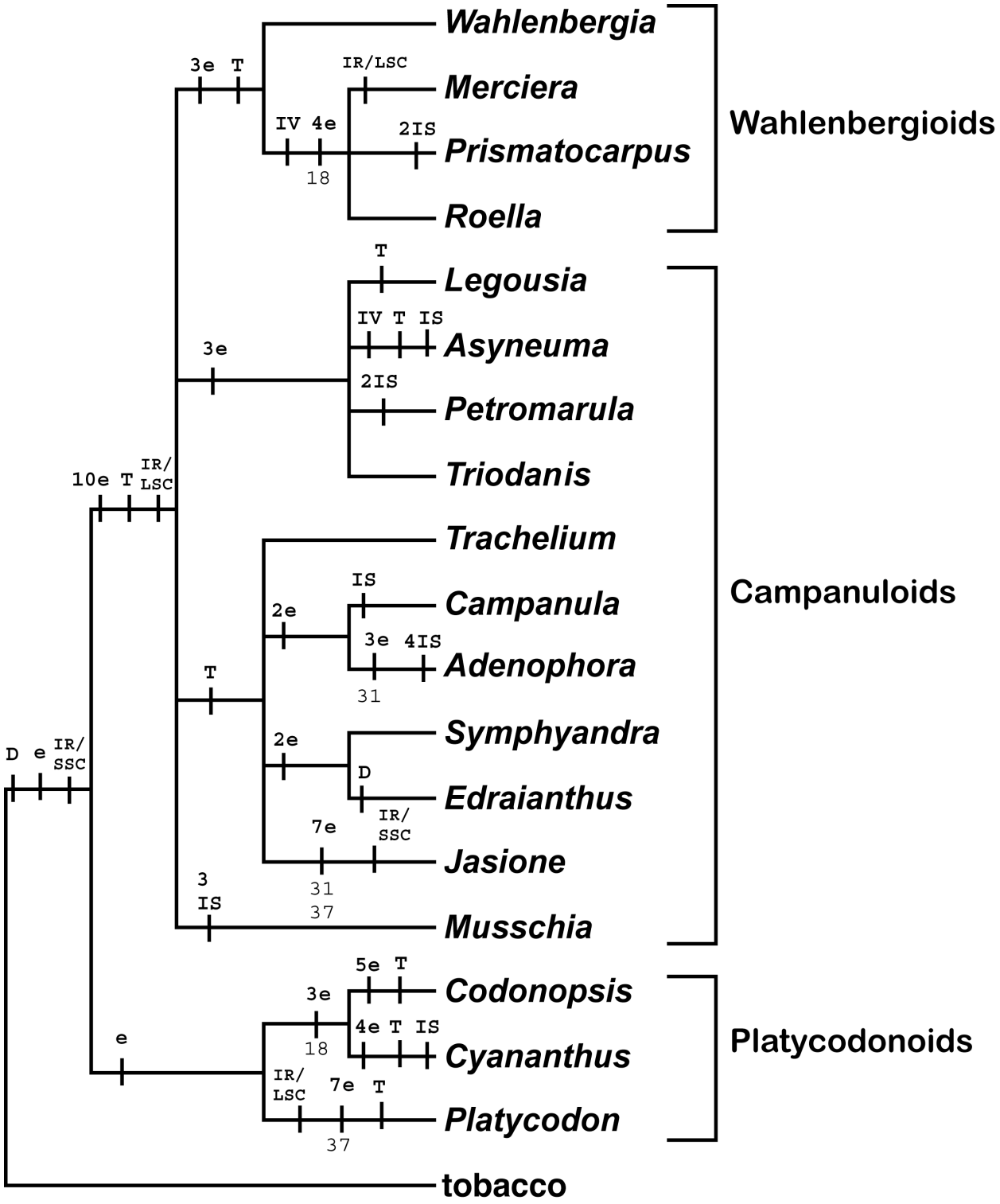

Repeats Data
atgacattagttaaatccaaaaaaaag[gc]agttaactaatgtcat
44 55681 55777 1 1.85e-15
atgacattagttaaatccaaaaaaaag[cg]agttaactaatgtcat
44 55777 55824 1 1.85e-15
atgacattagttaaatccaaaaaaaag[gc]agttaactaatgtcat
44 56153 56225 2 1.19e-13
agcctgat[gt]atgaggctgattatgagcctgat[gt]atgagcctgat
45 18846 18934 1 4.72e-16
tagta[at]atgttatttagtatatgttatttagtatatttagtatat
47 56186 56222 2 2.13e-15
atgagcctgattatgag[cg]ctgattatgag[gc]ctgattatgagcctgat
47 67589 67682 2 2.13e-15
ttcgtcttccaaac[ct][ta]ttttcagaaggttcgtcttcgtcttccaaac
47 67622 67682 2 2.13e-15
ttcgtcttccaaac[ct][ta]ttttcagaaggttcgtcttcgtcttccaaac
48 17086 17110 0 5.46e-20
tcaaattctggataattactattatcaaattctggataattactatta
49 18893 18930 0 1.37e-20
tatttagtatatgttatttagtatatgttatttagtatatttagtatat
49 18921 18967 0 1.37e-20
tatttagtatatttagtatatgttatttagtatatgttatttagtatat
50 68207 68495 2 3.76e-17
attt[ta]c[ca]ggttctaaaatgtcactagccttttggagcttatcatcatcat
51 56391 56416 0 8.53e-22
agactgaaaaatggaagttgaaaatagactgaaaaatggaagttgaaaata
51 68250 68262 2 9.79e-18
tcatcatcaggctcatcatcaggctcat[ca]atcaggctcat[ac]atcaggctca
53 67572 67638 0 5.33e-23
ttttcagaaggttcgtcttcgtcttccaaacctttttcagaaggttcgtcttc
53 67605 67758 0 5.33e-23
ttttcagaaggttcgtcttcgtcttccaaacctttttcagaaggttcgtcttc
53 56162 56210 2 6.61e-19
atgaggctgattatgagcctgat[gt]atgag[cg]ctgattatgagcctgattatgag
53 67539 67638 2 6.61e-19
Show Sequence
Show Mismatch as IUB Code
Save
Histogram
Close

Figure 8


Wahlenbergia
Merciera
Prismatocarpus
Roella
Wahlenbergioids
Legousia
Asyneuma
Petromarula
Triodanis
Trachelium
Campanula
Adenophora
Symphyandra
Edraianthus
Jasione
Musschia
Campanuloids
Codonopsis
Cyananthus
Platycodon
Platycodonoids
tobacco
3e T
IR/LSC
IV 4e
2IS
18
T
IV T IS
3e
2IS
10e T IR/LSC
IS
2e
3e 4IS
31
T
2e
D
7e
IR/SSC
31
37
3 IS
D e IR/SSC
5e T
3e
4e T IS
18
e
IR/LSC 7e T
37