8-2007

Chloroplast Comparative Genomics: Implications For Phylogeny, Evolution, and Biotechnology

Christopher Saski
Clemson University, csaski@clemson.edu

CHLOROPLAST COMPARATIVE GENOMICS: IMPLICATIONS FOR PHYLOGENY, EVOLUTION AND BIOTECHNOLOGY

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Genetics

---

by
Christopher Alan Saski
August 2007

Accepted by:
Jeffrey P. Tomkins, Committee Chair
Dr. Hong Luo
Dr. William R. Marcotte Jr.
Dr. Kerry Smith

ABSTRACT

Lack of complete chloroplast genome sequences is still a limiting factor determining phylogenetic relationships, discerning evolutionary forces, and extending chloroplast genetic engineering to useful crops. Therefore, the chloroplast genomes from six economically important crops were isolated and sequenced. The results will have an impact on chloroplast biology and biotechnology.

The complete soybean chloroplast genome was compared to the other completely sequenced legumes, *Lotus* and *Medicago*. The *rpl22* gene was found to be missing from all three legumes, a very informative phylogenetic marker. There is a single, large inversion changing the gene order in the legumes from the typical order found in *Arabidopsis*. Detailed analysis of repeat elements within the chloroplast genomes analyzed indicate they may play some functional role in evolution, and that the *psbA* and *rbcL* repeats indicate that the loss of an inverted repeat has only occurred once during the evolutionary history of the legumes. Ideal sites for integration of transgenes were also determined.

Next, the chloroplast genomes of the agriculturally important solanacaeae crops *Solanum lycopersicum* and potato were isolated and sequenced. Analysis of the complete chloroplast genome sequences revealed significant insertions and deletions (indels) within certain coding regions. Photosynthesis, RNA, and atp synthase genes are the least divergent and the most divergent genes are *clpP, cemA, ccsA,* and *matK*. The identified repeats characterized across the solanaceae are similar to the legumes, located in the same genes or intergenic regions indicating a possible functional role. A comprehensive genome-wide analysis of all coding sequences and intergenic spacer regions was done for the first time in

chloroplast genomes. Analysis of RNA editing sites demonstrated they were less common than what was previously observed in tobacco and *Atropa*, suggesting a loss of editing sites and a possible increase in variation at the RNA level.

Finally, the complete chloroplast genome sequences of barley, sorghum, and creeping bentgrass, were identified and compared to six published grass chloroplast genomes to reveal that gene content and order are similar, but two microstructural changes have occurred. First, the expansion of the inverted repeat at the small single copy/inverted repeat boundary that duplicates a portion of the 5' end of *ndhH* is restricted to three genera of the subfamily Pooideae (*Agrostis*, *Hordeum*, and *Triticum*). Second, a 6bp deletion in *ndhK* is shared by creeping bentgrass, barley, rice, and wheat, and this event supports the sister relationship between the subfamilies Erhartoideae and Pooideae. Repeat analysis revealed many dispersed repeats shared among the grasses, as well as repeats that flank a major genome rearrangement common only to the grasses suggesting this repeat had a functional role in the genome rearrangement. Examination of simple sequence repeat markers identified 16-21 potential SSRs. Distances based on intergenic spacer regions were analyzed as well as RNA editing sites. Phylogenetic trees based on DNA sequences of 61 protein-coding genes of 38 taxa using both maximum parsimony and likelihood methods provide moderate support for a sister relationship between the subfamilies Erhartoideae and Pooideae.

DEDICATION

I dedicate this manuscript to my wife and parents for all their love, support, inspiration, and dedication.

ACKNOWLEDGMENTS

I would like to Acknowledge Dr. Jeff Tomkins as my advisor. I acknowledge Dr. Henry Daniell and Dr. Robert Jansen for insightful discussions, motivation, data interpretation, and for assisting with scope and direction in this study. I would also like to acknowledge my graduate committee; Dr. William R. Marcotte Jr, Dr. Hong Luo, and Dr. Kerry Smith.

TABLE OF CONTENTS

Table of Contents (Continued) Page

Table of Contents (Continued) Page

LIST OF TABLES

List of Tables (Continued) Page

LIST OF FIGURES

List of Figures (Continued) Page

CHAPTER 1

INTRODUCTION

If there is one feature that distinguishes plant from animal life on our planet, it is not plants being primarily sessile, as a few animals share this trait, rather, it is the reliance of plants on solar energy to generate molecules with energy-rich bonds, the fuel that will be used by almost the entire biosphere (including plants themselves) to build other organized molecules and drive the rest of the processes that we know as life (Lopez-Juez and Pyke 2005). Chloroplasts are the sites of this wonderful process.

Endosymbiosis

Questions concerning the evolution of organelles have been a key force driving studies of organelle molecular biology (Daniell et al., 2004b). It is now widely accepted that the first plastids, derivatives of chloroplasts, arose from an endosymbiotic event between a photosynthetic bacterium (cyanobacteria) and a non-photosynthetic host (Howe et al., 1992). The green lineage among the descendants of this first photosynthetic eukaryote (there was a separate red lineage), eventually colonized the planet outside the oceans, around 450 million years ago (Willis et al., 2002, Lopez-Juez and Pyke 2005). The engulfed cyanobacteria turned into what we know as the chloroplast. Chloroplasts retained a small degree of their genetic autonomy, a large degree of their biochemistry, but lost some of their original functions and also acquired ones they did not possess when free-living (Timmis et al., 2004, Lopez-Juez and Pyke 2005). They needed to synthesize and accumulate their proteins, within themselves and in their surrounding cytoplasm, locate them to their correct destination, divide and propagate (Lopez-Juez and Pyke 2005). The chloroplast's ability to carry out

photosynthesis would determine the land plant's development and its need to adapt such development to environmental signals, such as light or the availability of raw materials (Lopez-Juez and Pyke 2005). The chloroplasts would also diversify into a variety of derivatives (Fig 1.1), that we now call other plastid types, to carry out other essential or specialized functions in other cells that were no longer photosynthetic, or merely to be transmitted more easily and economically in young, embryonic or undifferentiated cells (Waters et al., 2004).

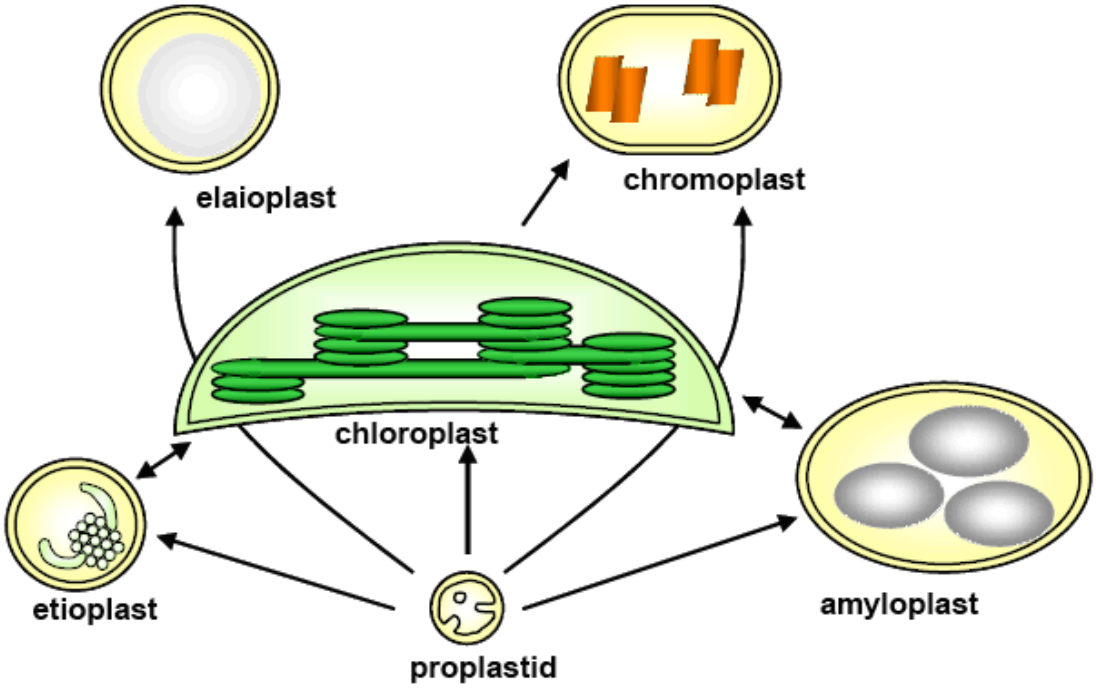

Fig. 1.1 Diversity of plastid types and their interconversions. Chloroplasts occupy the center of the figure to signify their evolutionary role as ancestors of all other plastid types (taken from Lopes-Juez and Pyke 2005)

Elaioplasts specialize in the storage of lipids. Chromoplasts are responsible for pigment synthesis and storage. Amyloplasts store starch through the polymerization of glucose. Etioplasts are chloroplasts that have not been exposed to light and are usually found in plants grown in the dark. If a plant is kept out of light for several days, its normal chloroplasts will actually convert into etioplasts. Proplastids are the progenitor of all plastid types. The chloroplasts or their derivatives therefore came under the control of developmental signals that affected the cells harboring them, or become influenced by the same environmental cues, to insure their function remained possible under a variety of conditions (Rodermel 2001, Lopez-Juez and Pyke 2005). Molecular research over the past three decades have revealed many prokaryotic features in the modern-day plant organelles, including some aspects of organelle division, genome organization and coding content, transcription, translation, RNA processing, and protein turn-over (Gray 2004). The confirmation of the basic endosymbiosis hypothesis (has raised many questions as to how evolution has shaped the modern day chloroplasts. It is still under debate as to whether there was a single (monophyletic) or multiple (paraphyletic) origin event for the plastid genome (Palmer 2003, Gray 2004). Complete chloroplast genome sequences from diverse taxa will aid in resolving this debate and provide additional support for the relationships among the land plants.

Chloroplasts and Other Plastid Types

Chloroplasts are the most noticeable feature of green cells in leaves and, excluding the vacuole, probably constitute the largest percentage of space within mesophyll cells (Lopez-Juez and Pyke 2005). Plastids are multifunctional and are used by the plant for critical biochemical processes other than photosynthesis, including starch synthesis, nitrogen

metabolism, sulfate reduction, fatty acid synthesis, DNA, and RNA synthesis (Zeltz et al. 1993). Each particular type of plastid carries identical plastid DNA (ptDNA) copies, which are attached to membranes (Kobayashi et al., 2002, Sato et al., 1993, Sato et al., 2001, Maliga 2004) in clusters called plastid nucleoids (Kuroiwa 1991, Maliga 2004). The number of plastids and ptDNA is highly variable depending on the cell type (Bendich 1987, Maliga 2004). In tobacco, the meristematic cells contain 10-14 proplastids, each containing 1-2 nucleoids per organelle, whereas leaf cells may contain 100 chloroplasts, with 10-14 nucleoids each, giving as much as 10,000 copies of the ptDNA per cell (Bendich 1987, Maliga 2004). The chloroplast genome generally has a highly conserved organization (Palmer 1991, Raubeson et al., 2005) with most land plant genomes composed of a single circular chromosome with a quadripartite structure that includes two copies of an inverted repeat (IR) that separate the large and small single copy regions (LSC and SSC) (Fig 1.2). The size of this circular genome varies from 35 to 217 kb but, the majority of plastid genomes from photosynthetic organisms are between 115-165 kb (Jansen et al. 2005). Compared to the nuclear and mitochondrial genomes, the plastid genome is quite conserved across taxa (Maier et al., 2004). However, due to comparisons of whole chloroplast genome sequence, differences in the general architecture (tobacco and *Arabidopsis*) have been reported (Hiratsuka et al., 1989, Doyle et al. 1992, Palmer and Stein 1986) and can mainly be attributed to evolutionary expansion/contraction or loss of the inverted repeat, genome rearrangements, dispersed repeats, and indels (Hiratsuka et al. 1989, Doyle et al. 1992, Palmer and Stein 1986, Maier et al., 2004). Since the inverted repeat is present in several algae, it seems likely that it is an ancient feature which has been later lost in individual branches during evolution (Palmer 1991). Characteristically, the IR-region contains a

complete rRNA operon. Duplicated rRNA operons are also observed in cyanobacterial genomes which argues for a selective pressure to increase rRNA gene number (Palmer 1991). Speculatively, the IR-organization may play a direct role in maintaining the conserved structure of the chloroplast chromosome and also in directly conserving genes encoded by the IR, as these genes characteristically have lower rates of nucleotide substitutions than those encoded in single copy regions (Curtis et al., 1984, Wolfe et al., 1987).

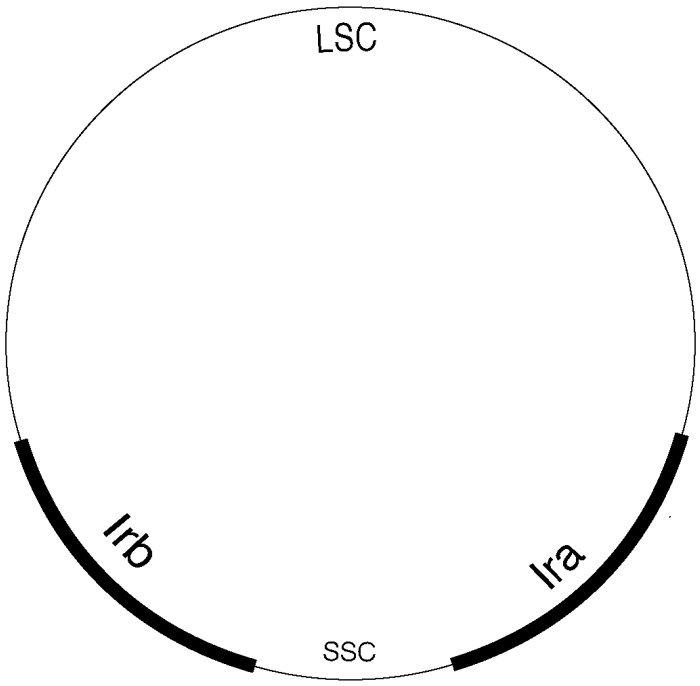

Fig. 1.2 Typical organization of a plastid chromosome in its circular monomeric form. Large and small single copy regions (LSC, SSC) are separated by the inverted repeats Ira and IRb (Jansen et al., 2005).

Gene Transfer

It has been noted that cyanobacterial genes for processes no longer needed inside the host are not found in present-day plant cells (e.g., motility-related genes) (Maier et al., 2004). The plastid genome is small (100-200 genes) when compared to the typical cyanobacterium composed of 3,000-4,000 genes (Maier et al., 2004). At first glance, it seems as if many of the cyanobacterial genes have been discarded. It became apparent that the plastid's proteome, despite its tiny genome, contained 1,000 to 5,000 proteins of comparable size to a cyanobacterial proteome (Martin et al., 1998, Rujan et al., 2001). Detailed analysis of homologies between modern plastid and nuclear genomes revealed substantial amounts of plastid-derived DNA in the nucleus (Maier et al 2004). This has been observed in Spinach (Timmis et al., 1983; Cheung et al., 1989), various chenopod species (Ayliffe et al., 1988), potato (du Jardin 1990), tomato (Pichersky et al., 1991), tobacco, (Ayliffe et al., 1992), rice, and *Arabidopsis* (Shahmuradov et al., 2003). These findings have set the stage to further study gene transfer to the nucleus. This information can provide invaluable phylogenetic markers such as the *rpl22* loss to the nucleus in the legumes (Gantt et al., 1991) that was discovered by chloroplast comparative genomics utilizing whole genome sequence.

Why do Plastids Have Genomes?

The chloroplast offers a particularly unfriendly environment for DNA. The chemistry of photosynthesis generates high concentrations of various oxygen species that are highly mutagenic (Allen et al., 1996). Whatever the selective pressures are that have reduced the plastid genome to its current size are unknown. The question still open is why this was not driven to completion. There are several hypothesis to address this question. First, it has been argued that several of the organelle encoded proteins are highly hydrophobic and hence

would not easily cross the plastid envelope when translated in the cytoplasm (von Heijne 1986; Palmer 1997). A previous described argument suggests the highly hydrophobic light-harvesting chlorophyll proteins are universally nuclear-encoded and the hydrophilic large subunit (*rbcL)* of RuBisCO, with few exceptions, is plastid-encoded (Maier et al., 2004). Additionally, other explanations for the maintenance of the plastid chromosome are that plastid proteins could be toxic in the cytosol (Martin et al., 1998). It has also been proposed that as gene transfer is an ongoing process, the last remnants of the plastid chromosome will eventually disappear over time (Herrmann 1997). The genes that appear to have remained are categorized as; rubisco subunit, photosystem proteins, cytochrome-related, ATP synthase, NADH dehydrogenase, ribosomal protein subunits, ribosomal RNAs, plastid-encoded RNA polymerase, and open reading frames with unknown function.

Phylogenetic Utility of Chloroplast Genomes

Most previous molecular phylogenetic studies of flowering plants have relied on one to several genes from the chloroplast, mitochondria, and/or nuclear genomes, though most of these analyses were based on chloroplast markers (RFLP and SSR) (Jansen et al., 2006). During the past few years there has been a rapid increase in the number of studies using complete genes and intergenic regions from completely sequenced chloroplast genomes for estimating phylogenetic relationships among angiosperms (Goremykin et al., 2003a, b, 2004, 2005, Leebens-Mack et al., 2005, Chang et al., 2006, Lee et al., 2006a, Jansen et al., 2006, Ruhlman et al., 2006, Bausher et al., 2006, Cai et al., 2006). These studies have resolved a number of issues regarding relationships among the major clades, including the identification of either *Amborella* alone or *Amborella* + *Nymphaeales* as the sister group to all other angiosperms, these studies also lend strong support for the monophyly of magnoliids,

monocots, and eudicots, the position of magnoliids as sister to a clade that includes both monocots and eudicots, the placement of *Vitaceae* as the earliest diverging lineage of rosids, and the sister group relationship between Caryophyllales and Asterids. However, some issues remain unresolved, including the monophyly of the eurosid I clade and relationships among the major clades of rosids (Jansen et al., 2006; Soltis et al., 2005). Completely sequenced chloroplast genomes provide a rich source of data that can be used to address phylogenetic questions at deep nodes in the angiosperm tree (Jansen et al., 2006; Goremykin et al., 2003a, b, 2004, 2005, Leebens-Mack et al., 2005, Chang et al. 2006, Lee et al., 2006a, Bausher et al., 2006, Cai et al., 2006). The use of DNA sequences from all of the shared chloroplast genes provides many more characters for phylogeny reconstruction compared to previous studies that have relied on only one or a few genes to address the same questions (Jansen et al., 2006). However, the whole genome approach can result in misleading estimates of relationships because of limited taxon sampling (Jansen et al., 2006, Leebens-Mack et al., 2005, Soltis et al., 2004, Stefanovic et al., 2004, Martin et al., 2005) and the use of incorrect models of sequence evolution in concatenated datasets (Jansen et al., 2006; Goremykin et al., 2005, Lockhart et al., 2005). Thus, there is a growing interest in expanding the taxon sampling of complete chloroplast genome sequences and developing new evolutionary models for phylogenetic analysis of chloroplast sequences (Jansen et al., 2006) to overcome these concerns. To date, there are more than 200 chloroplast genome sequences available; however only 26 are surprisingly from crop species. Table 1.1 includes a comprehensive list of crop chloroplast genomes sequenced and references.

Table 1.1 A list of crop chloroplast species completed to date.

Species	Reference	Accession number	Year completed
Citrus sinensis	Bausher et al., (2006)	NC_008334	2006
cucumis sativus	Unpublished	NC_007144	2005
Eucalyptus globules	Steane (2005)	AY780259	2005
Gossypium hirsutum	Lee et al., (2006)	DQ345959	2006
Helianthus annus	Timme et al., (2006)	DQ383815	2006
Lactuca sativa	Unpublished	NC_007578	2006
Medicago truncatula	Unpublished	AC093544	2001
Nicotiana tabacum	Shinozaki et al., (1986)	Z00044	1986
Oryza nivara	Masood et al., (2004)	NC_005973	2004
Oryza sativa	Hiratsuka et al., (1989)	NC_001320	1989
Panax schinseng	Kim and Lee (2004)	NC_006290	2004
Pinus thumbergii	Wakasugi et al., (1994)	NC_001631	1994
Populus trichocarpa	Unpublished	NC_008235	2003
Saccharum hybrid	Unpublished	NC_005878	2004
Saccharum officinarum	Asano et al., (2004)	NC_006084	2004
Solanum tubersoum	Unpublished	DQ231562	2005
Spinacia oleracea	Schmitz-Linneweber et al., (2001)	NC_002202	2000
Triticum aestivum	Ogihara et al., (2000)	AB042240	2001
Vitis vinifera	Jansen et al., (2006)	NC_007957	2006
Zea mays	Maier et al., (1995)	NC_001666	1995

RNA editing in Chloroplast genomes

Research in RNA editing has entered its second decade and brought to light an unanticipated breadth of examples of the process among diverse lower and higher eukaryotes (Smith et al., 1997). RNA editing is a co- or post-transcriptional process that modifies the sequence of an RNA transcript through nucleotide insertion, deletion, or modification to make it different from the DNA that encoded the RNA (Smith et al., 1997). In virtually all cases, the initial characterization has come from a comparison of a cDNA to the genomic sequence (Smith et al., 1997). Several higher plant chloroplast genomes have been sequenced and analyzed for editing, and generally have about 30 C-to-U editing sites (Kugita et al., 2003a, Kugita et al., 2003b, Maier et al., 1995, Surgiura 1995). All of the editing sites described for chloroplasts from vascular plants are C-to-U editing sites, and no U-to-C (reverse) edits have been identified. The function of C-to-U RNA editing generally causes a radical change in the amino acid specified by a codon, and would be predicted to perturb the structure and function of a protein (Mulligan 2004) and in many cases results in the restoration of conserved amino acid residues (Kotera et al., 2005). Editing has also been suggested to be a potential regulator of various steps in gene expression (Mulligan 2004). Knowledge of RNA editing in chloroplast genomes is particularly important for the identification of transcription start and stop sites, intron splicing, and phylogenetic analysis. This information will also have direct impacts in developing methods to better understand the mechanism behind RNA editing as well as heterologous gene expression in the plastid genome.

Chloroplast Molecular Markers

Since the first report on chloroplast DNA variation based on restriction patterns (Vedel et al., 1976), there has been increasing interest in chloroplast genomic sequence for the purposes of population genetics and phylogenetic studies (McCauley 1995; Morand-Prieur 2002). The use of chloroplast DNA (cpDNA) restriction fragment length polymorphisms (RFLP) as genetic markers in interspecific hybridization showed that most angiosperm species display maternal inheritance of the chloroplast genome (Reboud et al., 1993, Morand-Prieur 2002). It has been recently noted that there is little intraspecific variation among angiosperm chloroplast DNA (Morand-Prieur 2002) and that the highest frequency of mutations is found in the noncoding regions (Palmer 1992). It has been recently discovered that chloroplast simple sequence repeats are highly useful markers for size variations that are easy to analyze by using PCR and polyacrylamide gel electrophoresis (Powell et al., 1995, Morand-Prieur 2002). The complete tobacco chloroplast genome sequence has been mined for simple sequence repeats that resulted in high levels of intra- and interspecific diversity among solanaceous species (Powell et al., 1995, Provan et al., 1999, Bryan et al., 1999) the presence of which indicates the necessity for whole genome chloroplast sequence to develop polymorphic markers to reveal diversity at the intra- and interspecific level.

Plastids and Biotechnology

Plastid transformation involves transforming one or a few chloroplast DNA copies, followed by gradually diluting plastids carrying nontransformed copies on a selective medium (Maliga 2004). The most common integration site in chloroplast transformation is the transcriptionally active intergenic spacer region between *trnI*/*trnA*. This region is located in the inverted repeat near one of the two origins of replication. The plastid transformation

approach has been shown to have a number of advantages, most notably with regard to its high transgene expression levels (De Cosa et al., 2001), capacity for multi-gene engineering in a single transformation event (De Cosa et al., 2001, Lossl et al., 2003, Ruiz et al., 2003, Quesada-Vargas et al., 2005), and ability to accomplish transgene containment via maternal inheritance (Daniell 2002). Moreover, chloroplasts appear to be an ideal compartment for the accumulation of certain proteins, or their biosynthetic products, which would be harmful if accumulated in the cytoplasm (Daniell et al., 2001, Lee et al., 2003, Leelavathi et al., 2003, Ruiz et al., 2005). In addition, gene silencing has not been observed in association with this technique, whether at the transcriptional or translational level (DeCosa et al., 2001, Lee et al. 2003, Dhingra et al., 2004). Because of these advantages, the chloroplast genome has been engineered to confer several useful agronomic traits, including herbicide resistance (Daniell et al., 1998), insect resistance (McBride et al., 1995, Kota et al., 1999), disease resistance (DeGray et al., 2001), drought tolerance (Lee et al., 2003), salt tolerance (Kumar et al., 2004a), and phytoremediation (Ruiz et al., 2003). The chloroplast genome has also been utilized in the field of molecular pharming, for the expression of biomaterials, human therapeutic proteins, and vaccines for use in humans or other animals (Guda et al., 2000, Staub et al., 2000, Fernandez-San Milan et al., 2003, Leelavathi et al., 2003, Molina et al., 2004, Viitanen et al., 2004, Watson et al., 2004, Koya et al., 2005, Grevich et al., 2005, Daniell et al., 2005b, Kamarajugadda et al., 2006). Lack of complete chloroplast genome sequences is still one of the major limitations to extend this technology to useful crops. Chloroplast genome sequences are necessary for identification of spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes (Maier et al., 2004,

Daniell et al., 2005b). In land plants, about 40-50% of each chloroplast genome contains non-coding spacer and regulatory regions (Jansen et al., 2005). Identity between vector sequences and target sequence is necessary (DeCosa et al., 2001, Daniell et al., 2004b, Daniell et al., 2005b, Dhingra et al., 2004, Lee et al., 2006b), as transformation vectors with homologous sequence from another species have not yielded high frequency transformations so far even in tobacco, in which plastid transformation is highly efficient (Daniell et al., 2004b, Degray et al., 2001). Therefore, further genome sequencing projects of crop plant plastid chromosomes is one of the more pressing needs in this field to identify intergenic sequences as well as endogenous regulatory elements (Daniell et al., 2004b).

Our knowledge of the organization and evolution of chloroplast genomes has been expanding rapidly because of the large numbers of completely sequenced genomes published in the past decade. The use of information from whole chloroplast genome sequence has added to our understanding of chloroplast biology, the origins and relationships of land plants, and allowed development of useful traits to aid in worldwide needs. Many crop nuclear genomes have been mapped and/or partially sequenced, but there is limited or no information about their chloroplast genomes. The described studies were undertaken to characterize the complete chloroplast genomes of *Glycine max* (soybean), *Solanum lycopersicum* (tomato), *Solanum bulbocastanum* (potato), *Hordeum vulgare* (barley), *Sorghum bicolor* (sorghum), and *Agrostis stolonifera* (creeping bentgrass). The resulting information will give insight into molecular and evolutionary processes, relationships among plant taxa, and optimal sites for plastid transformation. The results obtained will also be the foundation for many future studies that will have direct impacts on our agriculture economy, national security, and planet overall

CHAPTER 2

THE COMPLETE CHLOROPLAST GENOME SEQUENCE OF *GLYCINE MAX* AND COMPARATIVE ANALYSIS WITH OTHER LEGUME GENOMES

Introduction

Glycine max (soybean) is a leguminous crop and is considered the most important source of vegetable protein. It is widely used as animal feed and for human consumption. The dry matter of soybeans contains about 20% oil and 35–40% protein. It is also the most widely planted genetically modified crop in the world, representing more than half of the soybean cultivated area worldwide (GMO Compass http://www.gmo-compass.org/eng/grocery_shopping/crops/19.genetically_modified_soybean.html). This includes glyphosate-tolerant cultivars, a trait that has been engineered via the nuclear genome but would offer better transgene containment if engineered via the chloroplast genome because the plastid genome of soybean is inherited maternally (Corriveau and Coleman, 1988). The primary goal of this study is to compare the chloroplast genome organization of *Glycine* with the two other completely sequenced legume chloroplast genomes (*Lotus japonicus* and *Medicago truncatula*) and with the model dicot, *Arabidopsis thaliana*. In addition to examining gene content and gene order, the distribution and location of repeated chloroplast sequences among legumes and *Arabidopsis* will be analyzed and assessed for their possible role in evolution of the chloroplast genome. Genetic markers will be mined for to assist plant geneticists. Intergenic spacer and regulatory sequences will be evaluated for use in future studies in chloroplast genetic engineering.

Methodology

DNA Sources

The large-insert genomic library of *Glycine max*, PI 437654, was constructed by ligating size fractionated partial *Hin*dIII digests of total nuclear DNA with the pINDIGOBAC-536 vector (Tomkins et al., 1999, Luo et al., 2001). The average insert size of the library was 136 kb. BAC clones containing the chloroplast genome inserts were isolated by screening the library with a barley chloroplast probe (Tomkins et al., 1999). The first 96 positive clones from screening were pulled from the library, arrayed in a 96-well microtitre plate, copied, and archived. Clones were then subjected to *Hin*dIII fingerprinting and high resolution agarose gels to verify relatedness. *Not*I digests and CHEF gels were used to determine average insert size. BAC-end sequences were determined and localized on the chloroplast genome of *Arabidopsis thaliana* to deduce the relative positions of the candidate clones, then one BAC clone that covered the entire chloroplast genome was chosen for the subsequent sequencing analysis.

DNA Sequencing and Data Assembly

The nucleotide sequence of the BAC clone was determined by the bridging shotgun method (Kaneko et al., 1995). The purified BAC DNA was subjected to hydroshearing, end repair, and then size-fractionated by agarose gel electrophoresis. Fractions of approximately 3.0–5.0 kb were eluted and ligated into the vector pBLUESCRIPT IIKS+. The libraries were plated and arrayed into 40 96-well microtitre plates, respectively, for sequencing reactions. Sequencing was performed using the Dye-terminator cycle sequencing kit (Perkin Elmer Applied Biosystems, USA). Sequence data from the forward and reverse priming sites of the shotgun clones were accumulated, equivalent to 8 times the size of the genome, roughly 150-152 kilobase pairs (Spielmann et al., 1988), and assembled using Phred-Phrap programs (Ewing and Green, 1998).

Genome Annotation

Annotation of the *Glycine* chloroplast genome was performed using DOGMA (Dual Organellar GenoMe Annotator, Wyman et al., 2004; http://evogen.jgi-psf.org/dogma). This program uses a FASTA-formatted input file of the complete genomic sequences and identifies putative protein-coding genes by performing BLASTX searches against a custom database of previously published chloroplast genomes. The user must select putative start and stop codons for each protein coding gene and intron and exon boundaries for intron-containing genes. Both tRNA and rRNA genes are identified by BLASTN searches against the same database of chloroplast genomes (Fig 2.1). The *Medicago* chloroplast genome sequence (NC_003119) has not been annotated so we also used DOGMA to annotate this genome.

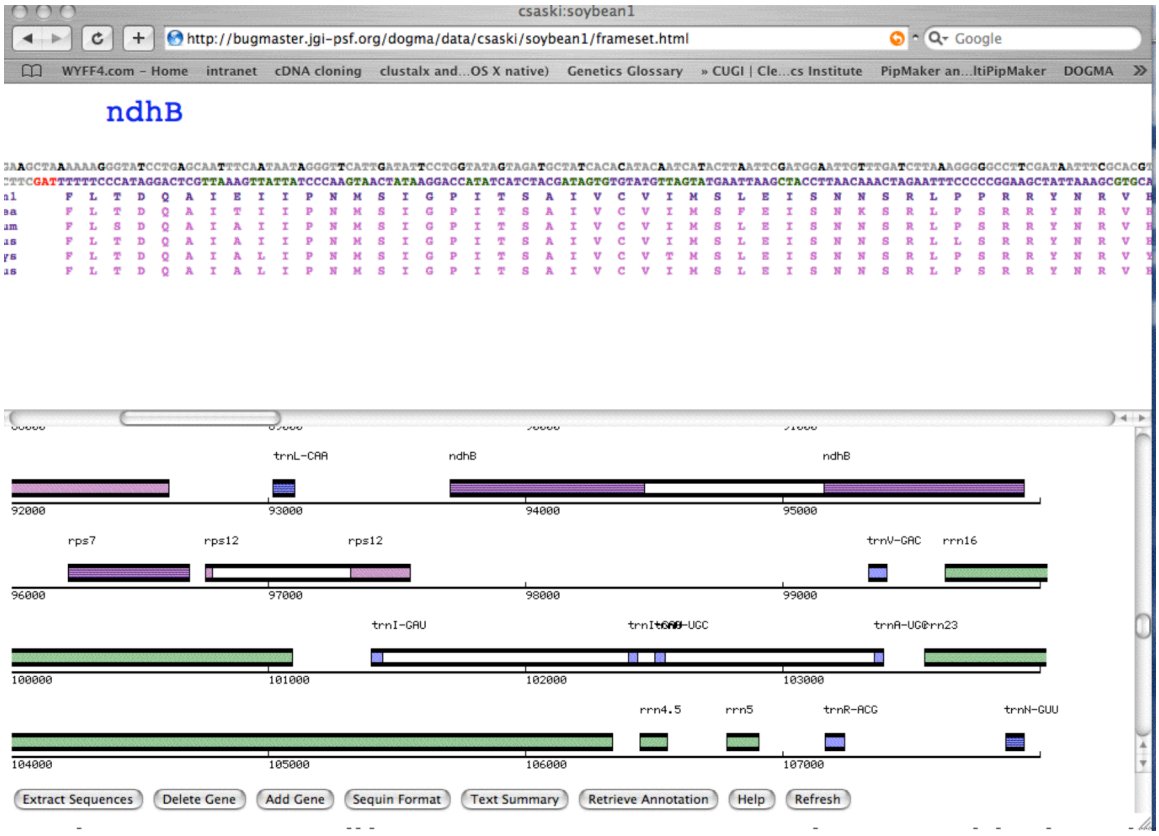

Fig 2.1 DOGMA. Dual Organellar GenoMe Annotator. Automates the annotation of extranuclear organelles (Dual Organellar GenoMe Annotator, Wyman et al., 2004; http://evogen.jgi-psf.org/dogma)

Molecular Evolutionary Comparisons

Gene content comparisons were performed using Multipipmaker (Schwartz et al., 2003). Two sets of comparisons were performed, one including four genomes (*Arabidopsis* [AP000423], and the three legumes *Glycine* [NC007942], *Lotus* [NC002694], and *Medicago* [AC093544]) using *Nicotiana* [NC001879] as the reference genome and a second that only included the three legumes using *Lotus* as the reference genome. Gene orders were examined by pairwise comparisons between the *Arabidopsis*, *Glycine*, *Lotus*, and *Medicago* genomes using PipMaker (Elnitski et al., 2002).

Repeat structure in legume chloroplast genomes was examined in two stages. First, REPuter (Kurtz et al., 2001) was used to identify the number and location of direct and inverted (palindromic) repeats in the three legumes and *Arabidopsis* using a minimum repeat size of 30 bp and a Hamming distance of 3 (sequence identity of 90%). Second, BLAST searches of repeats identified for *Medicago* were subject to BLAST searches against the complete chloroplast genomes of the other two legume genomes (*Glycine* and *Lotus*) and *Arabidopsis*. Blast hits that were 20 bp and longer with a sequence identity of ≥ 90% were identified and extracted from these results to determine which of the repeats were shared among the four genomes examined. To detect simple sequence repeats (SSRs) a modified version of the Perl script SSRIT was used (Temnykh et al., 2001). The modified script, CUGISSR (Jung et al., 2005), was used to search for SSRs ranging from di-to penta-nucleotide repeats.

Results

Size, gene content and organization of the Glycine chloroplast genome

The complete chloroplast genome size of *Glycine* is 152,218 bp (Fig. 2.2.). The genome includes a pair of inverted repeats of 25,574 bp (IRa and IRb) of identical sequence separated by a small single copy region of 17,895 bp, and a large single copy region of 83,175 bp. The IR extends from rps19 through a portion of ycf1 (Fig. 2.2). The *Glycine* chloroplast genome contains 111 unique genes, and 19 of these are duplicated in the IR, giving a total of 130 genes (Fig. 2.2). There are 30 distinct tRNAs, and 7 of these are duplicated in the IR. Nineteen genes contain one or two introns, and six of these are in tRNAs. The genome consists of 60% coding regions (52% protein coding genes and 8% RNA genes) and 40% non-coding regions, including both intergenic spacers and introns. The overall GC and AT content of the *Glycine* chloroplast genome is 34% and 66%, respectively. The AT bias is higher in the non-coding regions with 70% AT versus 62% AT in the coding regions.

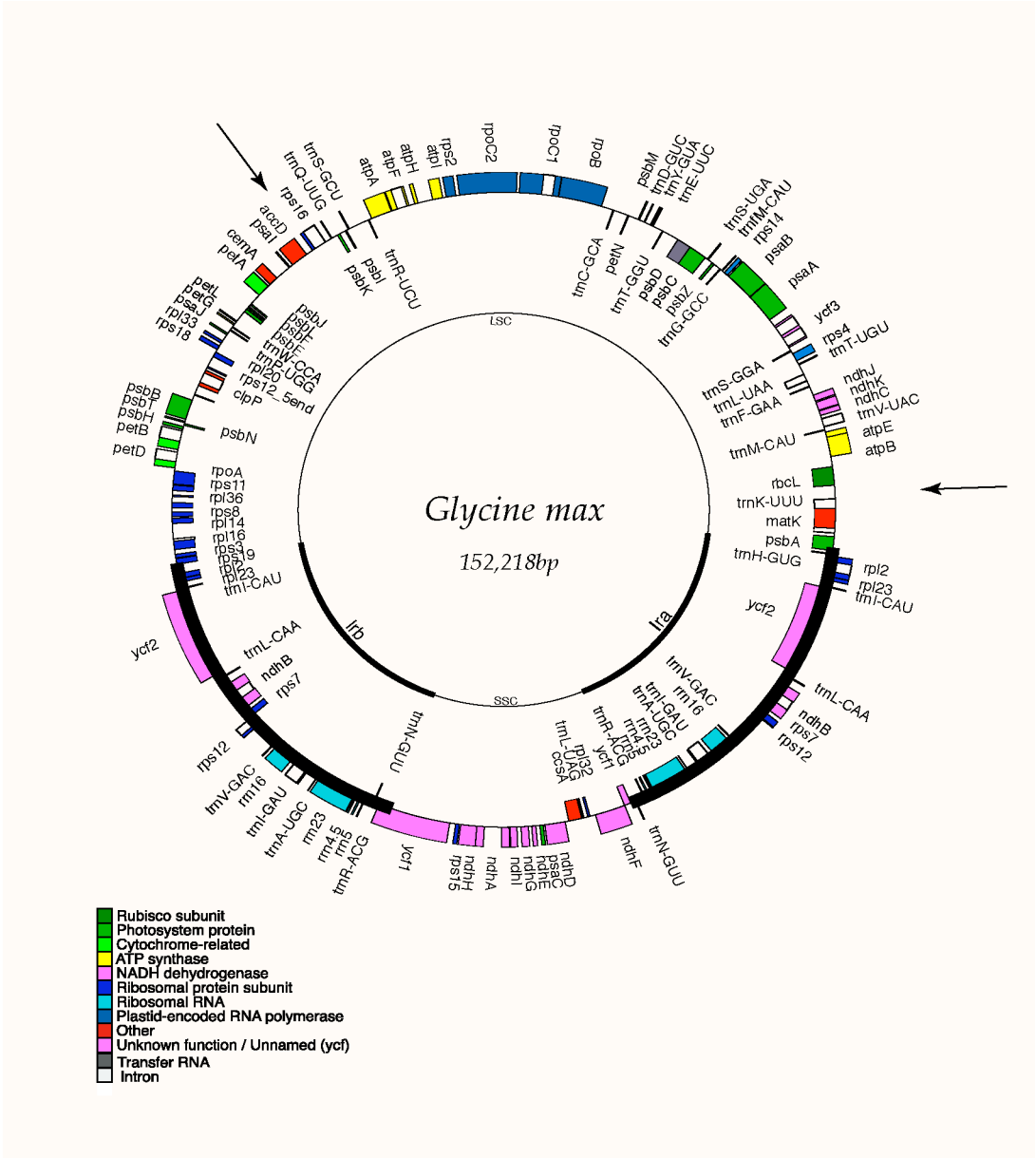

Fig 2.2 Gene map of *Glycine max* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb, 25,574 bp), which separate the genome into small (SSC, 17895 bp) and large (LSC, 83,175 bp) single regions. Genes on the outside of the map are genes transcribed in the clockwise direction and genes on the inside are transcribed counterclockwise. Arrows in bold indicate the 51 Kb inversion endpoints.

Comparison of genome organization among legumes and Arabidopsis

Gene content of the three sequenced legumes *Glycine*, *Lotus* [Kato et al., 2000; NC_002694] and *Medicago* [NC_003119] is nearly identical (Fig. 2.3A). *Medicago* does not have duplicate copies of the 19 genes in the IR because one copy of the IR has been lost (Palmer et al., 1987). A comparison of gene content between the three legumes and *Arabidopsis* shows that the *rpl22* gene is missing from all 3 legumes (see arrow 1 in Fig. 2.3A) and that *Medicago* is also missing *rps16* (see arrow 2 in Figs. 2.3 A-B).

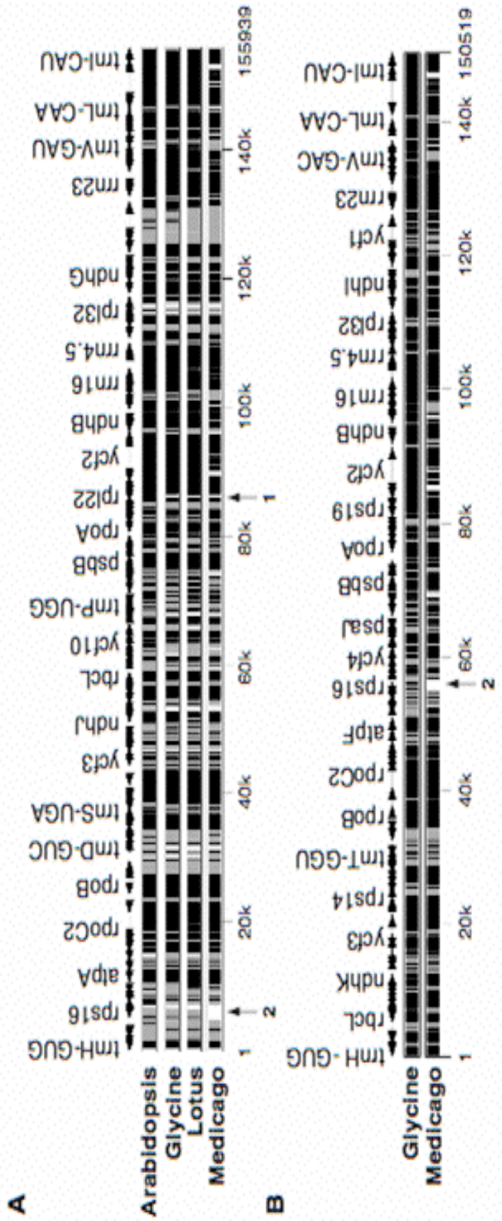

Fig 2.3 Multipipmaker alignments of legumes and *Arabidopsis* (A; using *Nicotiana* as reference genome) and legumes (B; using *Lotus* as a reference genome). Arrows indicate loss of *rpl22* (1) and *rps16* (ribosomal protein subunit) (2).

The gene order in *Glycine* differs from the gene order observed in the model dicot *Arabidopsis thaliana* by the presence of a single, large inversion of approximately 51 kb that reverses the order of the genes between rbcL and rps16 (see arrows in Fig. 2.2 also see Fig. 2.4). This same inversion is also present in *Lotus* and *Medicago* (Kato et al, 2000).

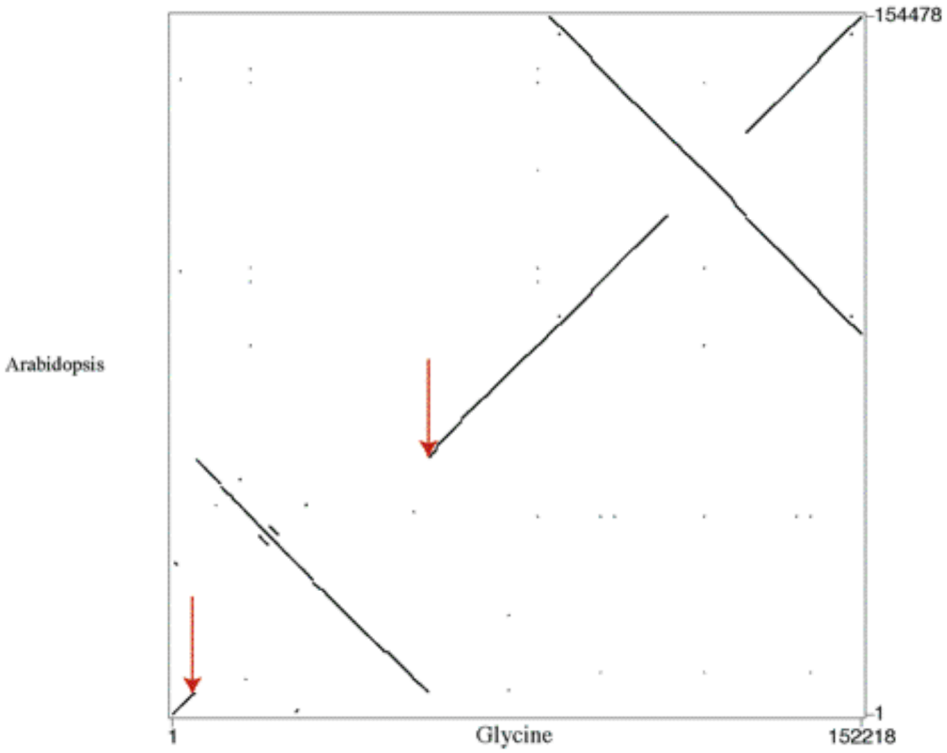

Fig 2.4 Pipmaker Dot plot illustrating the 51-Kb inversion in the legume chloroplast DNA when compared to the typical gene order of *Arabidopsis*. Arrows indicate 51 kb inversion endpoints.

Extent of the Inverted Repeat

The IR in *Glycine* is 25,574 bp long and includes 19 genes. At the IR/LSC junction the IR ends within the rps19 gene so that 68 bp of the 5' end of the gene is duplicated (Fig. 2.5). The IR/SSC junction is found within *ycf1* resulting in the duplication of 478 bp of the 5' end of this gene. Comparison of the IR region of the three completely sequenced legumes and *Arabidopsis* indicates that there is some contraction of the IR in the two legumes with an IR. At the IR/LSC boundary, the IR includes 68 and 1 bp of the *rps19* in *Glycine* and *Lotus*, respectively (Fig 2.5). Thus, the IR in both of these legumes has contracted relative to *Arabidopsis*, which has 113 bp of the 5' end of *rps19* duplicated (Fig 2.5). There has also been contraction of the IR in the legumes at the IR/SSC boundary relative to *Arabidopsis*. *Glycine* and *Lotus* have 478 bp and 514 bp of *ycf1* duplicated, whereas *Arabidopsis* has 1,027 bp duplicated in the IR. This contraction of the IR in these legumes accounts for the smaller size of their IR and larger size of the SSC (Fig. 2.5).

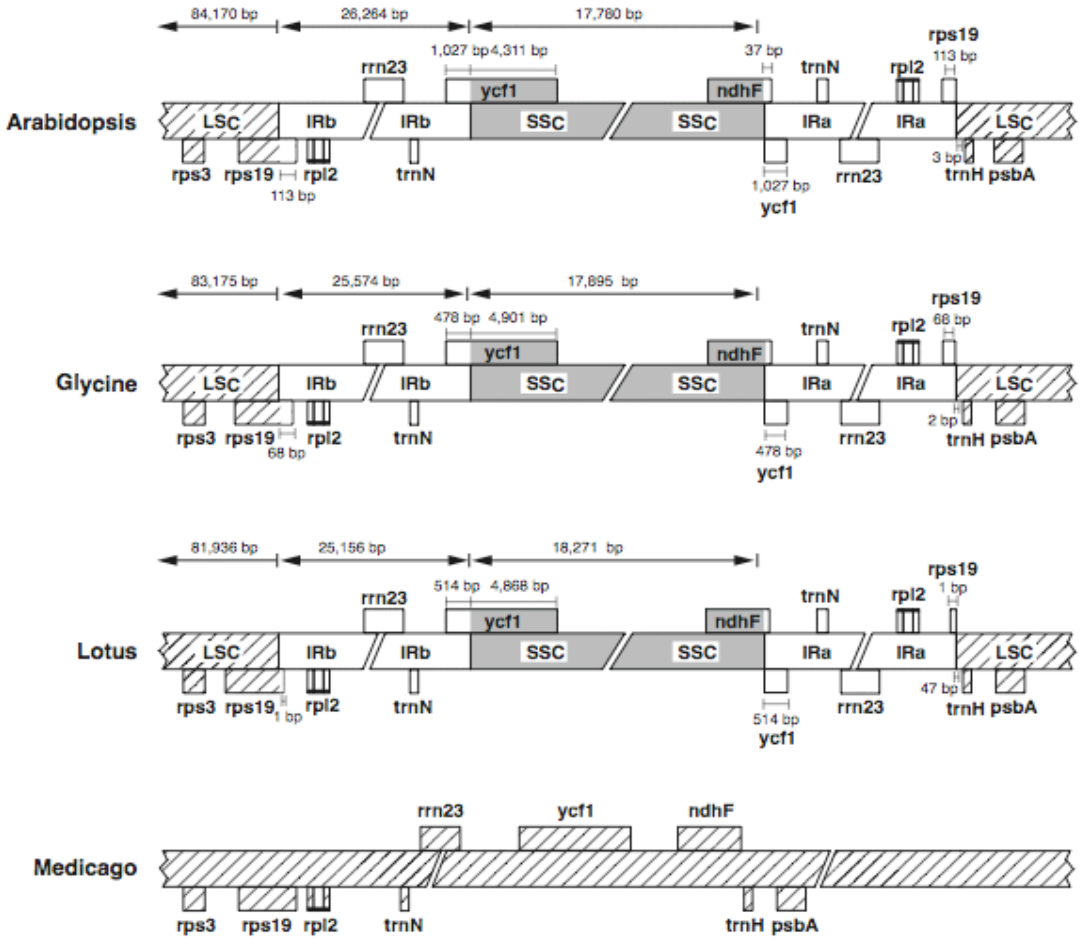

Fig. 2.5. Comparison of boundaries of IR, SSC, and LSC among the legume and *Arabidopsis* chloroplast genomes. IRa is missing in *Medicago*. Shaded regions indicate small single copy regions, cross-bars indicate large single copy region. *Medicago* is now considered all single copy.

In addition to the contraction of the IR boundary in legumes, IRa has been lost in *Medicago* (Fig. 2.5). This loss has resulted in *ndhF* (usually located in the SSC) being adjacent to *trnH* (usually the first gene in the LSC at the LSC/IRa junction). Loss of one copy of the IR in some legumes provides support for monophyly of six tribes (Palmer 1985, Wolfe 1988, Palmer et al., 1987b, Lavin et al., 1990). Wolfe (1988) identified duplicated sequences of portions of two genes, 40 bp of *psbA* and 64 bp of *rbcL*, in the region of the IR deletion between *trnH* and *ndhF* in the legume *Pisum sativum* and these duplications were later identified in another legume broad bean (*Vicia faba*, Herdenberger et al., 1990). Similar repeats in this region were found in other legumes without an IR, including two species of *Medicago* (Fig. 2.6). The *Medicago psbA* repeat has the same length of 40 bp and it has a high sequence identity with a segment of *psbA* at coordinates 446–485 in other legumes without the IR (Fig. 2.7A). The copies of the psbA repeat in Pisum and Vicia and in the two *Medicago* species have a 100% sequence identity with each other but the sequence identity between the *Pisum*/*Vicia* and *Medicago* repeats is 85% (Fig. 2.6). The sequence identity of this repeat compared to the complete, functional copy of *psbA* is 85% for *Pisum* and *Vicia* and 95% for the two *Medicago* species (Fig. 2.7A). The *rbcL* repeats are 39 bp long in the two *Medicago* species with a 95% sequence identity to each other (Fig. 2.6) and 90% sequence identity to coordinates 516 to 554 in the complete functional copy of *rbcL* (Fig. 2.7B). In *Vicia* and *Pisum* the *rbcL* repeat is 64 bp long with a 92% sequence identity to each other and 86–92% sequence identity to coordinates 516 to 579 in the complete functional copies of *Vicia* and *Pisum*, respectively (Fig. 2.7B).

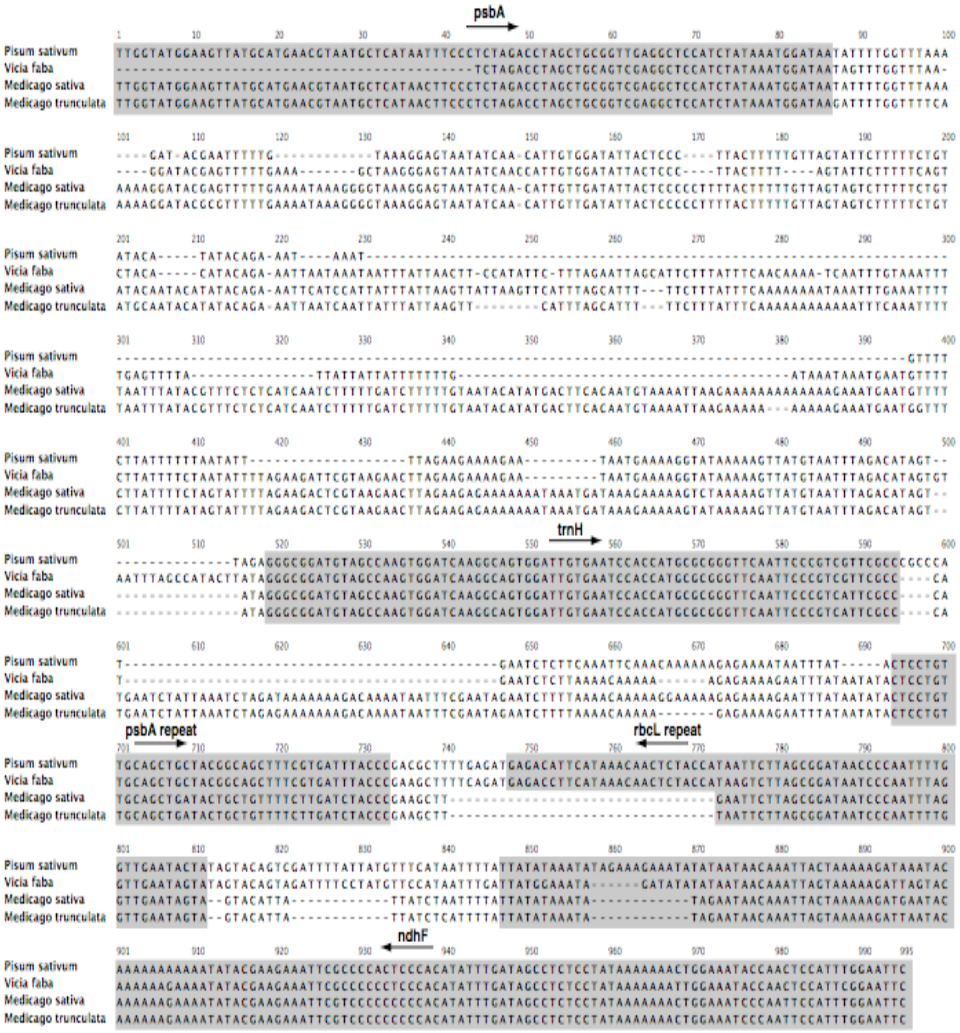

Fig 2.6. Sequence alignment of IR loss region between psbA and ndhF for *Medicago*, *Pisum*, and *Vicia*. Shaded regions show genes and repeat elements. Sequences for this figure were obtained from Genbank (P. sativum [M16899], Shapiro and Tewari, 1986; V. faba [X51471], Herdenberger et al., 1990; *M. sativa* [AY029748], D. Rosellini, unpubl.; *M. truncatula* [NC003119], Lin et al., unpubl.).

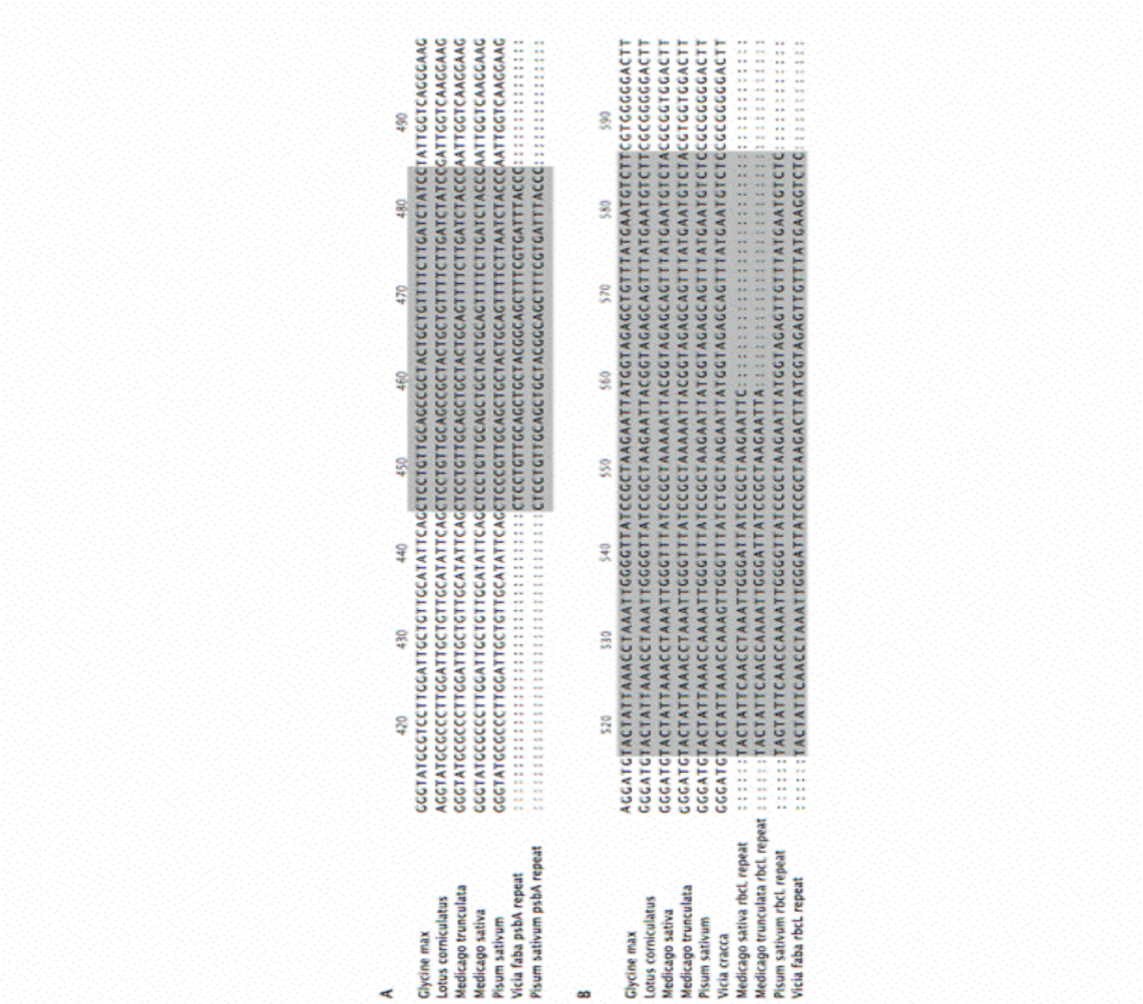

Fig 2.7. Sequence alignment of legume repeats for *psbA* (A) and *rbcL* (B) with functional copies of these genes. *psbA* sequences are from GenBank for *L. corniculatus* (AP002983), *M. truncatula* (AC093544), *M. sativa* (AY029748), *P. sativum* (M11005) and from the genome sequence of *G. max* generated in this study (NC_007942). *rbcL* sequences are from GenBank for *L. corniculatus* (AP002983), *M. truncatula* (AC093544), *M. sativa* (X04975), *P. sativum* (X03853) and from the genome sequence of *G. max* generated in this study (NC_007942). Sequences of the *psbA* and *rbcL* repeats for *P. sativum* and *V. faba* are from Shapiro and Tewari (1986, M16899) and Herdenberger et al. (1990, X51471), respectively. Colons in alignment indicate gap region

Repeat Analysis

Repeat analyses using REPuter found 67 to 191 direct and inverted repeats 30 bp or longer with a sequence identity of at least 90% among the three legume chloroplast genomes examined (Fig. 2.8) . *Medicago* has the largest number of repeats with 191 and *Lotus* has the fewest with only 67. The number of repeats in the legumes is higher than the 57 repeats identified in *Arabidopsis*. The majority of the repeats (54–81%) in all four genomes are between 30–40 bp in length. The longest legume repeats are in *Lotus* and *Glycine* and are 274 and 287 bp, respectively. The largest repeat in *Glycine* is a 287 bp sequence of *ycf2* that has 4 identical copies, 2 in each IR. The 2 copies in each IR are separated by 1,689 bp. The 4 copies of the 274 bp repeat in *Lotus*, which also represents a duplicated segment of *ycf2* in the IR, are separated by 1,963 bp in each IR. The two large repeats in *Glycine* and *Lotus* are very similar with 83% sequence identity at the nucleotide level.

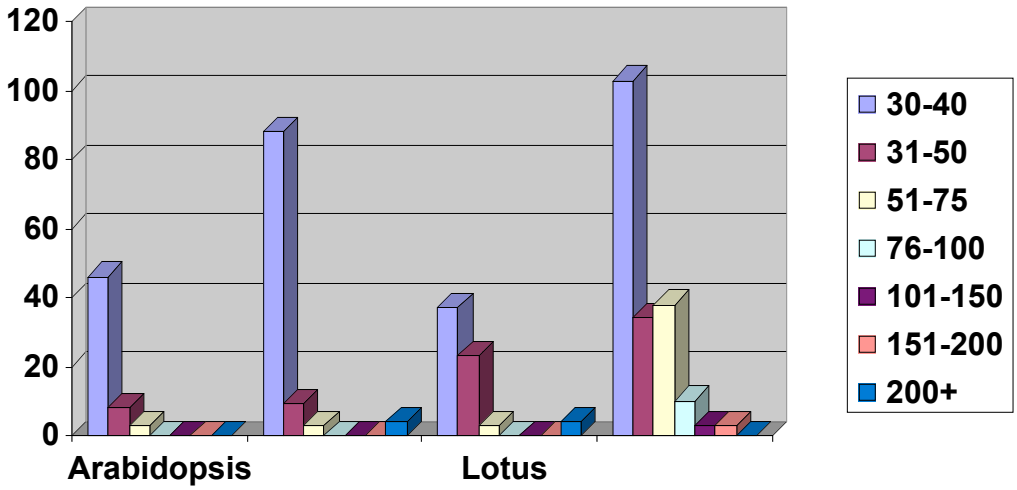

Fig 2.8. Histogram showing the number of repeated sequences ≥ 30 bp long with a sequence identity ≥ 90% in the three legume and *Arabidopsis* genomes using REPuter (Kurtz et al., 2001).

BlastN (Altschul et al, 1997) comparisons of the 191 *Medicago* repeats against the chloroplast genomes of *Arabidopsis*, *Glycine*, and *Lotus* reveal that 13 of the *Medicago* repeats show a sequence identity greater than 90% with sequences 30 bp or longer (Table 2.1). Five of the *Medicago* repeats are located in intergenic spacers or introns (repeats 3–7 in Table 2.1) and the remaining eight repeats are found in four genes, *psaA*, *psaB*, *ycf1* and *ycf2*. Many of the *Medicago* repeats are also found in *Arabidopsis*. One of these is repeat 3, which represents a portion of the *psbA* gene that is found in the intergenic spacer (IGS) between *trnH* and *ndhF* and in *psbA* of *Medicago* but is only found in *psbA* of *Arabidopsis*, *Glycine*, and *Lotus* (see section on IR extent above for more details). Two repeats are restricted to legumes (repeats 10 and 13) and these are located in *ycf2*. The number of *Medicago* repeats shared with only one other genome is 1 for *Arabidopsis* (repeat 6), 2 for *Lotus* (repeats 2 and 7), and 1 for *Glycine* (repeat 8).

Table 2.1. *Medicago* repeats in other legume chloroplast genomes and *Arabidopsis*. Only *Medicago* repeats that show a length > 20 bp and a sequence identity of > 90 % with the other genomes are listed. Length of *Medicago* repeats (in bp) and their locations (gene names and starting coordinates) are provided in column 1. The number of copies, length (bp), percent identity, and locations (gene or region names and starting coordinates) of the repeated sequences are listed for other genomes. IGS = intergenic spacer

Medicago repeat	*Glycine*	*Lotus*	*Arabidopsis*
29 bp, *ycf2*	4, 29, 93.1%, *ycf2*	4, 29, 93.1%, *ycf2*	2, 29, 93.1%, *ycf2*
32 bp, *psaA*/*psaB*	0	1, 32, 90.6%, *psaB*	0
40 bp, IGS *trnH* - *ndhF* and *psbA*	1, 37, 91.9%, *psbA*	1, 37, 91.9%, *psbA*	1, 37, 91.9%, psbA
41, *ndhA* intron and *ycf3* intron	1, 41, 92.7%, *rpl16* exon 2 1, 40, 92.5%, *ndhA* intron 1, 38, 94.7%, IGS *trnS* - *ycf3*	1, 41, 92.7%, IGS *trnS* - *ycf3* 1, 41, 92.7%, *ndhA* intron 1, 38, 94.7%, IGS *rpl16* - *rps3* 2, 38, 92.1%, IGS *rps12* - *ycf15*	1, 38, 92.%, IGS *trnS* - *ycf3* 2, 38, 94.7%, IGS *rps12* 3' end - *trnV*

Table 2.1 (Continued). *Medicago* repeats in other legume chloroplast genomes and *Arabidopsis*. Only *Medicago* repeats that show a length > 20 bp and a sequence identity of > 90 % with the other genomes are listed. Length of *Medicago* repeats (in bp) and their locations (gene names and starting coordinates) are provided in column 1. The number of copies, length (bp), percent identity, and locations (gene or region names and starting coordinates) of the repeated sequences are listed for other genomes. IGS = intergenic spacer

42, IGS *ycf15* - *rps12* 3' end and IGS *rps3* - *rpl16*	1, 42, 100%, *rpl16* exon 2 1, 42, 95.2%, IGS *ycf15* - *rps12* 3' end 1, 41, 95.2%, *rps12* 3' end exon 2 1, 39, 100%, *ndhA* intron 1, 39, 94.9%, IGS *trnS* - *ycf3*	2, 42, 97.6%, IGS *ycf15* - *rps12* 3' end 1, 40, 97.5%, *ndhA* intron 1, 40, 97.5%, IGS *rpl16* - *rps3* 1, 39, 97.4%, IGS *trnS* - *ycf3*	2, 42, 100%, IGS *trnV* - *rps12* 3' end 1, 40, 90%, *ndhA* intron 1, 39, 92.3%, IGS *trnS* - *ycf3*
42, IGS *ycf4* - *psaI* and IGS *psaI* - *accD*	0	0	1, 32, 93.8%, IGS *accD* - *psaI*
45, IGS *ycf1* - *trnN*	0	1, 20, 90%, IGS *trnV* - *ndhC*	0
48, *ycf1*	1, 21, 100%, *ycf1* 1, 22, 100%, *ycf1*		

Table 2.1 (Continued). *Medicago* repeats in other legume chloroplast genomes and *Arabidopsis*. Only *Medicago* repeats that show a length > 20 bp and a sequence identity of > 90 % with the other genomes are listed. Length of *Medicago* repeats (in bp) and their locations (gene names and starting coordinates) are provided in column 1. The number of copies, length (bp), percent identity, and locations (gene or region names and starting coordinates) of the repeated sequences are listed for other genomes. IGS = intergenic spacer

58, *psaB* and *psaA*	1, 52, 94.2%, *psaB* 1, 49, 91.8%, *psaA*	1, 52, 90.4%, *psaB* 1, 47, 95.7%, *psaA*	1, 58, 93.1%, *psaB* 1, 44, 95.4%, *psaA*
58, *ycf2*	2, 27, 92.6%, *ycf2*	2, 27, 92.6%, *ycf2*	0
61, *ycf2*	2, 41, 92.7%, *ycf2* 2, 39, 92.3%, *ycf2*	2, 41, 90.2%, *ycf2* 2, 41, 92.7%, *ycf2*	2, 39, 92.3%, *ycf2*
79, *psaB* and *psaA*	1, 76, 90.8%, *psaB*	1, 47, 95.7%, *psaA*	1, 76, 93.4%, *psaB* 1, 47, 95.7%, *psaA*
118, *ycf2*	2, 27, 92.6%, *ycf2* 2, 27, 96.3, *ycf2*	2, 27, 92.6%, *ycf2* 2, 27, 96.3, *ycf2*	0

The analyses identified 32 SSRs and these are composed of di- to penta- nucleotide repeating units (Table 2.2). Nearly 63% of all SSRs are di-nucleotide repeats and are composed primarily of AT or TA. The next most common SSR consists of tetra-nucleotide repeats and accounts for 19% of the SSRs with no common motif. The remaining 18% of the SSRs are composed of tri- and penta-nucleotide repeats. Of the SSRs identified, there are none within an open reading frame.

Table 2.2. Simple sequence repeats identified by CUGISSR in the soybean chloroplast genome. Table shows motif, number of repeated elements, location, and presence within an ORF.

Description	SeqLen	Motif	# Repeats	Start	Stop	INORF
Glycine max	152218	tct	4	2123	2134	N
		at	8	5159	5174	N
		at	9	5177	5194	N
		att	4	14613	14624	N
		tatc	3	18422	18433	N
		atag	3	18449	18460	N
		ta	8	24654	24669	N
		att	5	28630	28644	N
		aat	4	29628	29639	N
		ta	5	31739	31748	N
		ta	5	32799	32808	N
		ta	7	32834	32847	N
		at	5	33688	33697	N
		at	6	48408	48419	N
		ta	5	48433	48442	N
		ta	6	54290	54301	N
		at	5	65076	65085	N
		ta	5	67497	67506	N
		cttt	3	67677	67688	N
		ta	5	68067	68076	N
		at	5	68315	68324	N
		atca	3	78285	78296	N
		at	5	78336	78345	N
		ta	5	79502	79511	N
		ta	5	80708	80717	N
		cagaa	3	107701	107715	N
		at	5	116626	116635	N
		ttta	3	117184	117195	N
		at	6	118649	118660	N
		atca	3	119917	119928	N
		ta	5	122325	122334	N
		ttctg	3	127679	127693	N

Discussion

The *Glycine* genome has the typical organization for land plant chloroplast genomes with two identical copies of an inverted repeat that separate the large and small single copy regions. The size of the genome at 152,218 bp is also similar to most angiosperm chloroplast genomes that have two copies of the IR, which generally range in size from 134 – 164 kb (Jansen et al., 2005). The two IR containing legumes whose genomes have been sequenced, *Glycine* (reported here) and *Lotus* (Kato et al., 2000), are very similar in size with *Lotus* being 1,619 bp shorter than *Glycine*. Only a small portion of this difference in length can be attributed to the expansion of the IR in *Glycine* at the IR/LSC boundary (Fig. 2.5), a phenomenon common in flowering plants (Goulding et al., 1996). Therefore, most of this size variation is due to differences in sizes of intergenic spacer regions outside of the IR.

There is considerable variation in size of legume chloroplast genomes due to the loss of one copy of the IR from members of six related tribes (Palmer 1985, Palmer et al., 1987b, Lavin et al., 1990). A detailed examination of the IR loss region in Pea (*Pisum sativum*) and broad bean (*Vicia faba*) identified two repeated sequences of 40 and 64 bp in the region where the IR was deleted (Wolfe 1988, Herdenberger et al., 1990). These repeats showed a very high sequence identity to portions of two LSC genes, *rbcL* and *psbA* (Wolfe 1988). Wolfe suggested that the repeats could have been present prior to the IR loss and played a role in the deletion event (Wolf 1988). Alternatively, these repeats may have been formed as part of the IR deletion. In either case, Wolfe (1988) predicted that if other legumes that lost one copy of the IR share these repeats it would indicate that the IR deletion in legumes represents a single event. Examination of the IR region in the three legume chloroplast genomes (Fig. 2.6) clearly indicates that other legumes with only one copy of the IR have the

psbA and *rbcL* repeats. Thus, this IR loss occurred only once, and it provides an excellent phylogenetic marker supporting the monophyly of six tribes of legumes. The monophyly of this group of legumes is also supported by a sequence-based phylogeny of the plastid gene *matK* (Wojciechowski et al., 2004). The *psbA* repeats in *Pisum*, *Vicia* and the two *Medicago* species (Fig. 2.6) are identical in length and have a very high sequence identity (100% for *Pisum*/*Vicia* and 85% for *Pisum*/*Medicago*). In contrast, the *rbcL* repeat (Fig. 2.6) has diverged more in length (39 bp in *Medicago* vs 64 bp in *Pisum* and *Vicia*) but still has a very high sequence identity (94% for *Pisum*/*Vicia* and 95% for *Pisum*/*Medicago*). The sequenced legume genomes with both copies of the IR (*Glycine* and *Lotus*) do not have either the *psbA* or *rbcL* repeats suggesting that these repeats originated at or shortly after the time of the deletion event.

Gene content is highly conserved in most land plant chloroplast genomes (Palmer, 1991, Raubeson and Jansen, 2005). The *Glycine* genome contains 130 genes, 19 of which represent duplicate copies in the IR. The gene content is nearly identical to the completely sequenced *Lotus* chloroplast genome (Kato et al., 2000) and both of these legumes and *Medicago* lack the *rpl22* gene. The absence of *rpl22* from legume chloroplast genomes has been noted previously (Spielmann et al., 1988, Milligan et al., 1989, Gantt et al., 1991, Doyle et al., 1995). This gene represents an interesting case of gene transfer from the chloroplast to the nucleus. The nuclear encoded protein is imported back into the chloroplast by a transit peptide (Gantt et al., 1991). In addition to *rpl22*, the *Medicago* genome lacks a second ribosomal protein gene, *rps16*. Sequencing studies demonstrated the loss of this gene from *Pisum sativum* (Nagano et al., 1991) and an extensive survey of legumes using a filter hybridization approach suggested that there have been multiple independent losses of *rps16*

in legumes (Doyle et al., 1995). Additional losses of this gene in distantly related plant lineages [e.g., liverworts (Ohyama et al., 1986) and pine (Tsudzuki et al., 1992)] clearly indicate that this gene loss is not a very reliable phylogenetic marker.

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

With the exception of the IR, chloroplast genomes have very few repeated sequences (Palmer, 1991). However, a number of studies of rearranged chloroplast genomes have identified dispersed repeats [Chlamydomonas (green algae) (Maul et al., 2002), Pseudotsuga (Douglas-fir) (Hipkins et al., 1995), Trachelium (perennial herbs) (Cosner et al., 1997), Trifolium (clover) (Milligan et al., 1989), wheat (Bowman and Dyer, 1986; Howe, 1985), and Oenothera (primrose) (Hupfer et al., 2000, Sears et al.,1996, Vomstein and Hachtel, 1988)]. The most impressive example is Chlamydomonas in which it was estimated that the genome comprises more than 20% dispersed repeats. All of the genomes with repeated sequences other than the IR have inversions, and this correlation has been used to suggest that repeats may have mediated these changes (Palmer, 1991). The repeat analyses of the three legumes indicate that these genomes contain a substantial number of repeats (Fig. 2.8). The analyses was limited to repeats of 30 bp or longer with at least 90% sequence identity. Searches for shorter and/or more divergent repeats would likely identify many additional repeated

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

Because organellar genomes are often uniparentally inherited, chloroplast DNA polymorphisms have become a marker of choice for investigating evolutionary issues such as sex-biased dispersal and the directionality of introgression (Willis et al. 2005). They are also invaluable for the purposes of population-genetic and phylogenetic studies (Bryan et. al., 1999, Raubeson and Jansen 2005). Also, knowledge of mutation rates is important because they determine levels of variability within populations, and hence greatly influence estimates of population structure (Provan et. al., 1999). Mining for SSRs identified 32 di-penta nucleotide repeating units. These initial findings indicate a potential to test and utilize SSRs to rapidly analyze diversity in soybean germplasm collections.

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

In addition to providing insight into genome organization and evolution, availability of complete DNA sequence of chloroplast genomes should facilitate plastid genetic

engineering. Although many successful examples of plastid engineering in tobacco have set a solid foundation for various future applications, this technology has not been extended to many of the major crops. Stable plastid transformation has been recently accomplished via somatic embryogenesis using partially sequenced chloroplast genomes in soybean (Dufourmantel et al., 2004), carrot (Kumar et al., 2004a) and cotton (Kumar et al., 2004b; Daniell et al., 2005) and rice (Lee et al., 2005). Complete chloroplast genome sequences should provide valuable information on spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes and should help in extending this technology to other useful crops.

CHAPTER 3

COMPLETE CHLOROPLAST GENOME SEQUENCES OF *SOLANUM BULBOCASTANUM, SOLANUM LYCOPERSICUM* AND COMPARATIVE ANALYSIS WITH OTHER SOLANACEAE GENOMES

Introduction

Once thought to be poisonous, *Solanum lycopersicum* (*Solanum lycopersicum*) has become the second most commonly grown fruit crop in the world behind *Solanum bulbocastanum*. Traditional plant breeding has resulted in great progress in increasing yield, disease and pest resistance, environmental stress resistance and quality and processing attributes. However, *Solanum lycopersicum* plant breeding programs still strive to generate a better product. To assist in this goal, some plant breeding programs have been expanded to include molecular breeding and transgenic techniques. Tomato has long been recognized as an excellent genetic model for molecular biology studies. This has resulted in a flood of information including markers and genetic maps, identification of individual chromosomes, promoters and other nuclear genome sequences and identification of genes and their function. Although the *Solanum lycopersicum* genome is highly enabled through genetic/physical maps and a large database representation of genomic and expressed sequence, there is not much information on the chloroplast genome. Because of this reason segments of the tobacco chloroplast genome were used as flanking sequences to facilitate integration of transgenes into the *Solanum lycopersicum* chloroplast genome by homologous recombination, without knowing exact sequence identity (Ruf et al., 2001). This resulted in poor transformation efficiency (Ruf et al., 2001).

Solanum bulbocastanum (a mexican diploid species) is the most economically significant crop in the U.S. produce industry. With an annual farm value of $2.5 billion and per capita use of 140 pounds in 2001, potatoes rank first in value and consumption among all vegetables produced and consumed in the United States (USDA http://plants.usda.gov/java/profile?symbol=SOTU). Additionally, potato products such as french-fries and potato chips generate billions more in revenue for the food-processing and food service industries. Potatoes contain high vitamin C, high potassium, and are a good source of vitamin B6 and dietary fibers. Currently, exports account for 11% of US potato production in form of fresh, seed, frozen and dehydrated potatoes. However, there is not much information on the potato chloroplast genome. When the potato plastid genome was transformed, tobacco plastid flanking sequence were used to facilitate transgene integration by homologous recombination (Sidorov et al., 1999).

This study presents the complete sequence and analysis of the chloroplast genomes of *Solanum lycopersicum* and *Solanum bulbocastanum*. One goal of this research is to compare the genome organization of *Solanum bulbocastanum* and *Solanum lycopersicum* with the other two completely sequenced Solanaceae chloroplast genomes (tobacco and *Atropa*). In addition to examining gene content and gene order, the distribution and location of repeated sequences among members of the Solanaceae is determined. A second goal was to compare levels of DNA sequence divergence among chloroplast coding and non-coding regions. Intergenic spacer regions have been examined to identify ideal insertion sites for transgene integration and they are commonly used by plant systematists for resolving phylogenetic relationships among closely related species (Kelchner 2002). A final goal of this study is to examine the extent of RNA editing in Solanaceae chloroplast genomes by comparing the DNA sequences

with available expressed sequence. RNA editing is known to play an important role in several lineages of plants (Wolf et al., 2004, Kugita et al., 2003) but most of our knowledge about the frequency of this process in crop plants comes from studies in maize (Maier et al., 1995) and tobacco (Hirose et al., 1999).

Methodology

DNA Sources

The bacterial artificial chromosome (BAC) libraries of *Solanum bulbocastanum* and *Solanum lycopersicum* were constructed by ligating size fractionated partial *Hin*dIII digests of total cellular high molecular weight DNA with the pINDIGOBAC vector (Luo et al., 2001). The average insert size of the *Solanum bulbocastanum* and *Solanum lycopersicum* libraries are 177 kb and 155 kb, respectively. BAC related resources for these public libraries can be obtained from the Clemson University Genomics Institute BAC/EST Resource Center (www.genome.clemson.edu).

Chloroplast BAC clone identification/selection, sequencing protocols, sequence assembly, annotation, and pairwise comparisons among taxa were performed as described in chapter 2.

Repeat Structure

The repeat structure of the chloroplast genomes were examined in two stages. First, REPuter (Kurtz et al., 2001) was used to identify the number and location of direct and inverted (palindromic) repeats in the species of Solanaceae using a minimum repeat size of 30 bp and a Hamming distance of 3 (i.e., a sequence identity of ≥ 90 %). Second, the repeats identified for tobacco were blasted against the complete chloroplast genomes of all

four Solanaceae genomes. Blast hits of size 30 bp and longer with a sequence identity of ≥ 90% were identified to determine the shared repeats among the four genomes examined.

An aligned data set of all of the shared genes among the four Solanaceae chloroplast genomes was constructed by extracting these sequences from the annotated genomes either using DOGMA (Wyman et al., 2004) or the Chloroplast Genome Database (http://cbio.psu.edu/chloroplast/index.html). The sequences were aligned using ClustalX (Higgins et al., 1996).

Molecular evolutionary analyses were then performed on the aligned data matrix using MEGA2 (Molecular Evolutionary Genetics Analysis (Kumar et al., 2001)). Estimates of sequence divergence were based on the Kimura 2-parameter distance correction (Kimura, 1980).

Comparison of Intergenic Regions

Intergenic regions from four Solanaceae chloroplast genomes were compared using MultiPipMaker (Schwartz et al., 2003) (http://pipmaker.bx.psu.edu/pipmaker/tools.html). Also used was a program known as 'all_bz' that iteratively compares one pair of nucleotide sequences at a time until all possible pairs from all species have been compared. However, this program processes only one set of intergenic regions at a time. For genome-wide comparisons of corresponding intergenic regions from all species, the Guda lab (State University at Albany, NY) developed two programs (written in Perl). The first program iteratively creates a set of input files containing corresponding intergenic regions from each species and uses the 'all_bz' module, until all the intergenic regions in the chloroplast genome are processed. The second program parses the output from the above comparisons,

calculates percent identity by using the number of identities over the length of the longer sequence and generates results in tab-delimited tabular format.

Variations Between Coding Sequences and cDNAs

Each of the gene sequences from the *Solanum bulbocastanum* chloroplast genome was used to perform a BLAST search of expressed sequence tags (ESTs) from the NCBI Genbank. The retrieved EST sequences from *Solanum bulbocastanum*, *Solanum lycopersicum* and tobacco were then aligned with the corresponding gene for each species separately, using Clustal X. In the case of *Atropa*, no sequences were retrieved from the Genbank even though its chloroplast sequence has been completed and studies of RNA editing have been previously performed (Schmits-Linneweber et al., 2002). The aligned sequences were then screened and nucleotide and amino acid changes were detected using the Megalign software (DNAstar, Madison, WI). The following criteria were used for comparisons of the DNA and EST sequences: (1) when more than one EST sequence was retrieved using BLAST, a change was recorded only if all sequences had the same change (substitution); (2) changes were recorded based on the base substitutions, that is, if there was an indel that affected the DNA sequence, it was not considered; and (3) if a retrieved EST sequence was too different (more than three consecutive nucleotide substitutions in a given sequence), it was not used for the analysis. In most cases, EST sequences were not of the same length as that of the corresponding gene, so the length of the analyzed sequence was recorded. Once a variable site was detected, the sequence was translated using the Megalign program using the plastid/bacterial genetic code and differences in the amino acid sequence were recorded.

Results

Size, gene content and organization of the Solanum lycopersicum and Solanum bulbocastanum chloroplast genomes

The complete sizes of the *Solanum lycopersicum* and *Solanum bulbocastanum* chloroplast genomes are 155,460 and 155,372 (Fig. 3.1) bp, respectively. The genomes include a pair of inverted repeats of 25,613 bp (*Solanum lycopersicum*) and 25,588 bp (*Solanum bulbocastanum*), separated by a small single copy region of 18,361 bp (*Solanum lycopersicum*) and 18,381 bp (*Solanum bulbocastanum*) and a large single copy region of 85,873 bp (*Solanum lycopersicum*) and 85,815 bp (*Solanum bulbocastanum*). The difference in size of the two genomes is due partly to a slight expansion of the IR in *Solanum lycopersicum* resulting in a partial duplication *rps19*, a phenomenon that is quite common in chloroplast genomes (Goulding et al., 1996).

The *Solanum bulbocastanum* and *Solanum lycopersicum* chloroplast genomes contain 113 unique genes, and 20 of these are duplicated in the IR, giving a total of 133 genes (Fig. 3.1). There are 30 distinct tRNA genes, and 7 of these are duplicated in the IR. Seventeen genes contain one or two introns, and five of these are in tRNAs. The overall GC and AT content of the *Solanum bulbocastanum* and *Solanum lycopersicum* chloroplast genomes are 37.86% (*Solanum lycopersicum*), 37.88% (*Solanum bulbocastanum*) and 62.14% (*Solanum lycopersicum*), 62.12% (*Solanum bulbocastanum*), respectively.

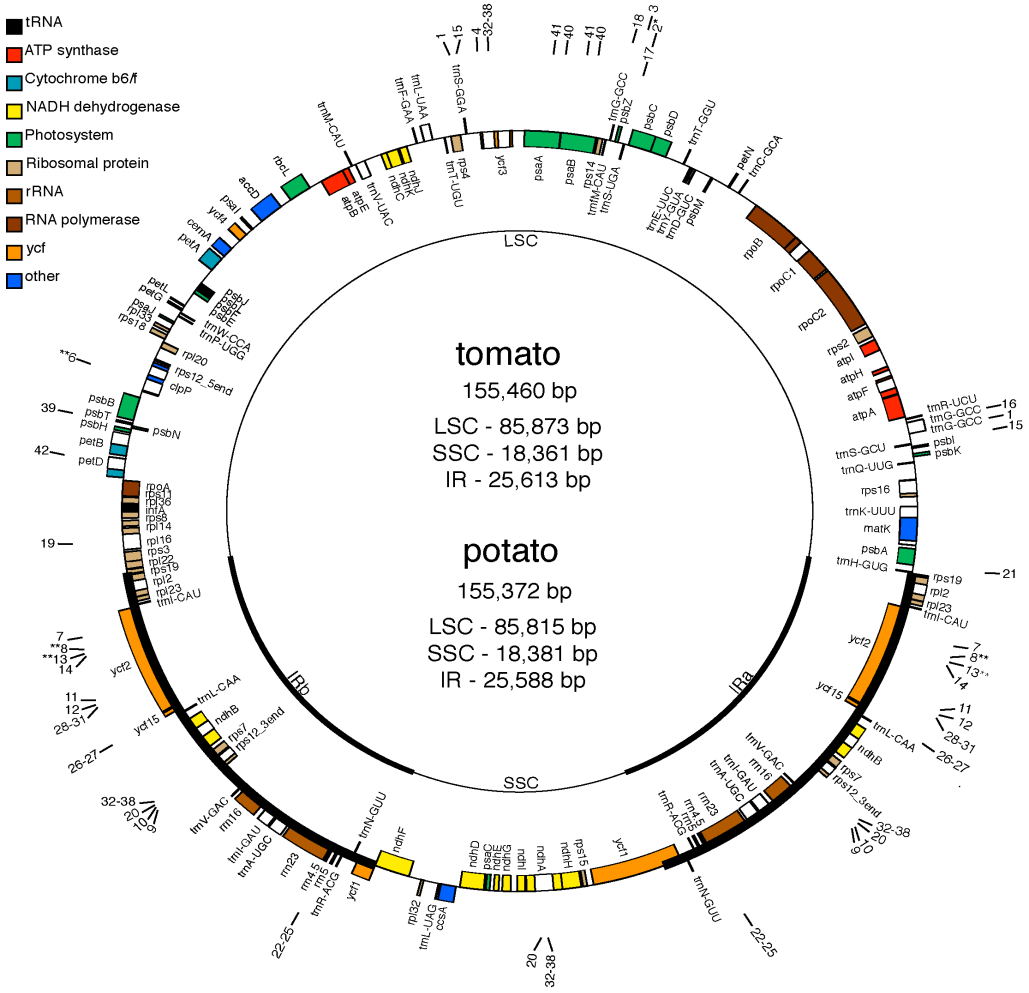

Fig 3.1. Gene map of *Solanum lycopersicum* and *Solanum bulbocastanum* chloroplast genomes. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into small (SSC) and large (LSC) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction. Numbered arrows around the map indicate the location of repeated sequences found in Solanaceae genomes (see Table 3.1 for details). Arrows with asterisks indicate the five groups of repeats that are not shared by all four Solanaceae genomes: * tobacco and *Solanum lycopersicum*, ** tobacco and *Atropa*, *** tobacco.

Gene content and gene order

Gene content of the four sequenced species of Solanaceae (*Solanum bulbocastanum* & *Solanum lycopersicum*, published here; tobacco [ref; NC_001879] and *Atropa* [NC_004561]) is identical. Similarly, the gene order is identical among all four sequenced Solanaceae genomes. However, there are significant additions or deletions of nucleotides within certain coding sequences. For example the ACACGGGAAAC sequence is uniquely present within the 16S rRNA gene of *Solanum bulbocastanum*, *Solanum lycopersicum* and *Atropa* but absent in tobacco or any other sequenced chloroplast genome (Fig. 3.2). Several deletions also occur within the coding sequence of *ycf2* in *Atropa*, *Solanum lycopersicum*, *Solanum bulbocastanum* and tobacco (Fig. 3.3). It should be noted that deleted nucleotides within the 16S rRNA and ycf2 are repeated sequences. In *Solanum lycopersicum ycf2* has two ribosome binding sites (GGAGG), whereas there is only one in all other Solanaceae members sequenced so far (Fig. 3.3).

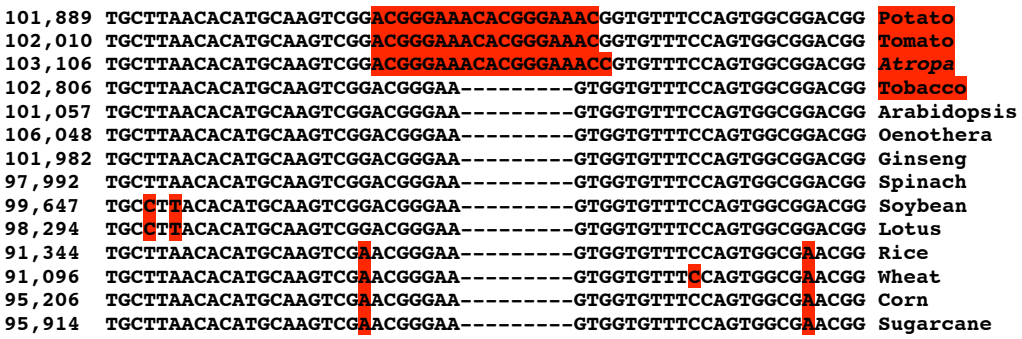

Fig. 3.2. Alignment of a portion of the 5' end of the 16S ribosomal RNA showing a nine bp insertion in *Atropa*, *Solanum bulbocastanum*, and *Solanum lycopersicum*. Nucleotides shown in red indicate base substitutions.

Fig. 3.3. Alignment of four regions of the *ycf2* gene among the four Solanaceae chloroplast genomes showing insertion and deletion events. Green indicates start codon, yellow shade indicates repeat sequence, red indicates nucleotide substitution. Ellipses indicate shine-delgarno sequence

Repeat Structure

REPuter found 33 to 45 direct and inverted repeats 30 bp or longer with a sequence identity of at least 90% among the four chloroplast genomes examined (Fig. 3.4). The majority of the repeats in all four genomes are between 30 to 40 bp in length. The longest repeats other than the inverted repeats are found in *Solanum lycopersicum* and consist of four 57 bp repeats not found in any of the other three genomes. Both tobacco and *Solanum bulbocastanum* both share a 50 and 56 bp repeat, whereas *Atropa* does not have a single repeat in the greater than 50 bp size range (excluding the IR).

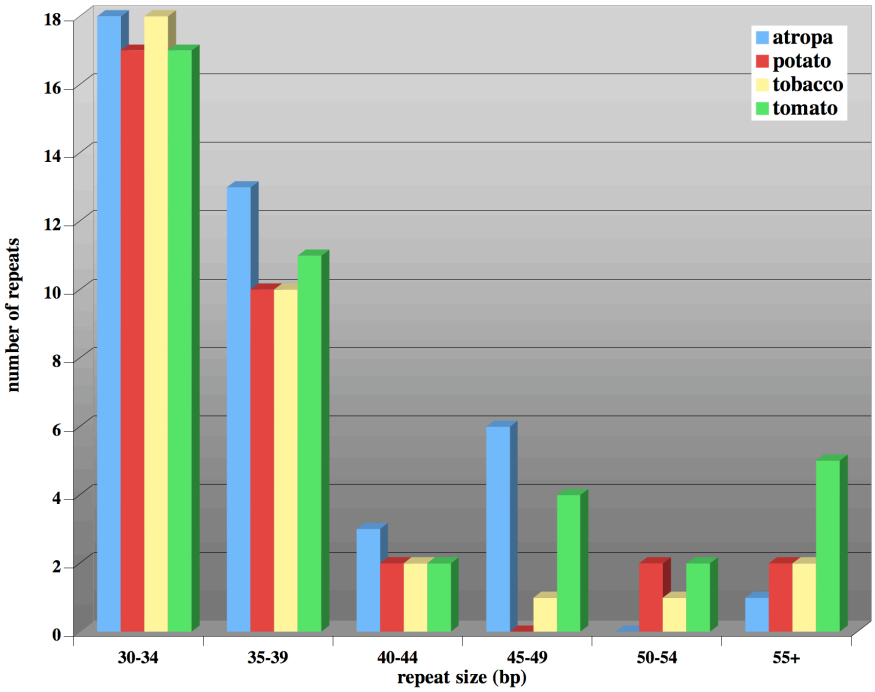

Fig. 3.4. Histogram showing the number of repeated sequences ≥ 30 bp long with a sequence identity ≥ 90% in the four Solanaceae chloroplast genomes using REPuter

BlastN comparisons of the tobacco repeats (excluding the inverted repeat) against the chloroplast genomes of *Atropa*, *Solanum bulbocastanum* and *Solanum lycopersicum* identified 42 repeats that show a sequence identity ≥ 90% with sequences ≥ 30 bp (Table 3.1, Fig. 3.1). Thirty-seven of the 42 repeats are found in all four Solanaceae chloroplast genomes and all of these are located in the same genes or intergenic regions.

Table 3.1. Tobacco repeats blasted against all four Solanaceae chloroplast genomes. Table includes blast hits at least 30 bp in size, a sequence identity ≥ 90%, and a bit-score of great than 40. Abbreviation for genomes are: N = *Nicotiana* (tobacco)77, A – *Atropa*51, P = *Solanum bulbocastanum*, T = *Solanum lycopersicum*; IGS = intergenic spacer. See Figure 1 for location of repeats on the gene map.

Tobacco Repeat	Size (bp)	Number of copies	Location	Genomes
1	30	2	IGS(1bp) - *trnS*-GCC	NAPT
2	30	1	IGS - (*psbC* - *trnS*-UGA)^{N}, Intron – (*clpP*#2 – *clpP*#3)^{T}	NT
3	30	1	IGS(1bp) - *trnS*-UGA	NAPT
4	30	1	Intron - (*ycf3* exon 2 - *ycf3* exon 3)	NAPT
5	30	2	*trnS*-GCU - IGS(1bp), *trnS*-GGA - IGS(1bp)	NAPT
6	30	1	Intron - (*clpP* exon 2 - *clpP* exon 3)	NA
7	30	2	*ycf2*	NAPT
8	30	2	*ycf2*	NA
9	30	2	IGS - (*rps12* 3'end - *trnV*-GAC)	NAPT
10	30	2	IGS - (*trnV*-GAC – *rps12* 3'end)	NAPT
11	30	2	*ycf2*	NAPT
12	30	2	*ycf2*	NAPT
13	30	2	*ycf2*	NA
14	30	2	*ycf2*	NAPT
15	31	2	IGS(2bp) - *trnS*-GCU, IGS(1bp) - *trnS*-GGA	NAPT
16	31	1	*trnG*-GCC - IGS(4bp)	NAPT
17	31	1	IGS(2bp) - *trnS*-UGA	NAPT
18	31	1	*trnG*-GCC - IGS(3bp)	NAPT
19	31	1	Intron - (*rpl16* exon 1 - *rpl16* exon 2)	NAPT
20	31	3	IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
21	32	2	IGS - (*trnH*-GUG - *psbA*)	N
22	34	4	IGS - (*rrn4.5* - *rrn5*)	NAPT
23	34	4	IGS - (*rrn4.5* - *rrn5*)	NAPT
24	34	4	IGS - (*rrn4.5* - *rrn5*)	NAPT
25	34	4	IGS - (*rrn4.5* - *rrn5*)	NAPT
26	35	4	IGS - (*ycf15* - *trnL*-CAA)	NAPT[1]
27	35	4	IGS - (*ycf15* - *trnL*-CAA)	NAPT[2]
28	37	4	*ycf2*	NAPT
29	37	4	*ycf2*	NAPT
30	37	4	*ycf2*	NAPT
31	37	4	*ycf2*	NAPT
32	39	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
33	39	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
34	39	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT

Table 3.1 (Continued). Tobacco repeats blasted against all four Solanaceae chloroplast genomes. Table includes blast hits at least 30 bp in size, a sequence identity ≥ 90%, and a bit-score of great than 40. Abbreviation for genomes are: N = *Nicotiana* (tobacco)77, A – *Atropa*51, P = *Solanum bulbocastanum*, T = *Solanum lycopersicum*; IGS = intergenic spacer. See Figure 1 for location of repeats on the gene map.

35	39	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
36	41	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (rps12 3'end - trnV-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
37	41	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (rps12 3'end - trnV-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
38	41	4	Intron - (*ycf3* exon 2 - *ycf3* exon 3), IGS - (*rps12* 3'end - *trnV*-GAC) x2, Intron - (*ndhA* exon 1 - *ndhA* exon 2)	NAPT
39	48	2	IGS(47bp) - *psbN*(1bp)	NAP[3]T
40	50	2	*psaB*, *psaA*	NAPT
41	50	2	*psaB*, *psaA*	NAPT
42	56	2	Intron - (*petD* exon 1 - *petD* exon 2)	NAPT[4]

Intergenic Spacer Regions

All intergenic spacer regions except those less than 11 bp across the four Solanaceae chloroplast genomes were compared (Fig. 3.5A, Table 3.2). Only four spacer regions (*rps11* - *rpl36*, *rps7* - *rps12 3'* end, *trn*I-GAU - *trnA*-UGC, *ycf2* - *ycf15*) have 100% sequence identity among all genomes (~2.5% of the spacer regions) and three of these regions are in the inverted repeat. Between *Solanum lycopersicum* and *Solanum bulbocastanum* 21 intergenic spacer regions have 100% sequence identity, whereas only 8 regions have 100% sequence identity between *Solanum lycopersicum* and *Atropa*, tobacco and *Solanum bulbocastanum*, *Atropa* and *Solanum bulbocastanum*, 9 regions between tobacco and *Solanum lycopersicum* and 10 regions between tobacco and *Atropa*. The number of intergenic spacer regions with 100% sequence identity reflects the close phylogenetic relationship among the four Solanaceae genomes (Bohs and Olmstead, 1997; Olmstead et al., 1999). It is noteworthy that one of the intergenic spacer regions that has 100% sequence identity between *Atropa* and *Solanum bulbocastanum* (trnI-CAU - ycf 2) has only 66-69% sequence identity among the other Solanaceae species examined. Similarly, *ycf4* - *cemA* has only 27 % identity between tobacco and *Atropa*, *Solanum bulbocastanum* and *Solanum lycopersicum*, whereas it has greater than 90% identity between other Solanaceae species examined. There are several deletions or insertions in the intergenic spacer regions between *trnQ* - *rps16*, *trnE* - *trnT*, *trnK* - *rps16*, *trnT* - *ycf* 5, *trnS* - *trnG*, *ycf2* - *trnI*, *ycf4* - *cemA*, *ycf15* - *trnL*.

Fig 3.5A

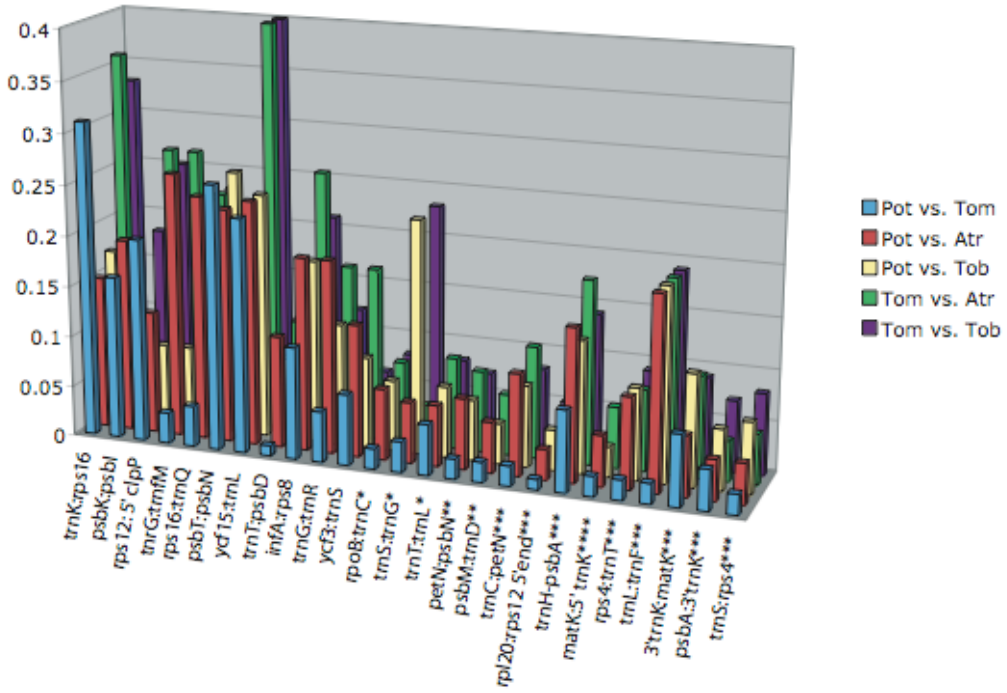

Figure 3.5. Histogram showing sequence divergence in pairwise comparisons among 4 Solanaceae chloroplast genomes for intergenic spacers (A) and coding regions (B). Pot = *Solanum bulbocastanum*, Tom = *Solanum lycopersicum*, Atr = *Atropa*, and Tob = tobacco. A. Comparisons of 21 of the most variable intergenic regions. *, **, and *** indicate the tier 1, tier 2, and tier 3 regions reported in Shaw et al. The plotted values were converted from percent identity to sequence divergence on a scale from 0 to 1. B.

Fig 3.5B

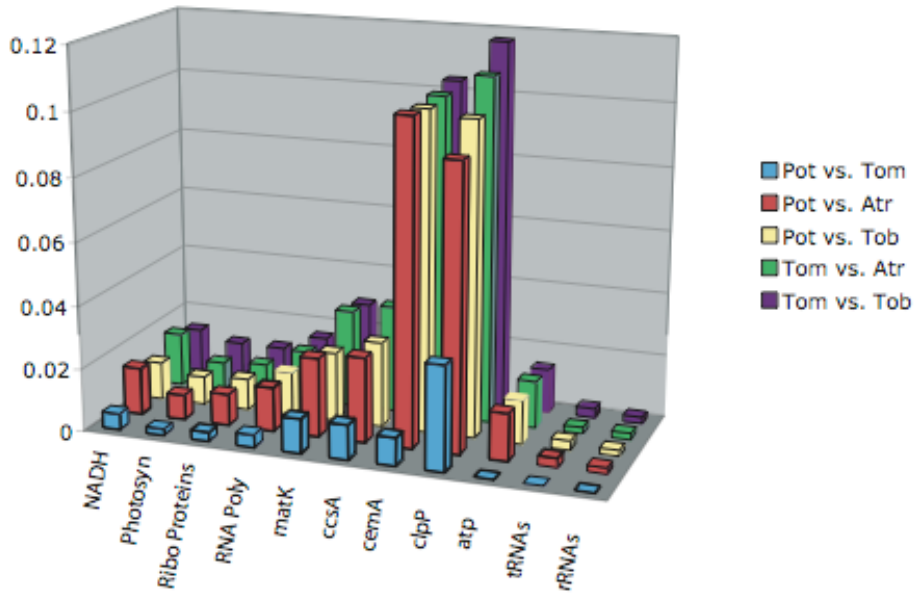

Figure 3.5 (Continued). Histogram showing sequence divergence in pairwise comparisons among 4 Solanaceae chloroplast genomes for intergenic spacers (A) and coding regions (B). Pot = *Solanum bulbocastanum*, Tom = *Solanum lycopersicum*, Atr = *Atropa*, and Tob = tobacco. A. Comparisons of 21 of the most variable intergenic regions. *, **, and *** indicate the tier 1, tier 2, and tier 3 regions reported in Shaw et al. The plotted values were converted from percent identity to sequence divergence on a scale from 0 to 1. B.

Table 3.2. Intergenic spacer regions that are 100% identical in *Atropa*, tobacco, *Solanum bulbocastanum* and *Solanum lycopersicum* or 100% identical to at least one other member of the Solanaceae. Names of genomes compared are abbreviated: Pot for *Solanum bulbocastanum*, Tom for *Solanum lycopersicum*, Atr for *Atropa*, and Tob for tobacco.

Intergenic ID	**Tob vs Atr**	**Tob vs Pot**	**Tob vs Tom**	**Atr vs Pot**	**Tom vs Pot**	**Tom vs Atr**
rps11:*rpl36*	100	100	100	100	100	100
*rps12*_3'end:*rps7*	100	100	100	100	100	100
trnA-UGC:*trnI*-GAU	100	100	100	100	100	100
ycf15:*ycf2*	100	100	100	100	100	100
trnV-GAC:*rrn16*	100	98	98	98	100	98
rrn4.5:*rrn5*	100	100	97	100	97	97
psbJ:*psbL*	96	96	96	100	100	100
trnA-UGC:*rrn23*	96	100	100	96	100	96
trnfM-CAU:*rps14*	100	97	97	97	100	97
trnN-GUU:*ycf1*	100	96	100	96	96	100
ycf1:*trnN*-GUU	100	96	100	96	96	100
rrn23:*trnA*-UGC	96	100	100	95	100	96
psbN:*psbH*	95	95	95	100	100	100
rpl23:*trnI*-CAU	97	97	97	97	100	97
rrn4.5:*rrn23*	100	95	95	95	100	95
rps8:*rpl14*	94	95	95	95	100	95
trnL-UAG:*ccsA*	95	94	94	95	100	95
trnD-GUC:*trnY*-GUA	94	94	94	94	100	94
ndhJ:*ndhK*	92	93	93	95	100	95
ndhD:*psaC*	93	93	93	94	100	94
rpoA:*rps11*	89	100	100	89	100	89
psbH:*petB*	95	92	92	92	100	92
rpoC2:*rpoC1*	95	92	92	91	100	93
rps14:*psaB*	95	91	91	91	100	92
trnI-CAU:*ycf2*	69	69	81	100	66	66

Sequence Divergence

The chloroplast genes were classified into 11 functional groups for comparisons of sequence divergence among coding regions (Table 3.3; Fig. 3.5B). Sequence divergence, which represents the proportion of nucleotide sites that differ, were estimated for all genes using the Kimura 2-parameter model 50. Overall, sequence divergence corresponds to the phylogenetic relationships among the four species of Solanaceae examined (Bohs and Olmstead, 1997, Olmstead et al., 1999, Spooner et al., 1993). For example, the two most closely related species, *Solanum bulbocastanum* and *Solanum lycopersicum*, have the lowest divergence values for all classes of genes. Comparisons of sequence divergence among functional groups indicates that the RNA, photosynthesis, and ATP synthase genes are the least divergent and that the most divergent genes are *cemA* (membrane protein), *clpP* (protease), *matK* (intron maturase), and *ccsA* (cytochrome related). The comparisons of the levels of sequence divergence between noncoding and coding regions (Figs. 3.5A-B) indicate that the noncoding regions are more divergent that coding regions.

Table 3.3 Comparisons of sequence divergence of Solanaceae chloroplast genes among the 11 different functional groups. Standard errors are in parentheses. Highly divergent genes do not contain standard error due to the amount of variation. Pairwise distances were calculated using the Kimura 2-parameter model (50). Names of genomes compared are abbreviated: Pot for *Solanum bulbocastanum*, Tom for *Solanum lycopersicum*, Atr for *Atropa*, and Tob for *Nicotiana*.

Gene group	**Length (bp)**	**Number of genes**	**Pot vs Tom**	**Pot vs Atr**	**Pot vs Tob**	**Tom vs Atr**	**Tom vs Tob**	**Atr vs Tob**
NADH	12102	11	0.005 (0.001)	0.015 (0.001)	0.012 (0.001)	0.017 (0.001)	0.014 (0.001)	0.013 (0.001)
Photosynthesis	14081	26	0.002 (0.000)	0.008 (0.001)	0.009 (0.001)	0.009 (0.001)	0.011 (0.001)	0.008 (0.001)
Ribosomal Protein	10207	22	0.003 (0.001)	0.010 (0.001)	0.010 (0.001)	0.010 (0.001)	0.011 (0.001)	0.009 (0.001)
RNA polymerase	10473	4	0.004 (0.001)	0.014 (0.001)	0.014 (0.001)	0.016 (0.001)	0.016 (0.001)	0.012 (0.001)
matK maturase	1530	1	0.011	0.025	0.022	0.031	0.029	0.017
ccsA-cytochrome synthesis	942	1	0.011	0.027	0.027	0.034	0.034	0.023
cemA- envelope membrane protein	690	1	0.009	0.102	0.101	0.102	0.104	0.010
clpP-Protease	621	1	0.033	0.090	0.099	0.109	0.117	0.026
ATP synthase genes	4968	6	0.000 (0.000)	0.015 (0.003)	0.014 (0.003)	0.015 (0.003)	0.014 (0.003)	0.015 (0.003)
tRNAs	2751	27	0.000 (0.000)	0.003 (0.001)	0.003 (0.001)	0.002 (0.001)	0.003 (0.001)	0.003 (0.001)
rRNAs	9064	4	0.000 (0.000)	0.002 (0.000)	0.002 (0.001)	0.002 (0.000)	0.002 (0.001)	0.002 (0.000)

RNA editing sites in the Solanum lycopersicum and Solanum bulbocastanum chloroplast transcripts

Based on the alignment of EST sequences retrieved from the NCBI Genbank with the coding regions from *Solanum bulbocastanum* and *Solanum lycopersicum*, 53 nucleotide substitution differences were observed in the *Solanum lycopersicum* sequence (Table 3.4) and 47 were observed in *Solanum bulbocastanum* (Table 3.5). However, with the exception of *rpl23*, all nucleotide substitutions occurred in different positions among both species. Of these substitutions, 11 were synonymous and 42 were nonsynonymous in *Solanum lycopersicum*, whereas *Solanum bulbocastanum* had 19 synonymous and 24 nonsynonymous substitutions. *Solanum bulbocastanum* had nine C-to-U conversions, five of which resulted in amino acid changes (Table 3.5). In *Solanum lycopersicum*, seven C-to-U conversions were observed, all of which resulted in an amino acid change (Table 3.4). Although most genes in both species experienced one and three nucleotide substitutions, four genes had more than five variable sites. These were *rpl36* and *rpoC2* in *Solanum lycopersicum*, with 7 and 10 nucleotide substitutions, respectively (Table 3.4), and *rpl16* and *ycf1* in *Solanum bulbocastanum*, with 5 and 7 substitutions, respectively (Table 3.5). In addition, an amino acid alteration was observed in the *Solanum lycopersicum ycf1*(unknown function) gene that results in a stop codon at position 604. There is a complete copy of *ycf1* and the truncated copy is at the IR/SSC boundary. It is the truncated copy that has the stop codon due to RNA editing. Thus there is still a full, functional copy of *ycf1*. Although there is evidence that *ycf1* is a necessary chloroplast gene, it is missing from all grass genomes (Maier et al., 1995).

Table 3.4. Differences observed by comparison of *Solanum lycopersicum* chloroplast genome sequences with EST sequences obtained by BLAST search in the NCBI Genbank.

Gene	Gene size (bp)	Sequence analyzed[a]	Number of variable sites	Variation type	Position(s)[b]	Amino acid change
atpA	1526	1-837	2	C-A	87	T-T
				G-A	653	G-E
atpB	1497	769-1497	2	C-A	954	D-E
				A-G	1062	R-R
atpF	555	322-555	1	G-A	408	A-A
atpH	246	29-246	1	A-C	141	G-G
ndhG	531	229-531	4	A-G	362	Y-C
				G-C	393	Q-H
				T-C	455	F-S
				T-G	494	V-G
ndhH	1182	692-1015	2	G-C	927	R-R
				T-G	928	F-V
psaB	2205	1778-2198	2	T-C	2138	F-S
				G-A	2146	G-S
psaJ	135	1-135	1	C-U	22	L-F
infA	105	1-105	1	C-U	46	Y-H
psbC	1423	756-1423	4	T-C	1310	F-L
				A-C	1323	H-P
				T-A	1324	
				A-U	1418	N-Y
rbcL	1436	469-1436	1	A-G	494	Y-C
rpl14	369	1-339	2	G-A	31	A-T
				T-C	254	V-A
rpl22	472	1-268	1	A-C	180	A-A
rpl23	282	1-282	2	C-U	71	S-F
				C-U	89	S-L
rpl36	114	1-114	7	T-G	20	V-G
				T-G	24	R-R
				T-C	31	C-R
				T-G	54	R-R
				T-A	77	I-N
				T-G	81	C-W
				T-G	82	S-A
rpoA	1014	1-594	3	C-U	65	T-I
				C-U	200	S-F
				A-C	594	I-I
rpoC2	4179	2392-3283	10	G-U	2409	Q-H
				G-A	2432	R-Q
				G-A	2518	V-I
				G-C	2606	R-P

Table 3.4 (Continued). Differences observed by comparison of *Solanum lycopersicum* chloroplast genome sequences with EST sequences obtained by BLAST search in the NCBI Genbank.

				G-U	2629	V-L
				C-A	2652	I-I
				T-A	2728	S-T
				G-A	2785	G-R
				G-A	2817	K-K
				T-G	3192	C-W
rps7F	468	109-468	1	C-G	137	A-G
rps12	258	1-258	1	C-U	107	S-L
rps18	306	163-306	1	T-G	223	L-V
ycf1	1140	10-628	2	A-U	603	N-K
				T-A	604	K-stop
ycf1R	3599	500-1094	1	A-G	751	K-E
ycf2	6837	981-1726	1	G-A	1704	K-K

Table 3.5: Differences observed by comparison of *Solanum bulbocastanum* chloroplast genome sequences with EST sequences obtained by BLAST search in the NCBI genbank.

Gene	Gene size	Sequence analyzed[a]	# variable sites	Variation type	Nucleotide position(s)[b]	Amino acid change
atpA	1525	435-1050	3	C-U	436	P-S
				G-A	651	G-G
				C-U	711	Y-Y
atpB	1497	564-1260	4	A-C	1158	E-D
				G-A	1246	E-R
				A-G	1247	
				G-A	1248	
atpH	247	1-247	3	G-U	16	A-S
				T-C	18	
				G-A	76	V-I
petB	648	20-648	2	G-U	405	G-G
				C-U	611	P-L
psaA	2253	829-1776	3	T-C	1530	G-G
				A-G	1725	G-G
				C-A	1726	P-T
psaC	247	1-177	3	T-C	147	V-V
				T-C	151	C-R
				G-A	156	K-K
psbA	1062	1-699	1	C-U	489	I-I
psbB	1527	856-1425	3	C-G	856	R-G
				C-U	1389	F-F
				T-C	1390	F-L
clpP	598	1-383	1	G-A	190	V-I
psbD	1062	321-534	1	T-G	532	A-A
rbcL	1436	886-1302	2	G-U	1255	A-S
				G-A	1300	G-R
rpl16	405	10-405	5	C-A	65	S-Y
				A-U	219	P-P
				C-U	226	L-L
				C-G	234	P-P
				A-C	243	T-T
rpl23	282	1-282	2	C-U	71	S-F
				C-U	89	S-L
rpl36	114	1-114	2	C-U	31	R-C
				G-U	73	L-V
rpoA	1014	298-798	4	G-A	420	T-T
				G-U	597	L-L
				T-C	780	L-L
				C-A	789	N-K
rps19	93	1-93	1	T-C	69	N-N
ycf1R	5669	647-1275	7	T-G	1080	F-L
				A-C	1195	K-Q
				A-U	1225	T-S
				T-G	1246	F-V
				A-G	1269	G-G

Table 3.5 (Continued): Differences observed by comparison of *Solanum bulbocastanum* chloroplast genome sequences with EST sequences obtained by BLAST search in the NCBI genbank.

				C-A	1273	Q-T
				A-C	1274	

Discussion

Evolutionary implications

The analysis of repeated sequences in Solanaceae chloroplast genomes revealed 42 groups of repeats shared among various members of the family (Table 3.1, Fig. 3.1). The fact that 37 of these 42 repeats are found in all four genomes examined suggests a high level of conservation for repeat structure. Furthermore, examination of the location of these repeats in the four genomes indicates that all of them occur in the same regions; either in genes, introns or within intergenic spacers. This high level of conservation of both sequence identity and location suggests that these elements may play a conserved functional role in the genome.

Except for the large inverted repeat, repeated sequences have generally been considered to be relatively uncommon in chloroplast genomes (Parmer, 1991). One extraordinary exception is Chlamydomonas, which was estimated to have a genome comprised of more than 20% dispersed repeats (Maul et al., 2002). Dispersed repeats have also been identified in several families of flowering plants, including Trachelium (Cosner et al., 1997) (Campanulaceae), Trifolium (Parmer et al., 1988) (Fabaceae), wheat (Bowman and Dyer, 1986; Howe, 1985) (Poaceae), and Oenothera (Hupfer et al., 2000; Sears et al., 1996; Vomstein and Hachtel, 1988) (Onagraceae). All of these genomes have gene order changes, suggesting that the repeats may have played a role in these alterations. The chloroplast genomes of Solanaceae are not rearranged yet they still have a substantial number of repeats. A similar comparison of repeat structure among three legume chloroplast genomes (Chapter 2) also identified a substantial number of repeat elements. Thus, it is becoming evident that chloroplast genomes contain a substantial number of repeated sequences other than the

inverted repeat. Additional studies are needed to assess the possible functional role of these repeat elements.

Intergenic spacer regions are the most widely used chloroplast markers for phylogenetic investigations at lower taxonomic levels in plants (Raubeson and Jansen, 2005, Shaw et al., 2005). Plant phylogeneticists have utilized these markers because IGS regions are considered more variable and therefore should provide more characters. The first genome-wide comparisons of the levels of sequence conservation in the intergenic spacer regions of four Solanaceae chloroplast genomes (Table 3.2, Fig. 3.5A) demonstrate a wide range of sequence divergence in different regions. Furthermore, comparisons of coding (Fig. 3.5B) and non-coding (Fig. 3.5A) regions generally support the contention that intergenic spacer regions are more variable and could provide more phylogenetically informative characters for phylogenetic studies at lower taxonomic levels. Shaw et al., 2003, recently compared the phylogenetic utility of 21 noncoding chloroplast DNA regions. In their study, they ranked these 21 regions into three tiers based on their phylogenetic utility with tier one being the most useful by calculating the number of potentially informative characters. Although the genome-wide comparisons are based on sequence divergence, the results agree with the relative ranking of these regions in the Solanaceae (Fig. 3.5A). However, these comparisons have identified several intergenic regions that have higher sequence divergence than the most variable tier 1 regions identified by Shaw et al. (Shaw et al., 2003). Thus, these genome-wide comparisons provide valuable new information for the plant systematics community about the potential phylogenetic utility of the chloroplast intergenic spacer regions.

Comparisons of DNA and EST sequences identified a substantial number of differences. Many of these differences are not likely due to RNA editing because previous studies of both *Atropa* (Schmitz-Linneweber et al., 2002) and tobacco (Hirose et al., 1999) have indicated that RNA editing events are exclusively C-to-U changes. Analyses of both *Solanum bulbocastanum* and *Solanum lycopersicum* sequences (Tables 3.4 and 3.5) showed a lower number of C-to-U changes than previously observed for these species (Hirose et al., 1999; Schmitz-Linneweber et al., 2002). In addition, none of the C-to-U conversions observed in *Solanum bulbocastanum* and *Solanum lycopersicum* were conserved with respect to the previous observations in tobacco and *Atropa*. It is more likely that the differences observed between the DNA and EST sequences are due to polymorphisms within these species, or even errors in the EST sequences. However, if future studies in the Solanaceae confirm that these differences are real and due to RNA editing then it is possible that there has been a loss of conserved editing sites in *Solanum bulbocastanum* and *Solanum lycopersicum*. Evolutionary loss of RNA editing sites has been previously observed and could possibly be due to a decrease in the effect of RNA-editing enzymes (Mulligan et al., 2004). Additionally, a considerable number of variable sites other than C-to-U conversions were observed in *Solanum lycopersicum* and *Solanum bulbocastanum*, suggesting that these chloroplast genomes may be accumulating considerable amounts of nucleotide substitutions, and some of the genes accumulate more variable sites than others. This has been previously observed in several chloroplast genes, such as *petL* and *ndhH* genes, which have a high frequency of RNA editing (Fiebig et al., 2004). This suggests that, even though the chloroplast genome is relatively highly conserved among species, much of its variability could also be accounted for at the transcript level. The evidence that *ycf1* is a necessary gene in dicots (Drescher et al., 2000) and missing in

monocots (Maier et al., 1995) is an interesting case of selection. The observation in this study identifies a case of RNA editing and partial gene duplication of *ycf1*. This gene is essential for cell survival in dicots (Drescher et al., 2000) and missing in monocots.

Implications for integration of transgenes

Several intergenic spacer regions have been used to integrate foreign genes into the *Solanum lycopersicum* and *Solanum bulbocastanum* plastid genes based on tobacco chloroplast sequence. These spacer regions are located between the following genes: *trnfM* and *trnG*, *rbcL* and *accD*, *trnV* and 3'-*rps12*, and 16S rRNA and orf 70B 35, 36, 56. Unfortunately, none of these regions have 100% sequence identity to the tobacco flanking sequence used in plastid transformation vectors. *Solanum bulbocastanum* plastid transformants were generated at 10-30 times lower frequencies than tobacco (Nguyen et al., 2005) and the intergenic spacer region between *rbcL* and *accD* region shows only 94% identity. Similarly, the *trnfM* and *trnG* intergenic spacer region used for *Solanum lycopersicum* plastid transformation has only 82% sequence identity, resulting in inefficient transgene integration. There are major deletions in the *Solanum lycopersicum* chloroplast genome in this intergenic spacer region when compared to tobacco, which was used for plastid transformation (Ruf et al., 2001). These studies point out the importance of choosing appropriate intergenic spacers for plastid transformation. The use of these regions in and *Solanum lycopersicum* or *Solanum bulbocastanum* with 100% sequence identity (Table 3.2) might have enhanced recombination efficiency and thereby increased the success of plastid transformation. Additionally, if species-specific vectors are used, then one could use any of the intergenic spacer regions for transgene integration.

In addition to providing insight into genome organization and evolution, availability of complete DNA sequence of chloroplast genomes should facilitate plastid genetic

engineering. Although many successful examples of plastid engineering in tobacco have set a solid foundation for various future applications, this technology has not been extended to many of the major crops. Complete native chloroplast genome sequences provide valuable information on spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes and will help in extending this technology to *Solanum lycopersicum* and *Solanum bulbocastanum*.

CHAPTER 4

COMPLETE CHLOROPLAST GENOME SEQUENCES OF *HORDEUM VULGARE*, *SORGHUM BICOLOR* AND *AGROSTIS STOLONIFERA*, AND COMPARATIVE ANALYSES WITH OTHER GRASS GENOMES

Introduction

Sorghum (*Sorghum bicolor*), with 25 species, is a member of the family Poaceae and tribe Andropogoneae (Garber 1950). Recent molecular phylogenetic analyses indicated that the genus may be paraphyletic (Spangler et al., 1999), and that it is comprised of three distinct lineages; Sorghum, Sarga, and Vacoparis (Spangler 2003). The genus Sorghum was redefined to include three species, *Sorghum bicolor*, *S. halepense*, and *S. nitidum*. *Sorghum bicolor*, cultivated grain sorghum, is the third most important cereal crop in the United States and the fifth most important crop in the world (Crop Plant Resources, 2000). Sorghum is well known for its capacity to tolerate conditions of limited moisture and to produce a harvest during periods of extended drought; circumstances that would impede production in most other grains (Crop Plant Resources, 2000). Sorghum is used for human nutrition and feed grain for livestock throughout the world (Carter et al. 1989). A more recent use of Sorghum is the production of ethanol, with one bushel producing the same amount of ethanol as one bushel of corn (National Sorghum Producers 2006). Some Sorghum varieties are rich in anti-oxidants and all varieties are gluten-free, an attractive alternative for those allergic to *Triticum aestivum* (US Grains Council 2006).

Of the various cereals, *Hordeum vulgare* L. (barley) is a major food, feed and malt crop. In 2005, *H. vulgare* ranked fourth in quantity produced and in area of cultivation of cereal crops in the world (http://faostat.fao.org/faostat/) demonstrating its broad consumption

and wide adoption in a variety of climates, from sub-arctic to sub-tropical. The United States is the eighth largest producer of *H. vulgare* in the world with current production estimated at 4.9 million acres. It is a short-season, early maturing crop grown on both irrigated and dry land production areas in the United States. Whole grain *H. vulgare* contains high levels of minerals and important vitamins, including calcium, magnesium, phosphorus, potassium, vitamin A, vitamin E, niacin and folate.

Among the non-food grasses, *Agrostis stolonifera* L. (creeping bentgrass) has attracted great attention in both academia and the biotech industry due to its social and economic importance. A. stolonifera is a wind-pollinated, highly outcrossing perennial grass used on golf courses worldwide. It can also enhance the natural beauty of the environment and increase the value of residential and commercial property, and provide many environmental benefits including preventing soil erosion, filtering water, and trapping dust and pollutants (Bonos et al. 2006). It has been extensively used, covering millions of acres globally, making it an economically valuable grass crop. Due to its aforementioned importance, transgenic *A. stolonifera* was produced conferring herbicide resistance (glyphosate) by engineering the CP4 EPSPS gene, which is one of the first transgenic, perennial, wind-pollinated crops grown outside of a typical agronomic environment (Wipff and Fricker 2001, Watrud et al. 2004, Reichman et al., 2006). Unfortunately, pollen-mediated transgene flow has been reported in several studies (Wipff and Fricker 2001, Watrud et al. 2004, Reichman et al., 2006) limiting its commercialization and demonstrating the requirement of effective containment strategies to protect the environment and to engineer this plant with environmentally friendly approaches like chloroplast engineering or cytoplasmic male sterility.

The agronomic, economic and/or social importance of *H. vulgare*, *S. bicolor* and *A. stolonifera* has made them the focus of numerous genetic studies attempting to improve these crop species. Much of this work has been restricted to investigations of nuclear genomes for these species (USDA, Cheng et al., 2004). This has resulted in very limited information on the organization and evolution of chloroplast genomes of *H. vulgare*, *S. bicolor* and *A. stolonifera*. This study aims to enhance our understanding of the chloroplast genome organization, evolution, and relationship among the grasses facilitating the improvement of those crops by chloroplast genetic engineering.

In this chapter, the complete sequence of the chloroplast genomes of *H. vulgare*, *S. bicolor* and *A. stolonifera* are presented. One goal is to compare the genome organization of *H. vulgare*, *S. bicolor* and *A. stolonifera* with six other completely sequenced grass chloroplast genomes; *Oryza sativa, O. nivara, Saccharum hybrid, S. officinarum, T. aestivum*, and *Z. mays*. In addition to examining gene content and gene order, the distribution and location of repeated sequences among these genomes are determined, including potential microsatellite markers. A second goal is to compare levels of DNA sequence divergence of non-coding regions. Intergenic spacer regions have been examined to identify ideal insertion sites for transgene integration, and to assess the utility of these regions for resolving phylogenetic relationships among closely related species (Kelchner 2002, Shaw et al., 2005, Timme et al., 2007). A third goal of this study is to examine the extent of RNA editing in the *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes by comparing the DNA sequence with available expressed sequence tag (EST) sequences. RNA editing is a co- or post-transcriptional process that occurs in organelles and changes the coding information in mRNAs (Kugita et al. 2003, Wolf et al. 2004). Most of our knowledge about the frequency of this process in crop plants

comes from studies in *Z. mays* (Maier et al., 1995) and *Nicotiana tabacum* (Hirose et al., 1999), and additional comparative studies are needed in other plant species to understand the extent of RNA editing in chloroplast genomes. A final goal is to assess phylogenetic relationships between *H. vulgare, S. bicolor, A. stolonifera* and other completely sequenced angiosperm chloroplast genomes.

Methodology

DNA Sources

Bacterial artificial chromosome (BAC) libraries of *H. vulgare* cv Morex (Yu et al., 2000) and *S. bicolor* cv BTX623 (CUGI, unpublished) were constructed by ligating size fractionated partial *Hin*dIII digests of total cellular, high molecular weight DNA with the pINDIGOBAC536 vector. The average insert size of *H. vulgare* (HV_MBa) and *S. bicolor* (SB_BBc) libraries was 106 kb and 120 kb, respectively.

The *A. stolonifera* L. cultivar Penn A-4 was supplied by HybriGene, Inc. (Hubbard, OR). Prior to chloroplast isolation, plants were kept in dark for two days to reduce levels of starch. Chloroplasts from young leaves were isolated using the sucrose step gradient method of Palmer (1986) as modified by Jansen et al. (2005). About 10 g of leaf tissue was homogenized in Sandbrink isolation buffer using pre-chilled tissue blender bursts at high speed for five seconds to get sufficient quantities of chloroplasts. The homogenate was filtered using four layers of cheesecloth and one layer of miracloth (Calbiochem, Cat# 474855) without squeezing. The filtrate was transferred to pre-chilled centrifuge tubes and centrifuged at 1000 g for 15 min at 4ºC. Pellets were resuspended in 7 ml of ice-cold wash buffer and gently loaded over the step gradient consisting of 18 ml of 52% sucrose, over-layered with 7 ml of 30% sucrose. The sucrose step gradient was centrifuged at 25,000 rpm

for 30-60 min at 4° C in a SW-27 rotor (Beckman). The chloroplast band from the 30%-52% interface was removed using a wide bore pipette, diluted with 10 volumes wash buffer, and centrifuged at 1,500 g for 15min at 4° C. Purified chloroplast pellets were resuspended in a final volume of 2 ml. The entire chloroplast genome was amplified by Rolling Circle Amplification (RCA) using the Repli-g RCA kit (Qiagen, Inc.) following the methods described in (Jansen et al., 2005). RCA was performed at 30° C for 16 hr; the reaction was terminated with final incubation at 65°C for 10 min. Digestion of the RCA product with the restriction enzymes BstXI, EcoRI and HindIII verified successful genome amplification, as well as DNA quality for sequencing.

Chloroplast BAC clone identification/selection, sequencing protocols, sequence assembly, annotation, and pairwise comparisons among taxa were performed as described in Chapter 2.

Molecular Evolutionary Comparisons

Gene content comparisons were performed with Multipipmaker (Schwartz et al., 2003). Comparisons included nine genomes: O. sativa (NC_001320, Hiratsuka et al., 1989), *O. nivara* (NC_005973, Shahid-Masood et al. 2004), *S. officinarum* (NC_006084, Asano et al. 2004), *Saccharum hybrid* (NC_005878, Calsa et al., unpublished), *T. aestivum* (NC_002762, Ogihara et al. 2000), *Z. mays* (NC_001400, Maier et al., 1995), *H. vulgare* (EF115541, current study), *S. bicolor* (EF115542, current study) and *A. stolonifera* (EF115543, current study) using *O. sativa* as the reference genome. Gene orders were examined by pair-wise comparisons between the above genomes using PipMaker (Elnitski et al. 2002).

Shared and unique repeats were identified for *H. vulgare*, *S. bicolor* and *A. stolonifera* genomes and compared to other grass genomes using Comparative Repeat Analysis (CRA,

Holtshulte and Wyman unpublished, http://bugmaster.jgi-psf.org/repeats/). This program filters the redundant output of REPuter (Kurtz et al., 2001) and identifies shared repeats among the input genomes. For repeat identification, the following constraints were set in CRA: a minimum repeat size of 30 bp and a Hamming distance of 3 (i.e., a sequence identity of ≥ 90 %). *Oryza sativa* was used as the reference genome. Blast hits 30 bp and longer with a sequence identity of ≥ 90% were identified to determine the shared repeats among the seven genomes examined. To detect SSRs, the Perl script CUGISSR (Jung et. al. 2005), was used to search for SSRs ranging from di-to penta-nucleotide repeats.

Intergenic spacer regions from seven grass chloroplast genomes were compared using MultiPipMaker (Schwartz et al. 2003, http://pipmaker.bx.psu.edu/pipmaker/tools.html). As described in Chapter 3, two Perl scripts that utilize the all_bz module for intergenic comparisons were used to calculate percent identity estimates.

Each of the genes from the *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes were used to perform a BLAST search of expressed sequence tags (ESTs) from the NCBI Genbank. The retrieved EST sequences from *A. stolonifera*, *H. vulgare* and *S. bicolor* were then aligned with the corresponding annotated gene for each species separately, using Clustal X. The aligned sequences were then screened and nucleotide and amino acid changes were detected using the Megalign software and the plastid/bacterial genetic code. Due to variation in length between an EST and the corresponding gene, the length of the analyzed sequence was recorded.

Phylogenic Analysis

The 61 genes included in the analyses of Goremykin et al. (2003a, 2004, 2005), Leebens-Mack et al. (2005), Chang et al. (2006), Lee et al. (2006), Jansen et al. (2006), and Ruhlman et al. (2006) were extracted from the chloroplast genome sequence of *A. stolonifera*, *H. vulgare* and *S. bicolor* using DOGMA (Wyman et al. 2004). The same set of 61 genes was extracted from chloroplast genome sequences of 35 other sequenced genomes. All 61 protein-coding genes of the 38 taxa were translated into amino acid sequences, aligned using MUSCLE (Edgar 2004) followed by manual adjustments for gaps, and then nucleotide sequences of these genes were aligned by constraining them to the aligned amino acid sequences. A Nexus file with character sets for phylogenetic analyses was generated after nucleotide sequence alignment was completed. The complete nucleotide alignment is available online at Chloroplast Genome Database (Cui et al., 2006, http://chloroplast.cbio.psu.edu).

Phylogenetic analyses using maximum parsimony and maximum likelihood were performed with PAUP* version 4.10b10 (Swofford 2003) and GARLI version 0.942 (Zwickl 2006, http://www.bio.utexas.edu/grad/zwickl/web/garli.html), respectively. Phylogenetic analyses excluded gap regions to avoid alignment ambiguities in regions with variation in sequence lengths. All MP searches included 100 random addition replicates and TBR branch swapping with the Multrees option. Non-parametric bootstrap analyses (Felsenstein 1985) were performed for MP analyses with 1000 replicates with TBR branch swapping, one random addition replicate, and the Multrees option. Modeltest 3.7 (Posada and Crandall 1998) was used to determine the most appropriate model of DNA sequence evolution for the combined 61-gene dataset. For maximum likelihood analyses in GARLI, two independent runs were performed using the default settings (see Garli manual at

http://www.bio.utexas.edu/grad/zwickl/web/garli.html). Non-parametric bootstrap analyses (Felsenstein 1985) were performed in GARLI for maximum likelihood analyses using default settings.

Results

Size, gene content and organization of the H. vulgare, S. bicolor and A. stolonifera chloroplast genomes

The complete sizes of the *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes are 136,462, 140,754 bp and 136,584 bp, respectively (Fig. 4.1). The genomes include a pair of inverted repeats of 21,579 bp (*H. vulgare*), 22,782 bp (*S. bicolor*) and 21,649 bp (*A. stolonifera*) separated by a small single copy region of 12,704 bp (*H. vulgare*), 12,502 bp (*S. bicolor*) and 12,740 bp (*A. stolonifera*) and a large single copy region of 80,600 bp (*H. vulgare*), 82,688 bp (*S. bicolor*) and 80,546 bp (*A. stolonifera*).

The *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes contain 113 different genes, and 18 of these are duplicated in the IR, giving a total of 131 genes (Fig. 4.1). There are 30 distinct tRNA genes, and 7 of these are duplicated in the IR. Sixteen genes contain one or two introns, and six of these are in tRNAs. The *H. vulgare* chloroplast genome consists of 56.7% coding regions that include 48% protein coding genes, 8.7% RNA genes and 43.3% non-coding regions, containing both intergenic spacer regions and introns. The *S. bicolor* chloroplast genome is composed of 52.1% coding regions that include 43.4% protein coding genes, 8.7% RNA genes and 47.9% non-coding regions. The *A. stolonifera* chloroplast genome is composed of 53.6% coding regions that include 44.7% protein coding genes, 8.9% RNA genes and 46.4% non-coding regions. The overall GC and AT content of the *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes are 38.31% (*H. vulgare*), 38.50%

(*S. bicolor*), 38.45% (*A. stolonifera*) and 61.69% (*H. vulgare*), 61.50% (*S. bicolor*) and 61.55% (*A. stolonifera*), respectively.

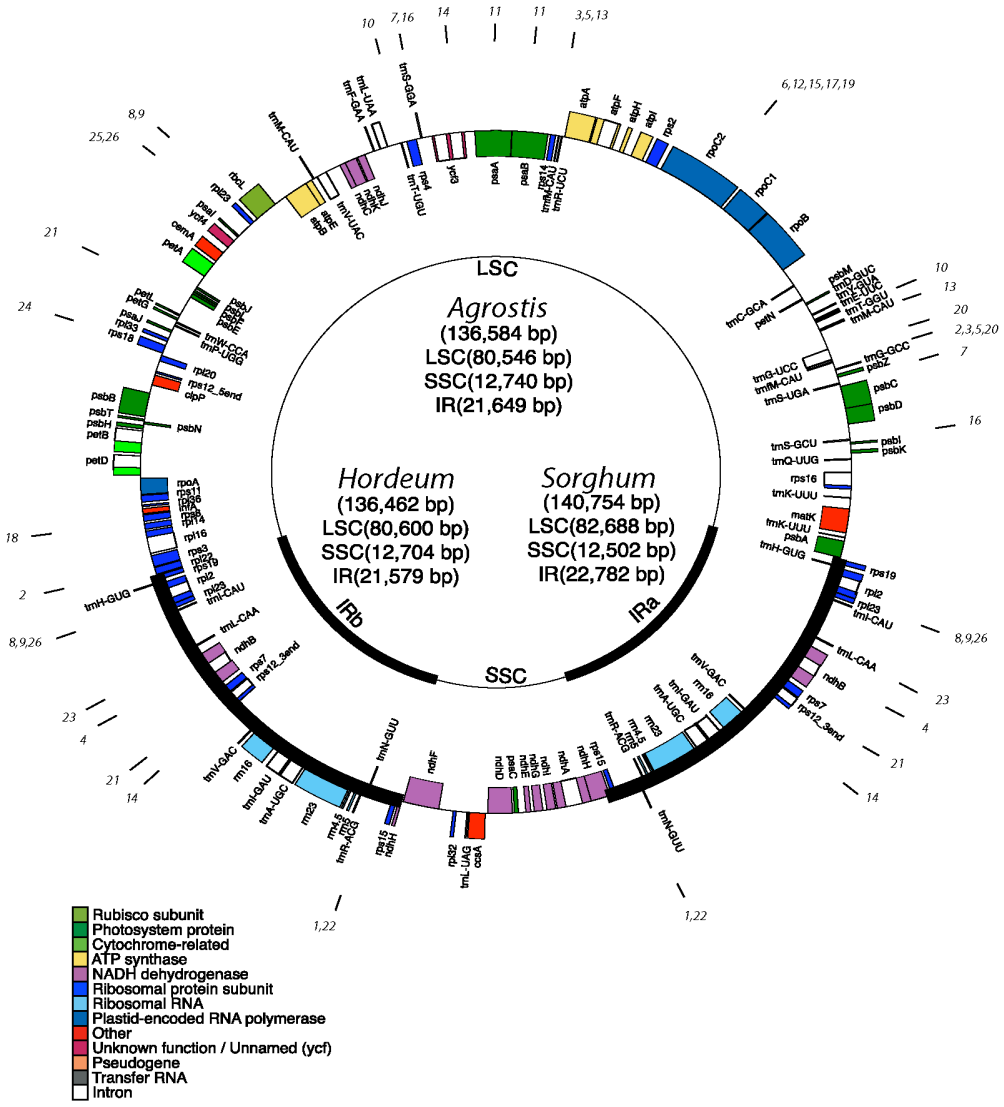

Fig 4.1. Gene map of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera* chloroplast genomes. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into small (SSC) and large (LSC) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction. Demarcations on the outside of the map indicate repeat number and location.

Gene Content and Order

Gene content and order of the *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes are similar to the other six sequenced grass chloroplast genomes (*O. sativa*, *O. nivara*, *Saccharum hybrid*, *S. officinarum*, *T. aestivum*, and *Z. mays*). Like other grass chloroplast genomes, the IR in *H. vulgare*, *S. bicolor* and *A. stolonifera* has expanded to include rps19. However, the extent of the IR at the SSC/IRa boundary differs between two of the genomes with the IR of *H. vulgare* and *A. stolonifera* expanded to duplicate a portion of ndhH, a feature that is shared with the *T. aestivum* chloroplast genome (Ogihara et al., 2000). This expansion includes 207 bp (69 amino acids) in *H. vulgare*, 174 bp (58 amino acids) in *A. stolonifera*, and 96 bp (32 amino acids) in *T. aestivum*. The *H. vulgare*, *S. bicolor* and *A. stolonifera* genomes also share the loss of introns in *clpP* and *rpoC1* with other grasses. There are insertions and deletions (indels) of nucleotides within several coding sequences. For example, CAAAAC is uniquely present within *matK* of *S. bicolor*, but absent in the rest of the grasses examined (Figure 4.2). There is also a 6 bp deletion in the *ndhK* gene in *H. vulgare*, *A. stolonifera*, *T. aestivum* and both species of *Oryza* (Figure 4.2).

ndhK

```
551 AGGATCGAACTCTATGTCAAAGTCAAAAGAAAAATAGATCTTTTACTACC S.hybrid
551 AGGATCGAACTCTATGTCAAAGTCAAAAGAAAAATAGATCTTTTACTACC S.officinarium
551 AGGATCGAACTCTATGTCAAAGTCAAAAGAAAAATAGATCTTTTACTACC S.bicolor
551 AGGATCGAACTCTATGTCAAAGTCAAAAGAAAAATAGATCTTTTACTACC Z.mays
551 AGGATCGAACTCTATCTCAAA------ATAAAAAAAGATGTTTTACTACC H.vulgare
551 AGGATCGAACTCGATCTCAAA------ATAAAAATAGATGTTTTACTACC A.stolonifera
551 AGGATCGAACTCTATCTCAAA------ATAAAAATAGATGTTTTACTACC T.aestivum
551 AGGATCGAACTCTATCTCAAA------AGAAAAATCGATGTTTTACTACC O.sativa
551 AGGATCGAACTCTATCTCAAA------AGAAAAATCGATGTTTTACTACC O.nivara
```

matK

```
1420 TTTTTTCTTTGATGTTCACCAAAAC------AACTCTTTTTTCTTTCAGT H.vulgare
1421 TTTTTTCTTTGATGTTCACCAAAAC------AAGCCTTTTTTCTTTCCGT A.stolonifera
1514 TTTTTTCTTTGATGTTCGCCAAAAC------AACTTACTTTTCTTTCCGG T.aestivum
1482 TTTTTTCTTTGATGTTCACCAAAAC------AATTCACTTTTCTTTCCAT S.officinarium
1514 TTTTTTCTTTGATGTTCACCAAAAC------AATTCACTTTTCTTTCCAT S.hybrid
1461 TTTTTTCTTTGATGTTCACCAAAACCAAAACAATTCACTTTTCTTTCCAT S.bicolor
1514 TTTTTTCTTTGATGTTCGCCAAAAC------AACTTACTTTTCTTTCCGT O.sativa
1514 TTTTTTCTTTGATGTTCGCCAAAAC------AACTTACTTTTCTTTCCGT O.nivara
1514 TTTTTTCTTTGATGTTCACCAAAAC------AATTCACTTTTCTTTCCAT Z.mays
```

Fig. 4.2 Alignment of a portion of the *ndhK* and *matk* genes illustrating a deletion within *H. vulgare*, *T. aestivum*, *A. stolonifera* and both *O. sativa* chloroplast genes of *ndhK* and an insertion unique to *S. bicolor* in the *matK* gene.

Repeat Structure

Repeat analyses identified 19 to 37 direct and inverted repeats 30 bp or longer with a sequence identity of at least 90% among the nine chloroplast genomes examined (Figure 4.3, Table 4.1). With one exception of 91 bp repeat, all other repeats range in size between 30 and 60 bp, and 78.4% are in the direct orientation while 21.6% are inverted. The longest repeats other than the inverted repeats found in *H. vulgare* and *S. bicolor* are 540 and 524 bp, respectively. BlastN comparisons of the *O. sativa* repeats against the chloroplast genomes of the eight other grasses identified 26 shared repeats ≥ 30 bp with a sequence identity ≥ 90% (Table 4.1*).* *H. vulgare* and *T. aestivum* share four repeats (31, 32, 36, and 38 bp) not found in any other genomes. Both *Oryza* species share 41 and 59 bp repeats. *Zea mays* has the most repeats with 37 and *A. stolonifera* has the fewest with 19. Seventeen of the 26 repeats are found in all eight chloroplast genomes and all of these are located in the same genes or intergenic spacer regions.

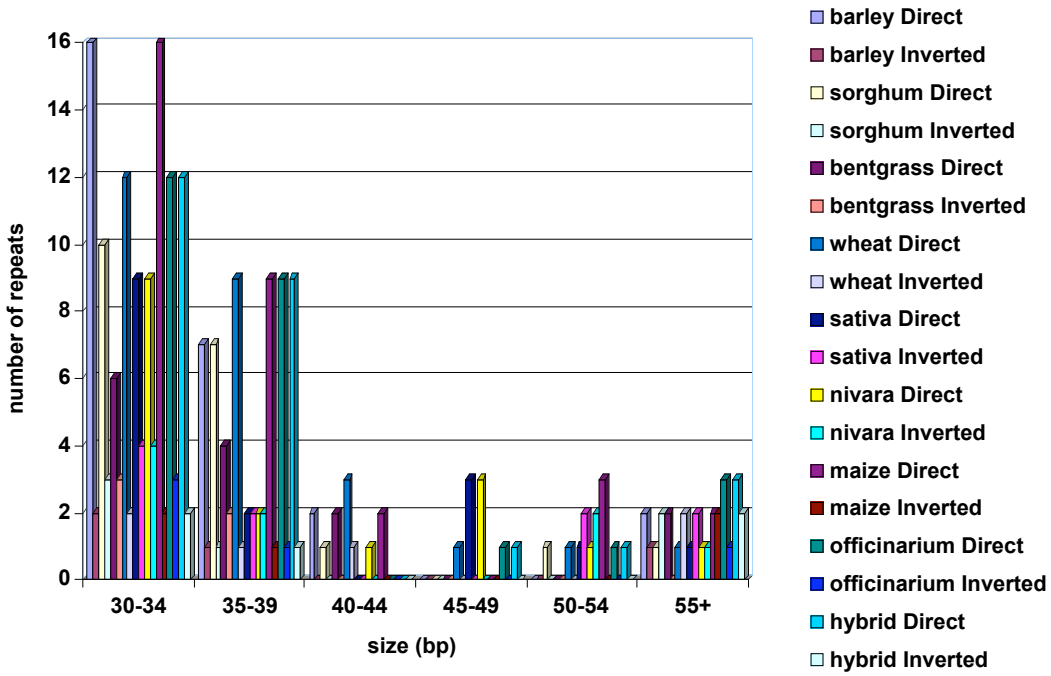

Fig. 4.3. Histogram showing the number of repeated sequences ≥ 30 bp long with a sequence identity ≥ 90% in nine grass chloroplast genomes

Table 4.1 Oryza sativa repeats blasted against all eight chloroplast genomes. Includes blast hits at least 30 bp in size, a sequence identity ≥ 90%, and a bit-score of great than 40. Sb = *Sorghm bicolor*, On = *Oryza nivara*, Ta = *Triticum aestivum*, Hv = *Hordeum vulgare*, Sh = *Saccharum hybrid*, So = *Saccharum officinarum*, Zm = *Zea mays*, As = *Agrostis stolonifera*.

Repeat Number	Size (bp)	Number of copies	Orientation	Location	Genomes
1	30	2	Direct	IGS – (*trnN*-GUU-*rps15*)	Sb,So,Sh,On,Zm
2	30	2	Direct	*rps3*	Sb,On,Ta,Hv,Sh,So,Zm,As
3	30	2	Direct	IGS – (*trnM*-CAU-*trnG*-UCC), *trnM*-CAU	Sb,On,Ta,Hv,Sh,So,Zm,As
4	30	2	Direct	Intron – (*ndhB*)	Sb,On,Hv,Sh,So,Zm,As
5	31	3	Direct	IGS – (*trnG*-GCC – *trnM*-CAU), IGS – (*trnM*-CAU – *rps14*)	Sb,On,Ta,Hv,Sh,So,Zm,As
6	31	2	Direct	*rpoC2*	Sb,On,Sh,So,Zm, As
7	32	2	Inverted	*trnS*-UGA	Sb,On,Ta,Hv,Sh,So,Zm,As
8	32	3	Inverted	*rpl23*	Sb,On,Ta,Hv,Sh,So,Zm,As
9	32	3	Inverted	*rpl23*	Sb,On,Ta,Hv,Sh,So,Zm,As
10	33	2	Inverted	*trnT*-GGU	Sb,On,Ta,Hv,Sh,So,Zm,As
11	34	2	Direct	*psaB, psaA*	Sb,On,Ta,Hv,Sh,So,Zm,As
12	34	2	Direct	*rpoC2*	Sb,On,Ta,Hv,Sh,So
13	34	2	Direct	*trnfM-CAU*	Sb,On,Ta,Hv,Sh,So,Zm,As
14	36	3	Inverted	Intron – (*ycf3* Exon1 – *ycf3* Exon2), IGS – (*trnV*-GAC – *rps12*_3end)	Sb,On,Ta,Hv,Sh,So,Zm,As
15	36	3	Direct	*rpoC2*	Sb,On,Ta,Hv,Sh,So,Zm,As
16	36	2	Inverted	*trnS*-GCU	Sb,On,Ta,Hv,Sh,So,Zm,As
17	37	2	Direct	*rpoC2*	Sb,On,Ta,Hv,Sh,So,Zm,As
18	45	3	Direct	*rps8*	Sb,On,Ta,Hv,Sh,Zm,As
19	45	2	Direct	*rpoC2*	Sb,On,Ta,Sh,So,Zm,As

Table 4.1 (Continued) Oryza sativa repeats blasted against all eight chloroplast genomes. Includes blast hits at least 30 bp in size, a sequence identity ≥ 90%, and a bit-score of great than 40. Sb = *Sorghm bicolor*, On = *Oryza nivara*, Ta = *Triticum aestivum*, Hv = *Hordeum vulgare*, Sh = *Saccharum hybrid*, So = *Saccharum officinarum*, Zm = *Zea mays*, As = *Agrostis stolonifera*.

20	47	2	Direct	IGS – (*trnG*-GCC – *trnfM*-CAU), Intron – (*trnfM*-CAU – *trnG*-UCC	On,Ta
21	50	3	Inverted	IGS - (*psbE – petL*), Intron – (*rps12_3end – rps7*)	Sb,On,Ta,Hv,Sh ,So,Zm,As
22	52	2	Direct	IGS – (*trnN*-GUU-*rps15*)	Sb,On,Ta,Hv,Sh ,So,Zm,As
23	52	4	Inverted	IGS – (*ndhB-trnL*-CAA)	Sb,On,Ta,Hv,Sh ,So,Zm,As
24	56	2	Direct	*rps18*	Sb,On,Sh,So,Z m,As
25	59	2	Inverted	IGS –(*psaI-rpl23)*	On
26	91	3	Inverted	*rp123* (69 bp) – IGS (*rp123 – accD*), *rp123* (79 bp) – IGS (*rp123 – rp12*)	Sb,On,Ta,Hv,Sh ,So,Zm,As

Previous studies of grass chloroplast genomes identified three inversions relative to the established consensus chloroplast gene order identical to that found in tobacco (Hiratsuka et al., 1989, Doyle et al., 1992, Palmer and Stein 1986). Because inversions are often associated with repeated sequences (Palmer 1991) the inversion endpoint regions were examined for repeats. Shared repeats flanking the endpoints of the largest 28 kb inversion of grasses were identified. Repeat analyses identified a 21 bp direct repeat in *O. sativa* that contains the motif GTGAGCTACCAAACTGCTCTA and flanks the inversion endpoints. This repeat has a Hamming distance of 2, and is shared by all the other grasses examined. Repeat analyses at the endpoints of the two other grass inversions failed to identify any shared repeats at the settings used in this analysis.

Simple sequence repeat analyses identified 16-21 SSRs per chloroplast genome and these are composed of di- to penta- nucleotide repeating units (Table 4.2). Nearly 50% of all SSRs are tetra-nucleotide repeats with no common motif. The next most common SSR consists of di-nucleotide repeats and accounts for 30% of the SSRs with a predominant motif of TA or AT. The remaining 20% of the SSRs are composed of tri- and penta-nucleotide repeats. Of the SSRs identified, the same di-nucleotide repeat (AT) is located within the coding region of the gene rpoC2 in all chloroplast genomes examined.

Table 4.2. Simple sequence repeats in the nine grass chloroplast genomes examined. Table shows motif, number of repeated elements, location, and presence within an ORF.

		# SSRs	Motif	# Repeats	Start	Stop	INORF	ORF ID
A.stolonifera	140754	17	ttat	4	11012	11023	Y	*rpoC2*
			aat	3	24240	24251	N	
			at	2	25539	25548	N	
			tcct	4	42353	42364	N	
			cttat	5	47561	47575	N	
			aaat	4	65509	65520	N	
			agaa	4	68264	68275	N	
			ta	2	84762	84771	N	
			aacg	4	98980	98991	N	
			caa	3	105494	105505	N	
			aaca	4	105501	105512	N	
			atta	4	105588	105599	N	
			aata	4	107654	107665	N	
			ct	2	114612	114623	N	
			tcgt	4	117977	117988	N	
			ta	2	132196	132205	N	
H.vulgare	**136462**	**21**	at	5	26364	26373	Y	*rpoC2*
			at	7	56573	56586	N	
			ta	6	15124	15135	N	
			ta	5	85456	85465	N	
			ta	5	132669	132678	N	
			tc	5	115218	115227	N	
			aat	4	25059	25070	N	
			aat	4	64188	64199	N	
			taa	4	50799	50810	N	
			ttc	4	65709	65720	N	
			aaca	3	106006	106017	N	
			aacg	3	99520	99531	N	
			aaga	3	72365	72376	N	
			aata	3	108235	108246	N	
			agaa	3	68964	68975	N	

Table 4.2 (Continued). Simple sequence repeats in the nine grass chloroplast genomes examined. Table shows motif, number of repeated elements, location, and presence within an ORF.

			taga	3	116703	116714	N	
			tcct	3	43301	43312	N	
			tcgt	3	118602	118613	N	
			tcta	3	101420	101431	N	
			ttca	3	64665	64676	N	
			ccata	3	44175	44189	N	
O.sativa	**134525**	**16**	ag	5	3223	3232	N	
			at	5	25478	25487	Y	*rpoC2*
			ct	5	36589	36598	N	
			tc	5	113474	113483	N	
			aat	4	24183	24194	N	
			tct	4	80517	80528	N	
			tat	4	108670	108681	N	
			taaa	4	4152	4167	N	
			cttt	3	15220	15231	N	
			gtag	4	51285	51300	N	
			aata	3	55770	55781	N	
			agaa	3	68356	68367	N	
			ttta	3	71703	71714	N	
			aacg	3	98267	98278	N	
			aata	3	106600	106611	N	
			tcgt	3	116839	116850	N	
O.nivara	**134494**	**18**	ag	5	3222	3231	N	
			at	5	25412	25421	Y	*rpoC2*
			ct	5	36523	36532	N	
			tc	5	113440	113449	N	
			aat	4	24117	24128	N	
			tct	4	80469	80480	N	
			tat	4	108629	108640	N	
			taaa	4	4151	4166	N	
			cttt	3	15157	15168	N	
			gtag	4	51207	51222	N	

Table 4.2 (Continued). Simple sequence repeats in the nine grass chloroplast genomes examined. Table shows motif, number of repeated elements, location, and presence within an ORF.

			aata	3	55705	55716	N	
			agaa	3	68286	68297	N	
			ttta	3	71634	71645	N	
			aacg	3	98218	98229	N	
			aaca	3	104470	104481	N	
			aata	3	106559	106570	N	
			tcgt	3	116809	116820	N	
			aaagt	3	57560	57574	N	
S.officinarum	**141182**	**16**	at	5	28187	28196	Y	rpoC2
			ta	5	67037	67046	N	
			ta	5	88487	88496	N	
			tc	5	117973	117982	N	
			ta	5	135735	135744	N	
			ctt	4	82941	82952	N	
			aaag	3	6174	6185	N	
			tcct	3	45521	45532	N	
			gtag	4	54633	54648	N	
			agaa	3	70894	70905	N	
			aacg	3	102837	102848	N	
			attg	3	108384	108395	N	
			aata	3	111094	111105	N	
			atcc	3	117870	117881	N	
			tcgt	3	121382	121393	N	
			tataa	3	21020	21034	N	
S.*hybrid*	**141182**	**16**	ta	5	8930	8939	N	
			tc	5	38416	38425	N	
			ta	5	56179	56188	N	
			at	5	89814	89823	Y	*rpoC2*
			ta	5	128664	128673	N	
			ctt	4	3384	3395	N	
			aacg	3	23280	23291	N	
			attg	3	28827	28838	N	

Table 4.2 (Continued). Simple sequence repeats in the nine grass chloroplast genomes examined. Table shows motif, number of repeated elements, location, and presence within an ORF.

			aata	3	31537	31548	N	
			atcc	3	38313	38324	N	
			tcgt	3	41825	41836	N	
			aaag	3	67800	67811	N	
			tcct	3	107148	107159	N	
			gtag	4	116260	116275	N	
			agaa	3	132520	132531	N	
			tataa	3	82647	82661	N	
S.bicolor	**140754**	**16**	at	5	28526	28535	Y	*rpoC2*
			ct	5	53726	53735	N	
			ta	5	67248	67257	N	
			ta	5	88644	88653	N	
			tc	5	118078	118087	N	
			ta	5	135829	135838	N	
			tta	4	39073	39084	N	
			ctt	4	83099	83110	N	
			tcct	3	45723	45734	N	
			gtag	4	54852	54867	N	
			agaa	3	71090	71101	N	
			aacg	3	103001	103012	N	
			attg	3	108508	108519	N	
			aata	3	111197	111208	N	
			atcc	3	117975	117986	N	
			tcgt	3	121469	121480	N	
T.aestivum	134545	21	ag	5	3235	3244	N	
			tc	5	14936	14945	N	
			ta	5	14959	14968	N	
			at	5	26191	26200	Y	*rpoC2*
			at	6	41788	41799	N	
			at	5	56570	56579	N	
			tc	5	113634	113643	N	
			aat	5	24888	24902	N	

Table 4.2 (Continued). Simple sequence repeats in the nine grass chloroplast genomes examined. Table shows motif, number of repeated elements, location, and presence within an ORF.

			tat	4	47730	47741	N	
			ttc	4	64988	64999	N	
			tcct	3	43164	43175	N	
			ttca	3	63925	63936	N	
			ttct	3	64227	64238	N	
			agaa	3	68245	68256	N	
			aaga	3	71631	71642	N	
			aacg	3	97881	97892	N	
			aata	3	106646	106657	N	
			tcgt	3	117001	117012	N	
			ataga	3	17184	17198	N	
			ccata	3	44040	44054	N	
			tttat	3	44785	44799	N	
Z.mays	140384	19	at	5	27734	27743	Y	*rpoC2*
			at	5	48185	48194	N	
			ta	6	66388	66399	N	
			ta	5	87788	87797	N	
			tc	5	117222	117231	N	
			ta	5	134940	134949	N	
			tat	5	20596	20610	N	
			ctt	4	82245	82256	N	
			aaat	3	18157	18168	N	
			tcct	3	44968	44979	N	
			gtag	4	54086	54101	N	
			agaa	3	70272	70283	N	
			accg	3	74068	74079	N	
			aacg	3	102116	102127	N	
			attg	3	107643	107654	N	
			agat	3	110050	110061	N	
			aata	3	110340	110351	N	
			atcc	3	117119	117130	N	
			tcgt	3	120609	120620	N	

Intergenic Spacer Regions

The similarity and divergence of intergenic spacer regions from seven grass chloroplast genomes including *A. stolonifera, H. vulgare, Z. mays, O. sativa*, *S. bicolor, S. officinarum* and *T. aestivum* were analyzed as in Chapter 3. The results of these analyses are presented in, Figures 4.4 and 4.5, and Tables 4.3 and 4.4. These species were subdivided into two groups for comparative analyses based on their position in phylogenetic trees (Figs. 4.4, 4.5). The first group includes *O. sativa*, *T. aestivum*, *H. vulgare* and *A. stolonifera* and the second group contains *Z. mays*, *S. officinarum* and *S. bicolor*.

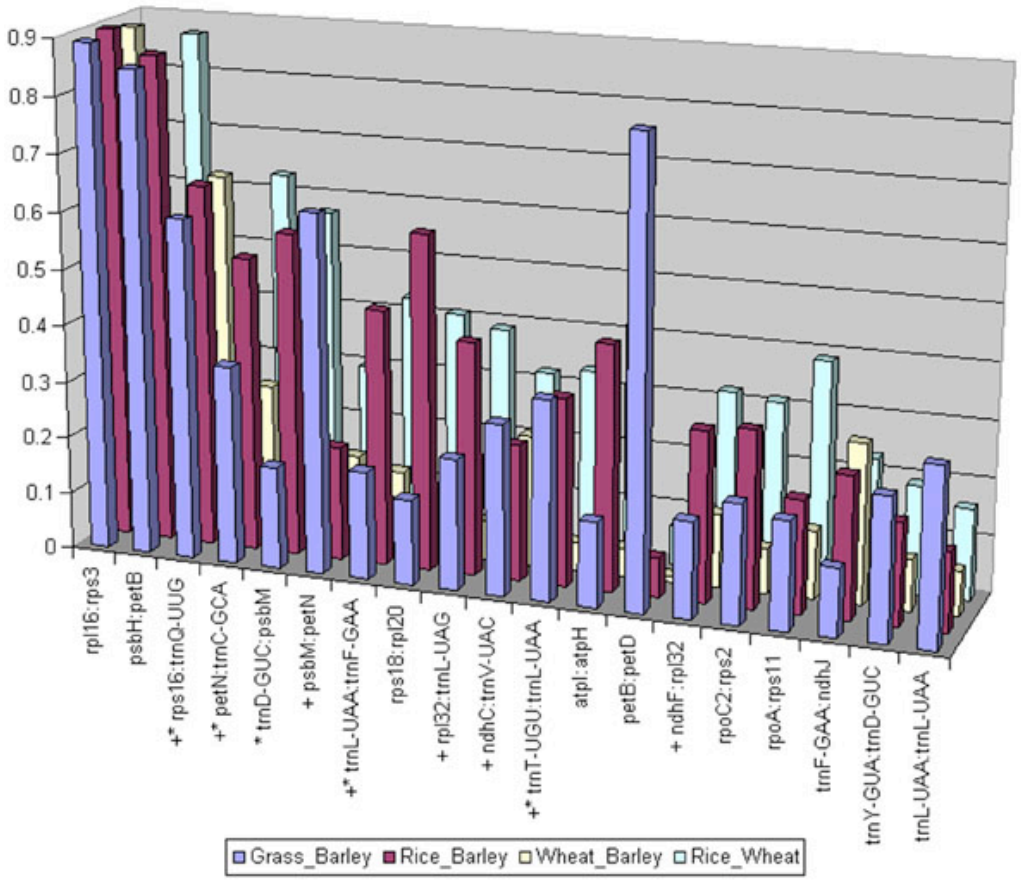

Fig. 4.4. Histogram showing pairwise sequence divergence of the intergenic spacer regions of rice (*Oryza sativa*), wheat (*Triticum aestivum*) barley (*Hordeum vulgare*) and bentgrass (*Agrostis stolonifera*) chloroplast genomes. Comparisons of 19 most variable intergenic regions with less than 80% average sequence identity. The values plotted in this histogram show percent sequence identities for all intergenic spacer regions. The plotted values were converted from percent identity to sequence divergence on a scale from 0 to 1 and included on the Y-axis. * indicate regions that are in the top 25 most variable intergenic spacer regions in Solanaceae, + indicate regions that are in the top 25 most variable intergenic spacer regions in Asteraceae (Timme et al. 2007).

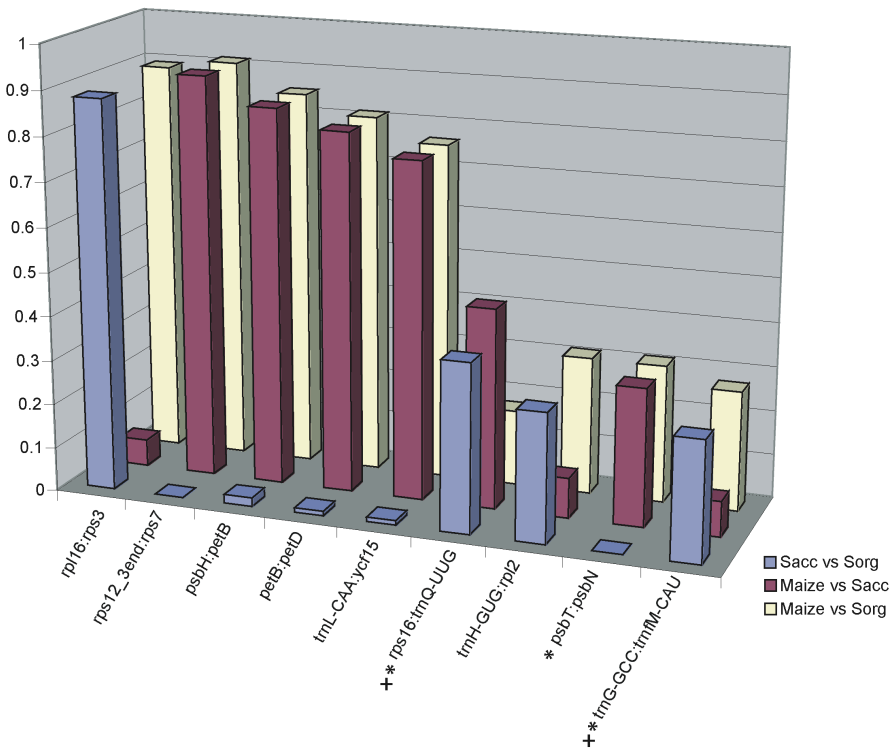

Fig.4.5. Histogram showing pairwise sequence divergence of the intergenic spacer regions of maize (*Zea mays*), sugarcane (*Saccharum officinarum*) and sorghum (*Sorghum bicolor*) chloroplast genomes. Comparisons of the nine most variable intergenic spacer regions with less than 80% average sequence identity. The values plotted in this histogram show percent sequence identities for all intergenic spacer regions. The plotted values were converted from percent identity to sequence divergence on a scale from 0 to 1 and included on the Y-axis. * indicate regions that are in the top 25 most variable intergenic spacer regions in Solanaceae, + indicate regions that are in the top 25 most variable intergenic spacer regions in Asteraceae (Timme et al. 2007).

Five intergenic spacer regions (*ndhD*:*psaC*, *psbJ*:*psbL*, *psbN*:*psbH*, *rrn23*:*trnA*-UGC, *trnA*-UGC:*rrn23*) have 100% sequence identity among *Z. mays*, *S. officinarum* and *S. bicolor*, whereas no spacer regions are identical among *O. sativa*, *T. aestivum*, *H. vulgare* and *A. stolonifera* despite of their close phylogenetic relationship. Divergence among *Z. mays*, *S. bicolor* and *S. officinarum* chloroplast genomes is much less because there are only nine intergenic spacer regions with less than 80% average sequence identity versus 19 among *O. sativa*, *T. aestivum, H. vulgare* and *A. stolonifera* (Figs. 4.4, 4.5). Only three of the intergenic regions in the two sets of comparisons have more than 80% average sequence divergence (*rpl16*:*rps3*, *psbH*:*petB*, and *rps12*_3end:*rps7*; compare Figs. 4.4, 4.5). Some spacer regions have indels resulting in extremely low sequence identity. For example, in *Z. mays*, deletion of a 558 bp intergenic region between *rps12* 3'end and *rps7* IGS has resulted in only 9% sequence identity between *Z. mays*:*S. bicolor* and *Z. mays*:*S. officinarum* comparisons. Nevertheless, this region shows 100% identity between *S. bicolor* and *S. officinarum*. Regions marked with asterisks or plus signs in Figures 4.4 and 4.5 are in the top 25 most variable intergenic spacers in Solanaceae (Chapter 3) and Asteraceae (Timme et al., 2007), respectively.

Table 4.3. Analysis of intergenic spacer regions of *O. sativa*, *T. aestivum*, *H. vulgare* and *A. stolonifera*. Intergenic spacer regions that are 100% identical in at least two of the four species are shown.

Intergenic_Region	*A. stolonifera*/ *H. vulgare*	*O. sativa*/ *H. vulgare*	*T. aestivum*/ *H. vulgare*	*A. stolonifera*/ *O. sativa*	*A. stolonifera*/ *T. aestivum*	*O. sativa*/ *T. aestivum*
trnA-UGC:*trnA*-UGC	100	99	99	99	98	98
trnH-GUG:*rpl2*	100	91	100	91	100	91
trnA-UGC:*trnI*-GAU	100	94	91	92	91	91
rpl23:*trnI*-CAU	97	97	100	97	97	97
trnI-CAU:*rpl23*	97	97	100	97	97	97
rrn4.5:*rrn23*	92	94	100	89	92	94
rrn23:*rrn4.5*	91	94	100	88	92	94
trnE-UUC:*trnY*-GUA	89	92	100	90	89	92
trnN-GUU:*trnR*-ACG	88	85	100	94	88	85
trnR-ACG:*trnN*-GUU	88	85	100	94	88	85
*rps12*_5end:*clpP*	86	80	100	78	86	80
ndhB:*rps7*	98	95	95	95	95	100
rps7:*ndhB*	98	94	94	94	94	100
trnQ-UUG:*psbK*	92	91	91	91	91	100
rps16:*trnQ*-UUG	40	36	36	56	56	100

Table 4.4. Analysis of intergenic spacer regions of *Z. mays*, *S. officinarum* and *S. bicolor*. Intergenic spacer regions that are 100% identical in at least two of the three species are shown below.

Intergenic spacer region	*Z. mays/S. officinarum*	*Z. mays*/*S. bicolor*	*S. officinarum* /*S. bicolor*
ndhD:psaC	100	100	100
psbJ:*psbL*	100	100	100
psbN:*psbH*	100	100	100
rrn23:*trnA*-UGC	100	100	100
trnA-UGC:*rrn23*	100	100	100
ndhB:*trnL*-CAA	100	99	99
trnL-CAA:*ndhB*	100	99	99
rps19:*trnH*-GUG	100	96	96
trnH-GUG:*rps19*	100	96	96
ndhB:*ndhB*	99	100	99
rps12:*trnV*-GAC	99	99	100
trnA-UGC:*trnA*-UGC	99	99	100
trnV-GAC:*rps12*	99	99	100
rrn16:*trnV*-GAC	98	98	100
trnN-GUU:*trnR*-ACG	98	98	100
trnR-ACG:*trnN*-GUU	98	98	100
trnV-GAC:*rrn16*	98	98	100
rpl23:*trnI*-CAU	97	97	100
rps2:*atpI*	97	97	100
rps7:*rps12*	97	97	100
rrn4.5:*rrn5*	97	97	100
trnI-CAU:*rpl23*	97	97	100
petG:*trnW*-CCA	96	96	100
ndhI:*ndhA*	95	100	95
psbC:*trnS*-UGA	95	95	100
rrn4.5:*rrn23*	95	95	100
rpl22:*rps19*	94	94	100
rpl36:*infA*	94	94	100
trnM-CAU:*atpE*	93	93	100
trnE-UUC:*trnY*-GUA	92	92	100

Table 4.4 (Continued). Analysis of intergenic spacer regions of *Z. mays*, *S. officinarum* and *S. bicolor*. Intergenic spacer regions that are 100% identical in at least two of the three species are shown below.

cemA:*petA*	91	91	100
ndhJ:*ndhK*	90	90	100
rps3:*rpl22*	89	89	100
trnA-UGC:*trnI*-GAU	86	86	100
psbT:*psbN*	69	69	100
rps12:*rps7*	9	9	100

Variations Between Coding Regions and cDNAs

Alignment of EST sequences and DNA coding sequences identified 15 nucleotide substitution differences in the *S. bicolor* chloroplast genome (Table 4.5), 25 in the *H. vulgare* genome (Table 4.6) and 1 in *A. stolonifera* (not shown). *S. bicolor* has six C-U conversions, five of which result in amino acid changes. *H. vulgare* also has six C-U conversions, all of which result in amino acid changes. Of these substitutions, 11 are non-synonymous and 4 are synonymous in S. bicolor. In *H. vulgare*, seventeen substitutions are non-synonymous and eight are synonymous. *S. bicolor* experienced 1-2 substitutions per gene while *H. vulgare* has 1-5 variable sites per identified gene. *H. vulgare* and *S. bicolor* share three variable positions in the *rpoC2*, *psaA*, and *atpB* genes (Tables 4.5, 4.6). At the time of the analysis of *A. stolonifera*, there were only 9018 EST sequences available for *A. stolonifera* to analyze potential RNA editing sites. Comparing the coding regions of the *A. stolonifera* chloroplast genome to available ESTs reveals only one potential editing site. This site is located within the *psbZ* gene at position 54 and suggests a C-U change, which does not result in a change in the amino acid. There are 89 ESTs that show support for a C-U change, and 5 that don't show the edit.

Table 4.5. Differences observed by comparison of S. bicolor chloroplast genome sequences with EST sequences obtained by BLAST search of NCBI GenBank

Gene	**Gene size**	**Sequence analyzed[a]**	**# variable sites**	**Variation type**	**Nucleotide position(s)[b]**	**Amino acid change**
rpoB	3231	1-2150	4	T-A	241	Y-N
				G-C	2048	S-T
				G-U	2050	E-L
				A-U	2051	E-L
clpP	651	265-651	5	G-A	337	A-T
				A-U	417	E-D
				T-C	508	S-P
				A-G	598	K-E
				G-A	630	P-P
rpl2	390	1-390	1	C-U	2	T-M
psaA	2253	117-894	3	G-C	81	A-A
				T-G	138	I-S
				C-A	396	F-L
ycf4	558	38-376	3	T-C	319	W-R
				T-C	342	R-R
				T-C	347	V-A
atpB	1497	1-670	3	C-U	490	R-C
				A-G	663	V-V
				T-C	669	N-N
ycf3	228	1-228	1	T-A	23	N-I
rpoC2	4434	3640-4315	1	C-U	4025	S-L
psaJ	129	1-129	1	T-G	72	G-G
petA	963	821-963	4	T-C	870	P-P
				C-U	883	R-C
				C-U	917	S-F
				C-U	949	V-I

Table 4.6. Differences observed by comparison of H. vulgare chloroplast genome sequences with EST sequences obtained by BLAST search of NCBI GenBank.

Gene	**Gene size**	**Sequence analyzed[a]**	**# variable sites**	**Variation type**	**Nucleotide position(s)[b]**	**Amino acid change**
rpoB	3231	1-2150	4	T-A	241	Y-N
				G-C	2048	S-T
				G-U	2050	E-L
				A-U	2051	E-L
clpP	651	265-651	5	G-A	337	A-T
				A-U	417	E-D
				T-C	508	S-P
				A-G	598	K-E
				G-A	630	P-P
rpl2	390	1-390	1	C-U	2	T-M
psaA	2253	117-894	3	G-C	81	A-A
				T-G	138	I-S
				C-A	396	F-L
ycf4	558	38-376	3	T-C	319	W-R
				T-C	342	R-R
				T-C	347	V-A
atpB	1497	1-670	3	C-U	490	R-C
				A-G	663	V-V
				T-C	669	N-N
ycf3	228	1-228	1	T-A	23	N-I
rpoC2	4434	3640-4315	1	C-U	4025	S-L
psaJ	129	1-129	1	T-G	72	G-G
petA	963	821-963	4	T-C	870	P-P
				C-U	883	R-C
				C-U	917	S-F
				C-U	949	V-I

Phylogenetic Analysis

The data matrix comprises 61 protein-coding genes for 38 taxa, including 36 angiosperms and two gymnosperm outgroups (Pinus and Ginkgo). The aligned sequences include 46,188 nucleotide positions but when the gaps are excluded to avoid ambiguities due to insertion/deletions there are 39,574 characters. Maximum Parsimony analyses resulted in a single most-parsimonious tree with a length of 62,437, a consistency index of 0.407 (excluding uninformative characters) and a retention index of 0.627 (Fig. 4.6). Bootstrap analyses indicate that 26 of the 35 nodes have bootstrap values ≥ 95%, 5 nodes have 80-94%, and 4 nodes have 50-79%. Maximum Likelihood analysis results in a single tree with a ML value of - lnL = 348086.2268 (Fig. 4.7). Support is very strong for most clades in the ML tree with ≥ 95% bootstrap values for 32 of the 35 nodes with and 60-69% support for the remaining three. The ML and MP trees only differ in the relationships among the rosids (compare Figs. 4.6, 4.7), although this difference is not strongly supported in the ML tree (63% bootstrap value). In the MP tree the eurosid II clade is sister to a clade that includes both members of eurosid I and Myrtales, whereas in the ML tree the eurosid II clade is sister to a clade that includes the Myrtales and one member of the eurosid I (Cucurbitales).

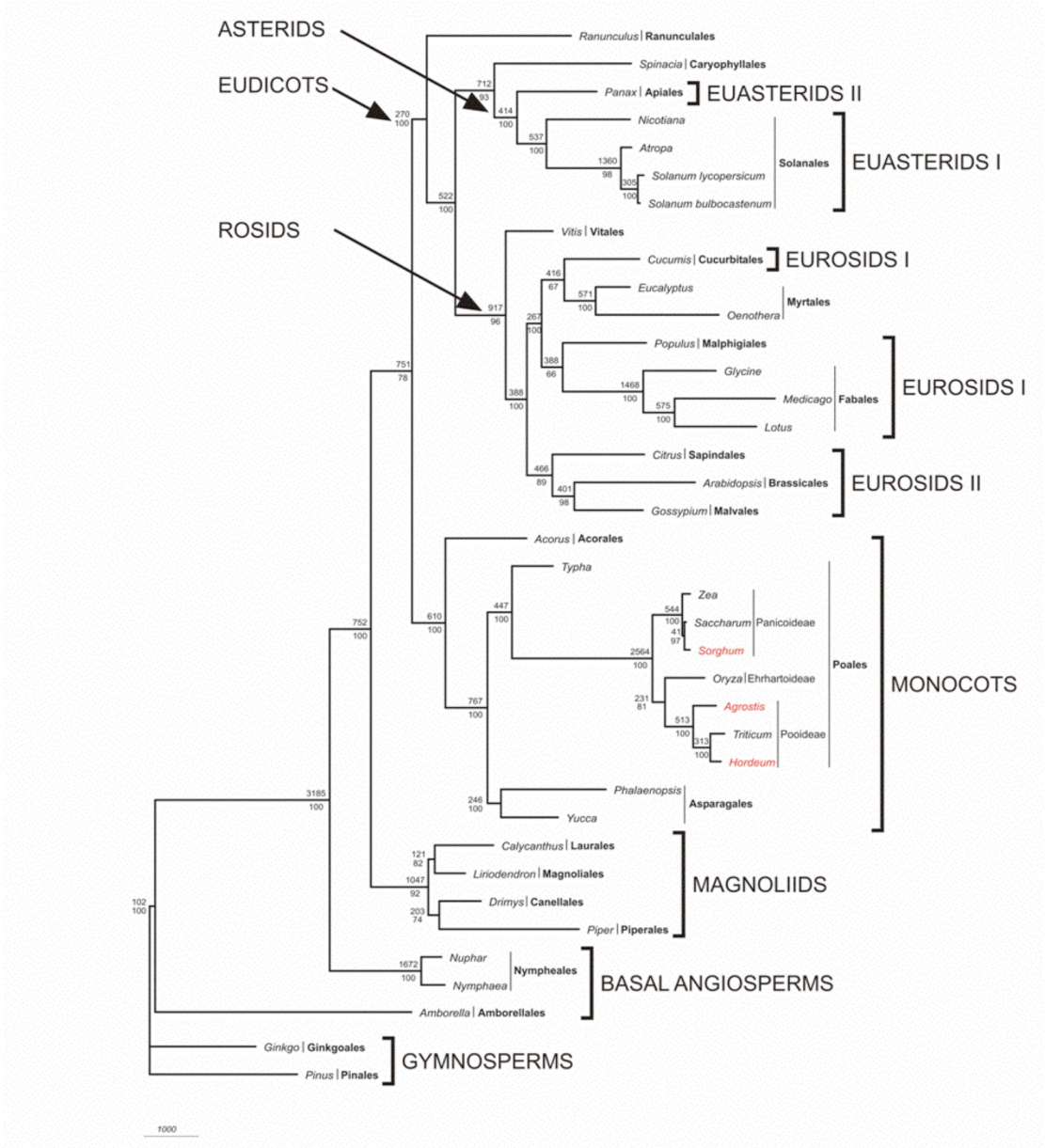

Fig 4.6. Phylogenetic tree of 38 taxa based on 61 plastid protein-coding genes using maximum parsimony. The tree has a length of 62,437, a consistency index of 0.407 (excluding uninformative characters) and a retention index of 0.627. Numbers above node indicate number of changes along each branch and numbers below nodes are bootstrap support values. Taxa in red are the new genomes reported in this study.

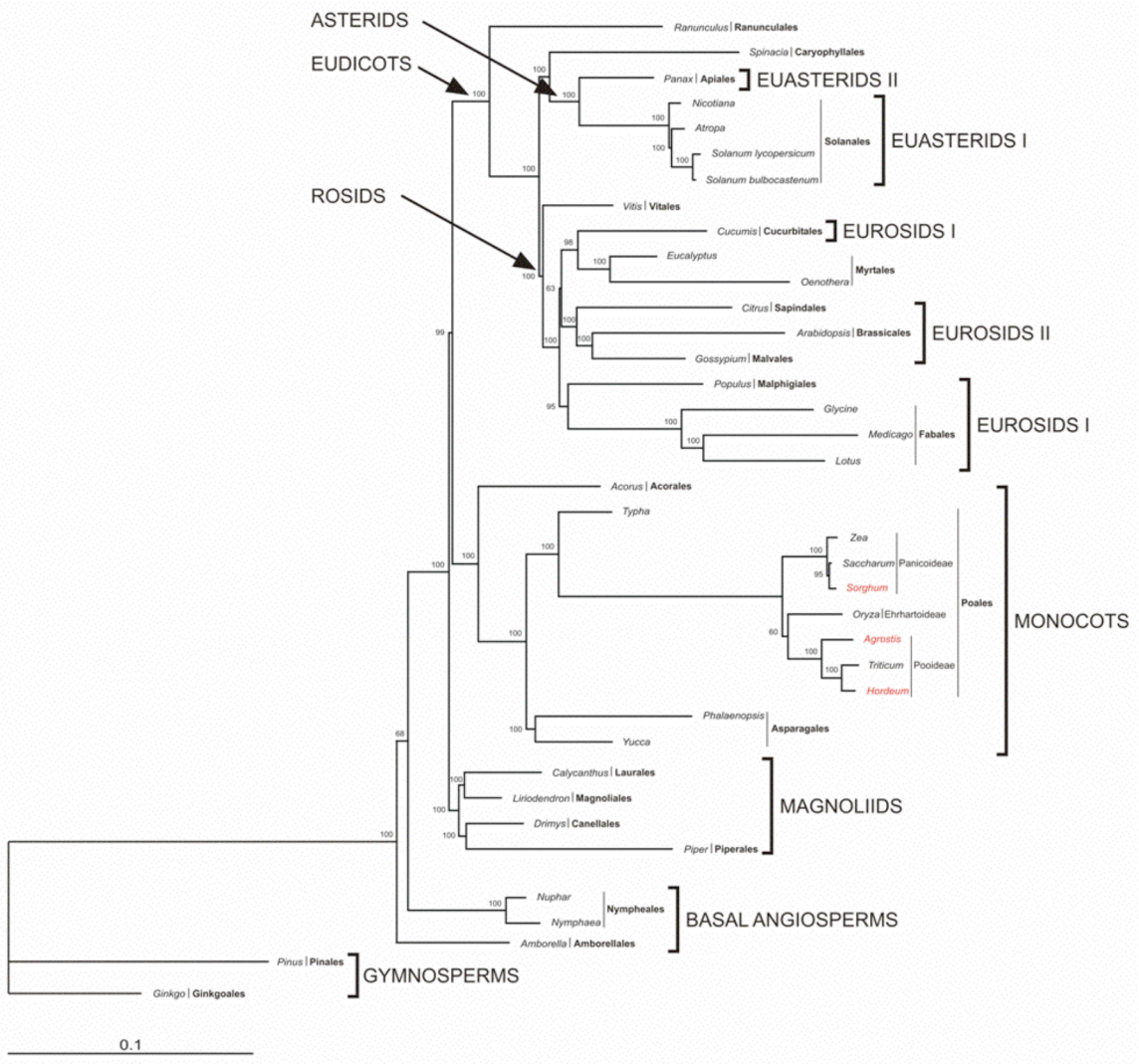

Fig 4.7. Phylogenetic tree of 38 taxa based on 61 plastid protein-coding genes using maximum likelihood. Taxa in red are the new genomes reported in this study

Discussion

Significance of transgene integration into grass chloroplast genomes

Although plastid transformation has been accomplished via organogenesis in a number of eudicots, two major obstacles have been encountered to extend plastid transformation technology to crop plants that regenerate via somatic embryogenesis: (i) the expression of transgenes in non-green plastids, in which gene expression and gene regulation systems are quite distinct from those of mature green chloroplasts, and (ii) our current inability to generate homoplastomic plants via subsequent rounds of regeneration, using leaves as explants. Despite these limitations, plastid transformation has recently been accomplished via somatic embryogenesis in several eudicot crops, including *Glycine max* L. Merr. (soybean), *Daucus carota* L. (carrot), and *Gossypium hirsutum* L. (cotton, Dufourmantel et al., 2004, 2005, Kumar et al., 2004a, b) and foreign genes have been expressed in high levels in non-green plastids, including proplastids and chromoplasts (Kumar et al., 2004a). Breakthroughs in plastid transformation of recalcitrant crops, such as *G. hirsutum* and *G. max*, have raised the possibility of engineering plastid genomes of other major crops via somatic embryogenesis. To date, only fragmentary data were reported for *O. sativa* plastid transformation (Khan and Maliga 1999). A promising step towards stable plastid transformation in *O. sativa* has been reported when stable integration and expression of the *aadA* and *sgfp* transgenes in their plastids was achieved (Lee et al., 2006b). Moreover, the transplastomic O. sativa plants generated viable seeds, which were confirmed to transmit the transgenes to the T1 progeny. Unfortunately, conversion of the transplastomic *O. sativa* plants to homoplasmy was not successful, even after two generations of continuous

selection. Thus, tissue culture and selection of transformed events continues to be a major challenge.

The success of chloroplast genetic engineering of crop plants is dependent, at least in part, on access to conserved spacer regions for inserting transgenes. The availability of sequences of complete chloroplast genomes for multiple crop plants in the grass family should facilitate plastid genetic engineering. Several studies have demonstrated that the use of intergenic spacer regions that have low sequence identities between the target genome and the flanking sequences in the chloroplast transformation vectors can result in substantially lower frequencies of transformants (Nguyen et al., 2005, Ruf et al. 2001, Sidorov et al., 1999). Given the low number of intergenic sequences that have high sequence identities among the seven sequenced chloroplast genomes (Tables 4.3, 4.4) it is unlikely that a single, highly conserved intergenic spacer (IGS) region will be appropriate throughout the grass family. Among Solanaceae chloroplast genomes, only four spacer regions have 100% sequence identity among all sequenced genomes and three of these regions are within the inverted repeat region (Chapter 3). Five intergenic spacer regions have 100% sequence identity among *Z. mays*, *S. officinarum* and *S. bicolor* chloroplast genomes. Thus the variation in the intergenic spacer region is quite similar between solanaceae and grass chloroplast genomes. However, not a single intergenic spacer region is identical among *O. sativa*, *T. aestivum* and *H. vulgare* chloroplast genomes. Thus, conservation of intergenic spacer regions is not uniform even within the same single family. However, it is noteworthy that the same intergenic spacer regions have very low sequence identity within Poaceae, Solanaceae and Asteraceae, as discussed below.

Organization and evolution of grass chloroplast genomes

The organization of chloroplast genomes is highly conserved in most land plants but alterations in gene content and order have been identified in several lineages (Raubeson and Jansen 2005). Notable rearrangements are known in two families with many crop species, a single 51-kb inversion common to most papilionoid legumes (Palmer et al., 1988, Doyle et al., 1996) and three inversions in the grasses (Quigley and Weil 1985, Howe et al., 1988, Hiratsuka et al., 1989, Doyle et al., 1992, Katayama and Ogihara 1996). The *H. vulgare*, *S. bicolor* and *A. stolonifera* chloroplast genomes contain all three of the inversions present in grasses.

Gene order and content of the sequenced grass chloroplast genomes are similar. However, two microstructural changes have occurred. First, the expansion of the IR at the SSC/IR boundary that duplicates a portion of the 5' end of *ndhH* is restricted to the three genera of the subfamily Pooideae (*Agrostis*, *Hordeum* and *Triticum*). These three genera form a monophyletic group in the phylogenetic trees based on DNA sequences of protein-coding genes (Figs. 4.6, 4.7) but the extent of the IR expansion differs in each of the three genera (32, 69, and 58 amino acids in wheat, barley, and bentgrass, respectively). Thus, it is not possible to determine if there have been three independent expansions or a single expansion followed by two subsequent contractions. Second, a 6 bp deletion in *ndhK* (Fig 4.2) is shared by *Agrostis*, *Hordeum*, *Oryza*, and *Triticum*, and this event supports the sister relationship between the subfamilies Erhartoideae and Pooideae (Figs. 4.6, 4.7).

Other than the inverted repeat, repeated sequences are considered to be relatively uncommon in chloroplast genomes (Palmer 1991). The analysis of the repeated sequences of grass chloroplast genomes revealed 26 groups of repeats shared among various members of the family (Table 4.2, Fig.4.3). Furthermore, 17 of the 26 repeats are shared among all

eight of the chloroplast genomes examined suggesting a high level of conservation of repeat structure among grasses. Examination of the location of these repeats suggests that all of them occur in the same location, either in genes, introns or within intergenic spacer regions. This high level of conservation of both sequence identity and location suggests that these elements may play a functional role in the genome, although we cannot rule out the possibility that this conservation may simply be due to a common ancestry. Because organellar genomes are often uniparentally inherited, chloroplast DNA polymorphisms have become a marker of choice for investigating evolutionary issues such as sex-biased dispersal and the directionality of introgression (Willis et al., 2005). They are also invaluable for the purposes of population-genetic and phylogenetic studies (Bryan et. al., 1999, Raubeson and Jansen 2005). Also, knowledge of mutation rates is important because they determine levels of variability within populations, and hence greatly influence estimates of population structure (Provan et. al., 1999). Based on mining for SSRs,16 to 18 SSRs within each of the nine genomes examined were identified (Table 4.2). These initial findings indicate a potential to test and utilize SSRs to rapidly analyze diversity in germplasm collections.

Previous studies of grass chloroplast genomes have identified three inversions in the family (Quigley and Weil 1985, Howe et al., 1988, Hiratsuka et al., 1989, Doyle et al., 1992, Katayama and Ogihara 1996). Analysis of the inversion endpoints indicate that there are shared repeats flanking the endpoints of the largest 28 kb inversion. This first inversion has endpoints between *trnG*-UCC and *trnR*-UCU at one end and *rps14* and *trnfM*-CAU at the other creating an intermediate form of the chloroplast genome prior to the second inversion when compared to *N. tabacum* (Hiratsuka et al., 1988, Doyle et al., 1992). Repeat analyses identified a 21 bp direct repeat in *O. sativa* that flanks the inversion endpoints, and this repeat

is shared by all other grasses examined. It is likely that the shared repeat facilitated this large inversion by intramolecular recombination. Two additional inversions, one largely overlapping the 28 kb event, subsequently gave rise to the gene order observed in *O. sativa* and *T. aestivum* (Hiratsuka et al., 1989). The endpoints of the second inversion (6 kb) occur between *trnS* and *psbD* on one end and *trnG*-UCC and *trnT*-GGU on the other (Doyle et al., 1992). The third inversion has endpoints between *trnG*-UCU and *trntT*-GGU and *trnT*-GGU and *trnE*-UUC. This inversion is quite small and accounts for the inverted orientation of *trnT*-GGU (Hiratsuka et al., 1989). The repeat analyses found no shared repeats that may have played a role in these two inversions. Chloroplast genome organization is also known from other monocots based on both gene mapping and complete genome sequencing (deHeij et al., 1983, Chase and Palmer 1989, Chang et al., 2006). Based on comparisons of four non-grass monocots (Spirodela oligorhiza (Lemnaceae), two orchids (*Oncidium excavatum* and *Phalaenopsis aphrodite*), and members of the Alliaceae (*Allium cepa* (monocot flowering plant), Asparagaceae (*Asparagus sprengeri*), and Amaryllidaceae (*Narcissus hybrid*) have the same gene order as tobacco. Thus, the inversions in *H. vulgare*, *S. bicolor* and *A. stolonifera* reported here are confined to the grass family as was previously suggested by Doyle et al., (1992).

Comparisons of DNA and EST sequences for *H. vulgare*, *S. bicolor* and *A. stolonifera* identified many differences (Tables 4.5, 4.6), most of which are not likely due to RNA editing. Previous investigations of RNA editing in chloroplast genomes in the angiosperms *N. tabacum* (Hirose et al. 1999) and *Atropa* (Schmitz-Linneweber et al. 2002) and in the fern *Adiantum* (Wolf et al. 2004) indicated that RNA edits only result in C-U changes. In the case of *H. vulgare*, *S. bicolor* and *A. stolonifera*, only seven differences in the DNA and EST sequences were C to U changes. Thus, these may be the result of RNA editing. The other

nine differences in *S. bicolor* and 19 differences in *H. vulgare* are likely due to either polymorphisms resulting from the use of different plants or cultivars or sequencing errors. In the case of *A. stolonifera*, only one C to U change was found. This could be attributed to the lack of available expression information since only 9018 EST sequences were available for *A. stolonifera* when the analysis was performed, suggesting a need for more comprehensive investigations into the chloroplast and nuclear transcriptomes.

Several recent comparisons of DNA and EST sequences for other crop species including *G. hirsutum* (Lee et al. 2006a), *Vitis vinifera* (Jansen et al. 2006), *Citrus sinensis* L. (Bausher et al. 2006), *Daucus carota* (Ruhlman et al. 2006), *Lactuca* and *Helianthus* (Timme et al., 2007), and *Solanum lycopersicum* and *S. bulboscastanum* (Chapter 3) have identified both putative RNA editing sites and possible sequencing errors. The much greater depth of coverage in the chloroplast genome sequences (generally 4-20X coverage) suggests that most of the differences other than changes from C to U are likely due to errors in EST sequences.

Phylogenetic studies at the inter- and intraspecific levels in plants have relied extensively on intergenic spacer regions of chloroplast genomes because the coding regions are generally too highly conserved at these lower taxonomic levels (Kelchner 2002, Raubeson and Jansen 2005, Jansen et al., 2005, Shaw et al., 2005). There have been many efforts to identify the most divergent intergenic spacers for phylogenetic comparisons at lower taxonomic levels with the hope that some universal regions could be found for angiosperms (Shaw et al. 2005, 2007, Timme et al. 2007). Only two previous studies have performed genome-wide comparisons among multiple, sequenced genomes in the families Asteraceae (Timme et al. 2007) and Solanaceae (Chapter 3). Comparison of the results in the Poaceae with these earlier studies indicates that there are considerable differences regarding

which intergenic spacer regions are most variable in these three families (Figs. 4.4, 4.5). Only three (Fig. 4.5) to five (Fig. 4.4) of the 25 most variable regions of Solanaceae are among the most variable intergenic spacers in grasses. The overlap in the regions with high sequence divergence between the Asteraceae and grasses is higher, with three (Fig. 4.5) to nine (Fig. 4.4) of the most variable IGS regions in the Poaceae among the 25 most variable regions in the Asteraceae. Overall, genome-wide comparisons among these three families indicate that there may be few universal IGS regions across angiosperms for phylogenetic studies at lower taxonomic levels. Thus, it will likely be necessary to identify variable IGS regions in chloroplast genomes for each family to locate the most appropriate markers for phylogenetic comparisons.

During the past three years there has been a rapid increase in the number of studies using DNA sequences from completely sequenced chloroplast genomes for estimating phylogenetic relationships among angiosperms (Goremykin et al., 2003a, b, 2004, 2005, Leebens-Mack et al., 2005, Chang et al., 2005, Lee et al., 2006a, Jansen et al., 2006, Ruhlman et al., 2006, Bausher et al., 2006, Cai et al., 2006). These studies have resolved a number of issues regarding relationships among the major clades, including the identification of either Amborella alone or Amborella + Nymphaeales as the sister group to all other angiosperms, strong support for the monophyly of magnoliids, monocots, and eudicots, the position of magnoliids as sister to a clade that includes both monocots and eudicots, the placement of Vitaceae as the earliest diverging lineage of rosids, and the sister group relationship between Caryophyllales and asterids. However, some issues remain unresolved, including the monophyly of the eurosid I clade and relationships among the major clades of rosids. The phylogenetic analyses reported here (Figs. 4.6, 4.7) with expanded taxon sampling are

congruent with these earlier studies so the discussion will focus on relationships among grasses.

This study has added complete chloroplast genome sequences for three genera of grasses representing two subfamilies (Pooideae and Erhartoideae, Grass Phylogeny Working Group 2001). This expands the number of sequenced grass genera to seven from three different subfamilies, Panicoideae, Pooideae and Erhartoideae. The phylogenetic trees (Figs. 4.6, 4.7) indicate that the Erhartoideae is sister to the Pooideae with weak to moderate bootstrap support (60 or 81% in ML and MP trees, respectively). The sister relationship of these subfamilies is also supported by a 6 bp deletion in *ndhK* (Fig. 4.2). This result is congruent with phylogenetic trees based on sequences of six genes (4 chloroplast and 2 nuclear, Grass Phylogeny Working Group 2001). This multigene tree, which included 68 genera of grasses, also provided only moderate bootstrap support (71%) for a close phylogenetic relationship between these two subfamilies. Furthermore, the clade including Pooideae and Erhartoideae also contained members of the Bambusioideae. Clearly, many additional chloroplast genome sequences are needed from the grasses to provide sufficient taxon sampling to generate a family-wide phylogeny based on whole genomes.

CHAPTER 5

CONCLUSIONS

The chloroplast is a plant organelle that contains the entire enzymatic machinery for photosynthesis. In addition to photosynthesis, several other biochemical pathways are compartmentalized within the chloroplasts, including biosynthesis of fatty acids, amino acids, pigments, vitamins, DNA, and RNA synthesis (Zeltz et al., 1993). The chloroplast genome generally has a highly conserved organization (Palmer 1991, Raubeson and Jansen 2005) with most land plant genomes composed of a single circular chromosome with a quadripartite structure that includes two copies of an inverted repeat that separate the large and small single copy regions. The size of this circular genome varies from 35 to 217 kb but among photosynthetic organisms the majority are between 115-165 kb (Jansen 2005).

Our knowledge of the organization and evoulution of chloroplast genomes has been expanding rabidly because of the large numbers of completely sequenced genomes published in the past decade. The use of information from chloroplast genomes is well established in the study of evolutionary patterns and processes in plants (Avise 1994, Raubeson and Jansen 2005). Comparative studies from the past indicate that chloroplast genomes of land plants are highly conserved in both gene order and gene content (Cosner et al., 1997). Several lineages of land plants have cp DNAs that have multiple rearrangements including Pinus (Wakasugi et al., 1994), and the angiosperm families Campanulaceae (Cosner et al., 1997), Fabaceae (Kato et al., 2000), Geraniaceae (Palmer et al., 1987a), and Lobeliaceae (Knox and Palmer 1998). In most of these studies, comparisons of gene content and order have been

made between distantly related taxa because only one genome sequence was available from groups with rearranged genomes.

Chloroplast genetic engineering offers a number of unique advantages, including a high-level of transgene expression (DeCosa et al., 2001), multi-gene engineering in a single transformation event (DeCosa et al., 2001), transgene containment via maternal inheritance (Daniell 2002), lack of gene silencing (Lee et al., 2003, position effect (Daniell et al., 2002), reduced pleiotropic effects (Lee et al., 2003, Daniell et al., 2001, Leelavathi et al., 2003) and undesirable foreign DNA (vector sequences) (Daniell et al., 2004a,b). Lack of complete chloroplast genome sequence is still one of the major limitations to extend this technology to useful crops. Chloroplast genome sequences are necessary for identification of spacer regions for integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for optimal expression of transgenes (Maier and Schmitz-Linneweber 2004, Daniell et al., 2005). In land plants, about 40-50% of each chloroplast genome contains non-coding spacer and regulatory regions. To expand our knowledge about crop chloroplast genomics and provide optimal sites for biotechnology application, our group revealed the complete chloroplast genome sequence for soybean, tomato, potato, barley, sorghum, and creeping bentgrass.

The chloroplast genome of *Glycine* is 152,218 basepairs (bp) in length, including a pair of inverted repeats of 25,574 bp of identical sequence separated by a small single copy region of 17,895 bp and a large single copy region of 83,175 bp. The genome contains 111 unique genes, and 19 of these are duplicated in the inverted repeat (IR). Comparisons of the *Glycine*, *Lotus* and *Medicago* confirm organization of legume chloroplast genomes based on previous studies. Gene content of the three legumes is nearly identical. The *rpl22* gene is

missing from all three legumes, and *Medicago* is missing *rps16* and one copy of the IR. Gene order in *Glycine*, *Lotus*, and *Medicago* differs from the usual gene order for angiosperm chloroplast genomes by the presence of a single, large inversion of 51 kilobases (kb). Detailed analyses of repeated sequences indicate that many of the *Glycine* repeats that are located in the intergenic spacer regions and introns occur in the same location in the other legumes and in *Arabidopsis*, suggesting that they may play some functional role. The presence of small repeats of *psbA* and *rbcL* in legumes that have lost one copy of the IR indicate that this loss has only occurred once during the evolutionary history of legumes (Chapter 2).

Analysis of the complete sequences of *Solanum lycopersicum*, *Solanum bulbocastanum*, tobacco, and *Atropa* chloroplast genomes reveals that there are significant insertions and deletions within certain coding regions or regulatory sequences (e.g., deletion of repeated sequences within 16S rRNA, *ycf2* or RBS in *ycf2*). RNA, photosynthesis, and ATP synthase genes are the least divergent and the most divergent genes are *clpP*, *cemA*, *ccsA* and *matK*. Repeat analyses identified 33 to 45 direct and inverted repeats ≥ 30 bp with a sequence identity of at least 90 %; all but five of the repeats shared by all four Solanaceae genomes are located in the same genes or intergenic regions, suggesting a functional role. A comprehensive genome-wide analysis of all coding sequences and intergenic spacer regions was done for the first time in chloroplast genomes. Only four spacer regions are fully conserved (100% sequence identity) among all genomes; deletions or insertions within intergenic spacer regions result in less than 25% sequence identity, underscoring the importance of choosing appropriate intergenic spacers for plastid transformation and providing valuable new information for phylogenetic utility of the chloroplast intergenic

spacer regions. Comparison of coding sequences with expressed sequence tags showed considerable amount of variation, resulting in amino acid changes; none of the C-to-U conversions observed in *Solanum bulbocastanum* and *Solanum lycopersicum* were conserved in tobacco and *Atropa*. It is possible that there has been a loss of conserved editing sites in *Solanum bulbocastanum* and *Solanum lycopersicum* (Chapter 3).

Comparisons of complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera* to six published grass chloroplast genomes reveal that gene content and order are similar but two microstructural changes have occurred. First, the expansion of the IR at the SSC/IRa boundary that duplicates a portion of the 5' end of *ndhH* is restricted to the three genera of the subfamily Pooideae (*Agrostis*, *Hordeum*, and *Triticum*). Second, a 6 bp deletion in *ndhK* is shared by *Agrostis*, *Hordeum*, *Oryza*, and *Triticum*, and this event supports the sister relationship between the subfamilies Erhartoideae and Pooideae. Repeat analysis identified 19-37 direct and inverted repeats 30 bp or longer with a sequence identity of at least 90%. Seventeen of the 26 shared repeats are found in all the grass chloroplast genomes examined and are located in the same genes or intergenic spacer regions. Examination of SSRs identified 16-21 potential polymorphic SSRs. Five intergenic spacer regions have 100% sequence identity among *Zea mays*, *Saccharum officinarum*, and *S. bicolor*, whereas no spacer regions were identical among *Oryza sativa*, *Triticum aestivum*, *H. vulgare* and *A. stolonifera* despite their close phylogenetic relationship. Alignment of EST sequences and DNA coding sequences identified six C-U conversions in both *S. bicolor* and *H. vulgare* but only one in *A. stolonifera*. Phylogenetic trees based on DNA sequences of 61 protein-coding genes of 38 taxa using both maximum parsimony and likelihood methods

provide moderate support for a sister relationship between the subfamilies Erhartoideae and Pooideae (Chapter 4).

Our knowledge of the organization and evolution of chloroplast genomes has been expanding rapidly because of the large numbers of completely sequenced genomes published in the past decade. The use of information gained from whole chloroplast genome sequence of soybean, tomato, potato, barley, sorghum, and creeping bentgrass has added to our understanding of chloroplast biology, the origins and relationships of land plants, and has laid the foundation for integrating useful traits via the chloroplast genome in these agriculturally and economically important crops.

APPENDIX

PUBLICATIONS RESULTING FROM THIS RESEARCH

CHAPTER 2

Saski C, Lee S-B, Daniell H Wood TC, Tomkins J, Kim H-G, Jansen RK (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol Biol 59:309–322. Copyright License # 1755331297265

CHAPTER 3

Daniell H, Lee SB, Grevich J, Saski C, Guda C, Tomkins J, Jansen RK (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor Appl Genet 112:1503-1518. Copyright License # 1760180797411

CHAPTER 4

Saski Christopher, Lee Seung-Bum, Fjellheim Sire, Guda Chittababu, Jansen Robert, Luo Hong, Tomkins Jeffrey, Rognli Od Arne, Daniell Henry, Clark Jihong Liu (2007) Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor*, and *Agrostis stolonifera*, and comparative analysis with other grass genomes. Theor Appl Genet. In press. Copyright License # 1755330934156

REFERENCES

Allen JF, and Raven JA (1996) Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J. Mol. Evol 42:482-492

Altschul, SF, Madden, TL, Schaffer, A.A., Zhang, JH, Zhang, Z, Miller, W, and Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25: 3389–3402.

Amer J Bot US Grains Council http://www.grains.org/page.ww?section=Barley%2C+Corn+%26+*Sorghum* &name=*Sorghum*. Cited 06 Nov 2006

APG II (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot J Linn Soc* 141:399-436

Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K (2004) Complete nucleotide sequence of the *sugarcane* (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA Res 11:93–99

Avise JC (1994) Molecular Markers, Natural History, and Evolution. Chapman & Hall, New York

Ayliffe MA, and Timmis JN (1992) Plastid DNA sequence homologies in the tobacco nuclear genome. Mol. Gen. Genet. 236:105-112

Ayliffe MA, Timmis JN, and Steele SN (1988) Homologies to chloroplast DNA in the nuclear DNA of a number of Chenopod species. Theor Appl Genet 75:282-285

Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H (2006) The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biology 6:21

Bendich A J (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? Bioessays 6:279-282

Bohs L, Olmstead RG (1997) Phylogenetic relationships in *Solanum* (Solanaceae) based on *ndhF* sequences. Syst Bot 22: 5–17

Bonos SA, Clarke, BB, Meyer WA (2006) Breeding for disease resistance in the major cool-season turfgrass. Annu Rev Phytopathol 44:213-234

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

Bryan GJ, McNicoll J, Ramsey G, Meyer RC, De Jong WS (1999) Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. Theor Appl Genet 99:859-867

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson J, dePamphilis CW, Jansen RK (2006) Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: Implications for the phylogeny of magnoliids. BMC Evol Biol 6:77

Carter PR, Hicks DR, Oplinger ES, Doll JD, Bundy LG, Schuler RT, Holmes BJ (1989) Grain *Sorghum* (Milo). Alternative Field Crops Manual

Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, Chen W-H, Cheng C-H, Lin C-Y, Liu S-M, Chang C-C, Chaw S-M (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol 23:279-291

Chase MW, Palmer JD (1989) Chloroplast DNA systematics of lilioid monocots: resources, feasibility, and an example from the Orchidaceae. Amer J Bot 76:1720-1730

Cheng M, Lowe BA, Spencer MT, Ye X, Armstrong CL (2004) Factors influencing *Agrobacterium*-mediated transformation of monocotyledonous species. In Vitro Cell Dev Biol 40:31-45

Cheung WY, and Steele Scott N (1989) A contigous sequence in spinach nuclear DNA is homologous to three separated sequences in chloroplast DNA. Theor Appl Genet 77:625-633

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

Cosner ME, Jansen RK, Palmer JD, Downie SR (1997) The highly rearranged chloroplast genome of Trachelium caeruleum (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr Genet 31: 419–429

Crop Plant Resources. (2000) *Sorghum*: *Sorghum bicolor*. http://darwin.nmsu.edu/~molbio/plant/sorghum.html (Accessed May 18, 2006)

Cui L, Veeraraghavan N, Richer A, Wall K, Jansen RK, Leebens-Mack J, Makalowska I, dePamphillis CW (2006) ChloroplastDB: the chloroplast genome database. Nucl Acids Res 34:D692-D696 [http://chloroplast.cbio.psu.edu/]

Curtis SE, and Clegg MT (1984) Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. 1:291-301

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

Daniell H, Carmona-Sanchez O, Burns BB (2004a) Chloroplast-derived vaccine antibodies, biopharmaceuticals, and edible vaccines in transgenic plants engineered via the chloroplast genome. Molecular Farming, S Schillberg Ed Wiley–VCH Verlag publishers, Germany, Chapter 8 pp 113–133

Daniell H, Chebolu S, Kumar S, Singleton M, Falconer R (2005) Chloroplast-derived vaccine antigens and other therapeutic proteins. Vaccine 23:1779-1783

Daniell H, Cohill P, Kumar S, Dufourmantel N, Dubald M (2004b) Chloroplast genetic engineering; In Daniell H, Chase C (eds) molecular biology and biotechnology of plant organelles. SpringerPublishers, Netherlands, pp 423-468

Daniell H, Datta R, Varma S, Gray S, Lee SB (1998) Containment of herbicide resistance through genetic engineering of the chloroplast genome. Nat Biotechnol 16:345-348

Daniell H, Dhingra A (2002) Multigene engineering: dawn of an exciting new era in biotechnology. Curr Opin Biotechnol 13:136-141

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

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

Daniell H, Lee SB, Pahchal T, Wiebe P (2001) Expression of the native cholera toxin B subunit gene and assembly as functional oligomers in transgenic tobacco chloroplasts. J Mol Biol 311:1001-1009

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

de Heij HT, Lustig H, Moeskops DM, Bovenberg WA, Bisanz C, Groot GSP (1983) Chloroplast DNAs of *Spinacia*, *Petunia*, and *Spirodela* have similar gene organization. Curr Genet 7:1-6

DeCosa B, Moar W, Lee S-B, Miller M, Daniell H (2001) Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. Nat Biotechnol 9:71–74

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

Dhingra A, Portis A Jr, Daniell H (2004) Enhanced translation of a chloroplast-expressed *rbcS* gene restores small subunit levels and photosynthesis in nuclear *rbcS* antisense plants. Proc Natl Acad Sci USA 101:6315-6320

Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proc Natl Acad Sci, USA 89:7723-7726

Doyle JJ, Doyle JL, Ballenger JA Palmer JD (1996) The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Mol Phylog Evol 5:429–438

Doyle JJ, Doyle JL, and Palmer JD (1995) Multiple independent losses of two genes and one intron from legume chloroplast genomes. Syst Bot 20:272–294

Drescher Anja, Ruf Stephanie, Tercillo Calsa Jr, Carreer Helaine, Bock Ralph (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Journal 22:97-104

Dufourmantel N, Pelissier B, Garçon F, Peltier JM, Tissot G (2004) Generation of fertile transplastomic soybean. Plant Mol Biol 55(4):479–89

Dufourmantel N, Tissot G, Goutorbe F, Garcon F, Jansens S, Pelissier B, Peltier G, Dubald M (2005) Generation and analysis of soybean plastid transformants expressing *Bacillus thuringiensis Cry1Ab* protoxin. Plant Mol Biol 58:659

Du Jardin P (1990) Homologies to plastid DNA in the nuclear and mitochondrial genomes of potato. Theor Appl Genet 79:807-812

Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Elnitski L, Riemer C, Petrykowska H, et al. (2002) PipTools: A computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. Genomics 80: 681–690

Ewing B, Hillier L, Wendl M, Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Gen Res 8:175-185

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791

Fernandez-San M A, Mingeo-Castel AM, Miller M, Daniell H (2003) A chloroplast transgenic approach to hyper-express and purify human serum albumin, a protein highly susceptible to proteolytic degradation. Plant Biotechnol J 1:71-79

Fiebig A, Stegemann S, Bock R (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucl Acids Res 32: 3615–3622

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, and Palmer JD (1991) Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J 18:2621-2630

Garber ED (1950) Cytotaxonomic studies in the genus *Sorghum*. University of California Publications in Botany 23:283–361

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003a) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol 20:1499-1505

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003b) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* - structural and phylogenetic analyses. Plant Syst Evol 242:119-135

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol 21:1445-1454

Goremykin VV, Holland B, Hirsch-Ernst KI Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol 22:1813-1822

Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252: 195–206

Grass Phylogeny Working Group (2001) Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Missouri Bot Gard 88:373-457

Gray Michael W (2004). Plastid genomes. In: Molecular Biology and Biotechnology of Plant Organelles. pp 115-150, Daniell, H. and Chase, C.D., Eds.,Springer publishers, Netherlands.

Grevich JJ, Daniell H (2005) Chloroplast genetic engineering: Recent advances and future perspectives. Crit Rev Plant Sci 24:83-108

Guda C, Lee SB, Daniell H (2000) Stable expression of biodegradable protein based polymer in tobacco chloroplasts. Plant Cell Rep 19:257-262

Hagemann R (2004) The Sexual Inheritance of Plant Organelles. In H Daniell, C Chase, eds, Molecular Biology and Biotechnology of Plant Organelles. Springer Publishers, Dordrecht, The Netherlands, pp 93–113

Herdenberger F, Pillay DTN, and Steinmetz A (1990) Sequence of the trnH gene and the inverted repeat structure deletion of the broad bean chloroplast genome. Nucl Acids Res 18:1297

Herrmann RG (1997) Eukaryotism, towards a new interpretation. In H. E. Schenk, K. W. Jeon, N. E. Muller, W. Schwemmler, K. W. Jeon (Eds), Eukaryotism and Symbiosis, pp 73-118, Springer: Berlin

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266: 383–402

Hipkins VD, Marshall KA, Neale DB, Rottmann WH and Strauss SH (1995) A mutation hotspot in the chloroplast genome of a conifer (Douglas-Fir, Pseudotsuga) is caused by variability in the number of direct repeats derived from a partially duplicated transfer-RNA gene. Curr Genet 27:572–579.

Hiratsuka J, Shimada H, Whittier R et al. (1989) The complete sequence of the *rice* (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185–194

Hirose T, Kusumegi T, Tsudzuki T, Sugiura M (1999) RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity. Mol Gen Genet 262:462–467

Howe CJ, Barker RF, Bowman, CM, Dyer TA (1988) Common features of three inversions in *wheat* chloroplast DNA. Curr Genet 13:343–349

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

Howe Christopher J, Barbrook Adrian C, Koumandou Lila V, Nisbet Ellen R, Symington Hamish A, Wightman Tom F (1992) Evolution of the Chloroplast Genome. Philos Trans R Soc Lond B Biol Sci 358:99-107

Hupfer H, Swaitek M, Hornung S, et al., (2000) Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome 1 of the five distinguishable Euoenthera plastomes. Mol Gen Genet 263:581–585

Iamtham S, Day A (2000) Removal of antibiotic resistance genes from transgenic tobacco plastids. Nat Biotechnol 18: 172–1176

Jansen RK, Kaittanis C, Saski C, Lee S-B, Tompkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: Effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol 6:32

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

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution, and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics 5:136-143.

Kamarajugadda S, Daniell H (2006) Choroplast derived anthrax and other vaccine antigens: their immunogenic and immunoprotective properties. Expert Rev Vaccines, 5:839-849.

Kaneko, T., Tanaka, A., Sato, S. et al. (1995) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. DNA Res. 2: 153–166.

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

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

Kelchner SA (2002) The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann Missouri Bot Gard 87:482–498

Khan M, Maliga P (1999) Fluorescent antibiotic resistance marker for tracking plastid transformation in higher plants. Nat Biotechnol 17:910-915

Kim K-J, Lee H-L (2004) Complete chloroplast genome sequence from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res 11:247-261

Kimura, M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

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

Kobayashi T, Takahara M, Miyagishima S, Kuroiwa H, Sasaki N (2002) Detection and localization of chloroplast encoded HU-like protein that organizes chloroplast nucleoids. Plant Cell 14:1579-1589

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

Kotera E, Tasaka M, and Shikanai T (2005) A pentatricopeptide repeat is essential for RNA editing in chloroplasts. Nature 433:326-330

Koya V, Moayeri M, Leppla SH, Daniell H (2005) Plant based vaccine: mice immunized with chloroplast-derived anthrax protective antigen survive anthrax lethal toxin challenge. Infect Immun 73:8266-8274

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

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003b) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucl Acids Res 31:2417–2423

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

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

Kumar S, Koichiro T, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17:1244–1245

Kuroiwa T (1991) The replication, differentiation, and inheritance of plastids with emphasis on the concept of organelle nuclei. Int. Rev. Cytol. 128:1-62

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

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

Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006a) The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. BMC Genomics 7:61

Lee SB, Kwon H, Kwon S et al. (2003) Accumulation of trehalose within transgenic chloroplasts confers drought tolerance. Mol Breed 11:1-13

Lee SM, Kang K, Chung H, Yoo SH, Xu XM, B. Lee SB, Cheong JJ, Daniell H,Kim M (2006b). Plastid transformation in the monocotyledonous cereal crop, rice (Oryza sativa) and transmission of transgenes to their progeny. Mol Cells 21:401-410

Leebens-Mack J, Raubeson LA, Cui L, Kuehl J, Fourcade M, Chumley T, Boore JL, Jansen RK, and dePamphilis CW (2005) Identifying the basal angiosperms in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol 22:1948-1963

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

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

Lockhart PJ, Penny D (2005) The place of Amborella within the radiation of angiosperms. Trends Plant Sci. 10:201–202

Lopez-Juez E, Pyke KA (2005) Plastids unleashed: their development and their integration in plant development. Int J Dev Biol 49:557-577

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

Luo M, Wang Y-H, Frisch D, Joobeur T, Wing RA, Dean RA (2001) Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (Fom-2). Genome 44:154–162

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

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

Maliga Pal (2004) Plastid Transformation in Higher Plants. Annu. Rev. Plant Biol 55:289-313

Martin W, and Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol. 118:9-17

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V (2005) Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203–209

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

Maul JE, Lilly JW, Cui L, et al., (2002) The Chlamydomonas reinhardtii plastid chromosome: Islands of genes in a sea of repeats. The Plt Cell 14:1–22

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller E, Harris EH and Stern DB (2002) The Chlamydomonas reinhardtii plastid chromosome: Islands of genes in a sea of repeats. The Plt Cell 14:1–22

McBride K, Svab Z, Schaaf D, Hogan P, Stalker D, Maliga P (1995) Amplification of a chimeric Bacillus gene in chloroplasts leads to an extraordinary level of an insecticidal protein in tobacco. Biotechnology 13:362-365

McCauley DE (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. Trends in Ecology and Evolution 10:198-202

Milligan BG, Hampton JN, Palmer, JD (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol 6:355–368

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

Morand-Prieur M E, Vedel F, Raquin C, Brached S, Sihachakr D, and Frascaria-Lacoste N (2002) Maternal inheritance of a chloroplast microsatllite marker in controlled hybrids between *Fraxinus excelsior* and *Fraxinus angustifolia.* Mol Ecol 11:613-617

Mulligan Michael R (2004) Chloroplast genetic engineering; In Daniell H, Chase C (eds) molecular biology and biotechnology of plant organelles. , SpringerPublishers, Netherlands, pp 423-468

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

National *Sorghum* Producers (2006) What is *Sorghum*? www.sorghum.growers.com/Sorghum-101. Cited 06 Nov 2006

Nguyen TT, Nugent G, Cardi T, Dix PJ (2005) Generation of homoplasmic plastid transformants of a commercial cultivar of potato (*Solanum tuberosum* L). Plt Sci 168: 1495-1500

Ogihara Y, Isono K, Kojima T, et al. (2000) Chinese spring *wheat* (*Triticum aestivum* L.) chloroplast genome: Complete sequence and contig clones. Plt Mol Biol Rep 18: 243-253

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

Olmstead RG, Sweere JA, Spangler RE, Bohs L, Palmer JD (1999) Phylogeny and Provisional Classification of the Solanaceae Based on Chloroplast DNA pp 111–137 in Solanaceae IV, Advances in Biology and Utilization, M Nee, D E Symon, JP Jessup, and JG Hawkes, eds Royal Botanic Gardens, Kew, pp111-137

Palmer J D (2003) The symbiotic birth and spread of plastids: how many times and whodunit? J. Phycol. 39:4-11

Palmer JD (1991) Plastid chromosomes: structure and evolution. In RG Hermann, ed, The molecular biology of plastids. Cell culture and somatic cell genetics of plants, vol. 7A, Springer-Verlag, Vienna, pp 5–53

Palmer JD (1992) Comparison of chloroplast and mitochondrial genome evolution in plants. Cell Organelles (ed Herrmann RG), pp 99-133, Springer Verlag, Vienna

Palmer JD (1997) Organelle genomes: going, going, gone! Science 275:790-791

Palmer JD, Nugent JM, Herbon LA (1987) Unusual structure of Geranium chloroplast DNA - A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families. Proc Natl Acad Sci USA 84:769–773

Palmer JD, Osorio B, Thompson WF (1988) Evolutionary significance of inversions in Legume chloroplast DNAs. Curr Genet 14:65–74

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10:823-833

Palmer, J.D. (1985) Evolution of chloroplast and mitochondrial DNA in plants and algae. In RJ MacIntyre, ed, Monographs in evolutionary biology: molecular evolutionary genetics. Plenum Press, New York, pp 131–240

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

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

Peeters NM, Hanson MR (2002) Transcript abundance supercedes editing efficiency as a factor in developmental variation of chloroplast gene expression. RNA 8:497-511

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

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

Pirchersky E, Logsdon J M Jr, McGrath J M, and Stasys R A (1991) Fragments of plastid DNA in the nuclear genome of *Solanum lycopersicum*: prevalence, chromosomal location, and possible mechanism of integration. Mol. Gen. Genet. 225:453-458

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: application to the population genetics of pines. PNAS 92:7759-7763

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

Provan J, Soranzo N, Wilson N, Goldstein D, Powell W (1999) A low mutation rate for chloroplast microsatellites. Genetics 153:943-947

Quesada-Vargas T, Ruiz ON Daniell H (2005) Characterization of heterologous multigene operons in transgenic chloroplasts: transcription, processing, translation. Plant Physiol 128:1746-1762

Quigley F, Weil JH (1985) Organization and sequence of five tRNA genes and of an unidentified reading frame in the *wheat* chloroplast genome: evidence for gene rearrangements during the evolution of chloroplast genomes. Curr Genet 9:495–503

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry R (ed) Diversity and evolution of plants-genotypic and phenotypic variation in higher plants. CABI Publishing, Wallingford, pp 45–68

Reboud X, Zeyll C (1993) Organelle inheritance in plants. Heredity 72:132-140

Reichman JR, Watrud LS, Lee EH, Burdick C, Bollman M, Storm M, King G, Mallory-Smith C (2006) Establishment of transgenic herbicide-resistant *creeping bentg*rass (*Agrostis stolonifera* L.) in nonagronomic habitats. Molecular Ecology 15:4243–4255

Rodermel S (2001) Pathways of plastid-to-nucleus signaling. Trends Plant Sci. 6:471-478.

Ruf S, Hermann M, Berger I, Carrer H, Bock R (2001) Stable genetic transformation of *Solanum lycopersicum* plastids and expression of a foreign protein in fruit. Nat Biotechnol 19: 870–875

Ruhlman T, Ahangari R, Devine A, Samsam M, Daniell H (2007) Expression of cholera toxin B-proinsulin fusion protein in lettuce and tobacco chloroplasts – oral administration protects against development of insulitis in non-obese diabetic mice.

Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell, D (2006) Complete plastid genome sequence of *Daucus carota*: Implications for biotechnology and phylogeny of angiosperms. BMC Genomics 7:224

Ruiz O, Hussein S, Terry N, Daniell H (2003) Phytoremediation of organomercurial compounds via chloroplast genetic engineering. Plant Physiol 132:1344-1352

Ruiz ON, Daniell H (2005) Engineering cytoplasmic male sterility via the chloroplast genome by expression of _-ketothiolase. Plant Physiol 138:1232-1246

Ruiz ON, Hussein H, Terry N, Daniell H (2003) Phytoremediation of organomercurial compounds via chloroplast genetic engineering. Plt Phys 32:1344–1352

Rujan T, and Martin W (2001) How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. Trends Genet. 17:113-120

Sato N, Albrieux C, Joyard J, Douce R, Kuroiwa T (1993) Detection and characterization of a plastid DNA-binding protein which may anchor plastid nucleoids. EMBO J. 12:555-561

Sato N, Ohta N (2001) DNA binding specificity and dimerization of the DNA binding domain of the PEND protein in the chloroplast envelope membrane. Nucleic Acids Res. 29:2244-2250

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283–290

Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Mol Biol Evol 19:1602-1612

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

Scott SE, Wilkenson MJ (1999) Low probability of chloroplast movement from oilseed rape (Brassica napus) into wild Brassica rapa. Nat Biotechnol 17: 390–392

Sears BB, Stoike LL, Chiu WL (1996) Proliferation of direct repeats near the Oenothera chloroplast DNA origin of replication. Mol Biol Evol 13:850–863

Shahid-Masood M, Nishikawa T, Fukuoka S, Njenga PK, Tsudzuki T, Kadowaki K (2004) The complete nucleotide sequence of wild *rice* (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated *rice*. Gene 340: 133-139

Shamudarov I A, Akbarova Y Y, Solovyev, V V, and Aliyev J A (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant. Mol. Biol. 52:923-934

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

Shaw J, Lickey EB, Beck JT et al. (2005) The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analyses. Amer J Bot 92:142–166

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. Amer J Bot, 94:275-288

Shinozaki K, Ohme M, Tanaka (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO J 5:2043–2049

Sidorov VA, Kasten D, Pang SZ, Hajdukiewicz PT, Staub JM, Nehra NS (1999) Technical advance: stable chloroplast transformation in potato: use of green fluorescent protein as a plastid marker. Plant J 19: 209–216

Smith HC, Gott JM, and Hanson MR (1997) A guide to RNA editing. RNA 3:1105- 1123

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu Y-Q, Chase MW, Farris JS, Stefanovi_ S, Rice DW, Palmer JD, Soltis PS (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483

Soltis, DE.;Soltis, PS.;Endress, PK.; Chase, MW (2005) Phylogeny and evolution of Angiosperms. Sunderland Massachusetts: Sinauer Associates Inc

Spangler RE, Zaitchik B, Russo E, Kellogg E (1999) Andropogoneae evolution and generic limits in *Sorghum* (Poaceae) using ndhF sequences. Syst Bot 24: 267–281

Spangler RE. (2003) Taxonomy of *Sarga*, *Sorghum* and *Vacoparis* (Poaceae: Andropogoneae). Australian Syst Bot 16: 279–299

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

Spooner DM, Anderson GJ, Jansen RK (1993) Chloroplast DNA evidence for the interrelationships of *Solanum lycopersicum*es, Potatoes, and Pepinos. Amer J Bot 8: 676–688

Staub JM, Garcia B, Graves J et al. (2000) High yield production of a human therapeutic protein in tobacco chloroplasts. Nat Biotechnol 18:333-338

Staub JM, Garcia B, Graves J (2000) High yield production of a human therapeutic protein in tobacco chloroplasts. Nat Biotechnol 18:333-338

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

Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian Blue Gum, Eucalyptus globulus (Myrtaceae). DNA Res 12:215-220

Stefanovic S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol Biol.4:35

Sugiura M (1995) The chloroplast genome. Essays Biochem 30:49-57

Swofford DL (2003) PAUP*: Phylogenetic analysis using parsimony (*and other methods), ver. 4.0 Sunderland MA: Sinauer Associates

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

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in *rice* (*Oryza sativa L.*): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441-1452

Timme RE, Kuehl JV, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. Genetics 94:302-312

Timmis J N, and Steele Scott N (1983) Sequence homology between spinach nuclear and chloroplast genomes. Nature 305:65-67

Timmis J.N., Ayliffe, M.A.., Huang C. Y., Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Rev. Genet. 5: 123-136

Tomkins, J.P., Mahalingam, R., Smith, H., Goicoechea, J.L., Knap, H.T., and Wing, R.A. (1999) A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. Plant Mol. Biol. 41: 25–32

Tregoning JS, Nixon P, Kuroda H, et al., (2003) Expression of tetanus toxin Fragment C in tobacco chloroplasts. Nucl Acids Res 31(4):1174–1179

Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka M, Shibata M, Wakasugi T and Sugiura M. (1992) Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequence of *the trnQ, trnK, psbA, trnI*, and *trnH* and the absence of *rps16*. Mol Gen Genet 232:206–214

Vedel F, Quetier F, Bayen M (1976) Specific cleavage of chloroplast DN and maternal inheritance of mitochondrial DNA in the genus *Actinidia*. Theor. Appl Genet 263:440-442

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

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

Von Heijne G (1986) Why mitochondria need a genome. FEBS Lett 198:1-4

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

Waters MT, Fray RG, and Pyke KA (2004) Stromule formation is dependent upon plastid size, plastid differentiation status and the density of plastids within the cell. Plant J 39:655-667

Watrud LS, Lee EH, Fairbrother A, Burdick C, Reichman JR, Bollman M, Storm M, King G, Van de Water PK (2004) Evidence for landscape-level, pollen-mediated gene flow from genetically modified *creeping bentg*rass with CP4 EPSPS as a marker. Proc Nat Acad Sci USA 101:14533–14538

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

Willis KJ, Macelwain JC (2002) The Evolution of Plants Oxford Univ Press, Oxford

Willis D, Hester M, Liu A, Burke J (2005) Chloroplast SSR polymorphisms in the Compositae and the mode of organellar inheritance in *Helianthus annuus*. *Theor Appl Genet* 110:941-947

Wipff JK, Fricker C (2001) Gene flow from transgenic *creeping bentg*rass (Agrostis stolonifera L.) in the Willamette valley, Oregon. Intl Turfgrass Soc Res J 9:224–242

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

Wolf PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339:89–97

Wolfe K H, Li W H, and Sharp P M (1987) Rates of nucleotide subsistution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA. 84:9054-9058

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

Wyman SK, Boore JL, Jansen RK (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255

Yu Y, Tomkins JP, Waugh R, Frisch DA, Kudrna D, Kleinhofs A, Brueggeman RS, Muehlbauer GJ, Wise RP, Wing RA (2000) A bacterial artificial chromosome library for barley (Hordeum vulgare L.) and the identification of clones containing putative resistance genes. Theor Appl Genet 101:1093–1099

Zeltz P, Hess WR, Neckermann K, Borner T, Kossel H (1993) Editing of the chloroplast *rpoB* transcript is independent of chloroplast translation and shows different patterns in barley and maize. EMBO J 12:4291-4296

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the *max*imum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin. [www.bio.utexas.edu/faculty/antisense/garli/Garli.html]