EFFECTS OF THE DOWNSTREAM BOX REGION ON HIGH-LEVEL EXPRESSION OF FOREIGN PROTEINS IN PLASTID-TRANSFORMED TOBACCO

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Benjamin N. Gray

August 2009

EFFECTS OF THE DOWNSTREAM BOX REGION ON HIGH-LEVEL EXPRESSION OF FOREIGN PROTEINS IN PLASTID-TRANSFORMED TOBACCO

Benjamin N. Gray, Ph.D.

Cornell University 2009

The potential for significant environmental benefits from lignocellulosic ethanol production via enzymatic cellulose hydrolysis have not yet been realized, in large part because of high enzyme production costs when using microbial protein production systems. Enzyme production *in planta* may lower these production costs, and the potential for high protein yields from plastid transformation makes this an attractive platform for cellulolytic enzyme production. The *Thermobifida fusca cel6A* and *bglC* genes, encoding an endoglucanase and a β-glucosidase, respectively, were inserted into the *Nicotiana tabacum* chloroplast for expression with various 14-amino acid downstream box (DB) fusions added to the N-terminus of each protein. The DB region, comprised of the 10-15 codons immediately downstream of the start codon, has previously been shown to be an important factor in determining foreign protein accumulation in chloroplasts and in other prokaryotic systems through an unknown mechanism.

Chloroplast expression of *cel6A* and *bglC* with the various DB fusions resulted in the accumulation of active protein varying over more than 2 orders of magnitude, from less than 0.1% of total soluble protein (%TSP) to over 10%TSP. Analysis of *cel6A* and *bglC* transcripts revealed differences in RNA processing, suggesting a feedback between the DB region and RNA degradation rates. Transcript abundance, however, did not appear to be the main driver of protein accumulation in the

transgenic these plant lines. Instead efficient translation would appear to stabilize properly processed transcripts.

Analysis of codon usage within the downstream box regions tested showed that high-level protein accumulation correlated with frequently used plastid codons. Moreover, an analysis of codon usage within highly expressed plastid ORFs revealed differential codon usage within the DB regions of highly expressed genes as compared with the overall codon usage within the chloroplast, similar to observations in *E. coli*. These differences in codon usage preferences were exploited to design codon-optimized DB regions for high-level foreign protein production in chloroplasts and other prokaryotic hosts.

BIOGRAPHICAL SKETCH

Benjamin N. Gray was born on April 17th, 1981 in Lancaster, Pennsylvania and attended Hempfield High School. He earned his bachelor's degree from the Johns Hopkins University, majoring in chemical and biomolecular engineering. He joined Dr. Beth Ahner's lab in the Department of Biological and Environmental Engineering at Cornell University in 2005 and worked in both Dr. Ahner's and Dr. Maureen Hanson's lab to complete his doctoral studies.

ACKNOWLEDGMENTS

I would like to thank Drs. Beth Ahner and Maureen Hanson for their guidance during my time at Cornell University. I would also like to thank Dr. Larry Walker and Dr. David Wilson, who served on my committee. I would like to thank the members of the Ahner and Hanson labs for their scientific discussions and camaraderie. In particular, Tim Vadas, Wade Heller, John Robbins, and Stephane Bentolila played important roles in providing technical guidance and making the time in the lab enjoyable. I would also like to thank Dr. David Wilson and Diana Irwin for their technical guidance, expertise, and years of work in cloning and characterizing the *T. fusca* genes that made this project possible. I would also like to thank my wonderful fiancée, Jess Walden, and the High Noon Athletic Club, for maintaining my sanity and making life at Cornell (and beyond) great.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH iii

ACKNOWLEDGMENTS iv

TABLE OF CONTENTS v

LIST OF FIGURES vi

LIST OF TABLES viii

CHAPTER 1 STRATEGIES FOR HIGH-LEVEL FOREIGN PROTEIN PRODUCTION IN HIGHER PLANT CHLOROPLASTS 1

CHAPTER 2 HIGH-LEVEL BACTERIAL CELLULASE ACCUMULATION IN CHLOROPLAST-TRANSFORMED TOBACCO MEDIATED BY DOWNSTREAM BOX FUSIONS 45

CHAPTER 3 STABILIZATION OF mRNA BY EFFICIENT DOWNSTREAM BOX FUSIONS TO ALLOW HIGH-LEVEL ACCUMULATION OF ACTIVE BACTERIAL BETA-GLUCOSIDASE IN TOBACCO CHLOROPLASTS 79

CHAPTER 4 CODON OPTIMIZATION OF THE GFP DOWNSTREAM BOX REGION FOR IMPROVED CELLULASE EXPRESSION IN TOBACCO CHLOROPLASTS AND IN *E. coli* 138

CHAPTER 5 DESIGN OF CODON-OPTIMIZED SYNTHETIC DOWNSTREAM BOXES FOR FOREIGN PROTEIN PRODUCTION IN TOBACCO CHLOROPLASTS 173

CHAPTER 6 EXTENSIVE HOMOLOGOUS RECOMBINATION BETWEEN INTRODUCED AND NATIVE REGULATORY PLASTID DNA ELEMENTS IN TRANSPLASTOMIC PLANTS 204

CHAPTER 7 CONCLUSIONS 243

APPENDIX TOBACCO CHLOROPLAST TRANSFORMATION 248

LIST OF FIGURES

Figure 1.1 Schematic Diagram of Typical Plastid Transformation Vectors 4
Figure 1.2 Schematic Diagram of ptrnI-RT Plastid Transformation Vector 14
Figure 1.3 Schematic Illustration of Translation of mRNAs with Efficient and Inefficient DB Regions 29

Figure 2.1 DB-Cel6A Plastid Transformation Vector Diagram 56
Figure 2.2 DNA Blot of DB-Cel6A Chloroplast Transformants 58
Figure 2.3 Immunoblot of DB-Cel6A Tobacco Protein 59
Figure 2.4 Chloroplast-produced Cel6A Activity Assays 61
Figure 2.5 Cellulose Affinity Purification of Cel6A 62
Figure 2.6 Photographs and Immunoblots from T1 Generation DB-Cel6A Tobacco 64
Figure 2.7 RNA Blot of *db-cel6A* RNA 66

Figure 3.1 DB-BglC Plastid Transformation Vector Diagram 92
Figure 3.2 DNA Blots of DB-BglC Chloroplast Transformants 93
Figure 3.3 Immunoblots of DB-BglC Tobacco Protein 95
Figure 3.4 RNA Blot of *db-bglC* RNA 96
Figure 3.5 RNA Blot of Polysome Fractionation of *nptII-bglC* RNA 98
Figure 3.6 5’ and 3’ Termini of *db-bglC* Monocistrons 100
Figure 3.7 Predicted RNA Secondary Structures at 5’ and 3’ Termini of *db-bglC* Monocistrons 101
Figure 3.8 Detection of Polyadenylated *db-bglC* Transcripts 103
Figure 3.9 Cellobiose Hydrolysis Assay with Chloroplast-Produced NPTII-BglC 105
Figure 3.10 Hydrolysis of Tobacco Lignocellulose 106
Figure 3.11 MolecularAnalysis of Marker-Free NPTII-BglC Tobacco 108
Figure 3.12 Confocal Microscopy of NPTII-BglC and Marker-Free NPTII-BglC Tobacco 110
Figure 3.13 DNA Blot with Marker-Free NPTII-BglC Tobacco DNA 112
Figure 3.14 Cre-Mediated Recombinations in Marker-Free NPTII-BglC Tobacco 114
Figure 3.S1 Phenotype of Marker-Free NPTII-BglC Tobacco 131
Figure 3.S2 Alignment of *loxP*-like Sites Causing Recombination in Tobacco Plastids 132

Figure 4.1 Serine Codon Usage in Tobacco Chloroplasts 147
Figure 4.2 Immunoblot of *E. coli*-produced GFP-Cel6A 149
Figure 4.3 GFP-Cel6A Plastid Transformation Vector Diagrams 151
Figure 4.4 DNA Blot of GFP-Cel6A Tobacco DNA 152

Figure 4.5 Immunoblot of GFP-Cel6A Tobacco Protein 154
Figure 4.6 RNA Blots of GFP-Cel6A Tobacco RNA 155
Figure 4.7 Comparison of GFP-Cel6A Tobacco Transcript and Protein Accumulation 160
Figure 4.S1 Secondary Structures of Codon-Altered GFP Downstream Box Regions 168
Figure 4.S2 *E. coli* and *N. tabacum* plastid CUFs 169

Figure 5.1 Diagrams of SynDB-Cel6A and SynDB-BglC Plastid Transformation Vectors 181
Figure 5.2 DNA Blots with SynDB-Cel6A and SynDB-BglC Tobacco 182
Figure 5.3 Photographs of HF(NtDB)Syn-Cel6A Regeneration 183
Figure 5.4 Immunoblots of SynDB-BglC and SynDB-Cel6A Protein 185
Figure 5.5 RNA Blots of SynDB-BglC and SynDB-Cel6A RNA 187
Figure 5.6 RNA Secondary Structure Predictions of Synthetic DBs 189
Figure 5.7 Proposed Alteration to HF(NtDB)Syn-Cel6A Plastid Transformation Vector 192

Figure 6.1 DNA Blots of WT and GFP-Cel6A tobacco 212
Figure 6.2 PCR Screening of GFP-Cel6A Tobacco for Recombinant DNA Species 214
Figure 6.3 Schematic Diagram of GFP-Cel6A Recombinant DNA Species 215
Figure 6.4 Schematic Diagram of 22XE2 Recombinant DNA Species 219
Figure 6.5 Schematic Diagram of MR220 Recombinant DNA Species 221
Figure 6.6 PCR Screening of UR-1 Tobacco for Recombinant DNA Species 223
Figure 6.7 *aadA*-Probed DNA Blot with MR220, TetC-Cel6A, 22XE2 and UR-1 Tobacco DNA 225
Figure 6.8 Secondary Recombination Event in UR-1 Tobacco 228
Figure 6.S1 Quantification of Recombinant DNA Species in DB-Cel6A and MR220 Tobacco 237

Figure 7.1 DB-Cel6A and DB-BglC Protein Accumulation in Transplastomic Tobacco 246

Figure A.1 Tobacco Seedling Spacing for Bombardment 249
Figure A.2 Tobacco Seedling Size When Ready for Bombardment 250
Figure A.3 Gene Gun, Showing Positions That can be used for Bombardment 251
Figure A.4 Bombarded Leaf Pieces Transferred to Selective Medium 253

LIST OF TABLES

Table 1.1 Comparison of Protein Expression Levels from the *N. tabacum* Nuclear and Plastid Genomes 6

Table 2.S1 List of Primer Sequences 74

Table 3.S1 List of Primer Sequences 129
Table 3.S2 5’ and 3’ Termini of *db-bglC* Transcripts 130

Table 4.1 Sequences of modified GFP DB Regions 148
Table 4.S1 Codon Usage Frequences in *E. coli* and *N. tabacum* Chloroplast 165
Table 4.S2 List of Primer Sequences 167

Table 5.1 HF(NtDB)Syn and LF(NtDB)Syn DB Sequences 177
Table 5.S1 Codon Usage Frequences in the *N. tabacum* Chloroplast 198
Table 5.S2 List of Primer Sequences 200

Table 6.1 Quantification of Recombinant DNA Species in T0 GFP-Cel6A Tobacco 217
Table 6.S1 List of Primer Sequences 234
Table 6.S2 Primer Pairs Used for PCR Amplification of Plastid DNA Species 235
Table 6.S3 Sizes of Plastid DNA Elements in Transformation Vectors 236

Chapter 1: Strategies for High-Level Foreign Protein Production in Higher Plant Chloroplasts

ABSTRACT

The expression of valuable foreign proteins in transgenic plants has been pursued by a number of researchers and biotechnology companies due to the potential for significant production cost savings relative to traditional cell culture protein production systems and to the relative ease of scale-up of plant-based protein production. Expression of foreign proteins from the plastid genome has been shown in a number of cases to result in significantly higher yields of foreign protein than expression of the same gene from the nuclear genome. The sequences important for foreign gene regulation in plastids are generally similar to bacterial regulatory sequences, but important differences exist. This review discusses the regulatory sequences required for high-level foreign protein production in the plastids of higher plants and techniques to optimize foreign protein production.

INTRODUCTION

Foreign protein expression *in planta* is a strategy that has been pursued for expression of a number of different proteins, including antibodies (reviewed in Peeters et al 2001), vaccines (reviewed in Rybicki 2009), industrial enzymes (reviewed in Hood and Woodard 2002), herbicide (e.g., Ye et al 2001) and insecticide (e.g., De Cosa et al 2001; Chakrabarti et al 2006) resistance proteins, and enzymes for the production of desirable secondary metabolites such as vitamin A in golden rice (Beyer et al 2002) or bioplastics (Lössl et al 2003, 2005; Mooney 2009). Major rationales for the expression of foreign proteins *in planta* include relative ease of scale-up and the potential for significant production cost savings relative to more traditional protein production systems such as Chinese hamster ovary (CHO) cells, insect cells, or microbes (Hood and Woodard 2002; Twyman et al 2003). In order to achieve the lowest possible production cost, high-level protein production, i.e., a high concentration of foreign protein in transformed tissue, is desired.

Plants contain three genomes, located in the nucleus, plastids, and mitochondria. Transformation of the mitochondrial genome and regeneration of transformed plants has not been achieved to date, but transformation of the nuclear genome is a relatively routine procedure in many plant species. Plastid transformation of tobacco (*Nicotiana tabacum*) is now a relatively routine procedure, and a number of other Solanaceous species including tomato (Ruf et al 2001) and potato (Sidorov et al 1999) as well as several species from outside the family *Solanaceae* (e.g., cotton, soybean, lettuce, and poplar; Kumar et al 2004; Dufourmantel et al 2004; Lelivelt et al 2005; Okumura et al 2006) have been successfully transformed. Plastid transformation of monocots has been achieved (Khan and Maliga 1999; Lee et al 2006), but homoplasmic transformants could not be recovered. It is likely that monocot plastid transformation and homoplasmic transformant regeneration will be achieved through new tissue culture techniques.

Typical nuclear transformation protocols involve the use of *Agrobacterium* sp. to insert a gene of interest into the nuclear genome at a random site and with variable copy numbers (Cluster et al 1996). Though various innovations to the standard *Agrobacterium*-mediated transformation protocol (e.g., Mallory et al 2002) have allowed for increased levels of foreign protein in transformed tissues, plant nuclear transformants generally produce foreign protein at a concentration of up to a few percent of total soluble protein (%TSP). This is due in part to variable copy number, to positional effects (i.e., expression of the foreign gene may be limited by insertion at a position in the nuclear genome that is unfavorable for efficient transcription or that disrupts a necessary native gene) and to silencing of highly-expressed genes by RNAi (Hobbs et al 1990, 1993; De Wilde et al 2000). Post-translational subcellular targeting of foreign proteins produced from genes inserted into the nuclear genome is also important in determining foreign protein accumulation (Ziegelhoffer et al 2001; Hood et al 2007). In contrast to plant nuclear transformation, plastid transformation is typically performed by either biolistic bombardment of plant tissue with the transformation vector (Svab and Maliga 1993) or by PEG-mediated transformation of protoplasts (Golds et al 1993). Plastid transformation is achieved by homologous recombination between the transformation vector and the plastid genome, resulting in integration of the gene(s) of interest at a predictable, pre-determined site (i.e., between the left and right flanking regions; Figure 1.1). The plastid genome is present in an extremely high copy number (100-10,000 copies per cell) relative to the nucleus (2 copies per cell), and there is no evidence that plastids possess RNAi machinery, resulting in a lack of silencing of highly-expressed genes. These characteristics have made it possible to express a number of foreign proteins at extremely high levels from the plastid genome of higher plants, with many reports of foreign protein yields of 5-15%TSP (reviewed in Maliga 2003), and some exceptional yields of 30%TSP or higher (Oey et al 2009a, 2009b; De Cosa et al 2001). The lack of positional effects due to targeted transformation of the plastid genome by homologous recombination results in extremely reproducible and heritable protein accumulation levels (Dufourmantel et al 2006), in contrast with nuclear transformants, where protein accumulation is quite variable among independently transformed plants and in progeny plants grown from the transformed plants' seed (Yin et al 2004).

The characteristics of the plastid genome discussed above have allowed for extremely high-level accumulation of valuable proteins expressed from the plastid genome. A number of proteins have now been expressed in both nuclear tobacco

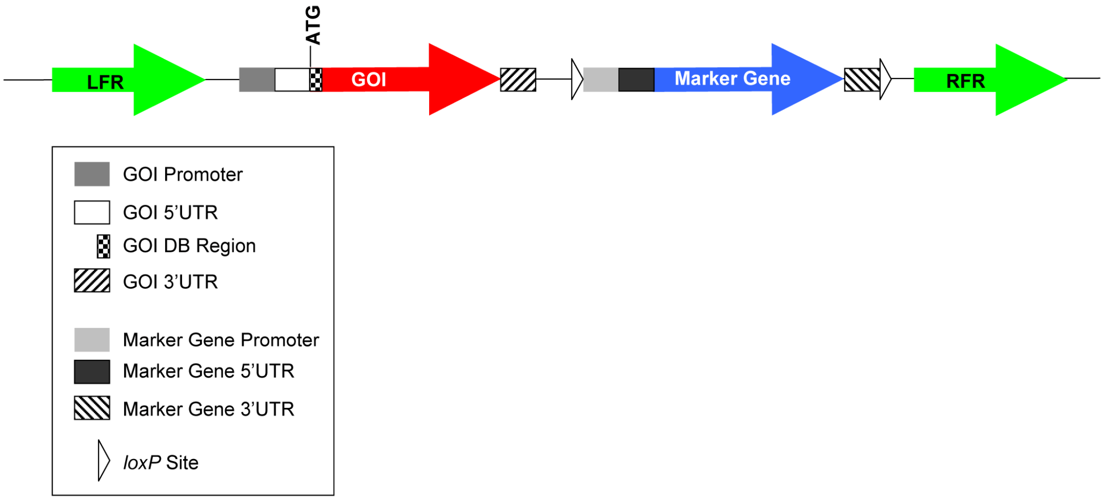


Figure 1.1. Schematic diagram of a typical plastid transformation vector. The left and right flanking regions (LFR and RFR, respectively) are shown, flanking the gene of interest (GOI) and marker gene to be inserted between the LFR and RFR in the plastid genome. *loxP* sites flank the marker gene expression cassette for future removal of the marker gene by Cre/*loxP* recombination.

transformants and in transplastomic tobacco lines, allowing for a comparison of protein yields from these two transformation strategies. Table 1.1 shows several of these comparisons, with plastid transformation often resulting in protein yields over an order of magnitude greater than nuclear transformation. It should be noted that further optimization of expression signals could improve the foreign protein yields of both nuclear and plastid transformants, but that with current technology, expression of foreign proteins from the plastid genome results in improved yields.

The bacteria-like transcription and translation machinery in the plastids of higher plants allows for comparisons to *Escherichia coli* and other well-studied bacterial protein expression systems. Many of the regulatory regions used for plastid expression of foreign proteins are analogous to their bacterial counterparts (e.g., 5’ and 3’ untranslated regions and bacteria-like promoters), and at least some plastid regulatory regions are active in *E. coli* (e.g., the *psbA* promoter; Brixey et al 1997), and vice versa (e.g., the *E. coli trc* promoter; Newell et al 2003). Nonetheless, important differences exist between higher plant plastids and *E. coli*, and in at least one case these differences have been exploited for high-level expression in plastids of proteins that are toxic when expressed in *E. coli* (Oey et al 2009a). This review will discuss strategies to achieve high-level foreign protein production in higher plant plastids, with an emphasis on the strategic use of gene regulatory regions.

TYPICAL FEATURES OF PLASTID TRANSFORMATION VECTORS

A schematic diagram of a typical plastid transformation vector is shown in Figure 1.1. The left and right flanking regions (LFR and RFR, respectively) are each ~0.75-1.5 kb regions of DNA from the plastid genome, and are used to target the insertion of the gene of interest (GOI) in an intergenic region between the LFR and RFR. The GOI is linked to a marker gene, typically encoding an antibiotic resistance protein, to allow

Table 1.1. Comparison of protein expression levels from the *N. tabacum* nuclear and plastid genomes

Protein	Nuclear Expression (%TSP)	Plastid Expression (%TSP)
Cel6A	0.1[a]	11[b]
Cel6B	0.02[a]	3[c]
GUS	3[d]	6.3[e]
EPSPS	0.04[f]	>10[f]

[a]Ziegelhoffer et al 1999; [b]Gray et al 2009a; [c]Yu et al 2007; [d]Herz et al 2005; [e]Mallory et al 2002; [f]Ye et al 2001

for selection of transformed plants. The most commonly used marker genes for plastid transformation of higher plants are *aadA*, conferring spectinomycin and streptomycin resistance (Svab and Maliga 1993), *aphA-6*, conferring kanamycin resistance (Huang et al 2002), and less commonly, *neo*, also conferring kanamycin resistance (Carrer et al 1993). The GOI and the marker gene are both regulated by 5' and 3' untranslated regions (5'UTR and 3'UTR, respectively) that largely regulate translation efficiency and RNA stability, respectively (discussed in greater detail below). Also shown is the downstream box (DB) region of the GOI, defined by the 10-15 codons immediately downstream of the ATG start codon. Advanced plastid transformation vectors often contain *loxP* sites flanking the marker gene expression cassette. Inclusion of these *loxP* sites allows for excision of the marker gene by Cre/*loxP* recombination after introduction of a plastid-targeted Cre gene by nuclear transformation (e.g., Corneille et al 2001). An alternate method of marker gene removal uses the phiC31 recombinase in conjunction with *attB/attP* sites flanking the marker gene (Kittiwongwattana et al 2007). While the inclusion of sequences for marker gene excision is not universal and is not required for expression of the GOI, inclusion of these sequences can allow for the generation of transplastomic plants lacking an antibiotic resistance gene that may be more acceptable to the public due to concerns over, e.g., horizontal transfer of antibiotic resistance genes. Furthermore, expression of the marker gene could represent a metabolic burden to the transplastomic plant. If this is the case, then removal of the marker gene could have the potential to improve expression of the GOI by relieving this metabolic burden, particularly under non-ideal growth conditions (e.g., under drought stress or conditions of high salinity).

GENE REGULATORY SEQUENCES IMPORTANT FOR HIGH-LEVEL PROTEIN EXPRESSION

Promoters

Plastid transcription is accomplished by the combined actions of two RNA polymerases recognizing different promoters, a T7-like single subunit nuclear-encoded polymerase (NEP) and a bacteria-like $\alpha_2\beta\beta'$ plastid-encoded polymerase (PEP). Transcription in undifferentiated plastids and in non-green tissues is performed primarily by the NEP, resulting in the production of ribosomal RNA and of mRNAs encoding ribosomal proteins that are included in the PEP, ultimately resulting in the accumulation of functional PEP. As chloroplasts develop, transcription of many plastid genes shifts to the PEP (reviewed in Hess and Börner 1999). Many plastid promoters have both PEP and NEP transcription start sites, and can be transcribed by either polymerase depending on growth conditions or when the expression of one of the two polymerases is eliminated (Allison et al 1996; Hajdukiewicz et al 1997).

Transcription of foreign genes inserted into the plastid genome is typically driven by plastid promoters included in the plastid transformation vector upstream of the GOI. The two promoters most often used in plastid transformation vectors are the ribosomal promoter (P*rrn*) and the *psbA* promoter (P*psbA*). P*rrn* contains both PEP and NEP transcription start sites, while P*psbA* contains only a PEP transcription start site (Allison et al 1996). It is conceivable that other less commonly used plastid promoters could be used to drive high-level expression of transgenes. A microarray study identified a number of highly expressed plastid genes and a number of plastid genes that are highly up-regulated in the light (Nakamura et al 2003). This study could be used as a basis for the identification of plastid promoters likely to be useful for high-level transgene expression. As an example, *clpP* was found to be highly expressed based on the data of Nakamura et al (2003). The use of P*clpP* for foreign gene transcription has been described in the patent literature (U.S. Patents 6,362,398; 6,624,296; and 7,129,397), though only one report exists in the scientific literature to our knowledge describing the use of P*clpP*, and this study was not designed to optimize transgene expression (Sriraman et al 1998). The *clpP* promoter or other plastid promoters may merit further study for the regulation of foreign genes in transplastomic plants.

The inclusion of promoters or other DNA sequences in the plastid transformation vector that are homologous to native plastid DNA has resulted in unintended and unwanted recombination events between the introduced and native copies of the plastid DNA element, making recovery of the desired transformation event more difficult. An example is the UR-1 line of tobacco, in which the P*psbA* sequence driving transcription of the *aadA* gene, rather than the LFR, mediated homologous recombination. This produced the UR-1 line of transplastomic tobacco lacking the GOI, but containing the *aadA* expression cassette (Gray et al 2009b). It may be possible to use non-plastid promoters that could give levels of mRNA comparable to or higher than the native plastid promoters typically used to drive foreign gene expression. Several groups have attempted to use bacterial (Newell et al 2003; Birch-Machin et al 2004; Mühlbauer and Koop 2005) or mitochondrial (Bohne et al 2007) promoters, and have found that these promoters are functional in plastids, but do not drive gene expression as well as either P*rrn* or P*psbA*. It is conceivable, however, that some as yet untested bacterial or mitochondrial promoters could work as well or better than the native plastid promoters typically used. Alternatively, computational approaches may be able to yield effective synthetic promoters to drive foreign gene expression, either constitutively or under certain growth conditions. One such synthetic promoter system has been described (Mühlbauer and Koop 2005), in which the native plastid P*rrn* promoter was altered to include *lac* operator sequences from the *Escherichia coli lac* operon. The novel promoters were used to drive IPTG-inducible expression of GFP. This approach resulted in transplastomic tobacco lines in which GFP expression was upregulated 20-fold following the spraying of a 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) solution on the plants.

In addition to the IPTG-inducible promoter developed by Mühlbauer and Koop (2005) described above, several hybrid transcription systems have been developed. These systems involve the use of a promoter recognized by T7 RNA polymerase, derived from the T7 bacteriophage, or of a promoter requiring the presence of a particular sigma factor not normally present in the plastid. Nuclear transformation is performed with transplastomic plants to introduce a plastid-targeted T7 RNA polymerase gene (McBride et al 1994) or sigma factor gene (Buhot et al 2006), regulated by an inducible promoter. The T7 RNA polymerase hybrid transcription system has been used successfully to demonstrate production of polyhydroxybutyric acid (PHB) in plastids (Lössl et al 2005). In this system, PHB production resulted in male sterility and growth reduction when PHB-synthesizing enzymes were expressed constitutively (Lössl et al 2003). By introducing a promoter recognized by T7 RNA polymerase to regulate the PHB-synthesizing enzymes and an inducible nuclear-encoded, plastid-targeted T7 RNA polymerase gene, fertile plants with normal growth characteristics were obtained (Lössl et al 2005). When T7 RNA polymerase production was induced, modest levels of PHB were produced and the growth problems associated with PHB production were observed, but in the absence of T7 RNA polymerase production, growth was normal. These studies demonstrate the potential utility of hybrid transcription systems for the expression of genes with deleterious effects on the plant. Plants can be grown in non-inducing conditions, repressing production of the RNA polymerase or sigma factor required for transcription of the toxic gene. At the desired time during plant growth, or even postharvest, the required RNA polymerase or sigma factor gene can be induced to start production of the toxic protein or secondary metabolite. By choosing the appropriate nuclear promoter, T7 RNA polymerase or the necessary sigma factor could be produced only in certain tissues or at certain stages of plant development to result in transcription of the gene of interest only in certain parts of the plant. As a hypothetical example, this could be useful if a certain protein is toxic for seed production when expressed in flowers. By choosing nuclear promoters not active in flowers to regulate the hybrid transcription factor, this protein could be expressed in transplastomic plants without affecting seed production. Disadvantages to these hybrid transcription systems are those associated with nuclear transformation discussed above, namely the potential for gene silencing or low expression levels of the nuclear-encoded RNA polymerase or sigma factor, which could result in limiting levels of these proteins, and the variable expression levels among independent nuclear transformants and among the progeny of primary transformants that could result in variable levels of plastid-produced protein. Additionally, it has been shown that the T7 RNA polymerase recognizes at least some NEP promoters, resulting in altered plastid gene transcription and a pale green phenotype in seedlings when the T7 RNA polymerase is expressed constitutively (Magee et al 2007). Nonetheless, like the IPTG-inducible plastid promoter developed by Mühlbauer and Koop (2005), these hybrid transcription systems could be valuable when attempting to express toxic or lethal proteins in the plastid. The potential for tissue-specific or developmental stage-specific expression may be of value for certain applications.

In the approaches described above, a promoter is included in the plastid transformation vector upstream of the foreign gene of interest to drive transcription of the foreign gene. An alternate method of foreign gene transcription takes advantage of the highly processive plastid RNA polymerase and inefficient termination at plastid 3’ untranslated regions (3'UTRs). In this approach, a promoterless foreign gene is inserted into the plastid genome downstream of a highly-transcribed plastid gene. Because transcription termination is inefficient in plastids (Stern and Gruissem 1987), the foreign gene is transcribed as part of a polycistron along with the gene(s) normally transcribed from the plastid promoter. By carefully choosing the insertion site in the plastid genome, this approach can result in high levels of mRNA and can give extremely high yields of foreign protein. An early description of this type of system demonstrated that a promoterless *uidA* gene inserted downstream of the plastid *rbcL* gene resulted in approximately four-fold higher GUS protein levels than a construct containing a heterologous ribosomal promoter inserted at the same site in the plastid genome, despite a greatly increased concentration of monocistronic *uidA* mRNA when P*rrn* was included upstream of the *uidA* ORF (Staub and Maliga 1995). Herz et al (2005) demonstrated high-level (~4%TSP) accumulation of GUS protein from a promoterless *uidA* gene inserted downstream of the plastid *psbA* gene. Herz et al (2005) also demonstrated that promoterless genes could be expressed when inserted into the *atpB/E* operon, but did not fully characterize these transplastomic lines. Chakrabarti et al (2006) inserted the *cry9Aa2* gene, encoding a Bt insect resistance protein, downstream of the plastid ribosomal promoter in the *trnI*/*trnA* intergenic region, resulting in high-level (10-20%TSP) accumulation of Bt protein in both soluble and insoluble fractions. Gray et al (2009a; Chapter 3) used a nearby insertion site in the *trnI*/*trnA* intergenic region to drive high-level (10-12%TSP) accumulation of a cellulase and a β-glucosidase. These experiments demonstrate that promoterless constructs designed for insertion of a foreign gene downstream of a highly-transcribed plastid gene can result in extremely high-level accumulation of the desired foreign protein. Promoterless constructs relying on read-through transcription have the advantage of avoiding undesirable recombination events between introduced and native promoter elements (Gray et al 2009b), and a promoterless construct has been shown in at least one instance (Staub and Maliga 1995) to be superior for protein production to a promoter-containing construct utilizing the same insertion site. I developed a versatile vector for read-through transcription of foreign genes inserted in the *trnI/trnA* intergenic region, ptrnI-RT (Figure 1.2). This vector contains a multi-cloning site for insertion of a gene of interest between the T7g10 5'UTR and the *psbA* 3'UTR (T*psbA*), as well as an *aadA* expression cassette flanked by *loxP* sites. A similar vector was used for high-level Cel6A (Gray et al 2009a) and BglC (Chapter 3) expression. It is likely that promoterless foreign genes can be inserted downstream of other highly-expressed plastid genes (e.g., *clpP*, as discussed above) to result in high-level foreign protein production.

5' Untranslated Regions

In plastids, as in other bacteria-like systems, the 5' untranslated region (5'UTR) of many genes contains a Shine-Dalgarno (SD) sequence (GGAGG) located approximately 10 nt upstream of the translation start codon that can base pair with ribosomal RNA to initiate translation. Unlike most bacterial 5'UTRs, plastid SD sequences are not strictly required, and many plastid 5'UTRs do not contain functional SD sequences, apparently using an alternate method to initiate translation (Hirose and Sugiura 2004). Some plastid 5'UTRs (e.g., *rbcL* and *atpE*) contain SD sequences that are essential for efficient translation initiation, while plastid 5'UTRs lacking functional SD sequences (e.g., *psbA* and *clpP*) may interact with nuclear-encoded, plastid-targeted proteins to regulate translation initiation.

Similar to the promoters typically used to drive transcription of foreign genes in plastids, 5'UTRs used to regulate foreign gene expression are often derived from highly-expressed plastid genes. The most commonly used plastid-derived 5'UTRs are

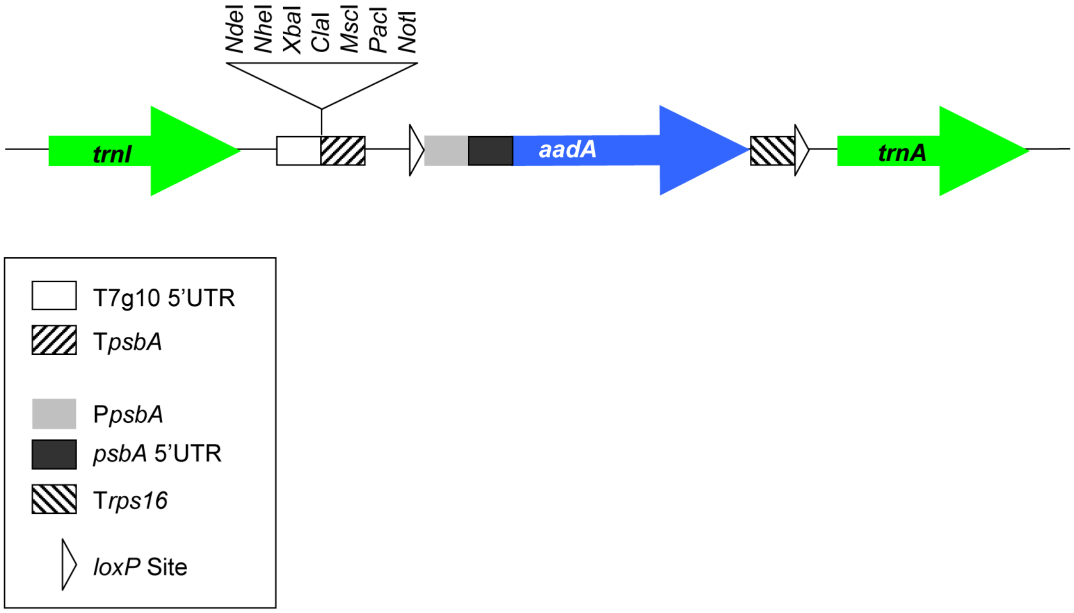


Figure 1.2. Plastid transformation vector ptrnI-RT. This versatile vector is designed for transgene insertion between the plastid *trnI* and *trnA* genes of the ribosomal RNA operon in the inverted repeat of the plastid genome. A multi-cloning site is included between the T7g10 5'UTR and *psbA* 3'UTR for transgene regulation, and an *aadA* expression cassette flanked by *loxP* sites is included for spectinomycin/streptomycin-based selection of plastid transformants.

from the *rbcL*, *psbA*, and *atpB* genes. Overexpression of some plastid-derived 5’UTRs can have deleterious effects, as was noted when the *clpP* 5’UTR was expressed at a high level to regulate the *neo* gene encoding the NPTII protein (Kuroda and Maliga 2002). In these plants, which accumulated modest levels of NPTII protein (0.3%TSP), a chlorotic phenotype was observed in young leaves that appeared to result from alterations in the normal splicing of *clpP* mRNA. The authors concluded that overexpression of the *clpP* 5’UTR resulted in competition for an RNA binding protein that normally interacts with the native *clpP* 5’UTR and is important for *clpP* transcript maturation. This study demonstrates the potential for unintended deleterious effects on plant health caused by overexpression of native plastid 5’UTRs.

Plastid 5’UTRs can also cause unintended effects on expression of the gene of interest, though these effects may not be deleterious to the health of the transplastomic plant. An example is the light-dependent accumulation of foreign proteins regulated by the *psbA* 5’UTR. Staub and Maliga (1994) observed that the accumulation of GUS protein from a *uidA* gene regulated by the *psbA* promoter and 5’UTR was up to 196 times higher in light-grown than in dark-grown seedlings, despite relatively minor changes (3 to 5-fold differences) in *uidA* mRNA levels. Light/dark cycling only weakly affected GUS accumulation from *uidA* genes regulated by the *rps16* or *rbcL* promoters and 5’UTRs, and deletion of the *psbA* 5’UTR while retaining the *psbA* promoter resulted in a loss of the light-altered GUS accumulation profile. These results strongly suggest that the *psbA* 5’UTR (and not the *psbA* promoter) is responsible for the effects of light/dark cycling on protein accumulation. Light-mediated effects on protein production could be either desirable or undesirable for foreign protein production, depending on the application, though it is typically desirable to produce high levels of foreign protein regardless of light conditions.

A number of examples of extremely high accumulation of foreign proteins from genes regulated by plastid-derived *atpB*, *psbA*, and *rbcL* 5'UTRs have been described. Examples include the AAD-GFP fusion protein produced at 8%TSP and 18%TSP using the *atpB* and *rbcL* 5'UTR regions, respectively (Khan and Maliga 1999), NPTII protein produced at 7%TSP and 11%TSP using the *atpB* and *rbcL* 5'UTR regions, respectively (Kuroda and Maliga 2001a), and human serum albumin produced at up to 11%TSP using the *psbA* 5'UTR (Fernández-San Millán et al 2003). Other plastid 5'UTRs have also been tested for transgene regulation, but not as thoroughly and without having achieved high-level foreign protein expression (e.g., 5'UTRs derived from the *rps19/rpl22*, *psaA/B*, and *psbD/C* operons; Herz et al 2005). It is possible that 5'UTRs that have been ineffective for high-level transgene expression thus far could mediate high-level foreign protein accumulation in the right context. An example of the variability associated with a given 5'UTR regulating a transgene is found in the work of Kuroda and Maliga (2001a), where NPTII protein accumulation varied from 0.3%TSP to 10.8%TSP under the control of the *rbcL* 5'UTR and from 4%TSP to 7%TSP under the control of the *atpB* 5'UTR, depending on the identity of the 5' coding region (the downstream box region, described below) fused to the *neo* ORF.

RNA stability can be affected in some cases by 5'UTR primary and secondary structure, as shown for *uidA* genes regulated by the *psbA* 5'UTR and the *rbcL* 5'UTR. Deletions in the psbA 5'UTR hairpin-loop structure resulted in up to 3-fold decreases in *uidA* mRNA (Zou et al 2003), consistent with the changes in *uidA* mRNA levels observed by Staub and Maliga (1994) as a result of light/dark cycling for a *uidA* gene regulated by the *psbA* 5'UTR. Shiina et al (1998) reported that *uidA* mRNA is stabilized against degradation in the dark when regulated by the *rbcL* 5'UTR. The native *rbcL* gene is transcribed at a lower rate in the dark than in the light, resulting in relatively stable levels of the native *rbcL* transcript. Use of the *rbcL* 5’UTR to regulate a transgene transcribed from a constitutive promoter could therefore result in elevated levels of foreign transcript in the dark as a result of the RNA-stabilizing properties of the *rbcL* 5’UTR. The role of plastid 5’UTRs in stabilizing or destabilizing RNA is not well-studied as compared with the role of the 5’UTR in promoting efficient translation, but the effects of 5’UTRs on mRNA stability appear to be minor compared with the effects of 5’UTRs on translation efficiency. Transcript stability appears to be regulated largely by the 3’UTR, as discussed below.

The use of 5’UTRs not derived from plastid genes can avoid undesired effects on plant health like those observed when overexpressing the *clpP* 5’UTR or on protein accumulation like those observed in light/dark cycling of foreign genes regulated by the *psbA* 5’UTR. One 5’UTR that has been found to be particularly effective for high-level foreign protein production in plastids is the T7g10 5’UTR, derived from the gene encoding the coat protein (gene 10) of the T7 bacteriophage and commonly used for expression of foreign proteins in *E. coli* (Olins et al 1988). The T7g10 5’UTR contains a consensus SD sequence that is likely to be important for translation initiation in plastids. Many of the highest-accumulating foreign proteins in plastids have been regulated by the T7g10 5’UTR (e.g., Oey et al 2009a, 2009b; Kuroda and Maliga 2001b; Tregoning et al 2003), demonstrating the utility of this 5’UTR for driving high-level foreign protein accumulation. Experiments in our lab have shown that 5’UTRs derived from the coat protein-encoding genes of bacteriophages other than T7 can function effectively in plastids (Yang, Gray, Hanson, and Ahner, unpublished). Additionally, the mitochondrial *atpA* 5’UTR has been been used to regulate a *neo* gene in plastids, mediating NPTII protein accumulation and thereby demonstrating that mitochondrial 5’UTRs can function in plastids, though high-level protein accumulation was not achieved with this 5’UTR (Bohne et al 2007). Further studies with 5’UTRs not derived from plastid genes should yield effective 5’UTRs without the potential for undesired recombination events that can be mediated by plastid-derived sequences (Gray et al 2009b).

The choice of 5’UTR can have major effects on foreign protein concentration, as illustrated in the work of Ye et al (2001). In this study, EPSPS protein accumulation increased approximately 100-fold by changing the 5’UTR from the *rbcL* to the T7g10 5’UTR. Only modest levels of EPSPS, however, were obtained even when the transgene was regulated by the T7g10 5’UTR (~0.2-0.3%TSP with the T7g10 5’UTR as compared with ~0.001-0.002%TSP with the *rbcL* 5’UTR). By adding a short downstream box (DB) fusion to the transgene, however, EPSPS accumulation of greater than 10%TSP was obtained, demonstrating both that the choice of 5’UTR can affect protein accumulation by several orders of magnitude and that the 5’ portion of the coding region can act in concert with the 5’UTR to regulate translation and further increase foreign protein accumluation.

Downstream Boxes

The downstream box (DB) region, defined by the 10-15 codons immediately downstream of the start codon, was first identified in *Escherichia coli* (Sprengart et al 1996). The DB region was found to have major effects on accumulation of foreign protein in *E. coli*, acting synergistically with the SD region upstream of the start codon to regulate protein accumulation. DB function was initially ascribed to base-pairing with ribosomal RNA, as has been shown for the SD region, but base pairing of the DB region with ribosomal RNA has effectively been ruled out by a number of later structural and biochemical studies (e.g., O’Connor et al 1999; La Teana et al 2000; Moll et al 2001). Experiments in *E. coli* suggest that DB-mediated effects on protein accumulation are regulated by the codon makeup of a given DB region, and not by the encoded amino acids or by the individual nucleotides in the DB region (Stenström and Isaksson 2002; Gonzalez de Valdivia and Isaksson 2004). A recent study of 154 *gfp* ORFs encoding the same amino acid sequence found that secondary structure near the 5' end of the ORF (i.e., the DB region) was the best predictor of GFP accumulation of the parameters considered (Kudla et al 2009). This suggests that the DB region may act through several different mechanisms. Clearly, further research is needed in order to gain a better understanding of DB function. The mechanism of DB function is still not well understood, but empirically it is well known that alterations in the DB region can have profound effects on protein accumulation, likely by altering translation efficiency.

Kuroda and Maliga (2001a) first reported that sequences like the DB region in *E. coli* appeared to function in tobacco chloroplasts. In this study, the authors investigated the accumulation of NPTII protein from a *neo* gene fused at its 5' terminus to the first 14 codons from the *rbcL* or *atpB* genes (i.e., the DB regions of these plastid genes) and regulated by the 5'UTRs originating from the plastid *rbcL* or *atpB* genes. The authors then made silent mutations in the *rbcL* and *atpB* DB regions while holding the rest of the coding region (i.e., the *neo* ORF), promoter (P*rrn*), and 5'UTRs constant. Silent mutations to the *rbcL* DB region resulted in a decrease in NPTII protein accumulation from 11%TSP to 0.3%TSP. Silent mutations to the *atpB* DB region resulted in a less dramatic, but still significant decrease in NPTII protein accumulation from 7%TSP to 4%TSP. These changes in NPTII protein accumulation occurred despite the lack of any changes to the amino acid sequence. Levels of *neo* mRNA were not significantly affected by these silent mutations, strongly suggesting a decrease in translation efficiency resulting from silent mutations in the *rbcL* and *atpB* DB regions as the cause of decreased NPTII accumulation. In these experiments, the DB regions of native plastid genes were altered in their native context, i.e., while preceded by their corresponding 5’UTR. This leaves open the possibility that the silent mutations to the *rbcL* and *atpB* DB regions affected an unknown aspect of translation resulting from an interaction between the 5’UTR and DB region requiring the native sequence of the DB regions tested, but this study clearly showed the importance of the DB region in regulating foreign protein accumulation in plastids. A follow-up study by Kuroda and Maliga (2001b) decoupled the effects of the DB region from the 5’UTR by using a single 5’UTR (the T7g10 5’UTR) to regulate a *neo* gene fused to either a “consensus” *E. coli* DB region, a “consensus” plastid DB region, or to a *Nhe*I restriction site resulting in a two amino acid N-terminal fusion. In these experiments, “consensus” DB regions were constructed by assuming that the DB could base pair with ribosomal RNA, and thus the *E. coli* and plastid “consensus” DB regions were changed to optimize the proposed base pairing with the respective ribosomal RNAs. NPTII protein accumulation was dramatically affected in these experiments as well, ranging from 0.2%TSP when fused to the plastid DB to 16%TSP when fused to the *E. coli* DB to 23%TSP when fused to the *Nhe*I restriction site. In this study, *neo* mRNA levels were affected by the DB fusions, with the lowest *neo* mRNA levels corresponding to the lowest NPTII protein levels and the highest *neo* mRNA levels corresponding to the highest NPTII protein levels, though a simple linear relationship between mRNA levels and protein levels was not observed. Alterations in RNA levels were likely a result of RNA degradation and not of differential transcription rates, as the same promoter was used in all three constructs studied in these experiments. This study decoupled the effects of DB fusions from the 5’UTR by holding the 5’UTR constant, but the changes to the DB region were non-silent, resulting in altered amino acid sequences of the encoded NPTII proteins. Kuroda and Maliga showed that plastid DB regions can greatly affect foreign protein accumulation as a result of both silent (2001a) and non-silent (2001b) changes to the

DB region, and that DB-mediated changes in foreign protein accumulation do not require the presence of a plastid-derived 5'UTR to occur (2001b). As is the case for *E. coli*, it does not appear that DB base-pairing with ribosomal RNA is the mechanism of DB-mediated effects on protein accumulation in plastids (2001b).

Downstream box fusions have been utilized to improve foreign protein production in plastids, with order of magnitude effects on foreign protein accumulation. In an early application of a DB fusion to improve plastid foreign protein production, Ye et al (2001) fused the first 14 codons from *gfp* to the ORF encoding EPSPS, which was regulated by the T7g10 5'UTR. Accumulation of EPSPS protein increased from 0.2-0.3%TSP when not fused to the GFP DB region to >10%TSP when fused to the GFP DB region, over a 30-fold increase in protein accumulation. This was particularly significant because Ye et al (2001) showed that altering the entire coding region of the EPSPS gene to include primarily plastid-preferred codons resulted in ~2-fold increases in protein accumulation. Given the labor-intensive nature of altering an entire ORF as compared with adding a short DB fusion, this study showed that fusion of an appropriate DB region to the gene of interest can be an extremely efficient way of improving protein accumulation. Major changes in protein accumulation as a result of DB fusions were also seen by Gray et al (2009a) when expressing Cel6A, a *Thermobifida fusca* endoglucanase. Fourteen amino acid DB regions from TetC, NPTII, and GFP were fused to the *cel6A* ORF. The GFP DB region was used in this study because it was shown to successfully enhance the accumulation of EPSPS when expressed in the plastid (Ye et al 2001). Neither the TetC nor the NPTII DB regions had been used previously, but the full-length TetC (Tregoning et al 2003) and NPTII (Kuroda and Maliga 2001b) genes had been expressed at high levels (25%TSP and 23%TSP, respectively) from the plastid genome. Because the DB region is important for achieving high-level foreign protein, it was hypothesized that the DB regions of these genes could be valuable for directing high-level accumulation of other foreign proteins. Somewhat surprisingly, accumulation of Cel6A protein varied over two orders of magnitude among the three DB regions tested, with GFP-Cel6A, NPTII-Cel6A, and TetC-Cel6A accumulating to ~0.1%TSP, 1%TSP, and 11%TSP, respectively. Monocistronic mRNA levels of the three *DB-cel6A* genes correlated with protein accumulation (i.e., higher monocistronic *tetC-cel6A* RNA levels were observed than *nptII-cel6A* RNA levels, which were in turn higher than *gfp-cel6A* monocistron levels). A similar phenomenon was observed previously when non-synonymous changes were made to DB regions (Kuroda and Maliga 2001b). This study showed that the DB regions of highly expressed genes can be used to direct high-level accumulation of other foreign proteins. In a follow-up study, the TetC, NPTII, and GFP DB regions were fused to the *bglC* ORF, encoding a *T. fusca* β-glucosidase, and inserted into the plastid genome (Chapter 3). These DB regions were again found to be useful for directing high-level foreign protein expression, but surprisingly, protein accumulation did not follow precisely the same trends that were observed when expressing Cel6A. In this study, GFP-BglC, TetC-BglC, and NPTII-BglC accumulated to <<0.3%TSP, 2%TSP, and ~11%TSP, respectively. It is not clear why NPTII-BglC accumulated to higher levels than TetC-BglC, while TetC-Cel6A accumulated to higher levels than NPTII-Cel6A. These studies show that, in the absence of a better mechanistic understanding of DB function, empirical optimization of the DB region is required in order to achieve high-level expression of the protein of interest. Further research on DB function in plastids and in *E. coli* should provide a better understanding of DB mechanism, streamlining the optimization of DB regions for foreign protein production in plastids. Nonetheless, the DB regions from ORFs that are capable of high-level expression in plastids (whether they are native plastid genes or foreign genes that have been expressed to high levels in plastids) can be useful in driving expression of other foreign genes and can serve as a useful starting point for DB optimization.

3' Untranslated Regions

Plastid 3' untranslated regions (3'UTRs), located immediately downstream of the stop codon, typically contain hairpin-loop structures that facilitate RNA maturation and prevent degradation of the RNA (Monde et al 2000a). In contrast to 5'UTRs and DB regions, most 3'UTRs do not play a major role in regulating translation efficiency (Eibl et al 1999), though translation is reportedly affected by the *petD* 3'UTR (Monde et al 2000b). Instead, most plastid 3'UTRs act primarily by regulating RNA processing events and by stabilizing RNA against degradation by ribonucleases, thereby allowing the RNA to accumulate to steady-state levels sufficient for high-level protein production. 3'UTRs used to regulate foreign genes in plastids are typically derived from plastid genes, with the *rps16*, *rbcL*, *psbA*, and *rpl32* 3'UTRs being commonly used.

Plastid 3'UTRs play an important role in regulating polycistron processing to generate monocistronic RNAs. In plastids, as in many bacterial systems, many genes are transcribed as part of a polycistronic operon. These polycistrons are often processed by intergenic endonucleolytic cleavage to produce monocistrons. The generation of monocistronic transcripts is necessary for translation of some plastid mRNAs, such as the maize *petD* transcript (Barkan et al 1994), while other plastid transcripts, such as the tobacco *psbB*, *petB*, and *petD* transcripts, can be translated in a polycistronic context (Barkan 1988). Following endonucleolytic cleavage in the intergenic regions of a polycistron, the mature monocistronic transcript is produced by exonucleolytic processing of the RNA. Exoribonucleases processively cleave nucleotides from the transcript until they reach a hairpin-loop structure in the UTR

(Monde et al 2000a). Unlike the case for many *E. coli* 3'UTRs, plastid 3'UTRs are poor terminators of transcription, functioning primarily post-transcriptionally to protect RNAs against exonucleolytic degradation. In contrast, many *E. coli* 3'UTRs are true transcription terminators, effectively stopping the procession of RNA polymerase (reviewed in Holmes et al 1983). The differences between *E. coli* and chloroplast transcription and RNA maturation have been exploited by Oey et al (2009a) to produce proteins in plastids that are toxic when produced in *E. coli*. Typically, this is a problem because many plastid promoters and 5'UTRs are recognized by *E. coli*, resulting in the inadvertent production of the protein encoded by the gene of interest during the cloning of the plastid transformation vector. Oey et al (2009a) overcame this problem by introducing *E. coli* terminators immediately upstream of the toxic genes of interest, preventing transcription of the genes in *E. coli* hosts. Because of the extremely processive nature of the plastid RNA polymerase, the *E. coli* terminators did not stop transcription of the toxic genes in plastids, allowing the transcription of these ORFs and translation to produce the *E. coli*-toxic proteins in transplastomic plants. Removal of the *E. coli* terminators by Cre/*loxP* recombination resulted in increased protein levels for one of the two *E. coli*-toxic proteins tested, but was not entirely necessary, as extremely high levels of protein (10-30%TSP) were observed even without excision of the *E. coli* terminators. Thus, although *E. coli* and other well-studied bacterial systems are useful for studying many aspects of plastid gene regulation, important differences between these systems exist. These differences can be exploited to allow for expression of proteins in plastids that cannot be expressed in *E. coli*.

While some plastid 3'UTRs have been shown to interact with RNA-binding proteins and affect translational efficiency (e.g., the *petD* 3'UTR; Monde et al 2000b), it appears that most of the functions performed by higher plant plastid 3'UTRs require only the stem-loop structure of the 3'UTR. There is at least one report of a bacterial 3'UTR (the *E. coli rrnB* 3'UTR) being used effectively in a plastid transformation vector, resulting in high-level accumulation of *gfp* mRNA (Newell et al 2003). The mitochondrial *atp9* 3'UTR has also been used effectively to regulate a foreign gene in plastids (Bohne et al 2007). Further, the hairpin-loop structure of the *loxP* sequence has been observed to regulate RNA maturation similarly to the *psbA* 3'UTR (Chapter 3). These results demonstrate that the use of 3'UTRs not derived from plastid genes (e.g., bacterial or mitochondrial 3'UTRs) or of synthetic sequences containing hairpin-loop structures can be effective in regulating transgenes for plastid expression. Because the 3'UTR has been shown to have only minor effects on foreign protein accumulation (Eibl et al 1999), it is recommended that 3'UTRs not derived from plastid genes be used in order to avoid unwanted recombination events between introduced and native copies of a plastid 3'UTR that could lead to plastid transformation without incorporation of the gene of interest (Gray et al 2009b).

Other Regulatory Elements

Zhou et al (2007) reported the identification of an intercistronic expression element (IEE) capable of mediating efficient processing of polycistronic RNAs to generate stable monocistronic transcripts. This IEE was derived from the intergenic region between the plastid *psbN* and *psbH* genes, normally transcribed as part of the plastid *psbB* polycistron. Inclusion of the IEE between the *yfp* and *nptII* ORFs in the plastid transformation vector resulted in the accumulation of monocistronic *yfp* mRNA that was translated far more efficiently than polycistronic transcripts, resulting in the accumulation of YFP protein. It should be noted that the IEE consists of a 50 nt sequence from a 111 nt region that serves as the 5'UTR of the divergently translated *psbN* and *psbH* transcripts following processing of the polycistronic primary transcript of the *psbB* operon containing *psbN* and *psbH*. It is likely that the full-length *psbN*/*psbH* intergenic sequence, or any of the various intergenic sequences in the many plastid polycistronic transcripts, could mediate faithful RNA processing of transcripts containing foreign genes of interest. Identification of the minimum sequence requirements for RNA processing could be useful, however, to minimize the length of plastid-derived sequence included in plastid transformation vectors. Plastid-derived sequences can mediate unwanted recombination events, complicating the recovery of the desired plastid transformant (Gray et al 2009b). Zhou et al (2007) and others (e.g., Barkan et al 1994; Chapter 3) have observed that RNA processing to generate monocistronic transcripts is an important aspect of the regulation of plastid protein production. The IEE identified by Zhou et al (2007) and other sequences capable of directing the processing polycistronic transcripts may be useful to increase foreign protein accumulation in plastids. Further research on the sequence and secondary structure requirements for efficient RNA processing should be a priority.

While not a true regulatory element in the same sense as promoters, UTRs, and IEEs that are discrete sequences, the codon usage of foreign genes to be expressed from the plastid genome should be considered. In *E. coli*, it has been shown that silent codon changes to generate an ORF using codons preferred by *E. coli* can have significant positive effects on protein accumulation (e.g., Makoff et al 1989; Daniell et al 2009). In at least some cases, this may be due to differences in the availability of tRNA species for translation of a given codon (Ikemura 1981). Similar methods have been explored for their effects on plastid gene expression. Higher plant plastid genomes are generally AT-rich, which could pose a problem for expression of GC-rich foreign genes. A number of foreign genes have been altered for plastid expression from their native GC-rich coding sequences to a more AT-rich ORF encoding the same polypeptide. Altering GC-rich sequences to more AT-rich sequences has resulted in ~1.5 to 2-fold gains in protein accumulation, regardless of the protein accumulation level. Optimization of coding sequences to more closely match plastid-preferred codons has generally given less improvement than has been observed in *E. coli*, suggesting that the plastid genome is better able to express ORFs not containing its preferred set of codons than *E. coli* (Daniell et al 2009). Ye et al (2001) expressed a naturally GC-rich gene encoding EPSPS in plastids, and also generated a synthetic EPSPS ORF containing primarily plastid-preferred codons. Altering the coding sequence of the EPSPS gene resulted in an increase of EPSPS protein accumulation from 0.001%TSP to 0.002%TSP when placed behind the *rbcL* 5'UTR, and from 0.2%TSP to 0.3%TSP when placed behind the T7g10 5'UTR. A similar ~2-fold change in protein accumulation was observed by Tregoning et al (2003), who expressed a GC-rich and an AT-rich version of the *tetC* gene in plastids, resulting in TetC accumulation of 10%TSP and 25%TSP, respectively. Thus, it appears that codon optimization of a GC-rich coding sequence to an AT-rich sequence more closely aligned with the codon usage of the plastid genome can result in ~2-fold increases in protein accumulation. This increase could be of great importance when high-level expression has already been achieved with the GC-rich gene, but would be of little use with poorly expressed proteins. Optimization of the 5' portion of a coding sequence (i.e., the DB region) can result in greater improvements in foreign protein production than optimization of the entire coding sequence (Ye et al 2001).

CONCLUSIONS

When expressing a foreign protein *in planta*, the key optimization parameter is the concentration of foreign protein in the harvested tissue. Foreign protein accumulation from plastid-expressed genes is affected by a number of discrete regulatory regions, e.g., promoters, 5' and 3'UTRs, DB regions, and IEEs, as well as by the codon makeup of the foreign gene. Of these regulatory regions, the 5’UTR and the DB regions (i.e., the regions flanking the start codon; Figure 1.1) can have the greatest effects on protein accumulation, suggesting that translation initiation may be limiting for protein production in the plastid.

Changes made to the 5’UTR (e.g., Eibl et al 1999; Ye et al 2001) and to the DB region (e.g., Kuroda and Maliga 2001a, 2001b; Ye et al 2001; Gray et al 2009a; Chapter 3) have resulted in protein accumulation improvements over several orders of magnitude. In at least one case, a foreign gene was initially expressed at 0.002%TSP when regulated by the *rbcL* 5’UTR and unfused to an effective DB region. Changing the 5’UTR to the T7g10 5’UTR and adding a GFP DB region to the gene resulted in accumulation of foreign protein at over 10%TSP (Ye et al 2001). This four order of magnitude improvement in protein accumulation was accomplished in two steps, first by a two order of magnitude improvement after changing the 5’UTR and then another two order of magnitude improvement by adding the GFP DB region. We propose on the basis of the numerous examples of tremendous improvements in foreign protein accumulation resulting from 5’UTR and DB optimization that translation initiation is limiting for high-level foreign protein production in plastids. We propose that the 5’UTR and DB region mediate efficient loading of ribosomes onto the mRNA and efficient translation of the 5’ region of the ORF of interest, respectively. By clearing the translation initiation site for the loading of a new ribosome, more protein can be produced from each transcript, as illustrated schematically in Figure 1.3. The effects of efficient 5’UTR and DB-mediated translation initiation are magnified by the stabilization of efficiently translated RNA against degradation relative to inefficiently translated RNA (Chapter 3). This establishes a positive feedback in which

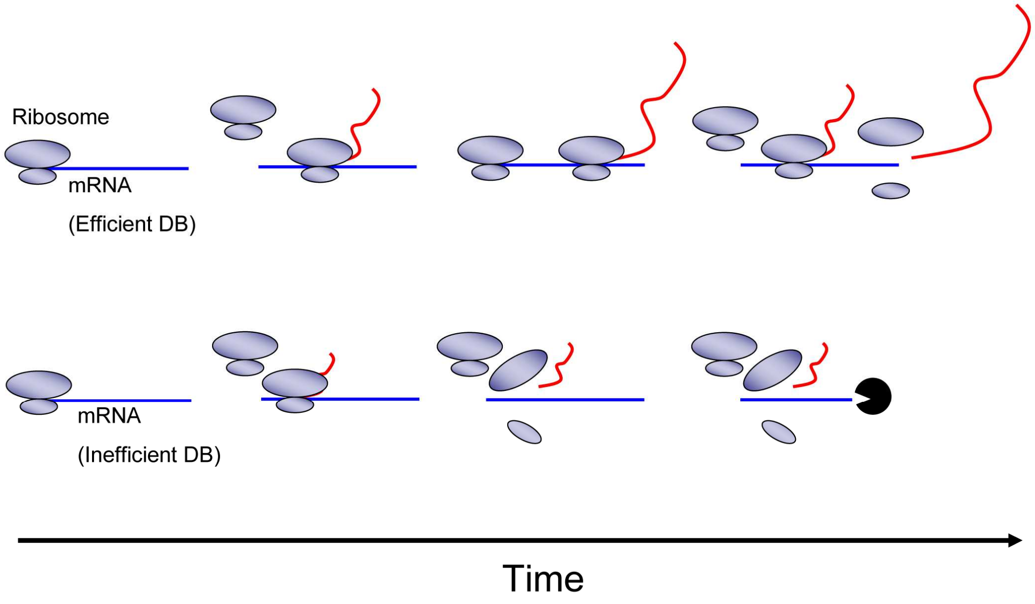


Figure 1.3. Schematic illustration of an mRNA containing an efficient DB region (top) as compared with translation of an mRNA containing an inefficient DB region (bottom). Translation of the mRNA containing an efficient DB region allows for rapid translation of the 5' portion of the ORF (second panel), clearing the start codon for the loading of another ribosome (third panel). This mRNA is efficiently translated, resulting in the production of full-length proteins (fourth panel). Translation of the mRNA containing an inefficient DB region results in slow translation of the 5' portion of the ORF (second panel), ultimately resulting in drop-off of the ribosomes and nascent polypeptide (third panel). This transcript is degraded by 3'-5' exonucleases in the absence of efficient translation (fourth panel).

efficiently translated transcripts become more abundant than less efficiently translated transcripts.

The T7g10 5'UTR is the most efficient 5'UTR tested to date for foreign protein production in plastids. Many of the highest reported yields of foreign protein in plastids used the T7g10 5'UTR, including the expression of one phage lytic protein at ~70%TSP (Oey et al 2009b) and another phage lytic protein at ~30%TSP (Oey et al 2009a), of NPTII at 23%TSP (Kuroda and Maliga 2001b), and of the tetanus toxin fragment C at 25%TSP (Tregoning et al 2003). Although some reports of extremely high levels of foreign protein accumulation exist using native plastid 5'UTRs (e.g., Kuroda and Maliga 2001a; Khan and Maliga 1999), direct comparisons of native plastid 5'UTRs with the T7g10 5'UTR have shown that the T7g10 5'UTR results in higher levels of foreign protein accumulation, even when high-level accumulation has been achieved using plastid 5'UTRs (Ye et al 2001; Tregoning et al 2003).

Methods for choosing effective DB regions for the ORF of interest are not well understood. In general, it appears that the DB regions of highly expressed genes, whether the genes are native plastid genes (Kuroda and Maliga 2001a) or foreign genes that have been expressed at high levels in plastids (Gray et al 2009a; Chapter 3), can mediate high-level expression of genes when fused to the ORF of interest. Unlike the situation for 5'UTR choice, where a single 5'UTR has been shown to be effective for a number of different genes (i.e., the T7g10 5'UTR), DB function is context-specific. Fusion of the TetC and NPTII DB regions to the *cel6A* (Gray et al 2009a) and *bglC* (Chapter 3) ORFs resulted in opposite effects, with the TetC DB region mediating high-level Cel6A accumulation, but only modest BglC accumulation, and the NPTII DB region mediating high-level BglC accumulation, but only modest Cel6A accumulation. Thus, the appropriate DB region for the ORF of interest must be determined empirically through a trial and error process in the absence of a better mechanistic understanding of DB function. While potentially tedious, finding the appropriate DB region for the gene of interest can make a tremendous difference in foreign protein accumulation. Several DB regions may need to be tested with the gene of interest until an effective DB fusion is found.

Changes to the promoter used to direct transcription of a foreign gene would seem intuitively to have the potential to greatly affect protein accumulation, as low-level mRNA accumulation could create a scenario in which transcript levels are limiting for protein production. Indeed, when extremely weak promoters are used (e.g., the *E. coli trc* promoter used by Newell et al 2003) this seems to be the case, and the use of a stronger promoter can greatly improve foreign protein accumulation. The inclusion of a strong promoter in the plastid transformation vector, however, is not necessary if an appropriate insertion site is used that can take advantage of read-through transcription from a native plastid promoter. This strategy has resulted in high-level accumulation of a number of proteins. In at least one case, a promoterless construct utilizing read-through transcription from the native *rbcL* promoter was ~4-fold more effective for foreign protein production than a similar construct that included the plastid ribosomal promoter for transcription of the foreign gene (Staub and Maliga 1995). Another study found that silent codon changes in the DB region of a gene transcribed by read-through transcription were approximately 15 times more effective at increasing protein accumulation than the inclusion of a ribosomal promoter in the transformation vector, despite significantly higher levels of mRNA when the promoter was included in the vector (Chapter 4). Plastid transformation constructs utilizing read-through transcription also lack the potential for unintended recombination events between the introduced and native copies of a plastid promoter (Gray et al 2009b). A caveat should be added that the highest reported foreign protein yields have come from vectors that included the ribosomal promoter (Oey et al 2009a, 2009b; De Cosa et al 2001), though it is conceivable that high-level expression of these proteins could also have been achieved using promoterless constructs. Despite this caveat, promoterless constructs utilizing read-through transcription are recommended in cases where constitutive expression of the foreign protein is desirable in order to avoid unintended recombination events with native plastid promoters (Gray et al 2009b). In cases where a toxic protein is to be expressed from plastids, an inducible promoter like the induction systems described by Mühlbauer and Koop (2005), McBride et al (1994), or Buhot et al (2006) may be the most practical method of protein production.

The 3'UTR used to regulate a foreign gene has little effect on protein accumulation, as long as an effective 3'UTR is used that can mediate the correct RNA processing events and stabilize the RNA against degradation. One study found ~1.2 to 1.3-fold changes in protein when the plastid *rbcL*, *rpl32*, or *psbA* 3'UTRs were used (Eibl et al 1999). In contrast, changes in the 5'UTR region in this study resulted in over 100-fold differences in protein accumulation. Thus, while it is necessary to include an effective 3'UTR containing a hairpin-loop secondary structure to regulate the foreign gene of interest, the sequence of the 3'UTR is not of particular importance. At least one example of a bacterial 3'UTR (the *E. coli rrnB* 3'UTR; Newell et al 2003) and one example of a mitochondrial 3'UTR (the *atp9* 3'UTR; Bohne et al 2007) functioning in plastids has been reported, and it is likely that other prokaryotic (e.g., bacterial or mitochondrial) 3'UTRs will also be effective in plastids. In addition, the *Chlamydomonas rbcL* 3'UTR has been used effectively for gene regulation in higher plant plastids (e.g., Herz et al 2005) and a hairpin-loop structure in the *loxP* sequence has been observed to function as a 3'UTR in plastids (Chapter 3). As with promoters, avoidance of sequences derived from the plastid genome of the target organism (e.g., the use of tobacco plastid 3'UTRs for tobacco plastid transformation) is recommended in order to avoid unwanted recombination events between an introduced and a native copy of a 3'UTR in the plastid (Gray et al 2009b).

Antibiotic resistance genes or other marker genes (e.g., herbicide resistance genes) must be used in order to select for transformation events. Following the generation of homoplasmic plants, however, these marker genes are no longer desirable. It is recommended that *loxP*, *attP/attB*, or other sequences suitable for the removal of the marker gene be included in plastid transformation vectors. This could allow for increased accumulation of foreign protein under certain growth conditions, though this has yet to be demonstrated in principle. Removal of antibiotic resistance genes should also help with public acceptance of transplastomic plants, as removal of the marker gene will alleviate fears of horizontal transfer of antibiotic resistance genes to, e.g., pests and soil bacteria.

Despite the highlighted references to high-level foreign protein accumulation in plastids, it should be noted that high-level foreign protein accumulation is far from assured from plastid transformants, even when efficient regulatory regions are used. Examples can be seen in the work of Gray et al, in which the T7g10 5'UTR was used to regulate *cel6A* (2009a) or *bglC* (Chapter 3) ORFs fused to three different DB regions. The TetC, NPTII, and GFP DB regions used in these experiments have all been used successfully to mediate high-level foreign protein expression, but only TetC-Cel6A protein and NPTII-BglC protein accumulated to high levels (~10%TSP) in transplastomic plants. Another example is in the work of Kuroda and Maliga (2001b), who expressed NPTII at levels between 0.2%TSP and 23%TSP, again dependent on the identity of the DB region fused to the *neo* ORF. In these cases, the mechanisms of high-level vs. low-level foreign protein accumulation are not entirely clear. One consideration that is outside the scope of this article, but that can have major effects on protein accumulation, is post-translational protein stability. Three of the highest reported plastid foreign protein yields have cited extreme protein stability as a major factor in the extremely high protein yields achieved (De Cosa et al 2001; Oey et al 2009a, 2009b). Susceptibility to protein degradation has also been cited as a cause for the low accumulation of VP6 protein (~0.6%TSP) in aging leaves of a transplastomic tobacco line that accumulated moderate levels of VP6 protein (3%TSP) in young leaves (Birch-Machin et al 2004). The guidelines presented in this article can be useful for the design of plastid transformation vectors for high-level foreign protein accumulation, but further research on the molecular mechanisms of high-level foreign protein accumulation is needed in order to reliably and predictably produce transplastomic plants with high-level foreign protein accumulation.

REFERENCES

Allison LA, Simon LD, Maliga P (1996) “Deletion of *rpoB* reveals a second distinct transcription system in plastids of higher plants” *EMBO J* **15**: 2802-2809.

Barkan A (1988) “Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs” *EMBO J* **7**: 2637-2644.

Barkan A, Walker M, Nolasco M, Johnson D (1994) “A nuclear mutation in maize blocks the processing and translation of several chloroplast mRNAs and provides evidence for the differential translation of alternative mRNA forms” *EMBO J* **13**: 3170-3181.

Beyer P, Al-Babili S, Ye X, Lucca P, Schaub P, Welsch R, Potrykus I (2002) “Golden Rice: introducing the beta-carotene biosynthesis pathway into rice endosperm by genetic engineering to defeat vitamin A deficiency” *J Nutr* **132**: 506S-510S.

Birch-Machin I, Newell CA, Hibberd JM, Gray JC (2004) “Accumulation of rotavirus VP6 protein in chloroplasts of transplastomic tobacco is limited by protein stability” *Plant Biotechnol J* **2**: 261-270.

Bohne AV, Ruf S, Börner T, Bock R (2007) “Faithful transcription initiation from a mitochondrial promoter in transgenic plastids” *Nucleic Acids Res* **35**: 7256-7266.

Brixey PJ, Guda C, Daniell H (1997) “The chloroplast *psbA* promoter is more efficient in *Escherichia coli* than the T7 promoter for hyperexpression of a foreign protein” *Biotechnol Lett* **19**: 395-400.

Buhot L, Horvàth E, Medgyesy P, Lerbs-Mache S (2006) “Hybrid transcription system for controlled plastid transgene expression” *Plant J* **46**: 700-707.

Carrer H, Hockenberry TN, Svab Z, Maliga P (1993) “Kanamycin resistance as a selectable marker for plastid transformation in tobacco” *Mol Gen Genet* **241**: 49-56.

Chakrabarti SK, Lutz KA, Lertwiriyawong B, Svab Z, Maliga P (2006) “Expression of the cry9Aa2 B.t. gene in tobacco chloroplasts confers resistance to potato tuber moth” *Transgenic Res* **15**: 481-488.

Cluster PD, O’Dell M, Metzlaff M, Flavell RB (1996) “Details of T-DNA structural organization from a transgenic Petunia population exhibiting co-suppression” *Plant Mol Biol* **32**: 1197-1203.

Corneille S, Lutz K, Svab Z, Maliga P (2001) “Efficient elimination of selectable marker genes from the plastid genome by the CRE-lox site-specific recombination system” *Plant J* **27**: 171-178.

Daniell H, Ruiz G, Denes B, Sandberg L, Langridge W (2009) “Optimization of codon composition and regulatory elements for expression of human insulin like growth factor-1 in transgenic chloroplasts and evaluation of structural identity and function” *BMC Biotechnol* **9**: 33.

De Cosa B, Moar W, Lee SB, Miller M, Daniell H (2001) “Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals” *Nat Biotechnol* **19**: 71-74.

De Wilde C, Van Houdt H, De Buck S, Angenon G, De Jaeger G, Depicker A (2000) “Plants as bioreactors for protein production: avoiding the problem of transgene silencing” *Plant Mol Biol* **43**: 347-359.

Dufourmantel N, Pelissier B, Garçon F, Peltier G, Ferullo JM, Tissot G (2004) “Generation of fertile transplastomic soybean” *Plant Mol Biol* **55**: 479-489.

Dufourmantel N, Tissot G, Garçon F, Pelissier B, Dubald M (2006) “Stability of soybean recombinant plastome over six generations” *Transgenic Res* **15**: 305-311.

Eibl C, Zou Z, Beck A, Kim M, Mullet J, Koop HU (1999) “In vivo analysis of plastid *psbA*, *rbcL* and *rpl32* UTR elements by chloroplast transformation: tobacco plastid gene expression is controlled by modulation of transcript levels and translation efficiency” *Plant J* **19**: 333-345.

Fernández-San Millán A, Mingo-Castel A, 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.

Golds T, Maliga P, Koop H-U (1993) “Stable plastid transformation in PEG-treated protoplasts of *Nicotiana tabacum*” *Bio/Technology* **11**: 95-97.

Gonzalez de Valdivia EI, Isaksson LA (2004) “A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*” *Nucleic Acids Res* **32**: 5198-5205.

Gray BN, Ahner BA, Hanson MR (2009a) “High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions” *Biotechnol Bioeng* **102**: 1045-1054.

Gray BN, Ahner BA, Hanson MR (2009b) “Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants” *Transgenic Res* doi:10.1007/s11248-009-9246-3.

Hajdukiewicz PT, Allison LA, Maliga P (1997) “The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids” *EMBO J* **16**: 4041-4048.

Herz S, Füssl M, Steiger S, Koop HU (2005) "Development of novel types of plastid transformation vectors and evaluation of factors controlling expression" *Transgenic Res* **14**: 969-982.

Hess WR, Börner T (1999) "Organellar RNA polymerases of higher plants" *Int Rev Cytol* **190**: 1-59.

Hirose T, Sugiura M (2004) "Functional Shine-Dalgarno-like sequences for translational intiation of chloroplast mRNAs" *Plant Cell Physiol* **45**: 114-117.

Hobbs SL, Kpodar P, DeLong CM (1990) "The effect of T-DNA copy number, position and methylation on reporter gene expression in tobacco transformants" *Plant Mol Biol* **15**: 851-864.

Hobbs SL, Warkentin TD, DeLong CM (1993) "Transgene copy number can be positively or negatively associated with transgene expression" *Plant Mol Biol* **21**: 17-26.

Holmes WM, Platt T, Rosenberg M (1983) "Termination of transcription in *E. coli*" *Cell* **32**: 1029-1032.

Hood EE, Woodard SL (2002) "Industrial proteins produced from transgenic plants" in *Plants as Factories for Protein Production* (Hood EE and Howard JA, eds.), Kluwer Academic Publishers.

Hood EE, Love R, Lane J, Bray J, Clough R, Pappu K, Drees C, Hood KR, Yoon S, Ahmad A, Howard JA (2007) "Subcellular targeting is a key condition for high-level accumulation of cellulase protein in transgenic maize seed" *Plant Biotechnol J* **5**: 709-719.

Huang FC, Klaus SM, Herz S, Zou Z, Koop HU, Golds TJ (2002) "Efficient plastid transformation in tobacco using the *aphA-6* gene and kanamycin selection" *Mol Genet Genomics* **268**: 19-27.

Ikemura T (1981) “Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system” *J Mol Biol* **151**: 389-409.

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

Kittiwongwattana C, Lutz K, Clark M, Maliga P (2007) “Plastid marker gene excision by the phiC31 phage site-specific recombinase” *Plant Mol Biol* **64**: 137-143.

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) “Coding-sequence determinants of gene expression in *Escherichia coli*” *Science* **324**: 255-258.

Kumar S, Dhingra A, Daniell H (2004) “Stable transformation of the cotton plastid genome and maternal inheritance of transgenes” *Plant Mol Biol* **56**: 203-216.

Kuroda H, Maliga P (2001a) “Sequences downstream of the translation initiation codon are important determinants of translation efficiency in chloroplasts” *Plant Physiol* **125**: 430-436.

Kuroda H, Maliga P (2001b) “Complementarity of the 16S rRNA penultimate stem with sequences downstream of the AUG destabilizes the plastid mRNAs” *Nucleic Acids Res* **29**: 970-975.

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

La Teana A, Brandi A, O’Connor M, Freddi S, Pon CL (2000) “Translation during cold adaptation does not involve mRNA-rRNA base pairing through the downstream box” *RNA* **6**: 1393-1402.

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

Lelivelt CL, McCabe MS, Newell CA, Desnoo CB, van Dun KM, Birch-Machin I, Gray JC, Mills KH, Nugent JM (2005) “Stable plastid transformation in lettuce (*Lactuca sativa* L.)” *Plant Mol Biol* **58**: 763-774.

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

Lössl A, Bohmert K, Harloff H, Eibl C, Mühlbauer S, Koop HU (2005) “Inducible trans-activation of plastid transgenes: expression of the *R. eutropha phb* operon in transplastomic tobacco” *Plant Cell Physiol* **46**: 1462-1471.

Makoff AJ, Oxer MD, Romanos MA, Fairweather NF, Ballantine S (1989) “Expression of tetanus toxin fragment C in *E. coli*: high level expression by removing rare codons” *Nucleic Acids Res* **17**: 10191-10202.

Mallory AC, Parks G, Endres MW, Baulcombe D, Bowman LH, Pruss GJ, Vance VB (2002) “The amplicon-plus system for high-level expression of transgenes in plants” *Nat Biotechnol* **20**: 622-625.

Magee AM, MacLean D, Gray JC, Kavanagh TA (2007) “Disruption of essential plastid gene expression caused by T7 RNA polymerase-mediated transcription of plastid transgenes during early seedling development” *Transgenic Res* **16**: 415-428.

Maliga P (2003) “Progress towards commercialization of plastid transformation technology” *Trends Biotechnol* **21**: 20-28.

McBride KE, Schaaf DJ, Daley M, Stalker DM (1994) "Controlled expression of plastid transgenes in plants based on a nuclear DNA-encoded and plastid-targeted T7 RNA polymerase" *Proc Natl Acad Sci USA* **91**: 7301-7305.

Moll I, Huber M, Grill S, Sairafi P, Mueller F, Brimacombe R, Londei P, Bläsi U (2001) "Evidence against an interaction between the mRNA downstream box and 16S rRNA in translation initiation" *J Bacteriol* **183**: 3499-3505.

Monde RA, Schuster G, Stern DB (2000a) "Processing and degradation of chloroplast mRNA" *Biochimie* **82**: 573-582.

Monde RA, Greene JC, Stern DB (2000b) "The sequence and secondary structure of the 3'-UTR affect 3'-end maturation, RNA accumulation, and translation in tobacco chloroplasts" *Plant Mol Biol* **44**: 529-542.

Mooney BP (2009) "The second green revolution? Production of plant-based biodegradable plastics" *Biochem J* **418**: 219-232.

Mühlbauer SK, Koop HU (2005) "External control of transgene expression in tobacco plastids using the bacterial lac repressor" *Plant J* **43**: 941-946.

Nakamura T, Furuhashi Y, Hasegawa K, Hashimoto H, Kazufumi W, Obokata J, Sugita M, Sugiura M (2003) "Array-based analysis on tobacco plastid transcripts: preparation of a genomic microarray containing all genes and all intergenic regions" *Plant Cell Physiol* **44**: 861-867.

Newell CA, Birch-Machin I, Hibberd JM, Gray JC (2003) "Expression of green fluorescent protein from bacterial and plastid promoters in tobacco chloroplasts" *Transgenic Res* **12**: 631-634.

O'Connor M, Asai T, Squires CL, Dahlberg AE (1999) "Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA" *Proc Natl Acad Sci USA* **96**: 8973-8978.

Oey M, Lohse M, Scharff LB, Kreikemeyer B, Bock R (2009a) “Plastid production of protein antibiotics against pneumonia via a new strategy for high-level expression of antimicrobial proteins” *Proc Natl Acad Sci USA* doi: 10.1073/pnas.0813146106.

Oey M, Lohse M, Kreikemeyer B, Bock R (2009b) “Exhaustion of the chloroplast protein synthesis capacity by massive expression of a highly stable protein antibiotic” *Plant J* **57**: 436-445.

Okumura S, Sawada M, Park YW, Hayashi T, Shimamura M, Takase H, Tomizawa K (2006) “Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts” *Transgenic Res* **15**: 637-646.

Olins PO, Devine CS, Rangwala SH, Kavka KS (1988) “The T7 phage gene 10 leader RNA, a ribosome-binding site that dramatically enhances the expression of foreign genes in Escherichia coli” *Gene* **73**: 227-235.

Peeters K, De Wilde C, De Jaeger G, Angenon G, Depicker A (2001) “Production of antibodies and antibody fragments in plants” *Vaccine* **19**: 2756-2761.

Ruf S, Hermann M, Berger IJ, Carrer H, Bock R (2001) “Stable genetic transformation of tomato plastids and expression of a foreign protein in fruit” *Nat Biotechnol* **19**: 870-875.

Rybicki EP (2009) “Plant-produced vaccines: promise and reality” *Drug Discov Today* **14**: 16-24.

Shiina T, Allison L, Maliga P (1998) “*rbcL* transcript levels in tobacco plastids are independent of light: reduced dark transcription rate is compensated by increased mRNA stability” *Plant Cell* **10**: 1713-1722.

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.

Sprengart ML, Fuchs E, Porter AG (1996) “The downstream box: an efficient and independent translation initiation signal in Escherichia coli” *EMBO J* **15**: 665-674.

Sriraman P, Silhavy D, Maliga P (1998) “The phage-type P*clpP*-53 plastid promoter comprises sequences downstream of the transcription initiation site” *Nucleic Acids Res* **26**: 4874-4879.

Staub JM, Maliga P (1994) “Translation of *psbA* mRNA is regulated by light via the 5’-untranslated region in tobacco plastids” *Plant J* **6**: 547-553.

Staub JM, Maliga P (1995) “Expression of a chimeric *uidA* gene indicates that polycistronic mRNAs are efficiently translated in tobacco plastids” *Plant J* **7**: 845-848.

Stenström CM, Isaksson (2002) “Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3’ side” *Gene* **288**: 1-8.

Stern DB, Gruissem W (1987) “Control of plastid gene expression: 3’ inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription” *Cell* **51**: 1145-1157.

Svab Z, Maliga P (1993) “High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene” *Proc Natl Acad Sci USA* **90**: 913-917.

Tregoning JS, Nixon P, Kuroda H, Svab Z, Clare S, Bowe F, Fairweather N, Ytterberg J, van Wijk KJ, Dougan G, Maliga P (2003) “Expression of tetanus toxin Fragment C in tobacco chloroplasts” *Nucleic Acids Res* **31**: 1174-1179.

Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) “Molecular farming in plants: host systems and expression technology” *Trends in Biotechnol* **21**: 570-578.

Ye GN, Hajdukiewicz PT, Broyles D, Rodriguez D, Xu CW, Nehra N, Staub JM (2001) "Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco" *Plant J* **25**: 261-270.

Yin Z, Plader W, Malepszy S (2004) "Transgene inheritance in plants" *J Appl Genet* **45**: 127-144.

Yu LX, Gray BN, Rutzke CJ, Walker LP, Wilson DB, Hanson MR (2007) "Expression of thermostable microbial cellulases in the chloroplasts of nicotine-free tobacco" *J Biotechnol* **131**: 362-369.

Zhou F, Karcher D, Bock R (2007) "Identification of a plastid intercistronic expression element (IEE) facilitating the expression of stable translatable monocistronic mRNAs from operons" *Plant J* **52**: 961-972.

Ziegelhoffer T, Will J, Austin-Phillips S (1999) "Expression of bacterial cellulase genes in transgenic alfalfa (*Medicago sativa* L.), potato (*Solanum tuberosum* L.) and tobacco (*Nicotiana tabacum* L.)" *Mol Breeding* **5**: 309-318.

Ziegelhoffer T, Raasch JA, Austin-Phillips S (2001) "Dramatic effects of truncation and sub-cellular targeting on the accumulation of recombinant microbial cellulase in tobacco" *Mol Breeding* **8**: 147-158.

Zou Z, Eibl C, Koop HU (2003) "The stem-loop region of the tobacco *psbA* 5'UTR is an important determinant of mRNA stability and translation efficiency" *Mol Genet Genomics* **269**: 340-349.

Chapter 2: High-Level Bacterial Cellulase Accumulation in Chloroplast-Transformed Tobacco Mediated by Downstream Box Fusions[1]

ABSTRACT

The *Thermobifida fusca* cel6A gene encoding an endoglucanase was fused to three different downstream box (DB) regions to generate cel6A genes with 14 amino acid fusions. The DB-Cel6A fusions were inserted into the tobacco (*Nicotiana tabacum* cv. Samsun) chloroplast genome for protein expression. Accumulation of Cel6A protein in transformed tobacco leaves varied over approximately two orders of magnitude, dependent on the identity of the DB region fused to the cel6A open reading frame (ORF). Additionally, the DB region fused to the cel6A ORF affected the accumulation of Cel6A protein in aging leaves, with the most effective DB regions allowing for high level accumulation of Cel6A protein in young, mature, and old leaves, while Cel6A protein accumulation decreased with leaf age when less effective DB regions were fused to the cel6A ORF. In the most highly expressed DB-Cel6A construct, enzymatically active Cel6A protein accumulated at up to 10.7% of total soluble leaf protein (%TSP). The strategy used for high-level endoglucanase expression may be useful for expression of other cellulolytic enzymes in chloroplasts, ultimately leading to cost-effective heterologous enzyme production for cellulosic ethanol using transplastomic plants.

INTRODUCTION

Though cellulosic ethanol is a promising fuel from an environmental standpoint, industrial production and commercialization of cellulosic ethanol has been slow, in large part due to the high cost of cellulases, the enzymes used for enzymatic cellulose

[1] Gray BN, Ahner BA, Hanson MR (2009) *Biotechnol Bioeng* **102**: 1045-1054.

hydrolysis (Lynd et al., 1996). One option for low-cost enzyme production is the use of transgenic plants as a heterologous protein production system (Twyman et al., 2003; Kusnadi et al., 1997; Danna 2001). Plant-based protein production can offer economic advantages over more traditional protein production platforms such as bacterial and fungal cultures, especially when the desired protein accumulates to high levels in transgenic plant tissues (e.g., greater than 10% of total soluble protein [TSP]). In this regard, chloroplast transformation offers an advantage over plant nuclear transformation, as the former technique often results in higher levels of foreign protein accumulation than the latter, improving the economics of production by increasing the protein concentration in harvested plant tissue (Maliga 2003). While nuclear transformants typically produce foreign protein up to 1%TSP in transformed leaf tissue, with some exceptional transformants producing protein at 5-10%TSP, chloroplast transformants often accumulate foreign protein at 5-10%TSP in transformed leaves, with exceptional transformants reaching as high as >40%TSP (Maliga 2003).

A major economic advantage of plant-based protein production over one that is microorganism-based is in the scale-up of protein expression. Whereas scale-up of microbial systems requires the purchase and maintenance of large fermentors and associated equipment, scale-up of plant-based protein production only requires the planting of more seed and harvesting of a larger area. Cellulase-expressing transgenic plants may offer significant capital cost savings over more traditional cellulase production via cellulolytic fungi or bacteria.

Enzymatic cellulose hydrolysis requires the concerted action of multiple cellulases with non-redundant activities. Cellulases are broadly grouped into two categories, the endoglucanases and exoglucanases, and are grouped into families based on amino acid similarity (Carbohydrate Active Enzymes Database). Endoglucanases act by randomly cleaving cellulose fibers to create glucose oligomers. Exoglucanases processively hydrolyze these glucose oligomers to produce mostly cellobiose. The experiments presented here focused on the identification of downstream box (DB) regions to direct high-level accumulation of Cel6A, an endoglucanase from *Thermobifida fusca*, in transformed tobacco chloroplasts. The DB region, defined by the 10-15 codons immediately downstream of the start codon, has been identified previously as an important regulator of translation efficiency in *Eschericia coli* (Sprengart et al., 1996) and in chloroplasts (Kuroda and Maliga 2001a, 2001b), which use prokaryotic-like translation machinery. The mechanism of translation enhancement by the DB region is unknown (O'Connor et al., 1999), but DB fusions have been used to increase foreign protein accumulation in *E. coli* and in tobacco chloroplasts. A DB fusion to the EPSPS gene allowed for more than a 30-fold improvement in foreign protein accumulation in tobacco chloroplasts (Ye et al., 2001). Downstream box fusions do not always result in increased foreign protein accumulation in chloroplasts; silent mutations in the native *rbcL* and *atpB* DB sequences decreased NPTII accumulation in chloroplast-transformed tobacco by approximately 35-fold and 2-fold, respectively (Kuroda and Maliga 2001b). Similarly, a downstream box designed to perfectly base-pair with a region of the ribosomal RNA that was termed the "anti-downstream box" (Sprengart et al., 1996) resulted in NPTII accumulation over 100-fold lower than that resulting from an NPTII gene lacking this downstream box fusion (Kuroda and Maliga 2001a).

The identification of DB regions that can predictably enhance foreign protein accumulation for many different proteins in chloroplasts would be of particular import to the expression of cellulases in transplastomic plants. The appropriate DB region could be fused to the coding regions of the various cellulases necessary for efficient cellulose degradation (i.e., endoglucanses, exoglucanases, and accessory enzymes) and then inserted into the chloroplast genome of the desired host plant. In the experiments presented here, a 100-fold difference in Cel6A protein accumulation was observed between the highest- and lowest-expressing transplastomic tobacco lines containing Cel6A genes fused to the three DB regions tested, demonstrating that fusion of the appropriate DB region to cellulase genes of interest can lead to high-level accumulation of these enzymes in transformed tobacco chloroplasts.

MATERIALS AND METHODS

Cloning and Plasmid Construction

Tobacco plastid DNA containing the *trnI* (tRNA-Ile) and *trnA* (tRNA-Ala) genes (nt 104500-106205 in Genbank entry Z00044) was PCR-amplified using primers ptDNA-fwd and ptDNA-rev (primer sequences are available as supplementary information), adding a *Sma*I site at the 5’ end of this DNA and amplifying a *Hind*III site from the native plastid DNA sequence. This PCR product was *Sma*I-*Hind*III digested and ligated into a pUC19 backbone to generate plasmid pPTDNA. Primers lox-PpsbA-fwd and PpsbA-aadA-rev were used to amplify the *psbA* promoter (PpsbA; nt 1610-1834 in Genbank entry Z00044) from tobacco plastid DNA and to add *Nsi*I and *Pst*I sites and a loxP recombination site to the 5’ end. Primers PpsbA-aadA-fwd and aadA-Trps16-rev were used to amplify the *aadA* gene from plasmid pCT08 (Shikanai et al., 2001), and these two PCR products were combined by overlap extension PCR to generate a P*psbA*-*aadA* fragment. Primers aadA-Trps16-fwd and Trps16-lox-rev were used to amplify the *rps16* terminator (T*rps16*) from tobacco plastid DNA (nt 4938-5096 in Genbank entry Z00044), adding a loxP recombination site and an *Nsi*I site at the 3’ end. Overlap extension PCR was used to add T*rps16* to the P*psbA*-*aadA* fragment generated above. The *aadA* cassette generated by this overlap extension PCR was digested by *Nsi*I and ligated into *Nsi*I-linearized pPTDNA to generate pPTDNA-aadA. An *Nde*I site was removed from the pUC19 backbone in pPTDNA-aadA using primers rmvNdeI1, rmvNdeI2, rmvNdeI3, and rmvNdeI4. PCR product rmvNdeI1- rmvNdeI4 was digested by *Aat*II and *Apa*I and ligated into a pPTDNA-aadA backbone generated by *Aat*II-*Apa*I digestion. The resulting plasmid was pPTDNA-aadA-NdeIdel.

Primers T7-fwd and T7-rev were used to amplify the T7g10 5'UTR (Kuroda and Maliga 2001a) from plasmid pNS6 (Spiridonov and Wilson 2001), adding *Pst*I and *Asc*I sites to the 5' end and an *Nhe*I site to the 3' end. Primers GFPCel6A-fwd and Cel6A-TpsbA-rev were used to amplify the *T. fusca cel6A* gene lacking its signal peptide from pGG86 (Ghangas and Wilson 1988), adding an *Nhe*I site and the first fourteen amino acids from green fluorescent protein (GFP; Ye et al., 2001) immediately downstream of the start codon and a *Not*I site immediately downstream of the *cel6A* stop codon. The *psbA* 3'UTR (TpsbA; nt 443-536 in Genbank entry Z00044) was amplified from tobacco plastid DNA using primers Cel6A-TpsbA-fwd and TpsbA-rev, introducing a *Not*I site at the 5' end of T*psbA* and a *Pst*I site at the 3' end of T*psbA*. The GFPCel6A-fwd/Cel6A-TpsbA-rev and Cel6A-TpsbA-fwd/TpsbA-rev PCR products were combined by overlap extension PCR using primers GFPCel6A-fwd and TpsbA-rev. This overlap extension PCR product was *Nhe*I digested and ligated to the *Nhe*I-digested T7-fwd/T7-rev PCR product. The resulting Cel6A cassette containing the *cel6A* gene flanked by the T7g10 5'UTR and T*psbA* was *Pst*I digested and ligated into *Pst*I-linearized pPTDNA-aadA-NdeIdel, resulting in plasmid pGFPCel6A. Plasmid pGFPCel6A was used as a template for amplification of *cel6A* genes containing 14-amino acid fusions from the NPTII gene (Kuroda and Maliga 2001a) and from the TetC gene (Tregoning et al., 2003) using primers NPTIICel6A-fwd/Cel6A-TpsbA-rev and TetCCel6A-fwd/Cel6A-TpsbA-rev, respectively. The resulting PCR products were *Nhe*I/*Not*I digested and ligated into the

*Nhe*I/*Not*I backbone of pGFPCel6A to generate pNPTIICel6A and pTetCCel6A, respectively. All plasmids were maintained in NEB-5-alpha *E. coli* (New England Biolabs, Ipswich, MA).

Plasmids pGFPCel6A, pNPTIICel6A, and pTetCCel6A were *Nhe*I-*Not*I digested and the resulting *cel6A* fragments were gel purified. The *cel6A* fragments were ligated into the *Nhe*I-*Not*I backbone of pNS6 (Spiridonov and Wilson 2001) to generate plasmids pGFPCel6AEC, pNPTIICel6AEC, and pTetCCel6AEC, respectively. These plasmids were maintained in NEB-5-alpha *E. coli* (New England Biolabs, Ipswich, MA) and were also transformed into BL21(DE3) *E. coli* cells (Invitrogen, Carlsbad, CA) for protein production.

Chloroplast Transformation

Tobacco chloroplasts were transformed by the particle bombardment method (Svab and Maliga 1993). Briefly, plasmid DNA was coated onto 0.6 μm gold beads (Bio-Rad, Hercules, CA). Two-week old tobacco seedlings (*Nicotiana tabacum* cv. Samsun) were bombarded with the DNA-coated beads. Leaves from bombarded seedlings were cultured on RMOP medium containing 500 mg/L spectinomycin (Svab and Maliga 1993). Newly generated shoots were screened via PCR for insertion of the *cel6A* gene at the anticipated site in the chloroplast genome, and positive transformants were transferred to MS medium containing 500 mg/L spectinomycin for rooting. Leaves from rooted plants were subjected to further rounds of tissue culture on RMOP with spectinomycin to obtain homoplasmic transformants. Homoplasmic transformants were transferred to pots and grown in a greenhouse to produce seed.

Southern Blotting

Leaf samples were flash frozen in liquid nitrogen, then finely ground in Eppendorf tubes. 2X CTAB buffer (2% hexadecyltrimethyl ammonium bromide, 1.4 M sodium chloride, 20 mM EDTA, 100 mM Tris pH 8.0, 0.2% β-mercaptoethanol) was added to the ground leaf samples and incubated for one hour at 65°C. DNA was extracted by two sequential phenol extractions followed by isopropanol precipitation. The isopropanol pellet was resuspended in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) and treated with RNase A (Invitrogen, Carlsbad, CA) for one hour at 37°C. DNA was isolated and RNase removed from this solution by phenol extraction. The aqueous phase of this phenol extraction was ethanol precipitated to isolate DNA, which was resuspended in H_2O.

Isolated DNA was completely digested by *Xho*I and then electrophoresed in 1% agarose. DNA was transferred from the agarose gel to a Hybond N+ membrane (Amersham Biosciences, Piscataway, NJ). Primers probe-fwd and ptDNA-rev were used to PCR amplify a portion of the *trnA* gene from wild-type tobacco DNA. This PCR product was used to synthesize a ^{32}P-labeled probe using the Ambion DECAprime II Random Primed DNA Labeling Kit (Ambion, Austin, TX) according to manufacturer's instructions. The ^{32}P-labeled probe was hybridized with the membrane, washed, and visualized using a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA).

SDS-PAGE and Immunoblotting

Tobacco leaf samples were frozen in liquid nitrogen and then finely ground in eppendorf tubes. Protein extraction buffer (20 mM Tris, pH 7.4, 1% Triton X-100, 0.1% SDS, 1mM PMSF, 0.01% β-mercaptoethanol) was added to ground leaf samples and vortexed. Supernatant was recovered following a five-minute centrifugation at

16,000 x *g*. The concentration of the protein contained in the supernatant was determined from a bovine serum albumin calibration curve using the Bio-Rad Protein Assay (Bio-Rad, Hercules, CA).

Protein samples were electrophoresed in 12% polyacrylamide gels, then transferred to nitrocellulose membranes (Pierce, Rockford, IL). Membranes were blocked by incubation with 5% milk in TBST (100 mM Tris, pH 7.6, 685 mM sodium chloride, 0.5% Tween-20), then incubated with anti-Cel6A antibody (kindly donated by David Wilson, Cornell University, Ithaca, NY) diluted 1:100,000 in 5% milk in TBST. Secondary antibody was horseradish peroxidase-conjugated anti-rabbit polyclonal antibody (Sigma, St. Louis, MO) diluted 1:25,000 in 5% milk in TBST. Membranes were incubated with SuperSignal West Dura Extended Duration Substrate (Pierce, Rockford, IL) and visualized on CL-Xposure film (Pierce, Rockford, IL). Purified Cel6A protein for quantitation was kindly donated by David Wilson (Cornell University, Ithaca, NY). Blots were quantified using Scion Image software (Scion Corporation, Frederick, MD).

Cel6A Production in E. coli

BL21(DE3) cells containing the pGFPCel6AEC, pNPTIICel6AEC, or pTetCCel6AEC plasmid (described above) were grown in LB medium containing kanamycin and Cel6A protein expression was induced with 0.1 mM IPTG. Induced cells were harvested by centrifugation and the spent cell culture medium was removed. Cells were resuspended in Tris (100 mM, pH 7.4) supplemented with 1 mM PMSF, then lysed in Tris (100 mM, pH 7.4) plus 1% SDS and 0.1% β-mercaptoethanol.

Cel6A Purification and N-terminal Sequencing

TetC-Cel6A protein was purified from tobacco leaf crude protein extract. Crude protein was extracted as described above from tobacco leaves transformed with pTetCCel6A. The crude leaf protein extract was incubated with CBind 200 cellulose resin (Invitrogen, Carlsbad, CA) and mixed to allow TetC-Cel6A protein to bind the cellulose. After Cel6A was allowed to bind the resin, the supernatant was removed. The cellulose resin was washed once with Tris (20 mM, pH 7.4), then washed twice with (20 mM, pH 7.4) plus 0.8 M NaCl. TetC-Cel6A was eluted in ethylene glycol. Buffer exchange and protein concentration was performed using a MacroSep column (30,000 MWCO; Pall, East Hills, NY), and purified TetC-Cel6A was re-suspended in Tris (20 mM, pH 7.4). Purity of the eluted TetC-Cel6A was assessed by Coomassie staining a 12% polyacrylamide gel.

GFP-Cel6A, NPTII-Cel6A, and TetC-Cel6A proteins were purified from the appropriate BL21(DE3) *E. coli* cell protein extract essentially as described above for purification of chloroplast-produced TetC-Cel6A, except that the resin was loaded into a chromatography column.

For N-terminal sequencing, eluted TetC-Cel6A was electrophoresed in a 12% polyacrylamide gel and transferred to a nitrocellulose membrane as described above. The nitrocellulose membrane was Ponceau stained and the TetC-Cel6A bands were excised from the membrane for sequencing. N-terminal sequencing of tobacco- and *E. coli*-produced TetC-Cel6A was performed at the Penn State University Core Facility (Hershey, PA).

Enzyme Activity Assays

Crude leaf protein extracts from T0 tobacco transformants were used to assess Cel6A enzyme activity against carboxymethyl cellulose (CMC). Two different amounts of total protein were added to 2% (w/v) CMC in Hepes buffer (50 mM, pH 7.0): 5 and 2.5 µg leaf protein extract from a TetC-Cel6A expressing plant, and 10 and 5 µg leaf protein extract from an NPTII-Cel6A expressing plant. Eighty microliter reactions were carried out in eppendorf tubes for sixteen hours at 50ºC while mixing. A blank control containing Hepes buffer with no CMC was included to account for any sugar present in the crude protein extract. Reducing sugar content was measured in 96-well plates using a DNS assay protocol adapted from Ghose (1987). A standard curve for quantification of Cel6A concentration in the crude protein extracts was generated by measuring reducing sugar release by known amounts of purified Cel6A protein added to a wild-type tobacco protein extract and incubated with 2% CMC.

RNA Blotting

T1 seeds were collected from self-pollinated T0 transformants. The seeds were planted in soil and transferred to individual pots in a greenhouse. Ninety-three days after planting, when the tobacco plants each had approximately 30 leaves, leaf samples were taken from young, mid-, and old leaves (i.e., approximate leaf numbers 28, 15, and 2, respectively) and frozen in liquid nitrogen for protein and RNA extraction. RNA was extracted from leaf samples using Trizol (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. RNA concentration was quantified based on spectrophotometric absorption at 260 nm. Three micrograms of total RNA were loaded in a 1% agarose gel for electrophoresis. Following electrophoresis, RNA was transferred to a Hybond N+ membrane (Amersham Biosciences, Piscataway, NJ). RNA was detected by hybridization with ^{32}P-labeled probes. Radiolabelled probes were generated using the DECAprime II Random Primed DNA Labeling Kit (Ambion, Austin, TX) to label PCR products. Primer pairs for these PCR products were Iprobe-fwd/Iprobe-rev (*trnI*) and C6probe-fwd/Cel6A-TpsbA-rev (*cel6A*).

Following hybridization with radiolabelled probes, the membrane was exposed to a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA) for detection. Isotope was removed from the membrane by exposure to a boiling solution of 0.1% SDS between each hybridization.

RESULTS

Chloroplast Transformation

Tobacco (*N. tabacum* cv. Samsun) chloroplast transformants were generated via particle bombardment using three plasmids containing the elements diagrammed schematically in Figure 2.1A. Each plasmid vector contained a gene coding for the mature Cel6A protein (with the native signal peptide removed) with an *Nhe*I site immediately downstream from the start codon, followed by the first 14 codons from TetC, NPTII, or GFP, respectively. The *cel6A* constructs are promoterless, relying on read-through transcription from the upstream Prrn promoter. The *aadA* gene is placed behind the *psbA* promoter to ensure high-level expression of *aadA* for antibiotic resistance. The *aadA* cassette, containing the *psbA* promoter, the *aadA* ORF, and the *psbA* 5'UTR and *rps16* 3'UTR, is flanked by loxP sites for future cre-mediated marker gene removal (Corneille et al., 2001).

Chloroplast transformants derived from the vectors diagrammed in Figure 2.1A were identified via PCR using primers trnIint-fwd and Cel6Aint-rev. Following several rounds of tissue culture regeneration, DNA was isolated from transplastomic plants and digested with *Xho*I. The schematic diagrams in Figures 2.1A and 2.1B show the locations of the relevant *Xho*I sites in transformed and wild-type tobacco chloroplasts, respectively, one internal to the *trnI* gene and the other downstream of the *trnA* gene. Homoplasmic plants were confirmed by Southern blotting, shown in

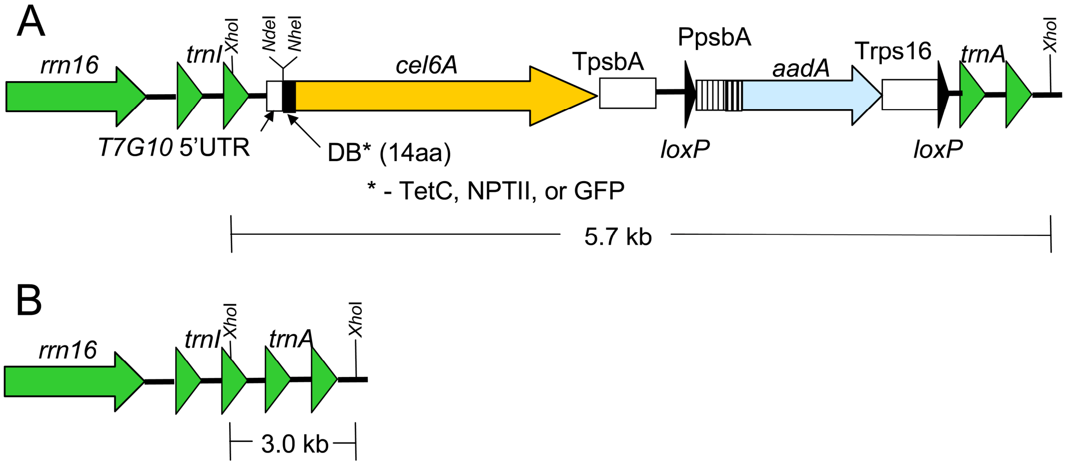


Figure 2.1. Schematic diagrams of tobacco chloroplast DNA. (A) Transformed chloroplast DNA, showing the 16S rDNA, *trnI*, and *trnA* genes along with the *Xho*I restriction sites relevant to Southern blot experiments. The T7g10 5'UTR is immediately upstream of the *cel6A* ORF. *Nde*I and *Nhe*I restriction sites are located at the 5' end of the DB, and the DB region is fused to the *cel6A* ORF. The *cel6A* gene is followed by the *psbA* 3'UTR (T*psbA*) and the *aadA* expression cassette. The *aadA* gene is flanked by the *psbA* promoter and 5'UTR (P*psbA*) and the *rps16* 3'UTR (T*rps16*). The entire *aadA* expression cassette is flanked by loxP sites. (B) Wild-type chloroplast DNA, showing the 16s rDNA, *trnI*, and *trnA* genes along with the *Xho*I restriction sites relevant to Southern blot experiments.

Figure 2.2. Wild-type tobacco showed a band at the expected size of 3.0 kb, while a *trnA* probe hybridized with a 5.7 kb band in transformed plants. Faint 3.0 kb bands in transformed plants are assumed to result from the transfer of chloroplast DNA to the nucleus (Ruf et al., 2000), though a low level of heteroplasmy cannot be completely ruled out. All progeny of the transformed plants grown from seed were resistant to spectinomycin and produced Cel6A protein. In the event that a few of the plastid genomes of the transplastomic plants examined were not transformed, it would be expected that completely homoplasmic plants would produce Cel6A protein at levels even higher than observed in plants containing some untransformed plastid genomes. Bands at approximately 7 kb and 2.5 kb are likely to result from unintended recombination events within transformed chloroplasts and represent a minor fraction of the chloroplast DNA. Transformed plants were transferred to soil and grown in greenhouse conditions to collect seed. Each lane in Figure 2.2 represents a transplastomic plant derived from a unique transformation event.

Protein Accumulation in T0 Transplastomic Transformants

Protein was extracted from the leaves of homoplasmic tobacco transformants for immunoblotting. Cel6A protein accumulation in young leaves of homoplasmic plants transformed with a given construct (i.e., pTetCCel6A, pNPTIICel6A, or pGFPCel6A) was consistent among plants derived from independent transformation events (data not shown). One plant transformed with each construct was therefore selected for further characterization. Figure 2.3 shows that TetC-Cel6A accumulated to significantly higher levels than NPTII-Cel6A, which in turn accumulated to significantly higher levels than GFP-Cel6A. Additionally, Cel6A protein concentration varied with leaf age. TetC-Cel6A protein concentration increased from approximately 3.5%TSP to 7.6%TSP as leaves aged, then decreased in the oldest leaves assayed. NPTII-Cel6A

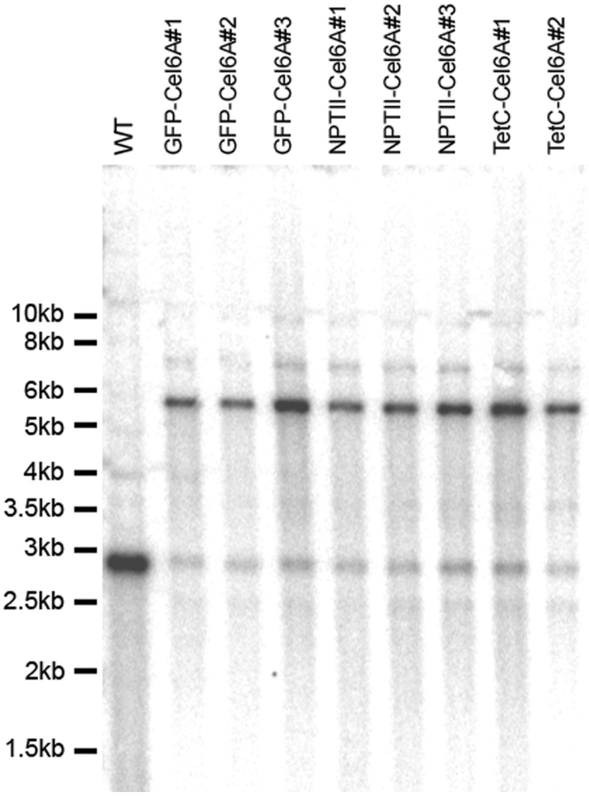


Figure 2.2. Southern blot showing *Xho*I-digested wild-type (WT) and transformed tobacco DNA. A *trnA*-specific probe hybridized with the expected 3.0 kb band in WT tobacco, and with a 5.7 kb band in transplastomic tobacco.

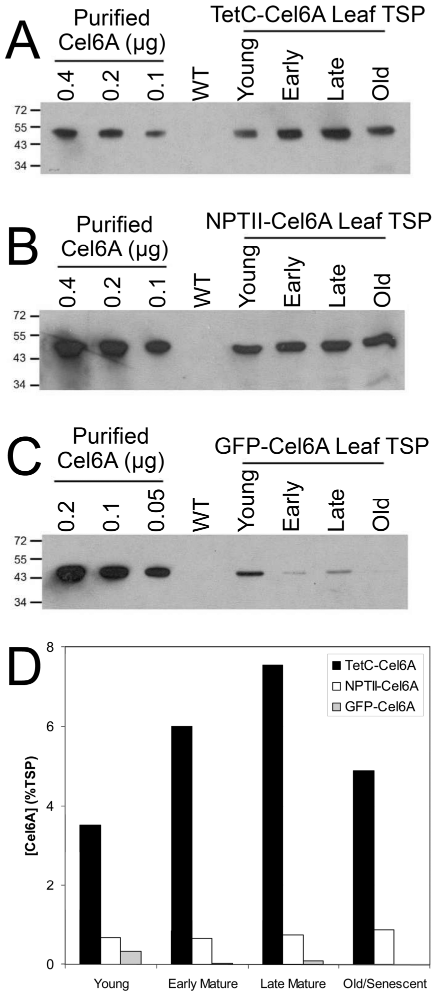


Figure 2.3. Immunoblots showing Cel6A protein accumulation in aging leaves. (A) TetC-Cel6A transformed plants. (B) NPTII-Cel6A transformed plants. (C) GFP-Cel6A transformed plants. (D) Immunoblots were quantified, showing Cel6A accumulation of up to 7.6%TSP in TetC-Cel6A transformed tobacco leaves, up to 0.9%TSP in NPTII-Cel6A transformed tobacco leaves, and up to 0.3%TSP in GFP-Cel6A transformed tobacco leaves.

protein concentration remained steady at approximately 0.7-0.9%TSP through plant development. GFP-Cel6A protein accumulated to approximately 0.3%TSP in young leaves, then dropped off quickly as leaves aged to levels that were below detection limits in the oldest leaves.

Activity assays using carboxymethylcellulose (CMC) as a substrate were also used to quantify Cel6A accumulation in aging leaves of T0 transplastomic plants. Figure 2.4 shows the results of these CMCase activity assays. Quantification of CMCase activity in T0 leaf protein extracts was in good agreement with immunoblot-based quantification of Cel6A protein accumulation, with no statistically significant difference in Cel6A accumulation as calculated by these two methods. This demonstrates that all or nearly all of the Cel6A protein produced in tobacco chloroplasts was active against CMC. The CMCase activity assay using protein extracted from transformed GFP-Cel6A-expressing tobacco was inconclusive, owing to the relatively low expression of GFP-Cel6A in these plants; error associated with quantification of CMCase activity (approximately ±0.5%TSP) is significantly larger than the accumulation of GFP-Cel6A (approximately 0.3%TSP), making the interpretation of tobacco chloroplast-produced GFP-Cel6A CMCase activity difficult. CMCase activity assays using purified Cel6A lacking any DB fusion, TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A indicated that CMCase activity was similar among these enzymes (data not shown). This indicates that the DB fusions to the Cel6A protein used here do not affect enzyme function.

TetC-Cel6A Purification and N-Terminal Sequencing

Tobacco chloroplast- and BL21(DE3) *E. coli*-produced TetC-Cel6A were purified to homogeneity from crude protein extracts by cellulose affinity purification. Figures 2.5A and 2.5B show Coomassie stained polyacrylamide gels with purified

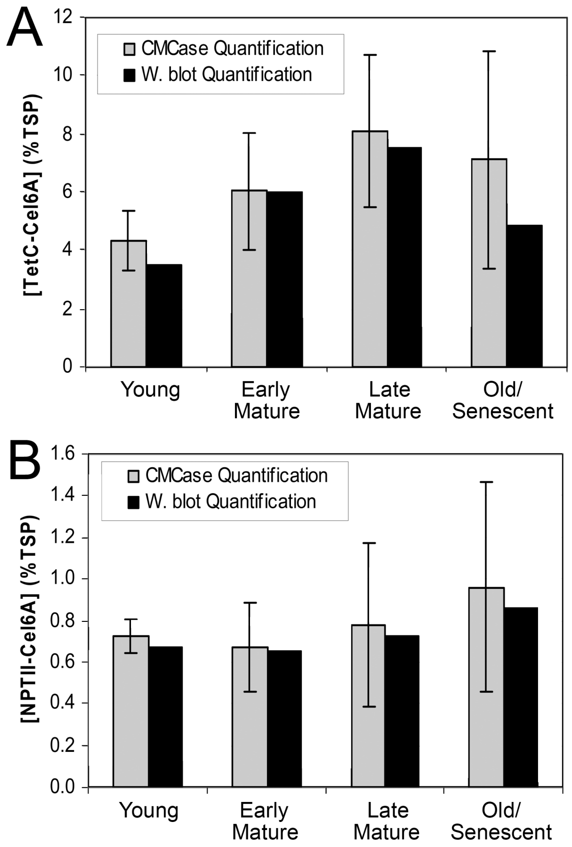


Figure 2.4. Comparison between Cel6A quantification from CMCase activity (gray bars) in tobacco leaves and immunoblot analysis (black bars, from Figure 2.3D). (A) TetC-Cel6A leaf extracts. (B) NPTII-Cel6A leaf extracts. Leaf protein extracts were used to digest 2% CMC and quantified against a standard curve generated by incubating known amounts of Cel6A with 2% CMC.

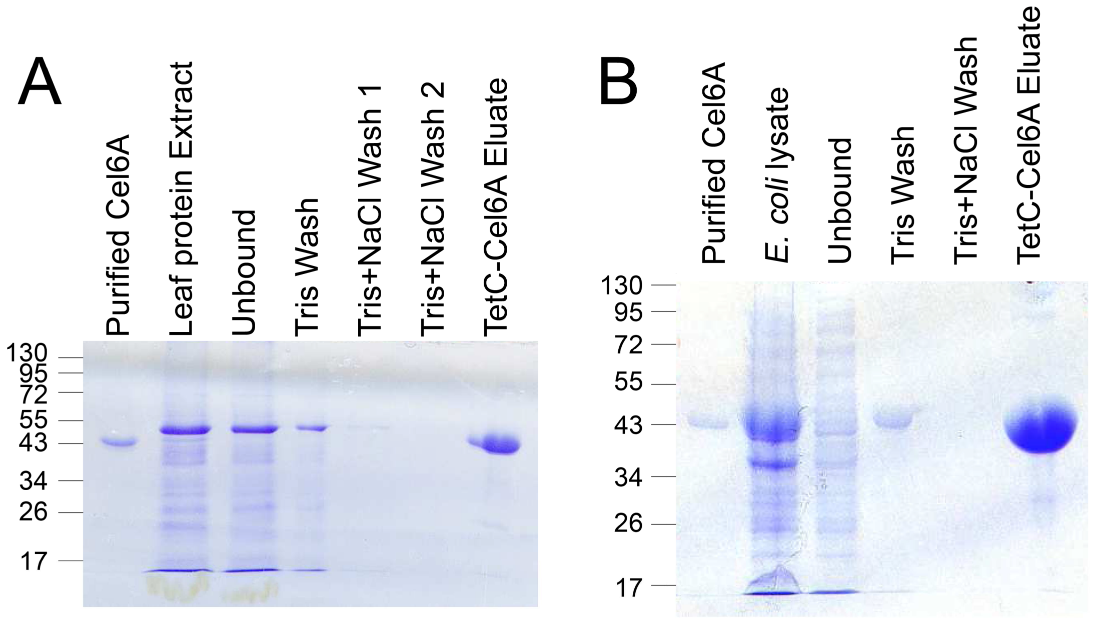


Figure 2.5. Coomassie-stained polyacrylamide gels showing cellulose-affinity purification of TetC-Cel6A. (A) Tobacco-produced TetC-Cel6A. (B) *E. coli*-produced TetC-Cel6A. Crude protein extracts were incubated with cellulose resin, then washed sequentially in Tris (20 mM, pH 7.4) and Tris (20 mM, pH 7.4) with NaCl (0.8 M) buffers. TetC-Cel6A was eluted in ethylene glycol. Ethylene glycol was removed by buffer exchange and eluted TetC-Cel6A was resuspended in Tris (20 mM, pH 7.4).

TetC-Cel6A from tobacco chloroplasts and from *E. coli*, respectively. This figure shows one-step purification of TetC-Cel6A from both protein expression platforms with very little contamination.

The five N-terminal amino acids of chloroplast- and *E. coli*-produced TetC-Cel6A were sequenced by Edman degradation. It was determined that f-Met was cleaved from both chloroplast- and *E. coli*-produced TetC-Cel6A, resulting in an N-terminal alanine residue.

Characterization of Protein Accumulation in T1 Generation Cel6A-Expressing Tobacco

Seed was collected from homoplasmic T0 transformants identified in Figure 2.2 and was planted in MS medium lacking antibiotic. Figure 2.6A shows WT, GFP-Cel6A, NPTII-Cel6A, and TetC-Cel6A seedlings grown from seed in MS medium. Cel6A-expressing plants are phenotypically indistinguishable from WT tobacco. No growth defects were observed in any of the Cel6A-expressing tobacco lines throughout the life cycle from germination to seed production.

Seed was also planted in soil and T1 generation plants were grown in greenhouse conditions to analyze Cel6A accumulation in aging leaves of T1 plants. Figure 2.6B shows the results of an immunoblot with protein extracted from aging leaves of T1 plants. Quantification of this immunoblot in Figure 2.6C shows that Cel6A protein accumulation in T1 plants agrees qualitatively with the protein accumulation in T0 plants, with TetC-Cel6A accumulating to higher levels than NPTII-Cel6A, which in turn accumulates to higher levels than GFP-Cel6A. In T1 plants, TetC-Cel6A accumulated to 7.6-10.7%TSP, NPTII-Cel6A accumulated to 0.8-1.0%TSP, and GFP-Cel6A accumulated to ≤0.1%TSP. Protein accumulation in the T1 TetC-Cel6A plant tested showed less variation with leaf age than in the T0 plant

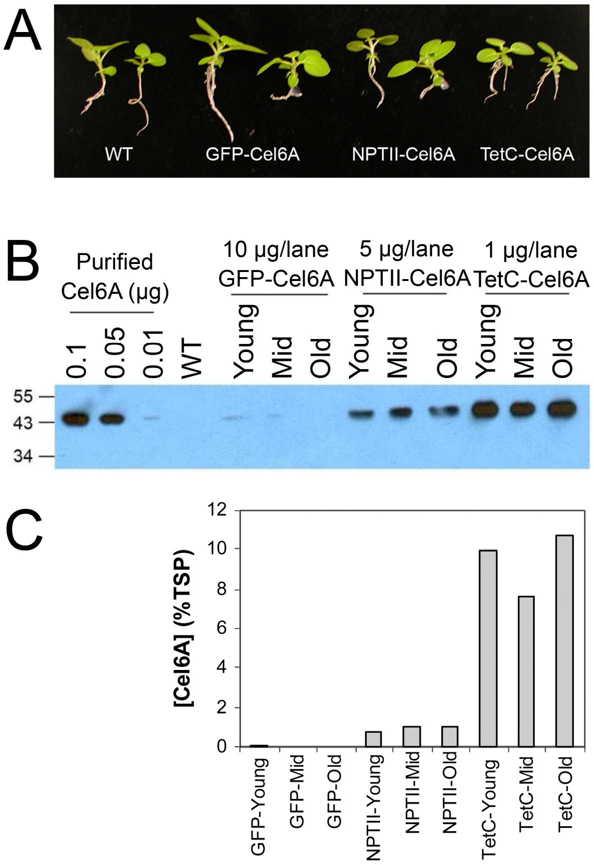


Figure 2.6: T1 generation of Cel6A-expressing tobacco. (A) T1 generation Cel6A-expressing tobacco seedlings planted in MS medium lacking antibiotic were phenotypically indistinguishable from wild-type tobacco. (B) Immunoblot with protein extracts from aging leaves of GFP-Cel6A, NPTII-Cel6A, and TetC-Cel6A transformed tobacco. (C) Quantification of the immunoblot in Figure 2.6B.

(Figures 2.3A, 2.3D, and 2.4A). GFP-Cel6A accumulation decreased with leaf age in the T1 plant tested, in agreement with GFP-Cel6A accumulation in the T0 plant tested (Figures 2.3C and 2.3D).

Characterization of cel6A mRNA in T1 Generation Cel6A-Expressing Tobacco

In order to determine whether the differences in Cel6A protein levels were reflected by RNA-level expression of the transgene, levels of c*el6A* mRNA were examined by probing RNA blots. Figure 2.7 shows an autoradiogram of a blot that used RNA extracted from the same leaves used for the immunoblot in figure 2.6B. Monocistronic *cel6A* transcript (at 1.3 knt) is 5 to 6-fold more abundant in TetC-Cel6A tobacco than in NPTII-Cel6A and GFP-Cel6A tobacco when band intensity was normalized to total RNA loaded. Dicistronic *trnI-cel6A* transcript (at 2.3 knt) was also most abundant in TetC-Cel6A tobacco. Tricistronic *16s rrn-trnI-cel6A* and an incompletely characterized transcript containing both *cel6A* and *trnI* (at 4.0 knt and approximately 3.0 knt, respectively) accumulated to similar levels, less than a two-fold difference when normalized to total RNA loaded (Figure 2.7) in TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A plants. There was a slight decrease in *cel6A* mRNA levels in aging leaves of TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A plants. This decrease correlated with decreasing protein levels in GFP-Cel6A tobacco, but not in NPTII-Cel6A or in TetC-Cel6A tobacco. RNA bands were identified on the basis of predicted transcript sizes and were confirmed on RNA blots with a *trnI*-specific probe (data not shown).

DISCUSSION

Transgenic plant-based expression of cellulolytic enzymes has potential as an economically attractive alternative to more traditional bacteria- and fungus-based

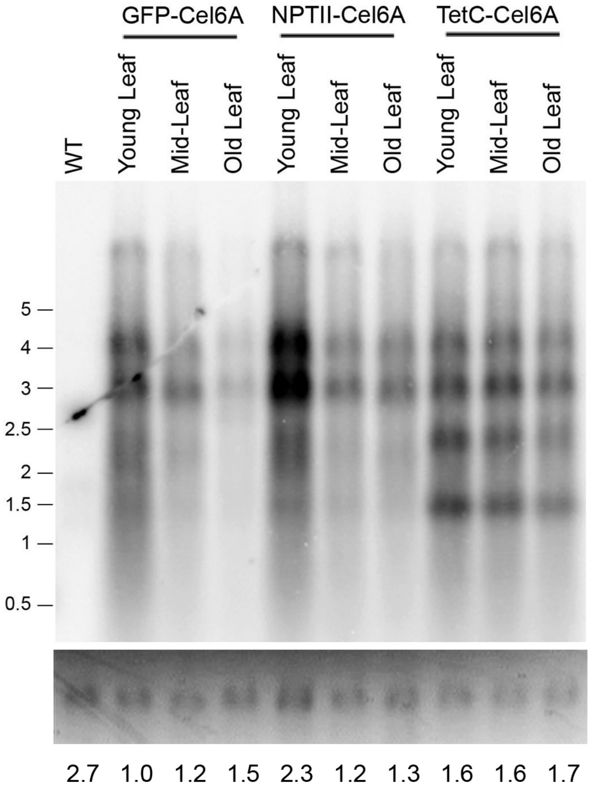


Figure 2.7: RNA blotting of *cel6A* mRNA from T1 generation Cel6A-expressing tobacco. Total RNA was hybridized with a radiolabelled *cel6A* probe, revealing differences in the accumulation of *cel6A* transcripts in GFP-Cel6A, NPTII-Cel6A, and TetC-Cel6A tobacco leaves. Major transcripts are seen at 4.0 knt (*16s rrn-trnI-cel6A*), 3.0 knt (unknown transcript containing both *trnI* and *cel6A*), 2.3 knt (*trnI-cel6A*), and 1.3 knt (*cel6A*). The ribosomal RNA bands from the ethidium-bromide stained agarose gel are shown below the RNA blot. Numbers below each lane indicate the relative RNA loading in that lane, normalized to the GFP-Cel6A young leaf (defined as 1.0).

enzyme expression. The experiments presented here showed high-level expression of Cel6A, an endoglucanase from the thermophilic bacterium *T. fusca*, in tobacco chloroplasts. Enzymatically active Cel6A protein accumulated at up to 10.7%TSP in the highest-expressing tobacco transformant analyzed. This line of transplastomic tobacco could represent a valuable platform for low-cost endoglucanase production. In combination with other essential enzymes for cellulose hydrolysis (i.e., exoglucanases and β-glucosidase), the tobacco chloroplast-produced Cel6A described here could be used for cellulosic ethanol production. Additionally, chloroplast-produced Cel6A could find application in textile processing, detergents, and/or enzymatic newspaper de-inking (Galante and Formantici 2003; Pèlach et al., 2003). Because of the high conservation of the chloroplast genome in vascular plants, we expect that the downstream box-Cel6A construct we have identified will also result in high-level Cel6A expression in the chloroplasts of other plant species.

The mature *cel6A* gene was expressed in tobacco chloroplasts as a fusion protein, with an *Nhe*I site added immediately downstream of the start codon, followed by the first 14 amino acids (i.e., the downstream box regions) from the TetC, NPTII, and GFP genes. The identity of the DB region fused to the *cel6A* ORF greatly affected Cel6A accumulation, with TetC-Cel6A accumulating to ~10%TSP, NPTII-Cel6A accumulating to ~1%TSP, and GFP-Cel6A accumulating to ~0.2%TSP in transformed tobacco leaves. The GFP DB region has been used previously to stimulate expression of the EPSPS gene in tobacco chloroplasts (Ye et al., 2001). Though the full-length TetC and NPTII genes have been expressed in tobacco chloroplasts, accumulating to 25%TSP and 23%TSP, respectively (Tregoning et al., 2003; Kuroda and Maliga 2001a), the DB regions from these genes have not been used to stimulate expression of other foreign proteins. We decided to test the downstream box regions of TetC and NPTII on a heterologous protein because of the efficient expression of the full-length

TetC and NPTII genes. The observed high-level accumulation of TetC-Cel6A and intermediate levels of NPTII-Cel6A suggests that that our strategy of selecting a DB from a highly expressed protein may be generally useful for choosing an appropriate DB region for testing chloroplast expression of foreign proteins.

Nevertheless, a DB region from a highly expressed protein does not always succeed in improving expression of a heterologous protein. Accumulation of GFP-Cel6A in transformed tobacco was relatively low, in contrast with a previous report that the GFP DB region stimulated expression of the EPSPS gene more than 30-fold (Ye et al., 2001). The only difference between the GFP DB region used by Ye et al. (2001) and the GFP DB region fused to *cel6A* is the presence of an *Nhe*I restriction site at the 5’ end of the GFP DB region that was included in the Cel6A expression constructs, but was not present in the EPSPS expression experiments of Ye et al. (2001). It is conceivable that the presence of the *Nhe*I site was detrimental to the function of the GFP DB region for Cel6A expression, though the *Nhe*I site was also included at the 5’ ends of the TetC and NPTII DB regions. In addition, an *Nhe*I site was included at the 5’ end of the NPTII gene for chloroplast expression, resulting in NPTII accumulating to 23%TSP (Kuroda and Maliga 2001a). More likely, it is possible that the GFP DB region was better suited for enhancement of some aspect of the expression of the EPSPS protein (e.g., protein folding, protein stability, and/or translation) that was that was less well-suited for Cel6A expression.

Fusion of an appropriate DB region to a foreign gene of interest is an important strategy for improving protein accumulation, though the selection process at this time is highly empirical. Codon optimization of foreign genes for chloroplasts has only resulted in modest improvements in protein accumulation (e.g., Ye et al., 2001). Fusion of the TetC DB region to the GC-rich (68% G+C in the mature gene) *cel6A* gene allowed for high-level accumulation of Cel6A protein in transformed chloroplasts. A better understanding of the mechanism of DB function will be required for less empirical optimization of the DB region.

Chloroplast transformation vectors typically include a promoter element upstream of the ORF encoding the foreign gene to be expressed driving transcription of the foreign gene. An alternative strategy used in the pTetCCel6A, pNPTIICel6A, and pGFPCel6A vectors is to utilize read-through transcription from a native chloroplast promoter. Foreign gene insertion between the chloroplast *trnI* and *trnA* genes results in a polycistronic message transcribed from the strong chloroplast ribosomal promoter upstream of the *trnI* gene. This strategy resulted in high level accumulation of TetC-Cel6A protein despite the lack of a promoter element directly upstream of the TetC-Cel6A ORF. This is consistent with an earlier report that read-through transcription of genes inserted between the chloroplast *trnI* and *trnA* genes is efficient and can lead to high-level foreign protein production (Chakrabarti et al., 2006). The use of read-through transcription from the native ribosomal promoter obviates the need for a heterologous promoter upstream of the gene(s) to be expressed, thereby lowering the chance of unintended recombination events with the native promoter element.

The good correlation between the cellulase activity in assays of tobacco protein extracts and the Cel6A quantification from denaturing immunoblots showed that all or nearly all of the Cel6A produced in transformed tobacco is able to hydrolyze the soluble substrate CMC. This confirmed that the catalytic domain of Cel6A was correctly folded and active in a crude leaf protein extract and that the downstream box fusions to the Cel6A gene did not disrupt CMCase activity. Chloroplast-produced TetC-Cel6A was purified from the crude leaf protein extract using a cellulose affinity column. Because efficient cellulose binding requires a properly folded and functional cellulose binding domain (CBD), this purification showed that the cellulose binding domain was also functional in chloroplast-produced TetC-Cel6A. High-level expression of full-length cellulase genes including both the catalytic domain and the cellulose binding domain is an important step toward the use of plant-produced cellulases for biomass degradation. Though one strategy to increase plant-based cellulase expression is truncation of the cellulase gene to remove the CBD (Ziegelhoffer et al., 2001), functional CBDs are required for efficient hydrolysis of insoluble cellulose produced in biomass (Din et al., 1994).

Cel6A produced by either tobacco chloroplasts or *E. coli* cells was purified in one step by elution from a cellulose resin, most likely due to the presence of the Cel6A CBD. This conclusion is based on previous observations that removal of the CBD from a cellulase greatly decreases cellulose binding capacity (Gilkes et al., 1988). This demonstrates the potential utility of the Cel6A CBD as a tag for purification of foreign proteins from plants or from *E. coli*. Use of the CBD as a purification tag is particularly attractive because of the relative low cost of cellulose resin and necessary reagents for cellulose affinity purification when compared with other commonly used purification techniques (e.g., Ni-NTA purification of His-tagged proteins or immunopurification of FLAG-tagged proteins). Though CBDs have been used previously for protein purification (Shoseyov et al., 2006), this is the first use, to our knowledge, of the Cel6A CBD for protein purification.

The observed differences in accumulation of chloroplast-produced TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A could be a result of differences in RNA processing, translation efficiency, protein stability, or some combination of these factors. Differences in transcription rates are unlikely, as all three *cel6A* genes are transcribed from the same ribosomal promoter. An unknown feedback mechanism that could affect transcription rates cannot be ruled out based on the experiments presented here. Differential accumulation of monocistronic *cel6A* transcript could be explained either by differences in processing of the polycistronic transcript or by differences in stability of the monocistronic transcript among the three *cel6A* constructs tested. Figure 2.7 shows that polycistron accumulation does not differ greatly among TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A plants, suggesting that the observed differences in monocistronic *cel6A* transcript accumulation are due to differences in RNA stability rather than to differences in RNA processing. Assuming approximately equal rates of transcription among the three constructs tested, greater RNA processing efficiency in TetC-Cel6A tobacco would be expected to cause a depletion of polycistronic *cel6A* transcripts in this line of tobacco, which is not observed. Based on the RNA blot shown in figure 2.7, it appears that polycistronic transcripts in all three lines of tobacco tested are processed with approximately equal efficiency, but that monocistronic TetC-Cel6A transcript is more stable than either NPTII-Cel6A or GFP-Cel6A monocistrons.

The mechanism behind differences in mRNA stability in chloroplasts is not entirely clear, though differential translation rates could play a role. Kuroda and Maliga (2001b) suggest that the nucleotide composition of the DB region may affect translation efficiency in tobacco chloroplasts based on a series of constructs containing silent mutations in the DB region. Research in *E. coli* has revealed that translation and RNA turnover are linked, with untranslated RNA being degraded more quickly than RNA that is being actively used for translation (Rapaport and Mackie 1994). The observed differences in *cel6A* monocistronic transcript accumulation among TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A plants are consistent with the hypotheses that polycistronic transcripts in TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A plants are all processed with equal efficiency, but that monocistronic TetC-Cel6A transcript is translated more efficiently than NPTII-Cel6A or GFP-Cel6A transcripts and that differential translation rates result in differential stability of the monocistronic *cel6A* transcripts.

Further experiments will be necessary to determine the mechanism behind the observed differences in Cel6A protein accumulation, but the chloroplast TetC-Cel6A experiments presented here demonstrate the potential for high-level accumulation of *T. fusca* Cel6A when fused to the appropriate DB region. The highest accumulation of TetC-Cel6A observed here, 10.7%TSP, is five to ten times higher than the Cel6A accumulation reported previously from chloroplasts transformed to express Cel6A behind the *rbcL* DB region (Yu et al., 2007). Further, TetC-Cel6A protein remained at a high concentration in older leaves of transformed tobacco, in contrast with chloroplast expression of rbcL-Cel6A (Yu et al., 2007). A report of nuclear Cel6A expression showed accumulation at only 0.1%TSP (Ziegelhoffer et al., 1999). Chloroplast expression of the TetC-Cel6A protein has therefore improved the accumulation of Cel6A protein over 100-fold and allowed for the accumulation of active enzyme in aging leaves, demonstrating the utility of chloroplast-based cellulase expression.

The TetC-Cel6A expressing tobacco lines reported here represent a promising step toward lowering the enzyme costs associated with cellulosic ethanol production. Chloroplast-produced Cel6A could be used, in conjunction with other cellulose degrading enzymes (e.g., exoglucanases and β-glucosidase) for efficient cellulose hydrolysis. Any or all of these enzymes could potentially be produced in transplastomic plants. A previous report of high-level (6%TSP) xylanase production in tobacco chloroplasts demonstrated the ability of chloroplasts to produce accessory enzymes involved in the degradation of lignocellulose (Leelavathi et al., 2003). Careful selection the appropriate DB regions for chloroplast expression of individual enzymes will enable the cost-effective production of enzymes by plants for cellulose hydrolysis.

ACKNOWLEDGEMENTS

The authors wish to acknowledge David Wilson for the gifts of purified Cel6A and anti-Cel6A antibody, and Julien P. Fey for his assistance in earlier chloroplast transformation experiments. BNG was the recipient of an NSF Graduate Research Fellowship. This work was supported by a USDA Grant (USDA NRI 2007-02133) to BAA and MRH.

APPENDIX

Supplementary Table 2.S1. Primers used in this study

Primer Name	Sequence
ptDNA-fwd	ATCCCGGGGTTTCTCTCGCTTTTGG
ptDNA-rev	TAAAGCTTTGTATCGGCTA
lox-PpsbA-fwd	ATGCATCTGCAGATAACTTCGTATAATGTATGCTATACGAAGTTATCCCGGGCAACCCACTAGC
PpsbA-aadA-rev	AACCGCTTCACGAGCCATGGTAAAATCTTGGTTTAT
PpsbA-aadA-fwd	ATAAACCAAGATTTTACCATGGCTCGTGAAGCGGT
aadA-Trps16-rev	TAATTGAATTTCGGTTGATTATTTGCCAACTACCTT
aadA-Trps16-fwd	AAGGTAGTTGGCAAATAATCAACCGAAATTCAATTA
Trps16-lox-rev	ATGCATAACTTCGTATAGCATACATTATACGAAGTTATACGGAATTCAATGGAAGC
trnIint-fwd	CTGGGGTGACGGAGGGAT
rmvNdeI1	CTGACGTCTAAGAAACCA
rmvNdeI2	TACTGAGAGTGCACCAAATGCGGTGTGAAA
rmvNdeI3	TTTCACACCGCATTTGGTGCACTCTCAGTA
rmvNdeI4	ATGGGCCCGCTATGCCAAAAGC
T7-fwd	CTGCAGGCGCGCCGGGAGACCACAACGGTTTCCCACTAGAAATAA
T7-rev	GCTAGCCATATGTATATC
GFPCel6A-fwd	ATGCTAGCGGCAAGGGCGAGGAACTGTTCACTGGCGTGGTCCCAATCAATGATTCTCCGTTCTAC
Cel6A-TpsbA-rev	ATAGACTAGGCCAGGATCGCGGCCGCTCAGCTGGCGGCGCAGGT
Cel6A-TpsbA-fwd	ACCTGCGCCGCCAGCTGAGCGGCCGCGATCCTGGCCTAGTCTAT
TpsbA-rev	ATGCTAGCTGCAGAAAAAGAAAGGAGCAATA
C6probe-fwd	GTAACGAGTGGTGCGACC
TetCCel6A-fwd	ATGCTAGCAAAAATCTGGATTGTTGGGTCGACAATGAAGAAGATATAAATGATTCTCCGTTCTAC
NPTIICel6A-fwd	ATGGCTAGCATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTAATGATTCTCCGTTCTAC
probe-fwd	ATAGTATCTTGTACCTGA
Cel6Aint-rev	TGCTGTGGTTGCCGCAGT
Iprobe-fwd	CACAGGTTTAGCAATGGG
Iprobe-rev	GAAGTAGTCAGATGCTTC

REFERENCES

Carbohydrate Active Enzymes Database (http://www.cazy.org). Coutinho PM, Henrissat B (1999) “Carbohydrate-active enzymes: an integrated database approach” in Gilbert HJ, Davies G, Henrissat B, Svensson B, eds. *Recent Advances in Carbohydrate Bioengineering*. Cambridge: The Royal Society of Chemistry. p 3-12.

Chakrabarti SK, Lutz KA, Lertwiriyawong B, Svab Z, Maliga P (2006) “Expression of the *cry9Aa2* B.t. gene in tobacco chloroplasts confers resistance to potato tuber moth” *Transgenic Res* **15**: 481-488.

Corneille S, Lutz K, Svab Z, Maliga P (2001) “Efficient elimination of selectable marker genes from the plastid genome by the CRE-lox site-specific recombination system” *Plant J* **27**: 171-178.

Danna KJ (2001) “Production of cellulases in plants for biomass conversion” in Romeo JT, Saunders JA, Matthews BF, eds. *Recent Adv Phytochem*. Oxford: Pergamon. p 205-231.

Din N, Damude HG, Gilkes NR, Miller RC Jr., Warren RAJ, Kilburn DG (1994) “C1-Cx revisited: intramolecular synergism in a cellulase” *P Natl Acad Sci USA* **91**: 11383-11387.

Galante YM, Formantici C (2003) “Enzyme applications in detergency and in manufacturing industries” *Curr Org Chem* **7**: 1399-1422.

Ghangas GS, Wilson DB (1988) “Cloning of the *Thermomonospora fusca* endoglucanase E2 gene in S*treptomyces lividans*: affinity purification and functional domains of the cloned gene product” *Appl Environ Microb* **54**: 2521-2526.

Ghose TK (1987) “Measurement of cellulase activities” *Pure Appl Chem* **59**: 257-268.

Gilkes NR, Warren RA, Miller RC Jr, Kilburn DG (1988) “Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis” *J Biol Chem* **263**: 10401-10407.

Kuroda H, Maliga P (2001a) “Complementarity of the 16S rRNA penultimate stem with sequences downstream of the AUG destabilizes the plastid mRNAs” *Nucleic Acids Res* **29**: 970-975.

Kuroda H, Maliga P (2001b) “Sequences downstream of the translation initiation codon are important determinants of translation efficiency in chloroplasts” *Plant Physiol* **125**: 430-436.

Kusnadi AR, Nikolov ZL, Howard JA (1997) “Production of recombinant proteins in transgenic plants: practical considerations” *Biotechnol Bioeng* **56**: 473-484.

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

Lynd LR, Elander RT, Wyman CE (1996) “Likely features and costs of mature biomass ethanol technology” *Appl Biochem Biotechnol* **57-58**: 741-761.

Maliga P (2003) “Progress towards commercialization of plastid transformation technology” *Trends Biotechnol* **21**: 20-28.

O’Connor M, Asai T, Squires CL, Dahlberg AE (1999) “Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA” *P Natl Acad Sci USA* **96**: 8973-8978.

Pèlach MA, Pastor FJ, Puig J, Vilaseca F, Mutjé P (2003) “Enzymic deinking of old newspapers with cellulase” *Process Biochem* **38**: 1063-1067.

Rapaport LR, Mackie GA (1994) "Influence of translational efficiency on the stability of the mRNA for ribosomal protein S20 in *Escherichia coli*" *J Bacteriol* **176**: 992-998.

Ruf S, Biehler K, Bock R (2000) "A small chloroplast-encoded protein as a novel architectural component of the light-harvesting antenna" *J Cell Biol* **149**: 369-378.

Shikanai T, Shimizu K, Ueda K, Nishimura Y, Kuroiwa T, Hashimoto T (2001) "The chloroplast clpP gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco" *Plant Cell Physiol* **42**: 264-273.

Shoseyov O, Shani Z, Levy I (2006) "Carbohydrate binding modules: biochemical properties and novel applications" *Microbiol Mol Biol R* **70**: 283-295.

Spiridonov NA, Wilson DB (2001) "Cloning and biochemical characterization of BglC, a beta-glucosidase from the cellulolytic actinomycete *Thermobifida fusca*" *Curr Microbiol* **42**: 295-301.

Sprengart ML, Fuchs E, Porter AG (1996) "The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*" *EMBO J* **15**: 665-674.

Svab Z, Maliga P (1993) "High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene" *P Natl Acad Sci USA* **90**: 913-917.

Tregoning JS, Nixon P, Kuroda H, Svab Z, Clare S, Bowe F, Fairweather N, Ytterberg J, van Wijk KJ, Dougan G, Maliga P (2003) "Expression of tetanus toxin Fragment C in tobacco chloroplasts" *Nucleic Acids Res* **31**: 1174-1179.

Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) "Molecular farming in plants: host systems and expression technology" *Trends Biotechnol* **21**: 570-578.

Ye G-N, Hajdukiewicz PTJ, Broyles D, Rodriguez D, Xu CW, Nehra N, Staub JM (2001) “Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco” *Plant J* **25**: 261-270.

Yu L-X, Gray BN, Rutzke CJ, Walker LP, Wilson DB, Hanson MR (2007) “Expression of thermostable microbial cellulases in the chloroplasts of nicotine-free tobacco” *J Biotechnol* **131**: 362-369.

Ziegelhoffer T, Will J, Austin-Phillips S (1999) “Expression of bacterial cellulase gene in transgenic alfalfa (Medicago sativa L.), potato (Solanum tuberosum L.) and tobacco (Nicotiana tabacum L.)” *Mol Breeding* **5**: 309-318.

Ziegelhoffer T, Raasch JA, Austin-Phillips S (2001) “Dramatic effects of truncation and sub-cellular targeting on the accumulation of recombinant microbial cellulase in tobacco” *Mol Breeding* **8**: 147-158.

Chapter 3: Stabilization of mRNA by Efficient Downstream Box Fusions to Allow High-Level Accumulation of Active Bacterial Beta-Glucosidase in Tobacco Chloroplasts

ABSTRACT

Cellulase production *in planta* has been proposed as a lower-cost alternative to microbial production, with plastid transformation as a preferred method due to high foreign protein yields. An important regulator of chloroplast protein production is the downstream box (DB) region, located immediately downstream of the start codon. Protein accumulation can vary over several orders of magnitude by altering the DB region, via an unknown mechanism that appears to affect translation efficiency, though the exact mechanism of DB function is not known. In this study, three DB regions were fused to the *bglC* ORF encoding a β-glucosidase from the thermophilic bacterium *Thermobifida fusca* and inserted into the tobacco (*Nicotiana tabacum*) plastid genome. More than a two order of magnitude of difference in BglC protein accumulation was observed, dependent on the identity of the DB fusion. Chloroplast-produced BglC was correctly folded and active. Fusion of an effective DB region to the *bglC* ORF resulted in stabilization of the monocistronic *bglC* transcript, primarily from 3'-5' exonuclease degradation. The antibiotic resistance gene *aadA* was removed from BglC-expressing tobacco by Cre/*loxP* recombination, surprisingly resulting in plants that grew normally in tissue culture, but that died in soil as a result of unintended Cre-mediated recombinations at two previously described *loxP*-like sites and at two novel *loxP*-like sites. These experiments demonstrate the potential utility of transplastomic plants as a vehicle for heterologous β-glucosidase production for the cellulosic ethanol industry and describe a previously underexplored mechanism of DB function for high-level protein production.

INTRODUCTION

Cellulosic ethanol has been promoted as a promising gasoline substitute with the potential to significantly reduce fossil fuel dependence and the environmental problems associated with fossil fuel usage. The preferred method of cellulosic ethanol production proceeds via enzymatic hydrolysis of a lignocellulosic substrate followed by fermentation of the resulting hydrolysate to produce ethanol. Typical microbial cellulase systems contain a complex mixture of endo- and exo-glucanases, β-glucosidase, accessory enzymes, and non-hydrolytic proteins for efficient cellulose hydrolysis (reviewed in Zhang and Lynd 2004). The most common industrial source of cellulases is the cell culture supernatant of the fungus *Trichoderma reesei*, due to the high specific activity of its enzymes and the high concentration of secreted protein from this microorganism. Although *T. reesei* cellulase preparations have high cellulase activity, they are often deficient in β-glucosidase activity (Juhász et al 2005). Multiple studies have found increased glucose and/or ethanol concentrations after cellulose hydrolysis using *T. reesei* cellulase preparations were supplemented with β-glucosidase, typically produced by *Aspergillus* sp. (e.g., Schell et al 1990; Lamed et al 1991; Spindler et al 1989). A low-cost source of β-glucosidase for supplementation of *T. reesei* cellulases would be of interest for industrial enzymatic cellulose hydrolysis.

Transgenic plants have been proposed as low-cost sources of foreign proteins, with projected order of magnitude cost savings relative to microbial protein production systems (Twyman et al 2003). Because production costs for foreign proteins expressed in transgenic plants are dependent on the concentration of foreign protein in transformed tissue, the highest achievable expression levels are desirable. To this end, chloroplast transformation has an advantage over nuclear transformation. Chloroplast transformation has resulted in reports of extraordinarily high levels of foreign protein, up to 70% of total soluble protein (%TSP; Oey et al 2008), and a number of proteins have been expressed at greater than 10%TSP from the chloroplast genome (reviewed in Maliga 2003), including a cellulase (Gray et al 2009a). In order to realize the potential for high-level foreign protein expression, foreign gene regulatory regions (e.g., 5' and 3' untranslated regions [UTRs], promoters, and terminators) must be chosen carefully. One regulatory region that has been shown to be important for foreign protein production in chloroplasts is the downstream box (DB) region (Ye et al 2001; Kuroda and Maliga 2001a, 2001b; Gray et al 2009a; Lenzi et al 2008), composed of the 10-15 codons immediately downstream of the start codon. Both silent and non-silent changes in this region have been shown to affect foreign protein production, though the mechanism of these effects is not well understood.

Besides the potential for high-level foreign protein accumulation, a second major advantage of chloroplast transformation over nuclear transformation is the potential for expression of multi-gene operons. The prokaryote-like transcription and translation machinery found in plastids is well-suited to expression of operons naturally found in many prokaryotic systems. This approach has resulted in accumulation of the Cry2Aa2 protein at approximately 45%TSP in transplastomic tobacco leaves (DeCosa et al 2001). Subsequent analysis of this transformant determined that the most actively translated transcript was the full-length polycistronic message, suggesting that processing of this transcript was not required for protein production (Quesada-Vargas et al 2005). This is in contrast with a report that polycistronic mRNA processing is required for efficient production of yellow fluorescent protein (YFP; Zhou et al 2007). Barkan (1988) demonstrated that the native plastid *psbB*, *petB,* and *petD* ORFs are translated from both monocistronic and polycistronic mRNA species, showing that mRNA processing is not strictly required for translation of native plastid transcripts. Our previous experiments demonstrated that accumulation of monocistronic *cel6A* mRNA correlated with Cel6A protein accumulation in transplastomic tobacco, though a causative relationship was not established (Gray et al 2009a).

Another important aspect of high-level foreign protein accumulation is the stability of the mRNA encoding the desired protein. In *E. coli*, it has been shown that decreased association of ribosomes with mRNA results in degradation of the mRNA, implying a feedback between translation efficiency and mRNA stability (Nilsson et al 1987; Rapaport and Mackie 1994). The downstream box region has been shown previously to influence both translation efficiency (Kuroda and Maliga 2001b) and RNA levels (Kuroda and Maliga 2001a; Gray et al 2009a) in chloroplasts, with less efficient translation correlating with degradation of the mRNA. RNA stabilization by an effective DB region creates a positive feedback in which the steady-state level of efficiently translated mRNA is higher than the steady-state level of an inefficiently translated transcript. If the RNA degradation machinery outcompetes the translation machinery for a given transcript, mRNA concentration could become limiting for protein production.

Our previous work attempted to optimize chloroplast expression of Cel6A, a *Thermobifida fusca* endoglucanase, by the fusion of three different DB regions to the *cel6A* ORF (Gray et al 2009a). The current report describes the fusion of these three DB regions, originating from the TetC, NPTII, and GFP genes, to the *bglC* ORF encoding a *T. fusca* β-glucosidase (BglC). The effects of DB fusion to the *bglC* ORF are explored at both the RNA and protein levels.

MATERIALS AND METHODS

DB-BglC Plasmid Vector Construction

The *bglC* open reading frame (ORF) was PCR-amplified from plasmid pNS6 (Spiridonov and Wilson 2001). In conjunction with reverse primer BglC-rev, forward primers TetCBglC-fwd, NPTIIBglC-fwd, and GFPBglC-fwd (primer sequences are shown in supplementary Table 3.S1) were used to generate modified *bglC* ORFs containing downstream box (DB) fusions from the TetC, NPTII, and GFP genes, respectively (Gray et al 2009a). These PCR reactions added *Nde*I and *Nhe*I sites at the 5' end of the *bglC* ORF and a *Not*I site at the 3' end. The resulting PCR products were *Nhe*I/*Not*I digested and ligated into the *Nhe*I/*Not*I backbone of plasmid pGFPCel6A (Gray et al 2009a) to generate pTetCBglC, pNPTIIBglC, and pGFPBglC, respectively.

Generation of Transplastomic Plants

Transplastomic tobacco was generated by the biolistic method essentially as described previously by Svab and Maliga (1993). Briefly, two-week old tobacco seedlings (*N. tabacum* cv. Samsun) grown in sterile MS agar medium were bombarded with 0.6 micron gold beads (Bio-Rad, Hercules, CA) coated with the appropriate plasmid DNA (i.e., pTetCBglC, pNPTIIBglC, or pGFPBglC). Two days after bombardment, leaves of bombarded seedlings were cut in half and transferred to RMOP agar medium containing 500 mg/L spectinomycin. Antibiotic resistant shoots containing the desired gene insertions were subjected to two to three additional rounds of regeneration on spectinomycin-containing RMOP medium to generate fully transformed shoots. These plants were transferred to MS medium containing 500 mg/L spectinomycin for rooting, then to soil for greenhouse growth and for seed collection.

DNA Blotting

DNA was extracted from tobacco leaves as described previously (Gray et al 2009a). Extracted DNA was thoroughly digested by *Xho*I, *Hind*III, *Xho*I/*Hind*III or *Xho*I/*Kpn*I and electrophoresed in a 1% (w/v) agarose gel. Following electrophoresis, DNA was transferred to a Hybond N+ membrane (Amersham Biosciences, Piscataway, NJ). A portion of the chloroplast *trnI* and *trnA* genes were amplified from WT tobacco DNA using primers Iprobe-fwd/Iprobe-rev and Aprobe-fwd/Aprobe-rev, respectively. These PCR products were radiolabeled with the DECAprime II Random Primed DNA Labeling Kit (Ambion, Austin, TX) according to the manufacturer's instructions and then hybridized with the DNA-containing membrane. Following hybridization, membranes were washed and exposed to a Phosphorimager screen for visualization (Molecular Dynamics, Sunnyvale, CA).

BglC Production and Purification

BL21(DE3) *E. coli* cells (Invitrogen, Carlsbad, CA) harboring the pNS6 plasmid (Spiridonov and Wilson 2001) were grown in LB medium containing 50 µg/mL kanamycin. BglC production was induced by adding 0.5 mM IPTG to the cell culture. Approximately 6 hours after IPTG induction, cells were harvested by centrifugation. Cells were re-suspended in Hepes (50 mM, pH 7.0) containing 1 mM PMSF, then lysed in Hepes (50 mM, pH 7.0) containing 0.5% (w/v) SDS and 1 mM dithiothreitol. Cell debris was pelleted by centrifugation following lysis and supernatant fluid was transferred to a fresh container. Supernatant protein was concentrated using a MacroSep column (MWCO 30,000; Pall, East Hills, NY) and then separated by ion affinity chromatography on a Q-Sepharose column (Sigma, St. Louis, MO). After loading onto the column, the protein was washed in three column volumes of Hepes (50 mM, pH 7.0) and eluted in a step gradient of sodium chloride (0-1 M NaCl, 0.1 M steps) in Hepes (50 mM, pH 7.0). Purity of the eluted protein was assessed by coomassie staining a 12% polyacrylamide gel. Purified BglC concentration was determined by measuring the spectrophotometric absorption at 280 nm.

SDS-PAGE and Immunoblotting

Protein was extracted from tobacco leaves as described previously (Gray et al 2009a). Protein samples were electrophoresed in 12% (w/v) polyacrylamide gels, then transferred to a nitrocellulose membrane (Pierce, Rockford, IL). The membrane was incubated in 5% (w/v) milk in TBST (100 mM Tris, pH 7.6, 685 mM sodium chloride, 0.5% [w/v] Tween-20), then exposed to primary antibody. Polyclonal anti-BglC antibody (kindly provided by David Wilson, Cornell University, Ithaca, NY) was diluted 1:250 in 5% (w/v) milk in TBST. Secondary antibody was horseradish peroxidase-conjugated anti-rabbit polyclonal antibody (Sigma, St. Louis, MO) diluted 1:25,000 in 5% (w/v) milk in TBST. Following incubation with secondary antibody, the membrane was incubated with SuperSignal West Dura Extended Duration Substrate (Pierce) and visualized on CL-Xposure film (Pierce). Immunoblots were quantified using Scion Image software (Scion Corporation, Frederick, MD).

Polysome Fractionation

Polysomes of NPTII-BglC tobacco were fractionated using a protocol adapted from Barkan (1988). Approximately 1 g young leaf tissue was finely ground in liquid nitrogen, then resuspended in polysome fractionation buffer (200 mM Tris, pH 9.0; 400 mM KCl; 200 mM sucrose; 35 mM $MgCl_2$; 25 mM EGTA; 2% [v/v] Triton X-100; 100 mM β-mercaptoethanol; 100 μg/mL chloramphenicol; 500 μg/mL heparin). Resuspended leaf tissue was centrifuged for 15 minutes at 10,000x*g* at 4ºC, then the supernatant liquid was transferred to a new container. This supernatant was overlaid onto a sucrose gradient consisting of 1.5 mL 1.75 M sucrose overlaid with 1 mL 0.5 M sucrose in cushion buffer (40 mM Tris, pH 8.5; 20 mM KCl; 10 mM $MgCl_2$; 100 mM β-mercaptoethanol; 100 µg/mL chloramphenicol; 500 µg/mL heparin). The sucrose gradient was centrifuged for 3 hours at 246,000x*g*. The supernatant was removed following centrifugation and the crude polysome pellet was resuspended in 200 µL polysome resuspension buffer (40 mM Tris, pH 8.5; 200 mM KCl; 30 mM $MgCl_2$; 5 mM EGTA; 100 µg/mL chloramphenicol; 500 µg/mL heparin). Crude polysomes were size fractionated in a 10-50% (w/v) sucrose step gradient (10% steps) in size fractionation buffer (50 mM Tris, pH 8.0; 20 mM KCl; 10 mM $MgCl_2$; 100 µg/mL chloramphenicol; 500 µg/mL heparin) by a 45 minute centrifugation at 237,000x*g*. Six fractions of approximately equal volume were collected from the sucrose gradient. RNA was precipitated from each fraction by ethanol precipitation, resuspended in 25 µL H_2O, and used for RNA blotting as described below.

RNA Blotting

Total RNA was extracted from tobacco leaves using Trizol reagent (Invitrogen) according to the manufacturer's instructions. RNA concentration was quantified by measuring the spectrophotometric absorption at 260 nm. RNA was electrophoresed in a 1% (w/v) agarose gel, then transferred to a Hybond N+ membrane (Amersham Biosciences). A *bglC*-specific radiolabeled probe was generated as described above from PCR product BglCint-fwd/BglC-rev. Following hybridization with the membrane, the membrane was washed and exposed to a Phosphorimager screen (Molecular Dynamics).

Circular RT-PCR and Determination of mRNA Processing Sites

Total leaf RNA was extracted as described above. Five micrograms of leaf RNA were simultaneously circularized and DNase-treated by the addition of 10 U T4 RNA ligase (Fermentas, Glen Burnie, MD) and 2.5 U DNase (Invitrogen). The RNA circularization/DNase reaction was allowed to proceed for 1 hour at 37ºC. Following RNA circularization, RNA was re-extracted using the Trizol reagent (Invitrogen) according to the manufacturer's instructions, then re-suspended in 25 µL RNase-free water. RNA concentration and quality were determined by measuring spectrophotometric absorbance at 260 and 280 nm. Gene-specific RT-PCR was performed with the Superscript III One-Step RT-PCR kit (Invitrogen), using 0.5 µg circularized RNA as a substrate and primers BglCint-fwd2 and BglCint-rev. Reverse transcription occurred at 50ºC for 30 minutes, followed by 40 PCR cycles consisting of 15 seconds denaturation at 94ºC, 30 seconds annealing at 55ºC, and 90 seconds extension at 68ºC. RT-PCR products were cloned into the pCR2.1 TOPO vector using the TOPO TA cloning kit (Invitrogen) according to the manufacturer's instructions. Plasmids were isolated from kanamycin-resistant clones for sequencing, which was performed at the Life Sciences Core Laboratory Center (Cornell University, Ithaca, NY). Sequences were manually aligned with known DB-BglC vector sequences to determine the 5' and 3' ends of DB-BglC mRNAs.

Amplification of Polyadenylated mRNAs

To amplify polyadenylated *bglC* mRNAs, 2.5 µg total RNA from young leaves of WT, NPTII-BglC, TetC-BglC, and GFP-BglC plants were thoroughly treated with DNase (Invitrogen). Following DNase treatment, polyadenylated *bglC* mRNAs were reverse transcribed using Sensiscript reverse transcriptase (Qiagen, Valencia, CA), with 30 ng DNase-treated RNA and primers BglCint-fwd and oligo(dT)$_{17}$. The reaction was also carried out without the addition of reverse transcriptase as a negative control. Following reverse transcription, polyadenylated *bglC* cDNAs were amplified using the reverse transcription products as a template and either BglCint-fwd/oligo(dT)$_{17}$ or BglCint-fwd2/oligo(dT)$_{17}$ as primer pairs. Biomix Taq polymerase (Bioline, Taunton, MA) was used according to the manufacturer's instructions for PCR amplification. Following PCR amplification, 7.5 μL of each 25 μL PCR reaction was electrophoresed in a 1.5% (w/v) agarose gel for DNA blotting, as described above. A ^{32}P-labelled probe was synthesized from PCR BglCint-fwd/BglC-rev for detection of polyadenylated *bglC* cDNAs by DNA blotting.

BglC Activity Assays

Protein was extracted from tobacco leaves as described previously (Gray et al 2009a). Protein extracts were incubated with 50 mM cellobiose (Sigma) for 10 minutes at 50ºC. Two different amounts of NPTII-BglC tobacco protein were assayed, containing 750 ng and 300 ng total protein, respectively. A standard curve was generated by adding known amounts of purified BglC protein (0-120 ng BglC) to WT tobacco protein and incubating with 50 mM cellobiose. Following the 10 minute incubation with cellobiose, all samples were transferred to a 95ºC heating block for 5 minutes to denature BglC and stop the reaction. Glucose was measured using a glucose assay kit (Sigma) essentially according to the manufacturer's instructions. The glucose assay kit protocol was modified to accommodate a 96-well plate format, and spectrophotometric absorbance at 540 nm was measured using a plate reader.

Tobacco Hydrolysis

Soluble protein was extracted from 12 g (fresh weight) WT tobacco leaf tissue as described above. Following extraction of the soluble protein, the leaf tissue was pre-treated for 24 hours at room temperature in 125 mM NaOH while mixing at 200 rpm. After 24 hours, the pre-treated leaf tissue was thoroughly washed in double-distilled water to remove NaOH. One gram of pre-treated and washed tobacco leaf tissue was added to each of four 15-mL centrifuge tubes. Two of the tubes received 2 mg WT protein and the other two tubes received 2 mg NPTII-BglC protein. Forty microliters of Spezyme CP cellulase (Genencor, Rochester, NY) were added to one of the tubes containing WT protein and to one of the tubes containing NPTII-BglC protein. Sodium acetate buffer (50 mM, pH 5.0) was added to all four tubes to a final volume of 2 mL. All tobacco enzyme mixtures were incubated at 50ºC for 24 hours while mixing. Samples were taken from the tubes immediately after adding enzyme to the tubes (time zero), as well as at 15 minutes, 30 minutes, 45 minutes, 1 hour, 4 hours, 6 hours, and 24 hours after the start of hydrolysis. Samples were stored at -80ºC until analysis. Glucose content of the samples was measured using a glucose assay kit (Sigma) as described above.

Cre-mediated Marker Gene Removal

A vector for plastid-targeted Cre expression was generated by PCR-amplifying the sequence encoding the RecA plastid targeting peptide (Kohler et al. 1997) from *Arabidopsis thaliana* genomic DNA using primers RecAsp-fwd/RecAsp-rev, adding a 5' *Xho*I site and 3' *Nde*I and *Xba*I sites, and the *cre* ORF from plasmid pCAGGS-creT (Kindly donated by Robert Weiss, Cornell University, Ithaca, NY; McDaniel et al 2003) using primers Cre-fwd/Cre-rev, adding a 5' *Nde*I site and a 3' *Xba*I site. The RecA plastid targeting peptide sequence was *Xho*I/*Xba*I digested and ligated into the *Xho*I/*Xba*I backbone of plasmid pENSG-YFP (Jakoby et al 2006) to generate pTARGET. The *cre* ORF was inserted behind the plastid targeting peptide sequence by *Nde*I/*Xba*I digesting PCR Cre-fwd/Cre-rev and ligating into the pTARGET

*Nde*I/*Xba*I backbone to generate pCPCRE. This plasmid was cloned into *Agrobacterium tumefasciens* (strain GV3101::pMP90RK). *Agrobacterium* cells harboring pCPCRE were used to infiltrate leaf pieces excised from T1 generation NPTII-BglC tobacco grown in sterile Magenta boxes (Magenta Corporation, Chicago, IL) on MS medium. Infiltrated leaf pieces were placed on MS104 agar medium containing glufosinate (Chem Service, West Chester, PA; 5 mg/L) and timentin (PlantMedia, Dublin, OH; 500 mg/L). Glufosinate-resistant shoots were transferred to MS agar medium containing glufosinate (5 mg/L) and timentin (500 mg/L) for rooting. DNA was extracted from rooted shoots for DNA blotting and PCR analysis.

Confocal Microscopy

Leaves from WT, NPTII-BglC, and NPTII-BglC/cre#1 plants grown in soil in standard greenhouse conditions were used for confocal microscopy in the Cornell University Microscopy and Imaging Facility (Cornell University, Ithaca, NY) with a Leica DMRE-7 (SDK) microscope containing a TCS-SP2 confocal scanning head (Leica Microsystems Inc., Bannockburn, IL). Cell and chloroplast sizes were quantified using MetaMorph software (Molecular Devices, Sunnyvale, CA).

Sequence Alignments

The *loxP*-like sequences observed in unintended Cre-mediated recombinations were aligned using the T-COFFEE multiple sequence alignment tool (http://www.ebi.ac.uk/Tools/t-coffee/index.html). A sequence logo was created using WebLogo3 (http://weblogo.berkeley.edu/). The results of the T-COFFEE sequence alignment and WebLogo creation were used to determine a consensus sequence for Cre-mediated recombination in plastids.

RESULTS

Identification of Fully Transformed DB-BglC Tobacco Transformants

Transplastomic tobacco lines containing the *bglC* ORF were generated via biolistic bombardment of tobacco seedlings with the plasmid vectors diagrammed in Figure 3.1A. These vectors are similar to those described previously (Gray et al 2009a), except that the *cel6A* ORF used previously was replaced with the *bglC* ORF. The *bglC* ORF was fused to the TetC, NPTII, and GFP DB regions, comprised of the first 14 codons from these genes, and inserted between the chloroplast *trnI* and *trnA* genes. DNA isolated from spectinomycin resistant shoots was initially *Xho*I digested and hybridized with a radiolabeled *trnI*-specific probe (Figure 3.2A). This DNA blot showed the expected 3.0 kb and 0.7 kb WT and transformed *Xho*I fragments, respectively, but also showed a faint 3.0 kb fragment in all of the DB-BglC transformed plants analyzed. Because these plants had been subjected to at least three rounds of tissue culture regeneration, it was suspected that these were fully transformed lines, and that the observed 3.0 kb *Xho*I fragments resulted from a nuclear copy of the *trnI* gene. Tobacco DNA was therefore digested with *Hind*III and hybridized with the same *trnI*-specific probe (Figure 3.2B). This blot showed the expected 7.7 kb and 10.6 kb *Hind*III fragments in WT and DB-BglC tobacco, respectively. Essentially no 7.7 kb *Hind*III fragments were detected in DB-BglC transformants, strongly suggesting that these tobacco lines were fully transformed and that the 3.0 kb *Xho*I fragments detected in DB-BglC tobacco originated primarily from an extraplastidic copy of *trnI*. NPTII-BglC#1 and NPTII-BglC#2 plants are derived from independent transformation events. Analysis at the DNA, protein, and phenotypic levels revealed no differences between these two plant lines (data not shown), and so NPTII-BglC#1 was used for all subsequent analyses described in this manuscript.

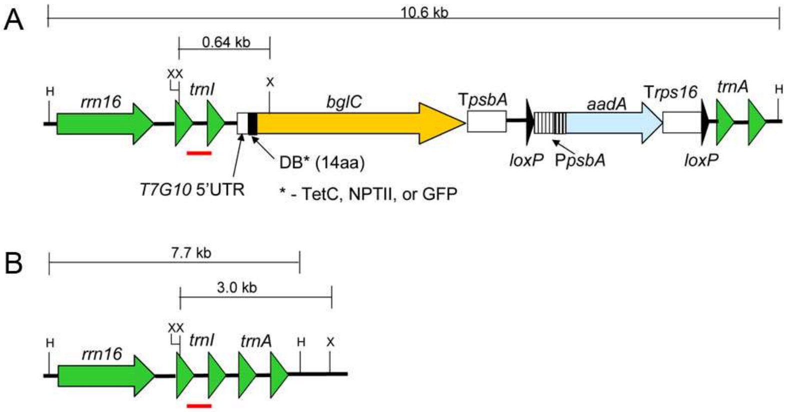


Figure 3.1: (A) Schematic diagram of the DB-BglC chloroplast transformation vectors. The three DB-BglC ORFs were inserted between the plastid *trnI* and *trnA* genes, to be transcribed from the native ribosomal promoter, located upstream of the 16s ribosomal DNA (*rrn16*). An *aadA* expression cassette, flanked by *loxP* sites, was linked to the *bglC* ORF for selection of transformed plants on spectinomycin. (B) Schematic diagram of the wild-type *trnI*/*trnA* region. The *Xho*I (X) and *Hind*III (H) sites relevant to DNA blotting experiments are shown, along with the predicted fragment sizes. The location of the *trnI* probe used for DNA blotting is shown below each schematic diagram.

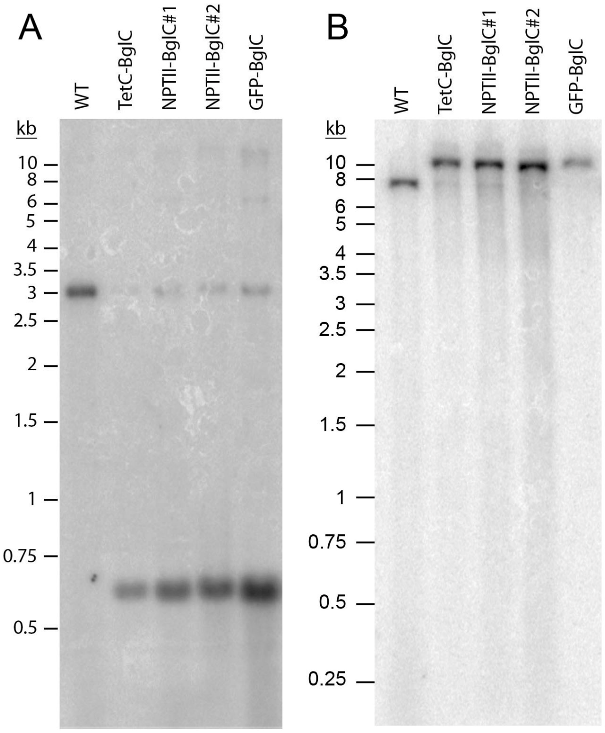


Figure 3.2: DNA blotting with WT and DB-BglC plants. (A) *Xho*I digested DNA, with 3.0 kb and 640 bp *trnI*-containing WT and DB-BglC fragments, respectively. (B) *Hind*III digested DNA, with 7.7 kb and 10.6 kb *trnI*-containing WT and DB-BglC fragments, respectively.

Chloroplast-Produced DB-BglC Protein Accumulation

When the plants each had approximately 30 leaves, soluble leaf protein was extracted from young, mature, and old (i.e., approximate leaf numbers 2, 15, and 28) NPTII-BglC, TetC-BglC, and GFP-BglC tobacco leaves for immunoblotting. Figure 3.3 shows that NPTII-BglC, TetC-BglC, and GFP-BglC accumulated to 8.0-11.6%, 1.6-2.6%, and <<0.3% of total soluble protein (%TSP), respectively. BglC protein concentration was stable as leaves aged, with similar BglC concentrations in young, mature, and old leaves for all three DB-BglC constructs tested.

Differential bglC *Transcript Processing and Abundance*

Total leaf RNA was isolated from aging leaves of T1-generation DB-BglC tobacco and separated by electrophoresis for RNA blotting. Figure 3.4 shows a complex pattern of *DB-bglC* transcript accumulation, with polycistronic RNAs transcribed from the native plastid ribosomal promoter located upstream of the *trnI*/*trnA* insertion site and various shorter processed RNAs produced from the primary polycistronic transcripts. The full-length polycistronic transcript is expected to contain the 16s, 23s, 4.5s, and 5s rRNA, *trnI*, *bglC*, *aadA*, and *trnA* transcripts. This RNA species is not observed by RNA blotting with a *bglC* probe, presumably because the primary transcript is rapidly processed. Major bands at approximately 3.0 knt and 1.7 knt, respectively, are likely dicistronic *bglC-aadA* and monocistronic *bglC* species, based on predicted sizes and hybridization with the *bglC* probe. The *trnI-bglC* dicistron is located at approximately 2.5 knt, and tricistronic *16s rrn-trnI-bglC*, *trnI-bglC-aadA* and *bglC-aadA-trnA* transcripts are located at approximately 4.3 knt, 3.7 knt and 3.8 knt, respectively. Larger polycistronic transcripts can be seen faintly, but make up a small fraction of all *DB-bglC* transcripts. A sub-ORF sized band at approximately 1.0 knt that accumulates in all three DB-BglC plant lines is likely an intermediate in the

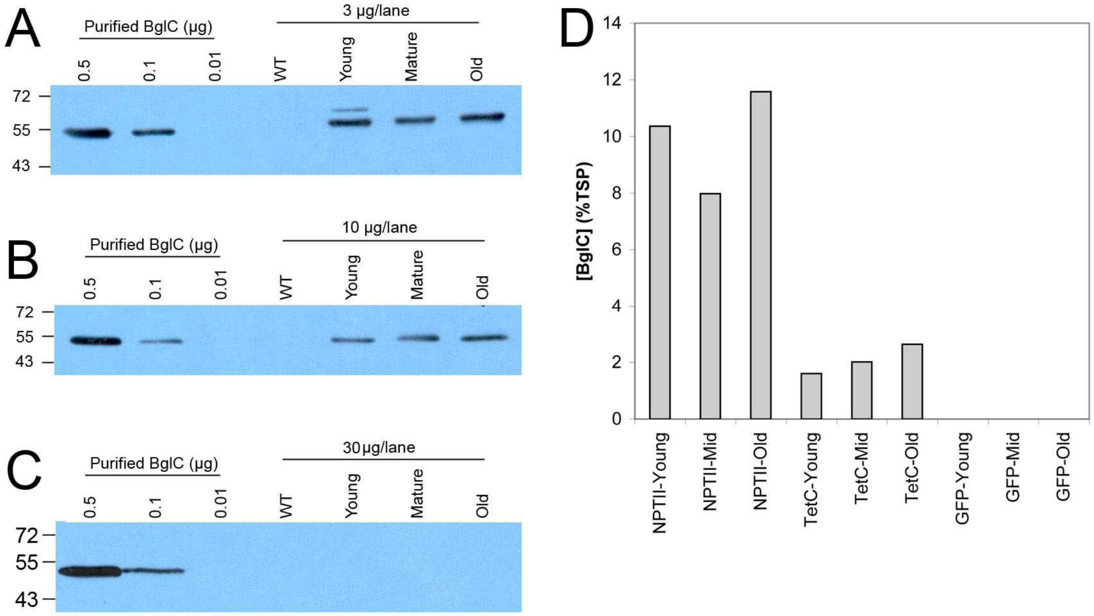


Figure 3.3: Immunoblotting with protein extracted from DB-BglC leaves. (A) NPTII-BglC. (B) TetC-BglC. (C) GFP-BglC. (D) Quantification of the immunoblots shown in A-C. NPTII-BglC accumulated to 10-12%TSP; TetC-BglC accumulated to 1.6-2.6%TSP; GFP-BglC accumulated to <<0.3%TSP.

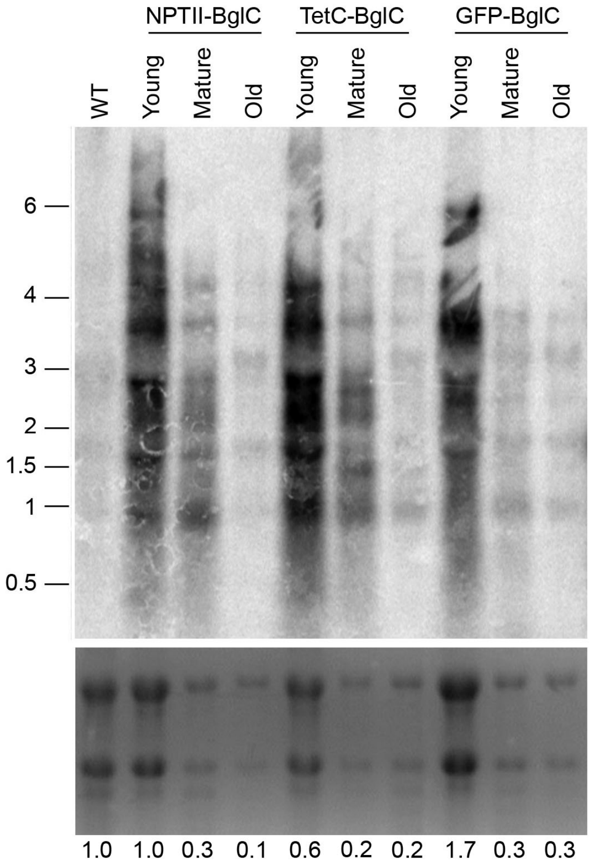


Figure 3.4: RNA blotting with RNA extracted from DB-BglC leaves, hybridized with a *bglC*-specific probe. Ethidium bromide-stained ribosomal RNA bands and relative loading quantifications are shown below the RNA blot.

bglC mRNA degradation pathway formed by endonucleolytic cleavage of the *bglC* monocistron.

No major quantitative or qualitative differences were seen in the *DB-bglC* transcripts produced by the three different DB-BglC constructs. All three plant lines showed a decrease in the concentration of *DB-bglC* transcripts in aging leaves, despite relatively little change in DB-BglC protein accumulation.

nptII-bglC *Transcript Association with Polysomes*

To determine which of the *nptII-bglC* transcripts observed above were actively translated, polysome-associated RNA extracted from NPTII-BglC plants was size fractionated on sucrose gradients and used for RNA blotting along with total RNA extracted from wild-type and from NPTII-BglC tobacco (Figure 3.5). Most *nptII-bglC*-containing transcripts were associated with polysomes, though some were more heavily associated with polysomes as judged by their enrichment in the bottom fractions of the sucrose gradient. Monocistronic *nptII-bglC* transcripts (~1.7 knt), dicistronic *bglC-aadA* (~3.0 knt) and *trnI-bglC* (~2.5 knt) transcripts, and tricistronic *trnI-bglC-aadA* (~3.7 knt) and *bglC-aadA-trnA* (~3.8 knt) transcripts were the main transcripts in the bottom fractions of the sucrose gradient. Notably, these transcripts are all of approximately equal abundance in total NPTII-BglC RNA, yet the monocistron is far more abundant in fractions 4-6 of the sucrose gradient than the polycistron. A major band of sub-ORF size (~1.0 knt) that is clearly visible in total NPTII-BglC RNA is faintly visible in fractions 1-3 of the sucrose gradient, indicating that this RNA species is weakly associated with ribosomes despite the lack of a full *nptII-bglC* ORF.

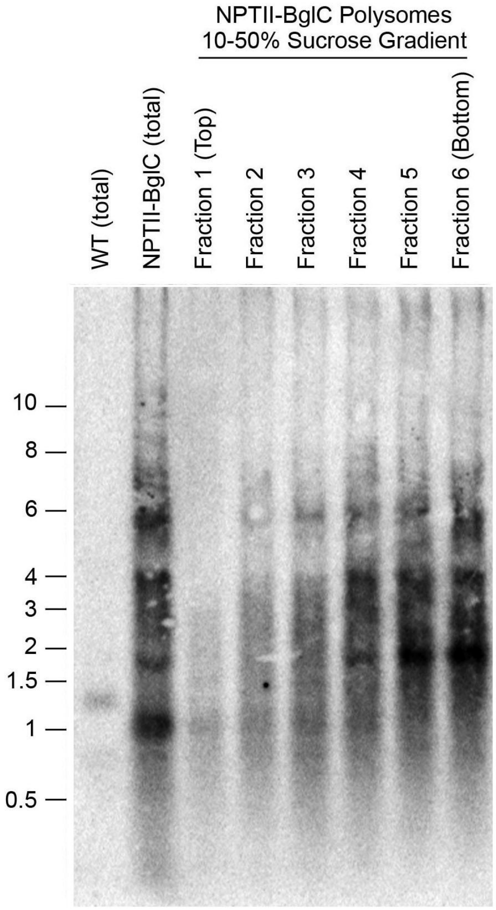


Figure 3.5: Polysome fractionation of *nptII-bglC* transcripts in a 10-50% (w/v) sucrose gradient. The monocistronic *nptII-bglC* transcript (~1.7 knt) was the primary *nptII-bglC* containing RNA species in the bottom of the gradient. Di- and tri-cistronic transcripts were also polysome associated. A sub-ORF sized RNA species (~1.0 knt) was faintly detectable in the top fractions of the sucrose gradient.

Mapping of 5' and 3' Ends of Monocistronic bglC *mRNAs*

Circular RT-PCR was performed to simultaneously determine the 5' and 3' ends of the monocistronic *DB-bglC* mRNAs produced in plants transformed with each of the three DB-BglC constructs. Figure 3.6A shows a schematic diagram of the *DB-bglC* mRNA ends determined with this method. The most common 5' end for monocistrons from all three DB-BglC constructs was at -124 (relative to the +1 start codon), in the intergenic region between *trnI* and *bglC* (Figure 3.6B). RNA from all three DB-BglC constructs contained a small number of untranslatable 5' ends within the *bglC* ORF. The distribution of observed 5' ends observed in monocistrons produced in each DB-BglC plant line was remarkably similar. Figure 3.6C shows a similar analysis of the observed 3' transcript ends, where greater differences were observed among the three DB-BglC constructs. Strikingly, *tetC-bglC* and *gfp-bglC* monocistrons were more likely to have a 3' end inside the *bglC* ORF, resulting in an mRNA that could not encode the full-length BglC protein, than *nptII-bglC* monocistrons. Over 80% of observed *nptII-bglC* monocistrons had a 3' end at least 51 nt downstream of the *bglC* stop codon. Over 75% of monocistrons in TetC-BglC tobacco contained 0-99 nt downstream of the *bglC* stop codon, and 75% of observed monocistrons from GFP-BglC tobacco retained fewer than 50 nt downstream of the stop codon. Significant fractions (19% and 42%, respectively) of observed *tetC-bglC* and *gfp-bglC* monocistrons were untranslatable due to 3' ends within the *bglC* ORF. The precise 5' and 3' termini determined by circular RT-PCR are shown in supplementary Table 3.S2.

Figures 3.7A and 3.7B show the lowest-energy predicted structures of the 5' and 3' termini, respectively, of *DB-bglC* monocistrons. The commonly observed -160 (within *trnI*) and -124 (in the intergenic region between *trnI* and *bglC*) 5' termini are marked in Figure 3.7A. Similarly, the commonly observed +1598 (T*psbA*) and +1645

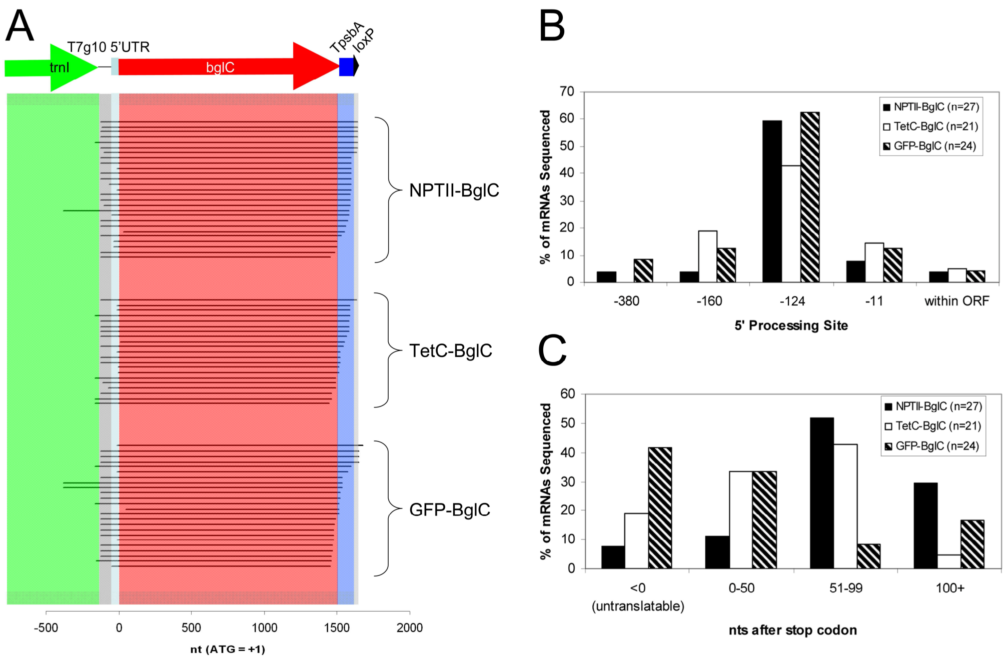


Figure 3.6: Identification of 5’ and 3’ ends of monocistronic *DB-bglC* mRNAs by circular RT-PCR. (A) Schematic diagram (to scale) showing the transcripts from each construct detected by circular RT-PCR. (B) Analysis of 5’ ends detected in each DB-BglC plant, relative to the AUG start codon at +1. (C) Analysis of 3’ ends detected in each DB-BglC plant.

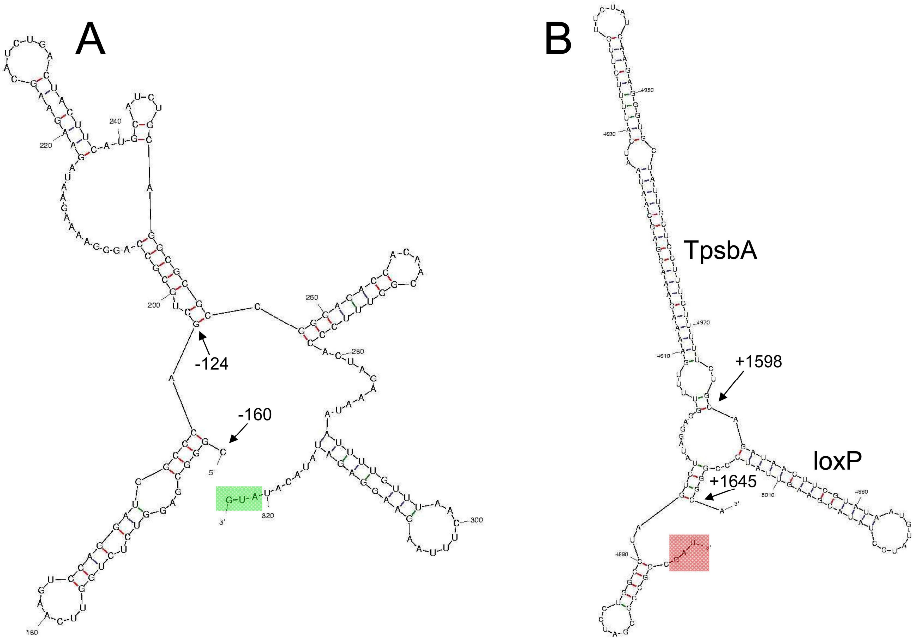


Figure 3.7: Lowest-energy structure predictions around the observed 5' and 3' ends of monocistronic *DB-bglC* transcripts. (A) Structure prediction of the 5' terminus, showing the commonly observed -160 and -124 5' termini and the AUG start codon (highlighted in green). (B) Structure prediction of the 3' terminus, showing the UAG stop codon (highlighted in red), the T*psbA* hairpin-loop, and the *loxP* hairpin-loop. Commonly observed +1598 and +1645 3' termini are indicated by arrows.

(*loxP*) 3’ termini are marked in Figure 3.7B. These figures show that the most commonly observed 5’ and 3’ termini are found at the ends of predicted hairpin structures, consistent with previous reports of chloroplast mRNA maturation by exonuclease trimming up to a hairpin structure (Hayes et al 1999; Monde et al 2000).

Identification of Polyadenylated bglC *transcripts*

It has been reported previously that chloroplast mRNAs are polyadenylated prior to their exonucleolytic degradation (Hayes et al 1999). Somewhat surprisingly, no polyadenylated transcripts were detected in the circular RT-PCR experiments described above, despite the detection of transcripts that appear to have been degraded by exonucleases. In order to determine whether *bglC* transcripts were polyadenylated in the chloroplast, reverse transcription was performed using forward primer BglCint-fwd and an oligo(dT)$_{17}$ primer as the reverse primer. Two primer pairs, BglCint-fwd/oligo(dT)$_{17}$ and BglCint-fwd2/oligo(dT)$_{17}$, were used for PCR amplification of polyadenylated *bglC* transcripts. Three polyadenylated *bglC* species were identified by this method (Figure 3.8). Two of the polyadenylation sites were located within the *bglC* ORF. A third polyadenylation site was located near the 3’ end of T*psbA*, though this site could also be located at the 3’ end of *loxP*; the exact polyadenylation site is difficult to determine more precisely than ±30 nucleotides.

All three polyadenylation sites were amplified from all three DB-BglC plants, though the relative abundances differed among the three constructs. NPTII-BglC plants contained the least polyadenylated *bglC* mRNA, and TetC-BglC plants contained more polyadenylated *bglC* mRNA than GFP-BglC plants. These differences were particularly pronounced for the polyadenylation sites within the *bglC* ORF, while polyadenylation downstream of T*psbA* appeared to be roughly equal among the three constructs.

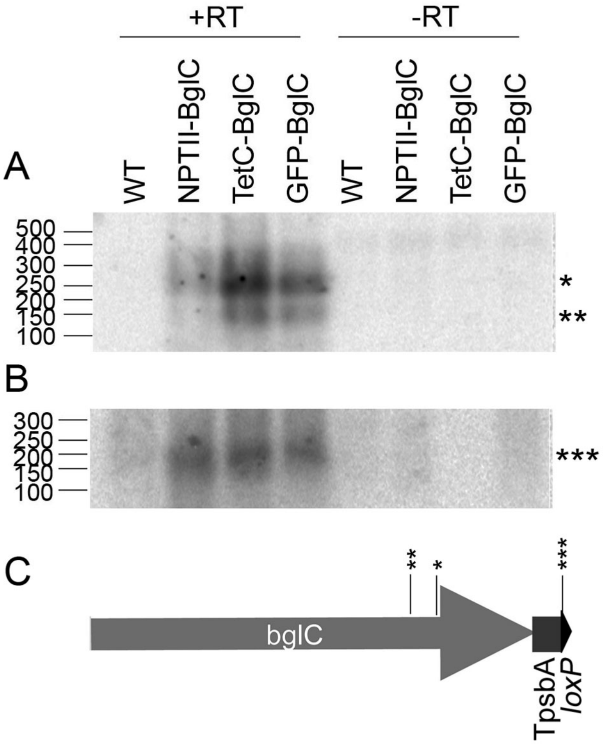


Figure 3.8: RT-PCR amplification and DNA blotting detection of polyadenylated *DB-bglC* transcripts. (A) PCR BglCint-fwd/oligo(dT)$_{17}$. (B) PCR BglCint-fwd2/oligo(dT)$_{17}$. (C) Schematic diagram (to scale) showing the observed polyadenylation sites in the *bglC* ORF and near the 3' end of T*psbA*. Both PCRs followed reverse transcription using primers BglCint-fwd and oligo(dT)$_{17}$.

Measurement of Chloroplast-Produced NPTII-BglC Activity

In order to determine whether chloroplast-produced NPTII-BglC protein was correctly folded and active against cellobiose, soluble NPTII-BglC tobacco leaf protein was extracted and incubated with cellobiose. Glucose concentration was measured after a 10 minute incubation at 50ºC. Figure 3.9 shows that NPTII-BglC tobacco leaf protein extract was able to produce glucose from cellobiose. Wild-type tobacco protein extract did not hydrolyze an appreciable amount of cellobiose, strongly suggesting that the cellobiose hydrolysis observed was a result of NPTII-BglC protein. Quantification of NPTII-BglC protein concentration against a calibration curve of known amounts of BglC added to WT tobacco protein was in good agreement with quantification of NPTII-BglC protein concentration from the immunoblot in Figure 3.3. This indicates that all, or nearly all, chloroplast-produced NPTII-BglC was correctly folded and active against cellobiose.

Hydrolysis of Tobacco Leaf Tissue by Chloroplast-Produced NPTII-BglC

In order to determine whether NPTII-BglC leaf protein extracts were suitable for use with commercial cellulase preparations for hydrolysis of complex lignocellulosic substrates, WT tobacco leaf tissue was pre-treated in NaOH, then hydrolyzed with Spezyme CP, a commercially produced cellulase preparation. Wild-type or NPTII-BglC tobacco leaf protein was added to the hydrolysis reaction and glucose concentration was measured at various time points during hydrolysis. Figure 3.10 shows that glucose concentration was significantly higher after 4 hours of hydrolysis at 50ºC when both Spezyme CP and NPTII-BglC protein were added to the hydrolysis reaction than when Spezyme CP was omitted or when WT tobacco protein was used. Glucose concentrations increased over the 24 hour testing period and were 3- to 8-fold

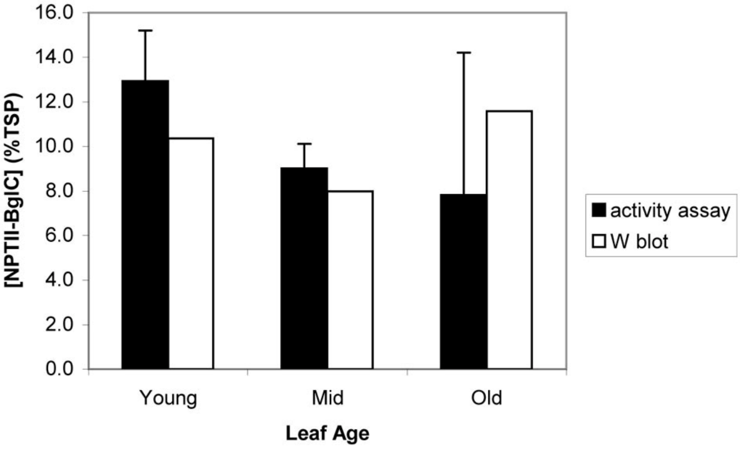


Figure 3.9: Cellobiose hydrolysis by chloroplast-produced NPTII-BglC. An NPTII-BglC leaf protein extract was incubated with cellobiose, and glucose concentration was assayed. Quantification of NPTII-BglC concentration was based on the incubation of known amounts of purified BglC with cellobiose. This quantification was in good agreement with quantification of NPTII-BglC concentration by immunoblotting, indicating that most or all of the chloroplast-produced NPTII-BglC was correctly folded and active against cellobiose.

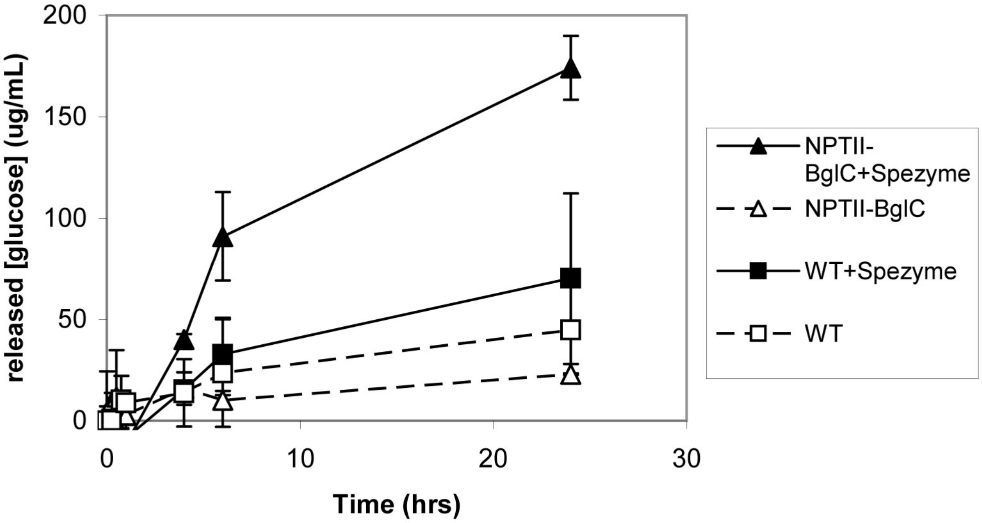


Figure 3.10: Hydrolysis of WT tobacco tissue by Spezyme CP (Genencor) with and without the addition of NPTII-BglC protein extract. Protein was extracted from WT tobacco leaf tissue, and the remaining tissue was pre-treated in NaOH, then hydrolyzed by the addition of Spezyme CP along with NPTII-BglC leaf protein (filled triangles) or with WT leaf protein (filled squares). Open triangles and squares show tobacco hydrolysis by NPTII-BglC and WT protein extracts, respectively, without the addition of Spezyme CP.

higher after 24 hours when both Spezyme CP and NPTII-BglC protein were added than in the other 3 hydrolysis mixtures tested.

Generation and Characterization of Marker-Free NPTII-BglC Tobacco

The *aadA* gene conferring spectinomycin and streptomycin resistance is not desirable following regeneration of homoplasmic transformants. To remove the *aadA* gene from fully transformed NPTII-BglC plants, leaf pieces from NPTII-BglC plants were infiltrated with *Agrobacterium* cells harboring the pCPCRE plasmid to insert a plastid-targeted *cre* gene into the nuclear genome of NPTII-BglC tobacco. Two independent transformants, NPTII-BglC/cre#1 and NPTII-BglC/cre#3, were regenerated following *Agrobacterium* infiltration. The *aadA* ORF, as well as the plastid regulatory DNA sequences P*psbA* and T*rps16* flanking the *aadA* ORF, were excised from both NPTII-BglC/cre#1 and NPTII-BglC/cre#3 tobacco lines, as confirmed by DNA blotting (Figure 3.11A). Sequencing of PCR BglCint-fwd/trnAint-rev showed that the *aadA* expression cassette was successfully removed, leaving only one *loxP* site (data not shown). Note that a minor plastid DNA species present in NPTII-BglC tobacco and formed by recombination between the native and introduced copies of P*psbA* (6.6 kb *Xho*I/*Kpn*I fragment; Gray et al 2009b) is not detected in either NPTII-BglC/cre line of tobacco, despite the presence of only one *loxP* site in this DNA species. This suggests that this minor DNA species is not actively maintained in the plastid, perhaps due to a lack of replication origin.

Immunoblotting was performed with NPTII-BglC tobacco containing the *aadA* gene and NPTII-BglC/cre#1 and NPTII-BglC/cre#3 tobacco lines from which *aadA* had been excised to test whether removal of the *aadA* gene affected the accumulation of NPTII-BglC protein. As shown in Figure 3.11B, NPTII-BglC protein accumulation

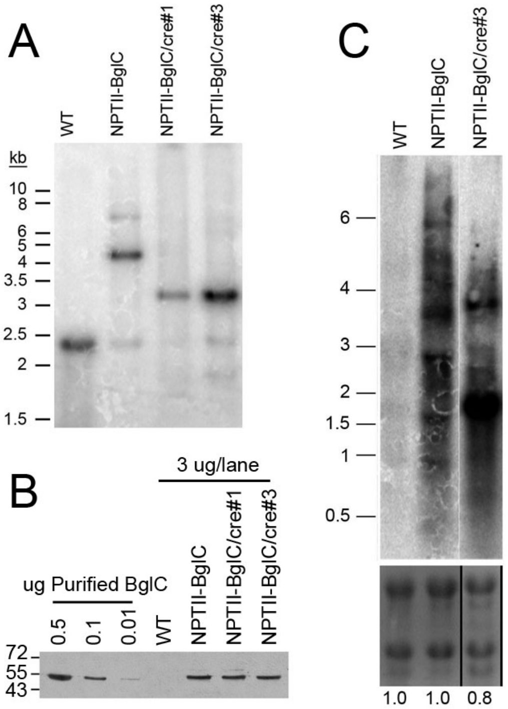


Figure 3.11: Analysis of NPTII-BglC/cre lines lacking the *aadA* expression cassette following Cre-mediated marker gene removal. (A) DNA blot following *Xho*I/*Kpn*I digestion of DNA. Cre-mediated removal of the *aadA* expression cassette was confirmed by the change in mobility of the *Xho*I/*Kpn*I fragment from 4.4 kb (NPTII-BglC) to 3.2 kb (NPTII-BglC/cre). (B) Immunoblot with leaf protein from NPTII-BglC, NPTII-BglC/cre#1, and NPTII-BglC/cre#3 tobacco lines. Accumulation of NPTII-BglC protein was unaffected by Cre-mediated marker gene removal, with NPTII-BglC accumulation of 10-12%TSP. (C) RNA blot with WT, NPTII-BglC, and NPTII-BglC/cre#3 RNA, probed by a *bglC*-specific radiolabeled probe. Ethidium bromide stained rRNA bands and loading quantification are shown below the RNA blot.

was unaffected by *aadA* excision, with high-level (10-12%TSP) accumulation of NPTII-BglC protein in both NPTII-BglC and NPTII-BglC/cre tobacco lines.

RNA blotting was also performed with NPTII-BglC/cre#3 tobacco (Figure 3.11C), surprisingly revealing that the accumulation of monocistronic *nptII-bglC* mRNA (~1.7 knt) was greatly increased in this line of marker-free NPTII-BglC/cre tobacco relative to NPTII-BglC tobacco containing the *aadA* expression cassette. A major band at ~3.3 knt likely derives from a tricistronic *trnI-nptII-bglC-trnA* transcript. Weak bands at ~2.5-2.6 knt likely derive from dicistronic *trnI-bglC* and *bglC-trnA* transcripts.

NPTII-BglC/cre plants appeared phenotypically normal in MS agar tissue culture, and were transferred to soil for greenhouse growth. Following transfer to soil, both NPTII-BglC/cre lines were unhealthy, with variegated leaves containing pale green and bleached sections (supplementary Figure 3.S1). Analysis of leaf protein extracted from these unhealthy leaves revealed a lowered Rubisco content as determined by Ponceau staining of nitrocellulose membranes, though immunoblotting showed that NPTII-BglC protein concentration was unchanged (data not shown). Confocal microscopy was performed with leaf sections from wild-type, NPTII-BglC, and NPTII-BglC/cre#1 tobacco leaves, revealing problems with NPTII-BglC/cre chloroplasts (Figure 3.12). NPTII-BglC/cre#1 leaves contained fewer chloroplasts per cell than either WT or NPTII-BglC leaves (20 ± 5 WT chloroplasts per cell; 17 ± 2 NPTII-BglC chloroplasts per cell; 9 ± 2 NPTII-BglC/cre#1 chloroplasts per cell; Figures 3.12A-C). Close examination of individual chloroplasts revealed that the grana stacks appeared to be degraded in many NPTII-BglC/cre#1 chloroplasts, in contrast with the appearance of WT and NPTII-BglC chloroplasts (Figures 3.12D-F). Both NPTII-BglC and NPTII-BglC/cre#1 cells were smaller than WT cells (WT cell area 24 ± 0.7 μm^2; NPTII-BglC cell area 12 ± 0.4 μm^2). Although NPTII-BglC cell

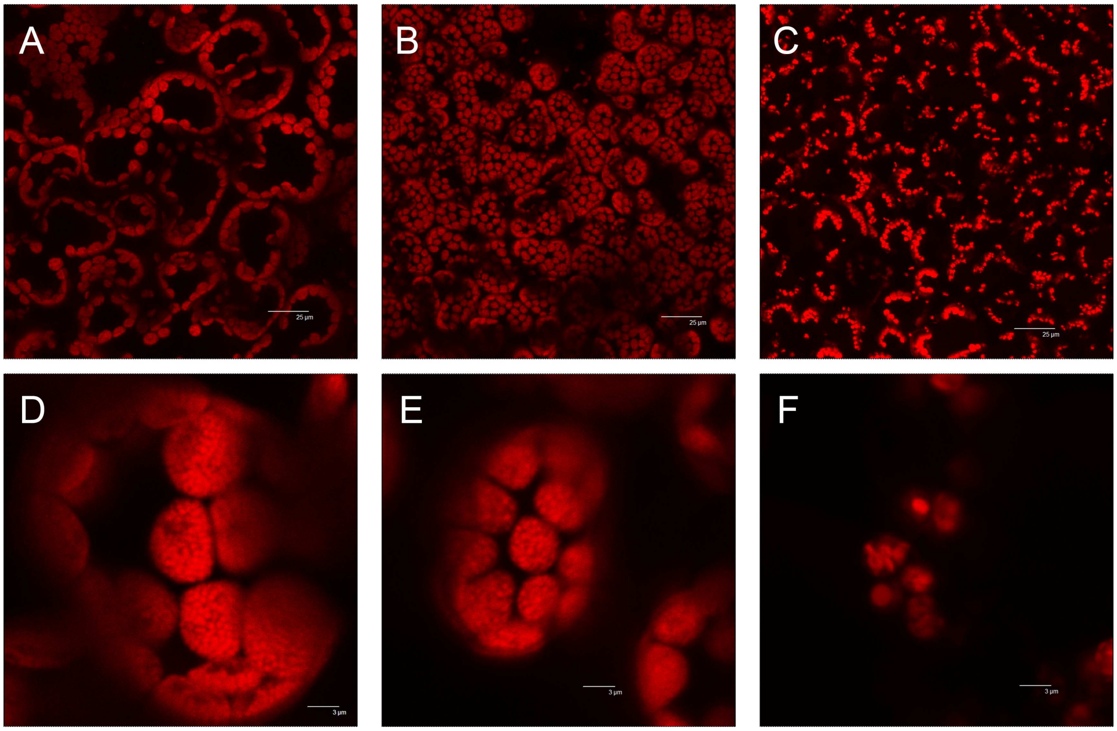


Figure 3.12: Confocal microscopy with soil-grown WT, NPTII-BglC, and NPTII-BglC/cre#1 tobacco. NPTII-BglC/cre#1 (C and F) tobacco contained fewer chloroplasts per cell than either WT (A and D) or NPTII-BglC (B and E) tobacco, and many of the NPTII-BglC/cre#1 chloroplast grana stacks appeared to be degraded.

sizes were smaller, the chloroplasts contained in each cell were also smaller (WT chloroplast area 0.7 ± 0.03 µm^2; NPTII-BglC chloroplast area 0.4 ± 0.04 µm^2), resulting in similar numbers of chloroplasts per cell. These unexpected differences in cell and chloroplast sizes were not pursued further.

DNA was extracted from unhealthy leaves of NPTII-BglC/cre#1 and NPTII-BglC/cre#3 plants growing in soil and used for DNA blotting. In contrast with the DNA blotting performed shortly after transformation with plastid-targeted Cre that showed primarily one band of the expected size, the DNA blot in figure 3.13 shows two major bands in NPTII-BglC/cre lanes, one of the expected size and one significantly smaller (approximately 0.9 kb *Xho*I/*Hind*III fragment, as compared with the expected 2.4 kb fragment). The two major *trnA*-containing bands in NPTII-BglC/cre lines were of approximately equal abundance, suggesting that they may have formed via flip-flop recombination of *loxP*, or *loxP*-like, sites in opposite orientation (Corneille et al 2003).

PCR was performed in an attempt to identify the unexpected bands observed by DNA blotting. Two *loxP*-like sites in the plastid genome that were able to recombine with genuine *loxP* sites were identified previously (Corneille et al 2003), and were suspected to be the cause of the unexpected bands. The *loxP*-like sites in P*psbA* and in *rps12* that were previously identified were PCR amplified from both NPTII-BglC/cre lines using primers trnKint-fwd/trnAint-rev and rps12-rev/trnAint-rev, respectively, and confirmed by sequencing (data not shown). Unintentional Cre-mediated recombination at the P*psbA* and *rps12 loxP*-like sites took place at precisely the same location in the 8-bp *loxP* spacer region that was observed by Corneille et al (2003). Neither of the unintended recombination products formed by Cre-mediated recombination at the P*psbA* or *rps12 loxP*-like sites is expected, however, to give rise to the 0.9 kb *Xho*I/*Hind*III fragments observed by DNA blotting. In order to identify

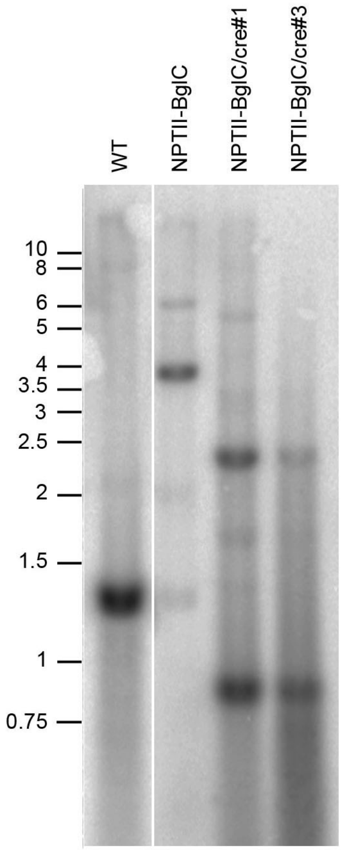


Figure 3.13: DNA blot with *Xho*I/*Hind*III digested DNA from unhealthy NPTII-BglC/cre tobacco lines following their transfer to soil. DNA blotting revealed, in addition to the expected 2.4 kb *Xho*I/*Hind*III fragment, a second major *Xho*I/*Hind*III fragment at approximately 0.9 kb in unhealthy NPTII-BglC/cre plants growing on soil.

other Cre-mediated recombinations, PCR was performed with primers 16s-rev/trnAint-rev and BglCint-rev/trnAint-rev using the DNA extracted from very unhealthy NPTII-BglC/cre leaves, and the resulting PCR products were sequenced. This sequencing revealed recombination between the *loxP* site remaining following Cre-mediated excision of the *aadA* expression cassette and two previously unreported *loxP*-like sequences, one in the NPTII DB region (Figure 3.14B) and one in the 16S *rrn*/*trnV* intergenic region (Figure 3.14C). Both of these recombinations occurred via flip-flop recombination of *loxP*-like sequences in opposite orientation, and recombination likely occurred in the 8 bp *loxP* spacer region, in agreement with previous observations of unintended Cre-mediated recombination in plastids (Corneille et al 2003). Note that recombination at the *loxP*-like site in the 16S *rrn*/*trnV* intergenic region may have occurred in either the spacer region or in the Cre-binding region; the exact site of recombination cannot be determined because of the long tract of homologous sequence between *loxP* and this *loxP*-like site at the junction between the *loxP* spacer and Cre-binding site. The unintended recombination at the *loxP*-like site in the NPTII DB region results in the formation of a DNA species that would give a 0.9 kb *Xho*I/*Hind*III fragment, as compared with the expected 2.4 kb *Xho*I/*Hind*III fragment following excision of the *aadA* expression cassette (Figure 3.14A), making this recombination likely to be the major unintended Cre-mediated recombination in NPTII-BglC/cre tobacco. Note that this DNA species gives a 1.9 kb *Xho*I/*Kpn*I *trnA*-containing fragment, faintly visible in NPTII-BglC/cre#3 tobacco DNA in the DNA blot shown in Figure 3.11A.

A sequence alignment of the genuine *loxP* sequence and the four *loxP*-like sequences observed to undergo Cre-mediated recombination with the genuine *loxP* sequence in transgenic plastids resulted in the identification of a consensus sequence and a number of bases apparently important for Cre recognition (supplementary Figure

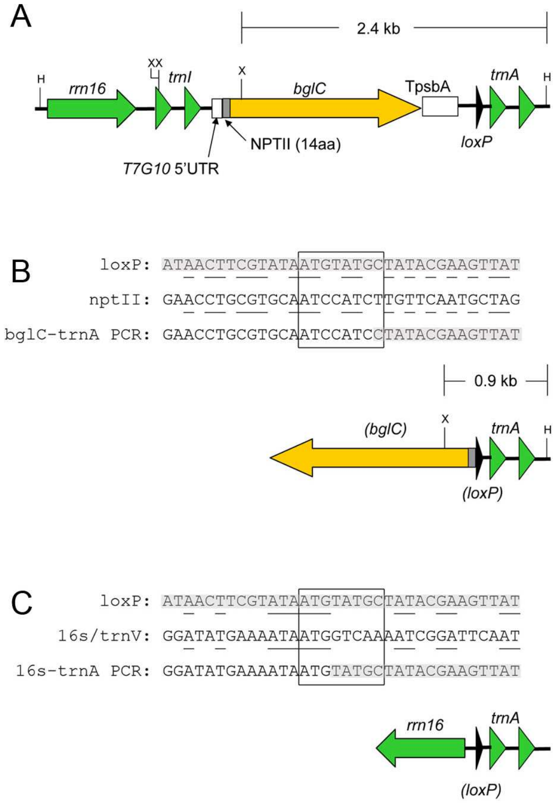

A
2.4 kb
H
rrn16
XX
trnI
X
bglC
TpsbA
trnA
H
NPTII (14aa)
loxP
T7G10 5'UTR
B
loxP: ATAACTTCGTATAATGTATGCTATACGAAGTTAT
nptII: GAACCTGCGTGCAATCCATCTTGTTCAATGCTAG
bglC-trnA PCR: GAACCTGCGTGCAATCCATCCTATACGAAGTTAT
0.9 kb
(bglC)
X
trnA
H
(loxP)
C
loxP: ATAACTTCGTATAATGTATGCTATACGAAGTTAT
16s/trnV: GGATATGAAAATAATGGTCAAAATCGGATTCAAT
16s-trnA PCR: GGATATGAAAATAATGTATGCTATACGAAGTTAT
rrn16
trnA
(loxP)

Figure 3.14: Identification of unintended Cre-mediated recombination events in NPTII-BglC/cre tobacco. (A) Schematic diagram of NPTII-BglC tobacco DNA following Cre-mediated excision of the *aadA* expression cassette. *Xho*I (X) and *Hind*III (H) sites relevant to DNA blotting are shown. (B) Sequencing of an unintended Cre-mediated recombination event at a *loxP*-like site in the NPTII DB region and schematic diagram showing the resulting DNA species, with *Xho*I (X) and *Hind*III (H) sites relevant to DNA blotting. (C) Sequencing of an unintended Cre-mediated recombination event at a *loxP*-like site in the 16s *rrn*/*trnV* intergenic region and schematic diagram showing the resulting DNA species. Bases common to the genuine *loxP* sequence and the *loxP*-like sequences resulting in unintended recombinations are underlined, and the 8 bp spacer region is boxed. Both of these recombinations were identified in both NPTII-BglC/cre#1 and NPTII-BglC/cre#3 tobacco. In (B) and (C), the *loxP* sites and the *bglC* ORF are labeled with parentheses to indicate that these features were lost as a result of Cre-mediated recombination.

3.S2). Only three of the 34 bases in the *loxP* sequence were strictly conserved among the four *loxP*-like sequences, though 13 of the 34 bases showed a strong preference for a particular nucleotide, and an additional 7 positions showed a strong preference for one of two nucleotides (e.g., A or T at position 3). The consensus *loxP* sequence reported here (ATWnnWTnSnATWnnATnnnTTATAYnMnnnTnY) differs significantly from the previously reported *loxP* consensus sequence of ATnACnnCnTATAnnnTAnnnTATAnGnnGTnAT (Corneille et al 2003), though it should be noted that this consensus sequence also did not match the two *loxP*-like sites reported by Corneille et al (2003). The *loxP* spacer region has been reported to require a TA dinucleotide sequence at its center (Missirlis et al 2006). This is in contrast with the observed unintended Cre-mediated recombinations in transgenic plastids, where a T was observed in 3 of 4 unintended recombinations, but the A reported previously to be important for Cre-mediated recombination was not observed in any of the 4 unintended recombinations (Supplementary Figure 3.S2).

Both NPTII-BglC/cre lines ultimately died on soil. Prior to their death, leaves from both NPTII-BglC/cre lines were surface sterilized and transferred to non-selective RMOP agar medium in an attempt to rescue these lines. NPTII-BglC/cre#1 was unable to be rescued due to persistent problems with contamination, but NPTII-BglC/cre#3 was successfully regenerated on RMOP medium to form new shoots that were transferred to MS medium for rooting. In MS or RMOP tissue culture medium, these shoots did not have the phenotype described above for marker-free NPTII-BglC/cre plants grown in soil. DNA blotting analysis using DNA extracted from NPTII-BglC/cre#3 plants rescued on tissue culture and re-rooted in MS agar revealed that the rescued plants contained only the expected plastid DNA species resulting in a 2.4 kb *trnA*-containing band, and no detectable 0.9 kb fragments (data not shown).

Currently, it is not known whether these rescued lines will be able to grow to maturity on soil for seed collection.

DISCUSSION

Previous DNA blots with DNA from plants transformed in the *trnI*/*trnA* intergenic region using the *Xho*I restriction enzyme were inconclusive as to whether the transformed plants were fully transformed or whether a small amount of WT plastid DNA remained, even after multiple rounds of tissue culture regeneration (Gray et al 2009a). A faint WT signal was detected by DNA blotting with DB-BglC DNA after *Xho*I digestion (Figure 3.2A), which could result either from a small amount of WT plastid DNA or from an extraplastidic copy of the *trnI*/*trnA* region. Plastid DNA sequences have been found in the *N. tabacum* nuclear genome, often with point mutations in the nuclear copies of plastid DNA resulting in differing restriction digest patterns (Ayliffe and Timmis 1992). I therefore hypothesized that if the faint WT bands observed following *Xho*I digestion resulted from extraplastidic copies of the *trnI*/*trnA* region rather than from a low level of heteroplasmy, then DNA blotting with the appropriate restriction enzyme should remove the faint WT band. The DNA blot shown in Figure 3.2B confirms that all of the DB-BglC tobacco lines described in this report are fully transformed, and suggests that an extraplastidic copy of the *trnI*/*trnA* region of plastid DNA has lost the relevant *Hind*III sites while retaining the relevant *Xho*I sites.

The DB-BglC chloroplast expression experiments described here demonstrate that the fusion of three different DB regions to an ORF of interest can result in variation of protein accumulation over more than two orders of magnitude (Figure 3.3). Similar variation in protein accumulation was seen in our previous DB-Cel6A chloroplast expression experiments (Gray et al 2009a). With both of these ORFs, the

GFP DB region resulted in the lowest observed expression, approximately 0.1%TSP or less, in contrast to the successful use of the GFP DB region to stimulate EPSPS production in chloroplasts that resulted in EPSPS accumulation of over 10%TSP (Ye et al 2001). Surprisingly, NPTII-BglC protein accumulated to the highest levels (10-12%TSP) in the constructs tested, while TetC-Cel6A accumulated to higher levels than NPTII-Cel6A (Gray et al 2009a). These results show that context is important to DB function and that, in the absence of a better understanding of DB function, empirical DB optimization is required for the gene of interest in order to ensure high-level protein production. Because the NPTII, TetC, and GFP DB regions tested in our Cel6A and BglC chloroplast expression experiments were identical at both the nucleotide and the amino acid levels, it is difficult to explain the opposite effects on protein production of the TetC and NPTII DB fusions to the *cel6A* and *bglC* ORFs. Speculatively, some feature of the mRNA secondary structure at the junction between the DB region and the *cel6A* and *bglC* ORFs could affect translation efficiency, though no obvious differences that could affect translation were predicted by m-fold software (data not shown). Alternatively, codon pair usage at the *DB-cel6A* or *DB-bglC* junction could explain the observed effects on Cel6A and BglC protein accumulation, particularly if codon pairs were formed at these junctions that were extremely favorable or unfavorable (Gutman and Hatfield 1989). Testing this hypothesis will require a calculation of codon pair usage in plastid ORFs.

RNA blotting revealed little difference in the accumulation of *bglC* transcripts among the three DB-BglC constructs (Figure 3.4). In our previous experiments expressing DB-Cel6A protein in transplastomic tobacco, accumulation of the monocistronic *cel6A* transcript correlated with accumulation of Cel6A protein (i.e., both monocistronic *tetC-cel6A* mRNA and TetC-Cel6A protein accumulated to the highest levels of the three constructs tested), suggesting a link between accumulation of the monocistron and protein production (Gray et al 2009a). Thus, it was surprising that no differences in *DB-bglC* transcript accumulation were observed by RNA blotting that could explain the observed differential DB-BglC protein accumulation.

Polysome fractionation was performed in order to determine the actively translated *nptII-bglC* transcripts (Figure 3.5). Both monocistronic and polycistronic *nptII-bglC* transcripts were found to be polysome-associated, but the monocistron was greatly enriched in the lower fractions of a sucrose gradient relative to the polycistronic transcripts, suggesting that although both monocistrons and polycistrons are translated to produce NPTII-BglC protein, translation of the *nptII-bglC* monocistron is more efficient than *nptII-bglC* translation from polycistronic transcripts. Although there were no obvious differences in monocistron accumulation among the three DB-BglC constructs tested, polysome fractionation showed that accumulation of the monocistron is an important step for efficient production of NPTII-BglC protein. These results are consistent with previous reports of efficient translation from polycistronic transcripts to produce foreign protein in transplastomic tobacco (Staub and Maliga 1995; Quesada-Vargas et al 2005), but in the case of *nptII-bglC*, RNA processing to produce monocistronic mRNAs appears to result in more efficient translation. Zhou et al (2007) observed a similar increase in translation efficiency of a *yfp* ORF following RNA processing to produce a monocistronic transcript.

Having established that processing of polycistronic transcripts to yield monocistrons can result in more efficient translation to produce NPTII-BglC protein, we wished to determine whether there were subtle differences among the three *DB-bglC* monocistrons. The precise 5’ and 3’ termini of monocistronic *nptII-bglC*, *tetC-bglC*, and *gfp-bglC* mRNAs were determined by circular RT-PCR, as shown in Figure 3.6. Mapping of mRNA ends showed that the 5’ end of the majority of monocistronic mRNAs for all three constructs was at -124 (relative to the +1 AUG start codon), in the intergenic region between *trnI* and the *bglC* ORF. The 3' termini of the three different bglC constructs, however, displayed striking differences. Most *nptII-bglC* monocistrons retained 100-150 nt after the stop codon of the *bglC* ORF. Monocistrons of *tetC-bglC* and *gfp-bglC* mostly retained 50-100 and <50 nt after their respective stop codons. These experiments were suggestive of differential 3'-5' exonucleolytic mRNA degradation, with transcripts encoding more abundant BglC proteins showing less degradation than transcripts encoding less abundant BglC proteins. Because the three DB-BglC constructs differed only at the 5' end of the *DB-bglC* ORF, these results suggest that the DB region plays a role in regulating mRNA turnover. The circular RT-PCR employed here to determine 5' and 3' RNA ends was not capable of detecting monocistronic RNAs that had been degraded more than approximately 40 nt past the start codon from the 5' end or approximately 50 nt past the stop codon from the 3' end due to the annealing sites of the primers used for these experiments.

RNA structure predictions showed hairpin-loop structures at both the 5' and 3' termini of the monocistronic *bglC* species (Figure 3.7). Particularly relevant are the predicted hairpins at the 3' terminus in T*psbA* and *loxP*, respectively. Many 3' termini were found immediately downstream of a predicted hairpin-loop structure, and *nptII-bglC* monocistrons were more likely to contain both the T*psbA* and the *loxP* hairpins. Many *tetC-bglC* and *nptII-bglC* monocistrons retained only the T*psbA* hairpin or a portion of the hairpin, while most *gfp-bglC* monocistrons (75%) contained neither the T*psbA* nor the *loxP* hairpins. The observations of some *tetC-bglC* and *gfp-bglC* monocistrons containing both the T*psbA* and *loxP* hairpins suggest that the observed differences are not a result of differential transcription or RNA processing, but of differential transcript degradation. Thus, it appears that many of the *tetC-bglC* and *gfp-bglC* monocistrons observed by RNA blotting had in fact been partially degraded by the loss of stabilizing 3' hairpin-loop structures, often into the *bglC* ORF resulting in an untranslatable transcript. These results are consistent with previous reports of 3'-5' RNA degradation in chloroplasts. A proposed model for chloroplast RNA maturation includes intergenic endonucleolytic cleavage followed by 3'-5' exonuclease processing until the exonuclease reaches a hairpin structure (Monde et al 2000).

A hypothesis to explain the RNA blotting, polysome fractionation, and circular RT-PCR observations described here follows. Transcription from the native 16s ribosomal RNA promoter generates a primary polycistronic transcript, followed by endonucleolytic cleavages to generate monocistronic rRNA, tRNA, and *DB-bglC* and *aadA* mRNA species. I propose on the basis of the polysome fractionation experiments (Figure 3.5) that all of the various *bglC*-containing polycistronic and monocistronic transcripts are actively translated, but that translation of the monocistronic mRNA is more efficient than translation in a polycistronic context. Further, I propose on the basis of the RNA blotting experiments (Figure 3.4) that transcription and processing of the primary transcript (i.e., generation of the monocistronic *DB-bglC* transcript) is approximately equal among the three DB-BglC constructs tested. The circular RT-PCR results (Figure 3.6) suggest that the three *DB-bglC* monocistrons are differentially degraded, primarily by 3'-5' exonucleases. Greater association of polyribosomes with the monocistronic *nptII-bglC* than with the monocistronic *tetC-bglC* or *gfp-bglC* transcripts results in a stabilization of the monocistronic *nptII-bglC* mRNA relative to the other two *DB-bglC* monocistrons tested. In the absence of translation, or if the monocistron is only weakly translated, the *loxP* and T*psbA* hairpins are cleaved by the combined actions of endo- and exo-nucleases as discussed above. When both of the 3' hairpins are removed from the monocistron, 3’-5’ exonucleases degrade the mRNA into the *bglC* ORF, resulting in an untranslatable transcript and finally in degradation of the mRNA. Similar results were observed by Kuroda and Maliga (2001a), who described lower accumulation of weakly-translated *neo* mRNAs (as detected by RNA blotting) fused to inefficient DB regions than of *neo* mRNAs fused to efficient DB regions.

Some *DB-bglC* transcripts were identified by circular RT-PCR that appeared to have been degraded from the 5’ end, suggesting a possible 5’-3’ exonuclease activity. The presence of 5’-3’ exonuclease activity in chloroplasts has been proposed previously and could result from a processive endonuclease like *E. coli* RNase E (Coburn and Mackie 1999). An RNase E homologue has been identified in higher plant chloroplasts with an activity similar to *E. coli* RNase E that could potentially account for this 5’-3’ degradation (Schein et al 2008). Whatever the enzyme responsible for the apparent 5’-3’ *bglC* transcript degradation, mRNA degradation in the 3’-5’ direction appears to be the prevailing mode of exonucleolytic *bglC* mRNA degradation.

In addition to the identification of *bglC* transcripts that were degraded by exonuclease degradation, an apparent endonucleolytic cleavage site was also detected by RNA blotting. Polyadenylation sites within the *bglC* ORF were identified by oligo(dT)$_{17}$ RT-PCR (Figure 3.8), consistent with endonucleolytic cleavage of the *bglC* transcript followed by polyadenylation of the mRNA fragments and exonucleolytic degradation of these polyadenylated fragments. This mechanism of chloroplast mRNA degradation has been observed previously (Klaff 1995; Bollenbach et al 2003; reviewed in Bollenbach et al 2004). The detection of *bglC* transcripts that were polyadenylated at sites internal to the *bglC* ORF, suggesting endonucleolytic cleavage, as well as the detection of *bglC* transcripts that appear to have been degraded by exonucleases without endonucleolytic cleavage suggests that there are at least three pathways to *bglC* transcript turnover: endonucleolytic cleavage followed by exonucleolytic degradation of the mRNA fragments, 3'-5' exonuclease degradation, and 5'-3' exonuclease degradation. TetC-BglC plants contained the largest amount of polyadenylated *bglC* transcripts, followed by GFP-BglC and then NPTII-BglC plants. The low abundance of polyadenylated *nptII-bglC* transcripts could be due to a low rate of *nptII-bglC* transcript degradation, consistent with the stabilization of *nptII-bglC* transcripts against 3'-5' degradation observed by circular RT-PCR. More polyadenylated *tetC-bglC* transcripts than polyadenylated *gfp-bglC* transcripts may accumulate as a result of rapid turnover of *gfp-bglC* transcripts by a combination of the three mRNA degradation pathways described above, though a differential rate of endonucleolytic cleavage of these transcripts and/or of polyadenylation of these transcripts cannot be ruled out on the basis of the experiments described here.

Because the nucleotide sequence of the vast majority of the three DB-BglC constructs is the same, it is not immediately clear what signal results in the degradation or stabilization of *bglC* mRNA. The only difference among the three constructs tested here is in the first 14 codons of the *bglC* ORF (i.e., in the DB regions). In *E. coli*, inefficient DB regions result in abortive translation in which the ribosomal subunits dissociate and the nascent polypeptide drops off from the transcript (Gonzalez de Valdivia and Isaksson 2005). Speculatively, abortive translation from inefficient DB regions (e.g., the GFP DB region studied here) could recruit the ribonucleases responsible for chloroplast mRNA degradation. Dissociation of the ribosome from the mRNA and accumulation of incomplete polypeptides could potentially serve as signals to recruit the ribonucleases, as in *E. coli*, where the introduction of premature stop codons results in dissociation of the ribosomes from mRNA and degradation of the mRNA species. This effect is particularly pronounced when premature stop codons are placed near the 5' end of an ORF (Nilsson et al

1987). Preferential degradation of weakly translated mRNAs would establish a positive feedback loop in which efficiently translated mRNA is stabilized against degradation, thus resulting in a larger pool of translatable mRNA. This hypothesis leaves open the question of what defines an efficient DB region in the context of a given ORF, but would explain the accumulation of both *nptII-bglC* monocistronic mRNA and NPTII-BglC protein. This type of positive feedback would also partially explain the order of magnitude differences in protein accumulation among the three DB-BglC constructs tested.

In *E. coli*, polynucleotide phosphorylase (PNPase) has been implicated in degradation of a weakly translated S20 RNA, likely as a result of an aspect of translation initiation and translation of the beginning of the ORF (Mackie 1989; Rapaport and Mackie 1994). A plastid-localized homologue of the *E. coli* PNPase has been found (Hayes et al 1996), suggesting that DB-mediated mRNA stabilization could function in part by inhibiting PNPase 3'-5' exonuclease activity. Bollenbach et al (2003) have reported that an endoribonuclease, CSP41a, initiates turnover of a number of chloroplast mRNAs. This is consistent with my data showing polyadenylation of sequences within the *bglC* ORF (Figure 3.8). I propose that PNPase, in conjunction with chloroplast endoribonucleases (e.g., CSP41a) and potentially other components of the chloroplast RNA degradation machinery, could be recruited by abortive translation events at inefficiently translated DB regions.

Differential rates of RNA degradation are not likely to be the sole cause of the observed differences in DB-BglC protein accumulation. Kuroda and Maliga (2001a) and Gray et al (2009a) observed differences in *neo* and *cel6A* mRNA concentrations, respectively, likely resulting from differential rates of RNA degradation when non-synonymous changes were made to the DB region fused to the ORF of interest. Synonymous changes to *rbcL* and *atpB* DB regions fused to the *neo* ORF, however, resulted in large differences in NPTII protein accumulation with no accompanying change in *neo* mRNA levels as measured by RNA blotting (Kuroda and Maliga 2001b). Similar results are described in this report for non-synonymous changes to the DB region fused to the *bglC* ORF, though a closer examination of *DB-bglC* monocistrons revealed that many of these transcripts were in fact partially degraded, demonstrating that the DB region may play a role in regulating plastid RNA degradation even when differences are not obvious by RNA blotting. Further, my observations of greatly increased *nptII-bglC* monocistron concentrations in NPTII-BglC/cre tobacco (Figure 3.11C), unaccompanied by a measurable change in NPTII-BglC protein concentration (Figure 3.11B), suggest that monocistronic *nptII-bglC* mRNA concentration in these plants is not limiting for protein production. It appears that increased rates of RNA degradation in plastids as a result of inefficient DB regions (Kuroda and Maliga 2001a; Gray et al 2009a; this report) are not a cause for decreased protein accumulation; rather, decreased translation efficiency (i.e., a decreased rate of protein production) appears to be a cause for an increased rate of RNA degradation. It is not known why inefficient DB regions sometimes result in complete breakdown of the mRNA (Kuroda and Maliga 2001a; Gray et al 2009a) and at other times result in only a partial degradation of the mRNA (this report). The lack of any observed change in *neo* mRNA concentration in the work of Kuroda and Maliga (2001b) may be due to partial degradation of weakly translated *neo* transcripts like the degraded *tetC-bglC* and *gfp-bglC* transcripts observed here, or may result from the use of plastid-derived *rbcL* and *atpB* 5'UTRs that may stabilize weakly translated transcripts.

My attempts to generate a marker-free line of NPTII-BglC tobacco by stable integration of a gene encoding plastid-targeted Cre resulted in successful removal of the *aadA* expression cassette. Marker-free NPTII-BglC/cre plants appeared healthy in tissue culture medium, but ultimately died when transferred to soil, apparently as a result of unintended chloroplast DNA recombination mediated by Cre. DNA blotting experiments shortly after transformation of NPTII-BglC tobacco with the *cre* gene showed the expected excision of the *aadA* expression cassette, demonstrating efficient and rapid removal of the target sequence located between two *loxP* sites (Figure 3.11A). Following transfer of NPTII-BglC/cre tobacco to soil, DNA blotting revealed the presence of a second major DNA species (Figure 3.13). PCR and sequencing analysis suggested that this major DNA species is the result of an unintended recombination event between the genuine *loxP* site remaining after the intended excision of *aadA* and a *loxP*-like site in the NPTII DB region (Figure 3.14B). These results suggest that Cre prefers the genuine *loxP* site as a substrate, but will cause recombination with *loxP*-like sites in the absence of a genuine *loxP* site. It is unknown whether at least one genuine *loxP* site is required for Cre-mediated recombination, or whether two *loxP*-like sites could recombine in the absence of a genuine *loxP* sequence. NPTII-BglC/cre#3 tobacco was rescued on tissue culture medium and regained a normal green phenotype. Surprisingly, the NPTII-BglC/cre#3 plastid genome lost the DNA species resulting from the unexpected Cre-mediated recombination with the *loxP*-like sequence in the NPTII DB region when this line of tobacco was rescued on tissue culture medium. This suggests that some aspect of growth on soil promotes Cre-mediated recombination, or that some aspect of growth on tissue culture medium (e.g., sucrose supplementation) suppresses Cre-mediated recombination.

A sequence alignment of the genuine *loxP* sequence with the four *loxP*-like sites observed to cause Cre-mediated recombination with the genuine *loxP* site in transgenic plastids revealed a consensus sequence differing in some respects from previously reported consensus *loxP* sequences (Supplementary Figure 3.S2; Corneille et al 2003; Missirlis et al 2006). Specifically, a TA dinucleotide sequence at the center of the *loxP* spacer region that was previously reported to be important for Cre-mediated recombination was not observed. Instead, the T in this dinucleotide was observed in 3 of 4 unintended recombinations, but the A was never observed. These results call into the question the necessity for this TA dinucleotide sequence for Cre-mediated recombination, at least in plastids. The consensus *loxP*-like sequence shown in Supplementary Figure 3.S2 could find utility in designing transgenes destined for insertion into the plastid genome when Cre/*loxP* removal of the antibiotic resistance gene will be performed. By making silent codon changes or altering regulatory DNA sequences, *loxP*-like sequences should be removed in order to avoid unintended Cre/*loxP* recombination like the recombination event between *loxP* and the *loxP*-like sequence in the NPTII DB diagrammed in Figure 3.14B.

NPTII-BglC protein accumulation was unaffected by Cre/*loxP* removal of the *aadA* expression cassette (Figure 3.11B). This suggests that the production of AAD protein does not represent a major metabolic burden on the plant under the growth conditions tested. The tissue culture and greenhouse growth conditions tested, however, were close to ideal for plant growth. It is possible that under stresses likely to be encountered in the field, a metabolic burden associated with *aadA* expression could become apparent and that NPTII-BglC accumulation following *aadA* removal could improve under these less ideal growth conditions. This hypothesis will require further testing following the successful generation of marker-free NPTII-BglC tobacco lines, by growing the plants under various stresses (e.g., heat or cold stress, high salinity, high or low light availability, nutrient deficiencies, or low water conditions).

Chloroplast-produced NPTII-BglC protein was assayed for activity against cellobiose (Figure 3.9), and was used in conjunction with Spezyme CP, a commercial cellulase preparation, to hydrolyze tobacco leaf tissue to glucose (Figure 3.10). These assays demonstrated that most, if not all, of the NPTII-BglC protein produced in chloroplasts is correctly folded and active. Importantly, NPTII-BglC was able to hydrolyze not only a purified cellobiose substrate, but also the complex mixture of glucose oligomers produced by enzymatic hydrolysis of tobacco leaf tissue. Commercial cellulase preparations are typically made from the supernatant liquid of *Trichoderma reesei* cell cultures, which have been shown to contain low levels of β-glucosidase activity relative to their cellulase activities (Juhász et al 2005). Supplementation of commercial cellulases with β-glucosidase has been shown to improve the performance of lab-scale cellulosic ethanol production (e.g., Schell et al 1990; Spindler et al 1989), and may be required for efficient glucose production on a commercial scale. A previous report of tobacco nuclear expression of a β-glucosidase from *Aspergillus niger* resulted in a maximum accumulation of 2.3%TSP when this β-glucosidase was targeted to the vacuole (Wei et al 2004). This study demonstrates the potential for a significant improvement in plant-based β-glucosidase production by expressing β-glucosidase genes from the plastid genome. The demonstration that chloroplast-produced NPTII-BglC can improve glucose production from a complex lignocellulosic substrate and the potential for low-cost foreign protein production in transgenic plants (Twyman et al 2003) suggests that transplastomic NPTII-BglC tobacco could be an important source of low-cost β-glucosidase for the cellulosic ethanol industry.

ACKNOWLEDGMENTS

B. Gray wishes to acknowledge David Wilson and Diana Irwin for their kind donation of anti-BglC antibodies for immunoblotting and the pNS6 plasmid for cloning the *bglC* gene. B. Gray also wishes to acknowledge Deborah Sills for her kind donation of Spezyme CP for tobacco hydrolysis. The authors wish to thank Robert Weiss for his kind donation of the pCAGGS-creT plasmid, Amir Sattarzadeh for his assistance with *Agrobacterium* cloning and leaf infiltration, and Carol Bayles for her assistance with confocal microscopy experiments. BNG was the recipient of an NSF Graduate Research Fellowship. This work was supported by a USDA Grant (USDA NRI 2007-02133) to BAA and MRH.

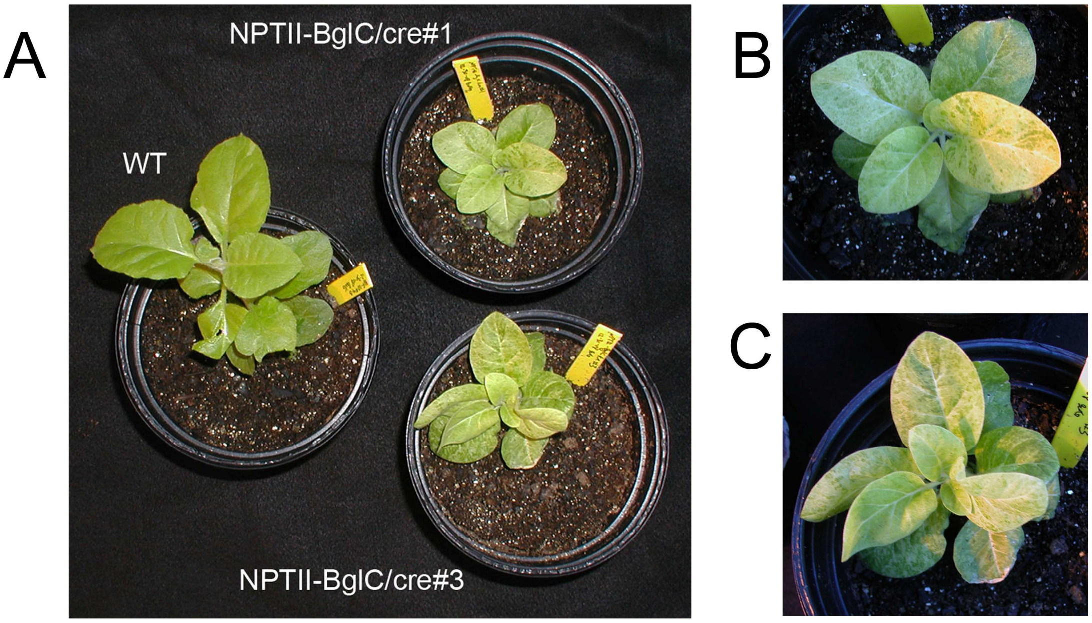
APPENDIX

Supplementary Table 3.S1. Primers used in this study

Primer Name	Primer Sequence
TetCBglC-fwd	CATATGGCTAGCAAAAATCTGGATTGTTGGGTCGACAATGAAGAAGATATAACCTCGCAATCGACGACT
NPTIIBglC-fwd	CATATGGCTAGCATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTACCTCGCAATCGACGACT
GFPBglC-fwd	CATATGGCTAGCGGCAAGGGCGAGGAACTGTTCACTGGCGTGGTCCCAATCACCTCGCAATCGACGACT
BglC-rev	ATGCGGCCGCTATTCCTGTCCGAAGAT
Iprobe-fwd	CACAGGTTTAGCAATGGG
Iprobe-rev	GAAGTAGTCAGATGCTTC
BglCint-fwd	TTCGTCCAGGACGGCGAC
Aprobe-fwd	ATAGTATCTTGTACCTGA
Aprobe-rev	TAAAGCTTTGTATCGGCTA
BglCint-fwd2	AAGGACAGCGGCTGGTGGT
BglCint-rev	GAGTCGTCGATTGCGAGGT
trnAint-rev	TCAGGTACAAGATACTAT
RecAsp-fwd	ATCTCGAGATGGATTCACAGCTAGTC
RecAsp-rev	ATTCTAGACATATGATCGAATTCAGAACTGAT
Cre-fwd	ATCATATGTCCAATTTACTGACC
Cre-rev	TCTAGACTAATCGCCATCTTCCAG
16s-rev	ATGTGTTAAGCATGCCGC
rps12-rev	CATTTATGAATTTCATAG
trnKint-fwd	AATCAACTGAGTATTCAA

Supplementary Table 3.S2. 5’ and 3’ termini of *db-bglC* transcripts relative to <u>A</u>TG (+1) and TA<u>G</u> (+1497) start and stop codons, respectively

TetC-BglC		**NPTII-BglC**		**GFP-BglC**	
5' terminus	3' terminus	5' terminus	3' terminus	5' terminus	3' terminus
-126	1647	-123	1649	-12	1674
-10	1599	-115	1647	-123	1649
-12	1598	-123	1646	-123	1647
-159	1597	-123	1646	-124	1644
-123	1594	-159	1645	-160	1594
-123	1591	-124	1644	-11	1570
-123	1589	-102	1635	-123	1538
-123	1578	-124	1600	-380	1535
15	1559	-123	1599	-380	1532
-124	1558	-12	1599	-123	1517
-10	1534	-123	1598	-123	1517
-123	1532	-123	1598	-159	1514
-124	1528	-65	1598	**45**	1513
-3	1525	-12	1598	-124	1509
-2	1525	-123	1597	-123	**1487**
-160	1502	-123	1595	-123	**1483**
-109	1500	-99	1590	-125	**1477**
-68	**1498**	-380	1589	-123	**1475**
-123	**1474**	-45	1583	-12	**1472**
-160	**1469**	-123	1577	-125	**1468**
-159	**1455**	-124	1562	-122	**1467**
		34	1559	-123	**1461**
		-124	1535	-153	**1460**
		-33	1500	-50	**1452**
		-32	1500		
		-124	**1487**		
		-124	**1459**		

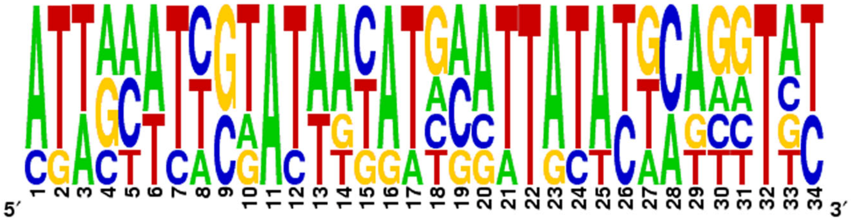


Supplementary Figure 3.S1: Photographs of the phenotype of NPTII-BglC/cre plants observed when grown in soil. (A) Photograph of WT, NPTII-BglC/cre#1, and NPTII-BglC/cre#3 tobacco grown in soil. (B) Close-up photograph of NPTII-BglC/cre#1 plant. (C) Close-up photograph of NPTII-BglC/cre#3 plant. Both NPTII-BglC/cre plants ultimately died on soil, though NPTII-BglC/cre#3 was rescued on tissue culture medium, where it regained a normal phenotype and appeared healthy.

A

```
loxP            ATAACTTCGTATAGCATACATTATACGAAGTTAT
NPTII           CTAGCATTGAACAAGATGGATTGCACGCAGGTTC
16S/trnV        ATTGAATCCGATTTTGACCATTATTTTCATATCC
psbA            ATTCAACAGTATAACATGACTTATATACTCGTGT
rps12           AGTATTTTCTATTATATTAGATATATTAGACTAT
                          *          *         *
Consensus       ATWnnWTnSnATWnnATnnnTTATAYnMnnnTnY
```

B



Supplementary Figure 3.S2: Sequence alignments of the genuine *loxP* sequence with *loxP*-like sequences observed to cause Cre-mediated recombination in transgenic plastids. (A) T-COFFEE alignment of the *loxP* sequence with the *loxP*-like sequences from the NPTII DB region (NPTII), 16s *rrn*/*trnV* intergenic region (16s/trnV), P*psbA* (psbA), and *3'rps12* ORF (rps12). Conserved bases are marked by an asterisk (*). The consensus sequence derived from this alignment is shown below the T-COFFEE output. (B) Weblogo output, showing the frequency of each nucleotide at each position. The most commonly observed nucleotide at each position is shown at the top of this diagram, with the letter size proportional the frequency that the nucleotide was observed at each position.

REFERENCES

Ayliffe MA, Timmis JN (1992) “Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA” *Theor Appl Genet* **85**: 229-238.

Barkan A (1988) “Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs” *EMBO J* **7**: 2637-2644.

Bollenbach TJ, Tatman DA, Stern DB (2003) “CSP41a, a multifunctional RNA-binding protein, initiates mRNA turnover in tobacco chloroplasts” *Plant J* **36**: 842-852.

Bollenbach TJ, Schuster G, Stern DB (2004) “Cooperation of endo- and exoribonucleases in chloroplast mRNA turnover” *Prog Nucleic Acid Res Mol Biol* **78**: 305-337.

Coburn GA, Mackie GA (1999) “Degradation of mRNA in *Escherichia coli*: an old problem with some new twists” *Prog Nucleic Acid Res Mol Biol* **62**: 55-108.

Corneille S, Lutz KA, Azhagiri AK, Maliga P (2003) “Identification of functional *lox* sites in the plastid genome” *Plant J* **35**: 753-762.

DeCosa B, Moar W, Lee SB, Miller M, Daniell H (2001) “Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals” *Nat Biotechnol* **19**: 71-74.

Gonzalez de Valdivia EI, Isaksson LA (2005) “Abortive translation caused by peptidyl-tRNA drop-off at NGG codons in the early coding region of mRNA” *FEBS J* **272**: 5306-5316.

Gray BN, Ahner BA, Hanson MR (2009a) “High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions” *Biotechnol Bioeng* **102**: 1045-1054.

Gray BN, Ahner BA, Hanson MR (2009b) “Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants” *Transgenic Res* doi:10.1007/s11248-009-9246-3.

Gutman GA, Hatfield GW (1989) “Nonrandom utilization of codon pairs in *Escherichia coli*” *Proc Natl Acad Sci USA* **86**: 3699-3703.

Hayes R, Kudla J, Schuster G, Gabay L, Maliga P, Gruissem W (1996) “Chloroplast mRNA 3’-end processing by a high molecular weight protein complex is regulated by nuclear encoded RNA binding proteins” *EMBO J* **15**: 1132-1141.

Hayes R, Kudla J, Gruissem W (1999) “Degrading chloroplast mRNA: the role of polyadenylation” *Trends Biochem Sci* **24**: 199-202.

Jakoby MJ, Weinl C, Pusch S, Kuijt SJ, Merkle T, Dissmeyer N, Schnittger A (2006) “Analysis of the subcellular localization, function, and proteolytic control of the Arabidopsis cyclin-dependent kinase inhibitor ICK1/KRP1” *Plant Physiol* **141**: 1293-1305.

Juhász T, Egyházi A, Réczey K (2005) “β-glucosidase production by *Trichoderma reesei*” *Appl Biochem Biotechnol* **121-124**: 243-254.

Klaff P (1995) “mRNA decay in spinach chloroplasts: *psbA* mRNA degradation is initiated by endonucleolytic cleavages within the coding region” *Nucleic Acids Res* **23**: 4885-4892.

Kohler RH, Cao J, Zipfel WR, Webb WW, Hanson MR (1997) “Exchange of protein molecules through connections between higher plant plastids” *Science* **276**: 2039-2042.

Kuroda H, Maliga P (2001a) “Complementarity of the 16S rRNA penultimate stem with sequences downstream of the AUG destabilizes the plastid mRNAs” *Nucleic Acids Res* **29**: 970-975.

Kuroda H, Maliga P (2001b) “Sequences downstream of the translation initiation codon are important determinants of translation efficiency in chloroplasts” *Plant Physiol* **125**: 430-436.

Lenzi P, Scotti N, Alagna F, Tornesello ML, Pompa A, Vitale A, De Stradis A, Monti L, Grillo S, Buonaguro FM, Maliga P, Cardi T (2008) “Translational fusion of chloroplast-expressed human papillomavirus type 16 L1 capsid protein enhances antigen accumulation in transplastomic tobacco” *Transgenic Res* **17**: 1091-1102.

Mackie GA (1989) “Stabilization of the 3’ one-third of *Escherichia coli* ribosomal protein S20 mRNA in mutants lacking polynucleotide phosphorylase” *J Bacteriol* **171**: 4112-4120.

Maliga P (2003) “Progress towards commercialization of plastid transformation technology” *Trends Biotechnol* **21**: 20-28.

McDaniel LD, Chester N, Watson M, Borowsky AD, Leder P, Schultz RA (2003) “Chromosome instability and tumor predisposition inversely correlate with BLM protein levels” *DNA Repair* **2**: 1387-1404.

Missirlis PI, Smailus DE, Holt RA (2006) “A high-throughput screen identifying sequence and promiscuity characteristics of the *loxP* spacer region in Cre-mediated recombination” *BMC Genomics* **7**: 73.

Monde RA, Schuster G, Stern DB (2000) “Processing and degradation of chloroplast mRNA” *Biochimie* **82**: 573-582.

Nilsson G, Belasco JG, Cohen SN, von Gabain A (1987) “Effect of premature termination of translation on mRNA stability depends on the site of ribosome release” *Proc Natl Acad Sci USA* **84**: 4890-4894.

Quesada-Vargas T, Ruiz ON, Daniell H (2005) "Characterization of heterologous multigene operons in transgenic chloroplasts. Transcription, processing, and translation" *Plant Physiol* **138**: 1746-1762.

Rapaport LR, Mackie GA (1994) "Influence of translational efficiency on the stability of the mRNA for ribosomal protein S20 in *Escherichia coli*" *J Bacteriol* **176**: 992-998.

Schein A, Sheffy-Levin S, Glaser F, Schuster G (2008) "The RNase E/G-type endoribonuclease of higher plants is located in the chloroplast and cleaves RNA similarly to the *E. coli* enzyme" *RNA* **14**: 1057-1068.

Schell DJ, Hinman ND, Wyman CE, Werdene PJ (1990) "Whole broth cellulase production for use in simultaneous saccharification and fermentation" *Appl Biochem Biotechnol* **24-25**: 287-297.

Spindler DD, Wyman CE, Grohmann K, Mohagheghi A (1989) "Simultaneous saccharification and fermentation of pretreated wheat straw to ethanol with selected yeast strains and β-glucosidase supplementation" *Appl Biochem Biotechnol* **20-21**: 529-540.

Spiridonov NA, Wilson DB (2001) "Cloning and biochemical characterization of BglC, a beta-glucosidase from the cellulolytic actinomycete *Thermobifida fusca*" *Curr Microbiol* **42**: 295-301.

Staub JM, Maliga P (1995) "Expression of a chimerical *uidA* gene indicates that polycistronic mRNAs are efficiently translated in tobacco plastids" *Plant J* **7**: 845-848.

Svab Z, Maliga P (1993) "High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene" *Proc Natl Acad Sci USA* **90**: 913-917.

Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) “Molecular farming in plants: host systems and expression technology” *Trends Biotechnol* **21**: 570-578.

Wei S, Marton I, Dekel M, Shalitin D, Lewinsohn E, Bravdo B-A, Shoseyov O (2004) “Manipulating volatile emission in tobacco leaves by expressing Aspergillus niger beta-glucosidase in different subcellular compartments” *Plant Biotechnol J* **2**: 341-350.

Ye GN, Hajdukiewicz PT, Broyles D, Rodriguez D, Xu CW, Nehra N, Staub JM (2001) “Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco” *Plant J* **25**: 261-270.

Zhang Y-H P, Lynd LR (2004) “Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems” *Biotechnol Bioeng* **88**: 797-824.

Zhou F, Karcher D, Bock R (2007) “Identification of a plastid intercistronic expression element (IEE) facilitating the expression of stable translatable monocistronic mRNAs from operons” *Plant J* **52**: 961-972.

Chapter 4: Codon optimization of the GFP downstream box region for improved cellulase expression in tobacco chloroplasts and in *E. coli*

ABSTRACT

The expression of foreign genes can be hampered by differences in codon usage between the source of the gene to be expressed and the expression host. Alteration of the coding region to include codons preferred by the expression host has been shown in a number of cases to improve protein expression levels. Studies have suggested that codon usage preferences may depend on the location within the ORF, such that a given codon could be detrimental for high-level protein production when placed at the 5' end of an ORF (i.e., in the downstream box region), but could have a positive effect on protein production when placed later in the ORF. I exploited these differential codon usage preferences to design codon-optimized downstream box regions derived from the *gfp* gene. These codon-optimized downstream box regions were fused to a cellulase gene for expression in *E. coli* and in tobacco chloroplasts. This strategy resulted in significant increases in cellulase accumulation in both protein production systems.

INTRODUCTION

Foreign protein expression in prokaryotic microorganisms such as *E. coli* has been exploited for the production of many valuable proteins. Transgenic plants can be used as alternate foreign protein production hosts, with the potential for significant production cost savings (Hood and Woodard 2002; Twyman et al 2003). In addition, scale-up of protein production in transgenic plants requires only the planting of more acreage, significantly easier and less expensive than the requirement for more and/or larger microbial fermentors for scale-up of protein production in a microbial system.

Foreign protein expression in transgenic plants can be accomplished by transformation of the nuclear genome via *Agrobacterium*-mediated transformation or by transformation of the plastid genome, typically by biolistic transformation (Svab and Maliga 1993) or by PEG-mediated transformation of protoplasts (Golds et al 1993). Transformation of the plastid genome has the advantage of the potential for extremely high yields of foreign protein (reviewed in Maliga 2003). In addition, many plastid regulatory sequences (e.g., promoters and 5’ and 3’ untranslated regions) are analogous to well-studied bacterial systems, allowing for comparisons between these systems.

One regulatory region that was originally discovered in *E. coli* is the downstream box (DB) region, defined by the 10-15 codons immediately downstream of the start codon (Sprengart et al 1996). The DB region was originally proposed to act by base-pairing with 16S ribosomal RNA in a manner analogous to the well-studied Shine-Dalgarno region located upstream of the start codon. Base-pairing of the DB region with ribosomal RNA has been ruled out by a number of structural and biochemical studies (e.g., O’Connor et al 1999; La Teana et al 2000; Moll et al 2001), but alterations in the DB region have been shown empirically to significantly affect foreign protein production in *E. coli*. The DB region has been shown to act in higher plant plastids in a manner similar to its action in *E. coli*. The DB region was first shown to be an important regulator of foreign protein accumulation in higher plant plastids by Kuroda and Maliga (2001a, 2001b), who showed that both silent and non-silent mutations in the DB region could result in changes in foreign protein concentration over more than two orders of magnitude. Ye et al (2001) fused the first 14 codons from *gfp* to the ORF encoding 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), resulting in more than a 30-fold increase in EPSPS accumulation relative to the expression of the *epsps* ORF with no DB fusion. More recently, the DB regions from the *tetc*, *neo*, and *gfp* genes were fused to the *T. fusca cel6A* (Gray et al 2009a) and the *T. fusca bglC* (Chapter 3) ORFs for expression in transplastomic tobacco, resulting in protein accumulation varying over more than two orders of magnitude in both cases.

Although DB base pairing with ribosomal RNA has been ruled out experimentally, the mechanism of DB function is still unknown. In tobacco plastids, the DB region has been implicated in regulating translation efficiency (Kuroda and Maliga 2001a). The mechanism of DB function seems to be dependent on the identity of the codons in the DB region, rather than on the identity of the individual nucleotides (Stenström and Isaksson 2002; Gonzalez de Valdivia and Isaksson 2004), though a recent study concluded that RNA secondary structure in the DB region was a more important determinant of protein accumulation than the codon content of the DB region (Kudla et al 2009). Bulmer (1988) observed that the DB regions of highly-expressed *E. coli* genes tended to have higher numbers of rare codons than poorly expressed genes, suggesting different selection pressures for codon usage in the DB region than in the rest of the ORF. The DB region is clearly context-specific, as illustrated in the work of Gray et al, who found that the TetC DB region mediated higher Cel6A accumulation than the NPTII DB region, while the NPTII DB region mediated higher BglC accumulation than the TetC DB region (Gray et al 2009a; Chapter 3). It appears that the DB region acts through multiple mechanisms, with the identity of the codons contained in the DB region and the extent of RNA secondary structure in the DB region being important determinants of DB function in specific contexts.

In the absence of a complete understanding of the mechanism of DB function, the DB region of the ORF of interest must be improved empirically through trial and error. While this method of DB alteration can result in high-level accumulation of the protein of interest in transplastomic plants (e.g., Ye et al 2001; Gray et al 2009a; Chapter 3), these alterations and the requirement to generate multiple lines of transplastomic tobacco to test the various DB fusions can be expensive and time consuming, and is not always successful in generating high-level protein accumulation. A less empirical method for DB optimization is therefore desirable. A method for optimizing the DB regions of foreign genes for expression in prokaryotic hosts is described here, based on the observation of biased codon usage such that some rare codons are overrepresented in the DB regions of highly expressed genes and can be beneficial in some cases when included in the DB region.

MATERIALS AND METHODS

Analysis of Codon Usage Frequencies

Tobacco chloroplast and *E. coli* codon usage frequencies (CUFs) were taken from the online Codon Usage Database (http://www.kazusa.or.jp/codon/). Downstream box CUFs were calculated in *E. coli* (EcDB CUFs) by counting codon occurrences in the first 14 codons of the *E. coli nusA*, *arcB*, *rpoD*, *torS*, *ligA*, *topA*, *fryA*, *mutY*, *recQ*, and *rpoS* ORFs. These ORFs were chosen because their encoded proteins were highly represented in the Integr8 Proteome Analysis database (http://www.ebi.ac.uk/integr8/). Tobacco chloroplast downstream box CUFs (NtDB CUFs) were calculated by counting codon occurrences in the first 14 codons of the *N. tabacum* chloroplast *rbcL, psaA, psaB, psaC, psbA, psbB, psbC, psbD, psbE,* and *psbF* ORFs.

Plasmid Construction

The EcTotGFP-Cel6A ORF was PCR amplified using primers EcTotGFP-Cel6A-fwd/Cel6A-rev (primer sequences are shown in supplementary Table 4.S2), with plasmid pGFPCel6A (Gray et al 2009a) as the DNA template. The EcDBGFP-Cel6A

ORF was PCR amplified using primers EcDBGFP-Cel6A-fwd/Cel6A-rev, with plasmid pGFPCel6A as the DNA template. Both of these PCR products were *Nhe*I/*Not*I digested and ligated into the *Nhe*I/*Not*I backbone of pGFPCel6AEC (Gray et al 2009a) to generate plasmids pGFPCel6AECTot and pGFPCel6AECDB, respectively. Plasmids pGFPCel6AEC, pGFPCel6AECTot, and pGFPCel6AECDB were maintained in DH5α (Invitrogen, Carlsbad, CA) *E. coli* cells, and were inserted into BL21(DE3) (Invitrogen) cells for protein production.

The strong ribosomal promoter (P*rrn*) was PCR amplified from WT tobacco DNA using primers Prrn-fwd/Prrn-rev. This PCR product was *Asc*I digested and ligated into the AscI backbone of pGFPCel6A (Gray et al 2009a) to generate plasmid pPrrn-GFPCel6A.

Plasmid pGFPCel6AECDB was *Nhe*I/*Not*I digested, and the EcDBGFP-Cel6A ORF was gel purified using the QiaQuick Gel Extraction Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The EcDBGFP-Cel6A ORF was then ligated into the *Nhe*I/*Not*I backbone of pGFPCel6A (Gray et al 2009a) to generate plasmid pcaGFPCel6A.

E. coli *Cell Growth and GFP-Cel6A Expression*

BL21(DE3) *E. coli* cells harboring the pGFPCel6AEC, pGFPCel6AECTot, and pGFPCel6AECDB plasmids were grown overnight at 37ºC in LB medium containing kanamycin (50 μg/mL). Overnight cell cultures were diluted by adding 500 μL of each overnight culture to 50 mL of fresh LB medium containing kanamycin (50 μg/mL). These 50 mL cultures were grown at 37ºC until they reached an OD600 of approximately 0.6. At this point, 50 μM IPTG was added to the cultures. The OD600 of each culture was measured at one hour intervals. Five hours after the addition of IPTG to the cultures, a 1 mL sample was taken from each culture. The cells were collected by centrifugation and spent LB medium was removed. Protein was collected from the cells for immunoblotting as described below.

Chloroplast Transformation

Tobacco (*Nicotiana tabacum* cv. Samsun) chloroplasts were transformed as described previously (Gray et al 2009a), using plasmids pPrrn-GFPCel6A and pcaGFPCel6A as transformation vectors. Transformed shoots were regenerated on RMOP medium containing spectinomycin (500 mg/L) and subjected to several rounds of tissue culture to generate fully transformed shoots containing the desired transgene insertion. These shoots were transferred to MS medium containing spectinomycin (500 mg/L) for rooting. Rooted plants were transferred to soil and grown in a greenhouse for seed collection.

DNA Extraction and DNA Blotting

DNA was extracted from tobacco leaves as described previously (Gray et al 2009a). Following extraction, DNA was thoroughly digested by *Xho*I and *Hind*III for DNA blotting. Digested DNA was electrophoresed in a 1% agarose gel, then transferred to a Hybond N+ nylon membrane (Amersham Biosciences, Piscataway, NJ). A PCR product from the plastid *trnA(UGC)* gene was generated from primers probe-fwd/ptDNA-rev using plasmid pGFPCel6A as a template. This PCR product was ^{32}P-labeled using the DecaPrime II Random Primed DNA Labeling kit (Ambion, Austin, TX) and used to probe the DNA blot. Following probing, the blot was washed and exposed to a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA) for visualization.

Protein Extraction and Immunoblotting

Protein was extracted from *E. coli* cells as described previously (Gray et al 2009a). Briefly, cell pellets were re-suspended in Tris (100 mM, pH 7.4) containing PMSF (1 mM). Cells were then lysed in Tris (100 mM, pH 7.4) containing SDS (1% w/v) and β-mercaptoethanol (0.01% v/v). Cell debris was removed by centrifugation and protein concentration was determined using the Bio-Rad protein assay (Bio-Rad, Hercules, CA) according to the manufacturer's instructions.

Protein was extracted from tobacco leaf tissue as described previously (Gray et al 2009a). Briefly, leaf samples were ground in liquid nitrogen. Protein extraction buffer containing Tris (20 mM, pH 7.4), Triton X-100 (1% v/v), SDS (0.1% w/v), PMSF (1 mM), and β-mercaptoethanol (0.01% v/v) was added to ground leaf tissue. Leaf tissue was pelleted by centrifugation and supernatant containing soluble protein was transferred to a fresh eppendorf tube. Protein concentration was determined using the Bio-Rad protein assay (Bio-Rad) according to the manufacturer's instructions.

Protein samples were electrophoresed in a 12% polyacrylamide gel, then transferred to a nitrocellulose membrane (Pierce, Rockford, IL). Membranes were incubated in 5% (w/v) milk in TBST as a blocking step, then incubated in anti-Cel6A antibody (kindly donated by David Wilson, Cornell University, Ithaca, NY) diluted 1:100,000 in 5% milk in TBST. Secondary antibody was horseradish peroxidase-conjugated anti-rabbit antibody (Sigma, St. Louis, MO). Immunoblots were incubated with SuperSignal West Dura Extended Duration Substrate (Pierce), then exposed to CL-Xposure film (Pierce). Bands were quantified by using Scion Image software (Scion Corporation, Frederick, MD).

RNA Extraction and RNA Blotting

RNA extraction and RNA blotting protocols were performed as described previously (Gray et al 2009a). RNA was hybridized with a radiolabeled PCR probe synthesized using primers C6probe-fwd and Cel6A-rev (*cel6A*) or aadA-fwd and aadA-rev (*aadA*). Following hybridization, the blot was washed and exposed to a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA) for visualization. Isotope was removed from the blot between hybridizations by exposure to a boiling solution of 0.1% (w/v) SDS.

RESULTS

Downstream Box CUFs Differ from Overall CUFs

Codon usage in the downstream box (DB) regions of highly-expressed *E. coli* and tobacco chloroplast genes was calculated by counting the occurrence of each codon in the DB region of ten highly-expressed genes. Because the DB region is likely to act at the level of translation, high expression was defined not by mRNA abundance, but by the abundance of the encoded protein. It was found that CUFs in the DB regions of genes encoding highly abundant *E. coli* proteins differed significantly from the CUFs of the *E. coli* genome as a whole (CUFs are shown in supplementary Table 4.S1 and supplementary Figure 4.S2A). A number of codons (e.g., CTG, AAA, and CAG) were overrepresented in the DB regions of highly expressed *E. coli* genes relative to their overall usage, while many other codons (e.g., GAT, GGC, CCG) were underrepresented in the same DB regions.

Similar to the results described above for *E. coli*, tobacco chloroplast codon usage in the DB regions of highly expressed genes differs from the overall codon usage in the tobacco chloroplast genome (CUFs are shown in supplementary Table 4.S1 and supplementary Figure 4.S2B). A number of codons were overrepresented (e.g., GAA, ATT, and TTT) or underrepresented (e.g., AAA, ATG, and CTT) in the DB regions of highly expressed genes relative to the chloroplast genome as a whole. As an example, the amino acid serine can be encoded by six different codons. In the tobacco plastid genome as a whole, UCU is the preferred codon, while AGC is the least-used serine codon. In the DB regions of highly expressed plastid genes, however, AGC is overrepresented, and is in fact the most highly used codon in these DB regions (Figure 4.1).

GFP-Cel6A Protein Accumulation in E. coli

In order to test whether the differences in codon usage frequencies between the DB regions of highly expressed genes and the coding regions of the genome as a whole could be exploited to optimize DB regions for high-level protein production, three constructs were prepared for GFP-Cel6A expression in BL21(DE3) *E. coli* cells. The three constructs contained an unmodified GFP DB region, a GFP DB region altered to match the codon usage of the *E. coli* genome as a whole (HF[EcTot]GFP), or a GFP DB region altered to match the codon usage of the DB regions of highly expressed *E. coli* genes (HF[EcDB]GFP). The three DB regions are shown in Table 4.1.

Five hours after IPTG induction of GFP-Cel6A expression, cells were collected and lysed for protein collection. Immunoblotting with these protein extracts revealed that GFP-Cel6A protein accumulated to higher levels in cells expressing the *HF(EcDB)gfp-cel6A* ORF than in cells expressing the *HF(EcTot)gfp-cel6A* ORF, which in turn accumulated to higher levels than in cells expressing the *gfp-cel6A* ORF (Figure 4.2). No GFP-Cel6A could be detected in the absence of IPTG induction (data not shown). Cell growth was unaffected by GFP-Cel6A expression as determined from the optical density of the cell cultures at 600 nm. GFP-Cel6A protein concentrations five hours after IPTG induction were approximately 20% of total

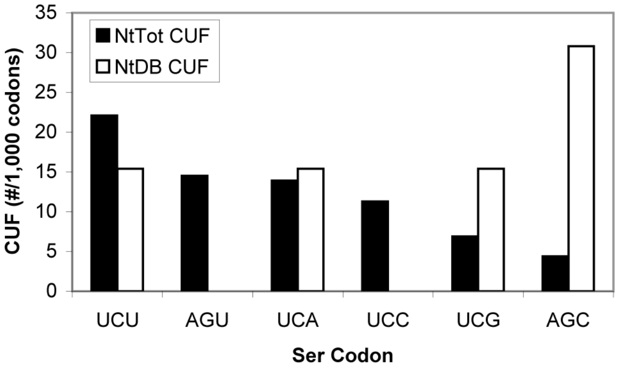


Figure 4.1: A comparison of serine codon usage in the tobacco chloroplast as a whole (black bars) and in the DB regions of highly-expressed tobacco chloroplast genes (white bars). Codon usage frequences in number of uses per 1,000 codons are shown on the y-axis.

Table 4.1. Sequences of modified GFP DB regions

DB Region	DB Sequence
GFP	GCT AGC GGC AAG GGC GAG GAA CTG TTC ACT GGC GTG GTC CCA ATC
HF(EcTot)GFP	GCT AGC GGC AAA GGC GAA GAA CTG TTT ACC GGC GTG GTG CCG ATT
HF(EcDB)GFP	GCT AGC GGT AAA GGT GAA GAA CTG TTT ACG GGT GTT GTT CCG ATT

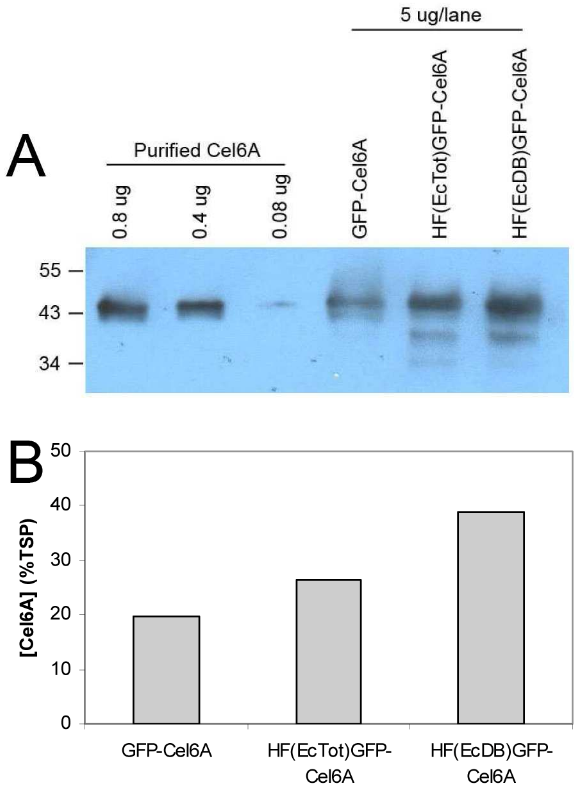


Figure 4.2: GFP-Cel6A protein accumulation in *E. coli*. (A) Immunoblot of intercellular protein from *E. coli* cells expressing the *gfp-cel6A*, *HF(EcTot)gfp-cel6A*, and *HF(EcDB)gfp-cel6A* ORFs. (B) Quantification of the immunoblot shown in (A), showing GFP-Cel6A protein concentration as a percentage of total soluble protein (%TSP).

soluble protein (%TSP), 27%TSP, and 39%TSP in *E. coli* expressing the *gfp-cel6A*, *HF(EcTot)gfp-cel6A*, and *HF(EcDB)gfp-cel6A* ORFs, respectively.

Tobacco Chloroplast Transformation

The tobacco chloroplast genome was transformed by biolistic bombardment with the plasmid vectors diagrammed in Figure 4.3. The plasmids pGFPCel6A (Gray et al 2009a), pPrrn-GFPCel6A, and pcaGFPCel6A all mediated the insertion of a *gfp-cel6A* ORF, regulated by the T7g10 5' untranslated region (5'UTR) and *psbA* 3'UTR and linked to an *aadA* expression cassette for selection on spectinomycin, between the plastid *trnI* and *trnA* genes. The pPrrn-GFPCel6A transformation vector includes a copy of the strong plastid ribosomal promoter (P*rrn*) upstream of the *gfp-cel6A* ORF. All three chloroplast transformation vectors contain *loxP* sites flanking the *aadA* expression cassette for future CRE-lox excision of the antibiotic resistance gene. The pGFPCel6A and pPrrn-GFPCel6A transformation vectors both contain the same *gfp-cel6A* ORF, with the unaltered GFP DB region contained in plasmid pGFPCel6AEC (Table 4.1). The pcaGFPCel6A transformation vector contains the HF(EcDB)GFP DB region included in the pGFPCel6AECDB plasmid (Table 4.1). These GFP DB regions encode the same amino acids, but the codon usage in the GFP DB region in pcaGFPCel6A more closely matches the codon usage in the DB regions of highly expressed plastid genes than the codon usage in the GFP DB region of pGFPCel6A (Table 4.1 and supplementary Table 4.S1).

Fully transformed GFP-Cel6A tobacco lines were generated previously (Gray et al 2009a). Both Prrn-GFP-Cel6A and caGFP-Cel6A tobacco lines were subjected to multiple rounds of tissue culture to create fully transformed tobacco lines, as confirmed by DNA blotting (Figure 4.4). Wild-type tobacco exhibited the expected 1.3 kb *Xho*I/*Hind*III restriction fragment, while transformed GFP-Cel6A and caGFP-

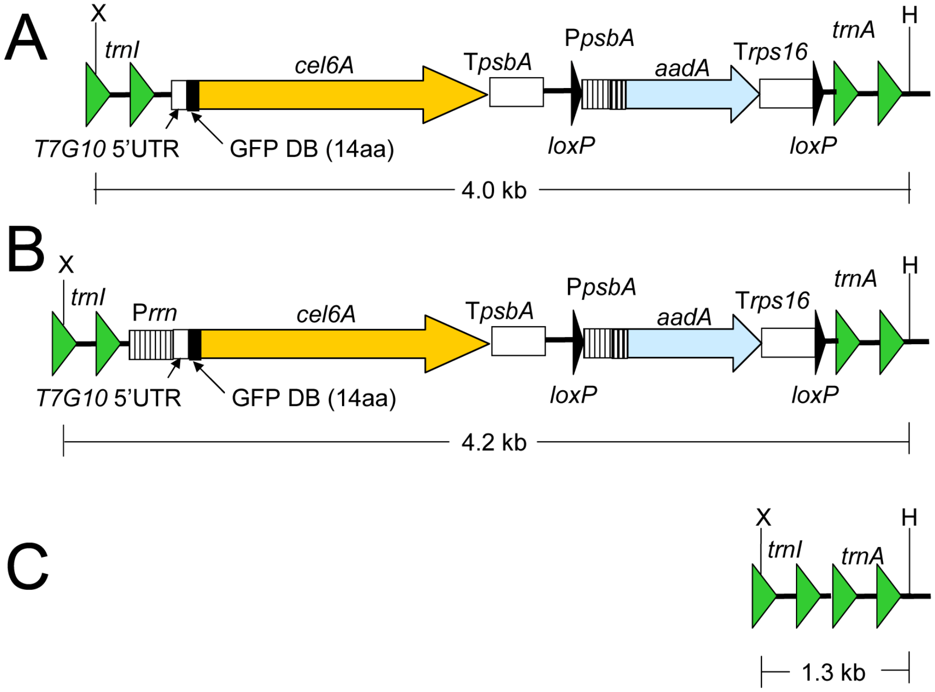


Figure 4.3: Schematic diagrams (not to scale) of pGFPCel6A, pPrrn-GFPCel6A, and pcaGFPCel6A plastid transformation vectors. (A) Diagram of pGFPCel6A and pcaGFPCel6A plastid transformation vectors. (B) Diagram of pPrrn-GFPCel6A transformation vector. (C) Diagram of *trnI/trnA* region of plastid DNA in the wild-type tobacco plastid genome. *Xho*I (X) and *Hind*III (H) sites relevant to DNA blotting experiments, along with predicted *Xho*I/*Hind*III fragment sizes, are shown.

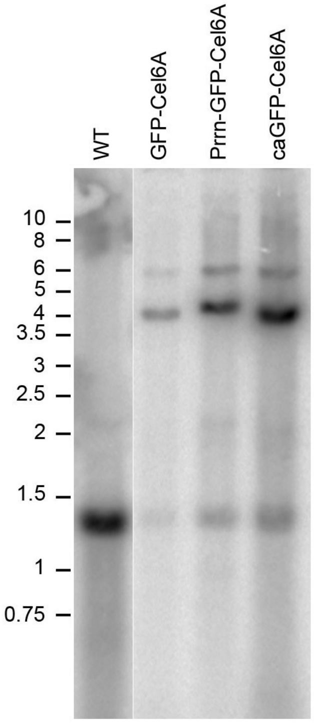


Figure 4.4: DNA blotting with DNA extracted from WT, GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco. The expected DNA fragment sizes are seen, demonstrating transformation of the *trnI/trnA* region of the plastid genome with the vectors diagrammed in Figure 4.3.

Cel6A tobacco lines exhibited a 4.0 kb fragment and Prrn-GFP-Cel6A tobacco exhibited a 4.2 kb fragment when hybridized with a *trnA*-specific probe. A minor restriction fragment at 5.6 kb in all three lines of transplastomic tobacco is the result of a homologous recombination event between the introduced and native copies of P*psbA* (Gray et al 2009b).

GFP-Cel6A Protein Accumulation in Transplastomic Tobacco

Differential GFP-Cel6A protein accumulation was observed among GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco, as revealed by immunoblotting with anti-Cel6A antibody (Figure 4.5). In GFP-Cel6A tobacco, GFP-Cel6A protein accumulated to approximately 0.01%TSP, slightly lower than observed previously (Gray et al 2009a). GFP-Cel6A accumulated to approximately 0.03%TSP in when the *gfp-cel6A* ORF was placed behind the P*rrn* promoter in Prrn-GFP-Cel6A tobacco, and to approximately 0.5%TSP in caGFP-Cel6A tobacco.

GFP-Cel6A Transcript Accumulation in Transplastomic Tobacco

In order to determine whether transcript abundance could partially explain the observed differences in protein abundance among the three transplastomic tobacco lines tested, RNA blotting was performed. Probing WT, GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco RNA for *cel6A* transcripts revealed that Prrn-GFP-Cel6A tobacco contained far more *cel6A* RNA in the form of both polycistronic and monocistronic transcripts than either GFP-Cel6A or caGFP-Cel6A tobacco (Figure 4.6A). The RNA blot shown in Figure 4.6A was stripped and re-probed with an *aadA*-specific probe, revealing that the vast majority of *aadA* transcript in all three transplastomic tobacco lines is monocistronic, consistent with its transcription from the P*psbA* promoter upstream of the *aadA* ORF (Figure 4.6B).

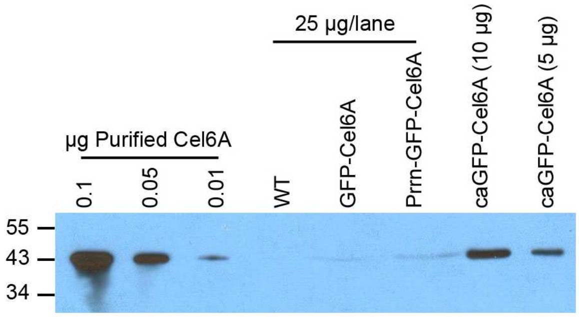


Figure 4.5: Immunoblotting of soluble protein extracted from GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco. Twenty-five µg of protein from WT, GFP-Cel6A and Prrn-GFP-Cel6A tobacco were loaded, while 10 µg and 5 µg of caGFP-Cel6A protein was loaded, as indicated by the lane labels.

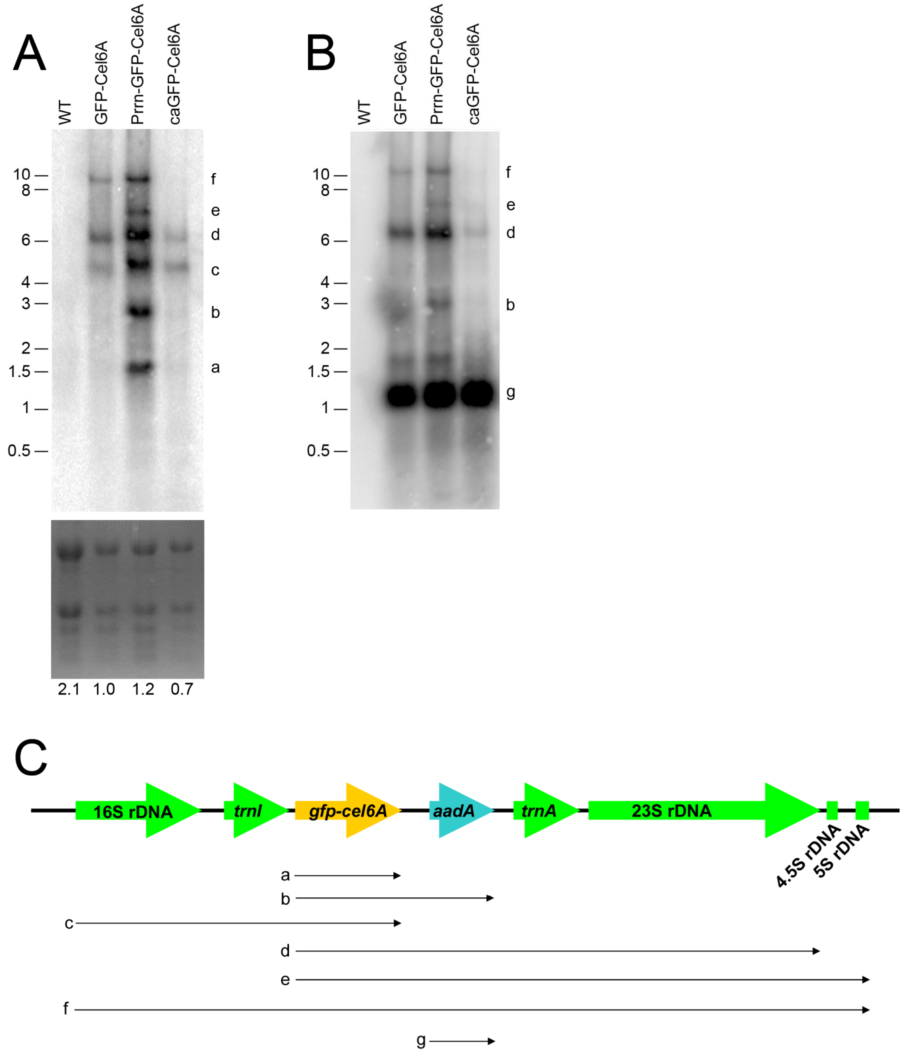


Figure 4.6: RNA blotting with RNA extracted from WT, GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco. (A) Blot probed with a *cel6A*-specific probe. Ethidium-bromide stained ribosomal RNA bands are shown below the blot, along with the relative loading in each lane. (B) The same blot shown in (A), hybridized with an *aadA*-specific probe. (C) Diagram (to scale) of transcripts observed in the RNA blots in (A) and (B). Lowercase letters a-g correspond to the band labels in (A) and (B).

The majority of *cel6A*-containing transcripts with an increased abundance in Prrn-GFP-Cel6A tobacco could have been transcribed from the P*rrn* promoter upstream of the *gfp-cel6A* ORF in this tobacco line (i.e., RNA species a, b, d, and e in Figure 4.6). Surprisingly, two polycistronic transcripts that were most likely transcribed from the native plastid P*rrn* promoter upstream of the 16S rDNA (i.e., RNA species c and f in Figure 4.6) were also more abundant in Prrn-GFP-Cel6A tobacco than in GFP-Cel6A or caGFP-Cel6A tobacco. The primary transcript produced from the native plastid P*rrn* promoter (RNA species f in Figure 4.6) is more abundant in GFP-Cel6A than in caGFP-Cel6A tobacco, though less abundant than in Prrn-GFP-Cel6A tobacco.

DISCUSSION

A novel method for the generation of optimized DB regions was tested in both *E. coli* and in tobacco chloroplasts. In both protein production systems, codon usage preferences were analyzed both for the genome as a whole and for the DB regions of highly expressed genes. A number of studies have reported that some codons are used more frequently than others to encode a given amino acid, and that these codon preferences are species-specific. Altering an ORF of interest to more closely match the codon usage preferences of the organism to be used for protein production has been shown in many cases to increase the accumulation of the protein of interest (e.g., Robinson et al 1984). While this type of optimization of codon usage has proven effective, codon usage gradients have also been reported, such that an amino acid could be preferentially encoded by one codon in most cases, but that a different codon could be preferred in the DB region, resulting in the inclusion of rare codons at the 5' ends (i.e., in the DB regions) of highly expressed ORFs in *E. coli* (Bulmer 1988). An analysis of ten highly expressed tobacco plastid genes found a similar phenomenon, as illustrated by the CUFs for serine codons (Figure 4.1). As shown in this figure, UCU is the most commonly used serine-encoding codon in the plastid genome as a whole, while AGC is the least commonly used serine-encoding codon in the plastid genome as a whole. In the DB regions of highly-expressed plastid genes, however, AGC is the most commonly used serine-encoding codon. Optimization of coding sequences to include plastid-preferred codons typically does not take into account this type of differential codon usage, with different preferred codons in the DB region than in the ORF as a whole.

In order to test the hypothesis that DB regions may prefer different codons than the genome as a whole, three GFP DB regions encoding the same amino acid sequence were constructed for GFP-Cel6A expression in *E. coli* (Table 4.1). The results of these experiments suggested that the inclusion of DB-preferred codons in the DB region of the ORF of interest, rather than codons preferred by the genome as a whole, can be beneficial for protein production in *E. coli*.

A recent study on GFP expression in *E. coli* concluded that RNA secondary structure in the GFP DB region was a more important factor in determining GFP accumulation than codon usage in the *gfp* ORF (Kudla et al 2009). While the three GFP DB regions that we tested in *E. coli* were altered to optimize the DB codon usage, and not the RNA secondary structure, there were differences in the predicted secondary structures of these three DB regions (supplementary Figure 4.S1). Predicted secondary structures of the GFP, HF(EcTot)GFP, and HF(EcDB)GFP DB regions did not correlate with GFP-Cel6A accumulation (ΔG = -11.4 kcal/mol, -10.4 kcal/mol, and -12.5 kcal/mol for GFP, HF[EcTot]GFP, and HF[EcDB]GFP, respectively). Kudla et al (2009) gave three examples of the DB regions of *gfp* ORFs giving high GFP accumulation and three examples of the DB regions of *gfp* ORFs resulting in low GFP accumulation. In the three GFP DB regions shown by Kudla et al to give low GFP accumulation in *E. coli*, several codons are used that are highly underrepresented in *E. coli* DB regions as compared with the genome as a whole (i.e., TTC, GGC, GGG, and GTG; supplementary Table 4.S1). I propose that, while secondary structure may play an important role in determining translation efficiency and hence in determining protein accumulation, the differential usage of codons in DB regions as compared with the genome as a whole may explain in part the results of Kudla et al (2009), where codon usage was not found to be a useful predictor of GFP accumulation. When codons that are often used, but are unfavorable in the context of a DB region, are included at the 5' end of an ORF, they may have a detrimental effect on translation efficiency. The GFP DB regions identified by Kudla et al (2009) to give high-level GFP accumulation in *E. coli* may be useful for driving high-level accumulation of other foreign proteins. Fusion of these DB regions to other ORFs of interest may be merited to test whether the sequences tested by Kudla et al act similarly when fused to other ORFs. It has been shown in tobacco plastids that a given DB region does not always give high-level protein accumulation, and that the optimal DB region may depend on the identity of the ORF to which it is fused (Gray et al 2009a; Chapter 3).

Based on the promising results in *E. coli*, the potential for improving GFP-Cel6A accumulation in tobacco plastids by altering the GFP DB region codon usage was tested. GFP-Cel6A accumulation in tobacco plastids was previously shown to be quite low relative to TetC-Cel6A or NPTII-Cel6A protein accumulation, and monocistronic *gfp-cel6A* RNA levels were also shown to be low in the GFP-Cel6A transplastomic tobacco line (Gray et al 2009a). Because the HF(EcDB)GFP DB region most closely matched the DB codon usage of highly expressed plastid genes, the HF(EcDB)GFP-Cel6A ORF was inserted into a plastid transformation vector to generate the caGFP-Cel6A line of transplastomic tobacco. Because the GFP DB codon usage was not fully optimized to match plastid codon usage, this construct is considered "codon-altered," rather than codon-optimized. As a second strategy to improve GFP-Cel6A protein accumulation in transplastomic tobacco, the strong plastid ribosomal promoter, P*rrn*, was inserted upstream of the *gfp-cel6A* ORF in the Prrn-GFP-Cel6A tobacco line (pcaGFPCel6A and pPrrn-GFPCel6A plastid transformation vectors are diagrammed in Figure 4.3). The inclusion of a P*rrn* sequence upstream of the *gfp-cel6A* ORF resulted in a dramatic increase in *gfp-cel6A* transcript abundance (Figure 4.6), but only a modest increase in GFP-Cel6A protein accumulation (Figure 4.5). In contrast, caGFP-Cel6A tobacco accumulated *gfp-cel6A* RNA to levels only slightly higher than GFP-Cel6A tobacco (Figure 4.6), but accumulated GFP-Cel6A protein to a concentration approximately 30-fold higher than GFP-Cel6A tobacco (Figure 4.5). The monocistronic transcript has been shown previously to be the most important *bglC* mRNA species for translation in tobacco plastids (Chapter 3), and RNA blotting experiments showed a correlation between Cel6A protein abundance and monocistronic *cel6A* transcript abundance (Gray et al 2009a), suggesting that monocistronic mRNA accumulation is important for Cel6A protein production in plastids. Monocistronic *gfp-cel6A* transcript concentration and GFP-Cel6A protein concentration were therefore quantified from RNA blots (Figure 4.6A) and from immunoblots (Figure 4.5). The relative abundances of *gfp-cel6A* monocistron and GFP-Cel6A protein are shown in Figure 4.7. This figure shows that there is an 8-fold increase in monocistronic *gfp-cel6A* transcript in Prrn-GFP-Cel6A tobacco relative to GFP-Cel6A tobacco, but only a 2-fold increase in protein abundance. In caGFP-Cel6A tobacco, the monocistronic *gfp-cel6A* transcript was only 2-fold more abundant than in GFP-Cel6A tobacco, but GFP-Cel6A protein accumulation was approximately 30-fold higher in caGFP-Cel6A tobacco than in GFP-Cel6A tobacco. These results suggest that, while the monocistronic *gfp-cel6A*

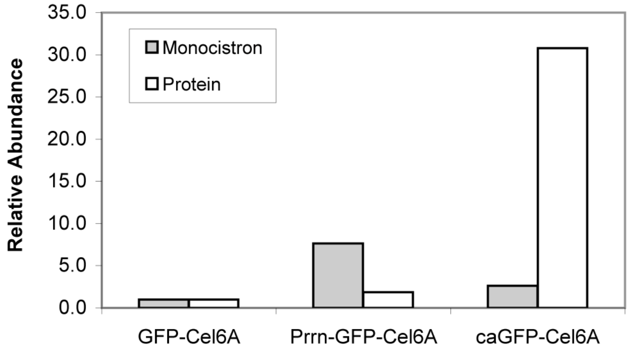


Figure 4.7: Comparison of monocistronic *gfp-cel6A* transcript abundance (gray bars) and GFP-Cel6A protein abundance (white bars) in GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco. The abundance of monocistron and protein abundance in GFP-Cel6A tobacco is defined as 1.0 in both cases. Monocistron accumulation in Prrn-GFP-Cel6A tobacco was 8-fold higher than in GFP-Cel6A tobacco, while protein accumulation in caGFP-Cel6A tobacco was 30-fold higher than in GFP-Cel6A tobacco.

transcript may be important for GFP-Cel6A protein production, RNA levels are not limiting for protein production. Instead, increasing translation efficiency through changes in codon usage in the GFP DB region is a far more effective way to improve GFP-Cel6A protein accumulation than increasing *gfp-cel6A* transcript levels. This suggests that translation efficiency is the limiting factor in high-level protein production in plastids, and that the DB region is an important regulator of translation efficiency.

Changes in the DB region of foreign genes expressed in tobacco plastids have been shown previously to affect translation efficiency, and therefore to affect foreign protein accumulation (e.g., Kuroda and Maliga 2001a, 2001b). Additionally, changes in the DB region of foreign genes expressed in tobacco plastids have been shown previously to affect transcript levels, with decreased RNA levels correlating with decreased protein levels (Kuroda and Maliga 2001b; Gray et al 2009a; Chapter 3). When the TetC, NPTII, and GFP DB regions were fused to the *T. fusca cel6A* ORF for tobacco plastid expression, GFP-Cel6A accumulated to one order of magnitude lower concentration than NPTII-Cel6A and two orders of magnitude lower concentration than TetC-Cel6A. Additionally, monocistronic *gfp-cel6A* transcript was markedly less abundant than *nptII-cel6A* or *tetC-cel6A* transcripts (Gray et al 2009a). It was hypothesized that the low level of *gfp-cel6A* monocistron could be limiting for GFP-Cel6A protein production, and that increasing the concentration of monocistronic *gfp-cel6A* transcript could increase GFP-Cel6A protein accumulation. The inclusion of a promoter upstream of the *gfp-cel6A* ORF, however, dramatically increased the concentration of both monocistronic and polycistronic *gfp-cel6A* transcripts, but had only a minor effect on GFP-Cel6A protein accumulation. This suggests that the low transcript levels of foreign genes whose protein products accumulate to low levels in transplastomic plants are not the cause of low protein levels, but are an effect of inefficient translation. Inefficient translation of the *bglC* ORF in transplastomic tobacco has been shown to cause partial RNA degradation, primarily in the 3’-5’ direction, though this degradation was not apparent by RNA blotting (Chapter 3). It is possible that some portion of the *gfp-cel6A* transcripts in Prrn-GFP-Cel6A tobacco detected by RNA blotting are also partially degraded, and thus untranslatable. When TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A were expressed in tobacco chloroplasts, amino acid changes at the N terminus of the Cel6A protein made it impossible to rule out differential protein degradation rates as the cause of differential protein accumulation (Gray et al 2009a). The use of silent mutations in the GFP DB region here eliminates the possibility of differential protein degradation rates resulting from changes to the N terminus of the protein, as the GFP-Cel6A polypeptides produced in GFP-Cel6A, Prrn-GFP-Cel6A, and caGFP-Cel6A tobacco are all identical at the amino acid level.

Silent mutations in the GFP DB region resulted in a 30-fold increase in GFP-Cel6A protein accumulation. While GFP-Cel6A protein accumulation was still modest, this 30-fold increase is significantly higher than the 1.5- to 2-fold increases in protein concentration typically seen after a GC-rich ORF is altered by the introduction of silent mutations to contain primarily plastid-preferred codons (e.g., Ye et al 2001; Tregoning et al 2003). This is particularly important when considering the time and labor required to synthesize a fully synthetic ORF. The *cel6A* ORF contains 411 codons. By making just 11 silent mutations to the GFP DB region, protein accumulation was increased 30-fold, with far less time and labor than would be required to synthesize a synthetic *cel6A* ORF containing plastid-preferred codons. The use of DB-preferred codons in the DB region, rather than codons preferred by the plastid genome as a whole (e.g., the use of AGC, rather than UCU, to encode serine in the DB region) could be important in generating large increases in protein accumulation in plastids. Standard methods of altering codon use of an ORF of interest to generate an ORF for expression in the host of choice do not take into account the differential codon usage preferences of the DB region. Optimization of the DB region in this way can allow for high-level expression of GC-rich genes such as the *T. fusca cel6A* and *bglC* genes from the AT-rich plastid genome, and could be more effective than standard codon optimization techniques.

The GFP DB region was chosen to be fused to the *cel6A* ORF because it had been shown previously to enhance the accumulation of EPSPS accumulation in transplastomic tobacco (Ye et al 2001). Fusion of the GFP DB region to the *cel6A* (Gray et al 2009a) or to the *bglC* (Chapter 3) ORFs, however, resulted in low-level protein accumulation. It is not clear why the GFP DB region worked well in improving EPSPS accumulation, but not Cel6A or BglC accumulation. These experiments demonstrate that, though the DB region has the potential to improve foreign protein accumulation in transplastomic plants, the mode of action of the DB region is context-dependent, and a given DB region can be effective when fused to one ORF, but not to another.

We report that DB optimization using codons used preferentially in the DB regions of highly expressed genes can lead to improved protein accumulation in both *E. coli* and in tobacco plastids. Though this type of differential codon preference has been described previously (Bulmer 1988), to our knowledge, this is the first report of the use of DB-preferred codons to stimulate foreign protein accumulation. The methods for DB region optimization described here can be used in chloroplasts with various DB fusions to ORFs of interest, or to optimize the codon usage within the native 5' region of the ORF of interest, to increase protein accumulation in *E. coli*, transplastomic tobacco, and likely in other prokaryotic protein production hosts.

ACKNOWLEDGEMENTS

The authors wish to acknowledge David Wilson for his kind donation of anti-Cel6A antibody. BNG was supported by an NSF Graduate Research Fellowship. This work was supported by a USDA grant (USDA NRI 2007-02133) to BAA and MRH.

APPENDIX

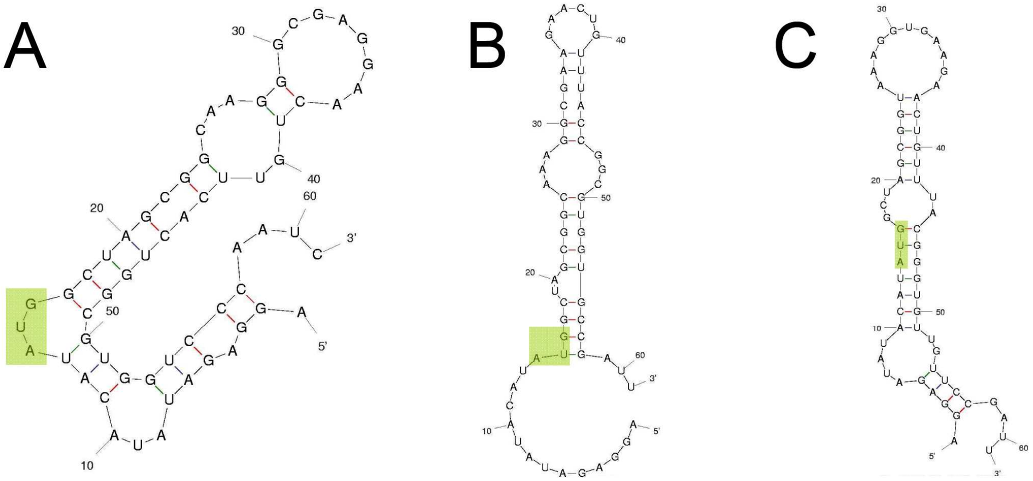
Supplementary Table 4.S1. Codon usage frequencies in the *E. coli* genome (EcTot), DB regions of highly expressed *E. coli* genes (EcDB), *N. tabacum* plastid genome (NtTot), and DB regions of highly expressed *N. tabacum* plastid genes (NtDB)

Codon	EcTot CUF	EcDB CUF	NtTot CUF	NtDB CUF
AAA	33.2	46.2	37.4	15.4
AAC	24.4	23.1	12.8	0.0
AAG	12.1	7.7	14.5	15.4
AAT	21.9	38.5	36.5	46.2
ACA	6.4	7.7	15.1	23.1
ACC	22.8	23.1	10.0	23.1
ACG	11.5	30.8	5.4	15.4
ACT	8.0	0.0	20.0	30.8
AGA	1.4	7.7	17.5	7.7
AGC	16.6	0.0	5.4	30.8
AGG	1.6	0.0	6.8	7.7
AGT	7.2	7.7	14.9	0.0
ATA	3.7	0.0	24.4	30.8
ATC	18.2	15.4	17.2	0.0
ATG	24.8	7.7	24.5	0.0
ATT	30.5	23.1	39.2	61.5
CAA	12.1	53.8	26.0	23.1
CAC	13.1	0.0	5.5	0.0
CAG	27.7	46.2	9.0	7.7
CAT	15.8	7.7	16.8	15.4
CCA	6.6	0.0	12.1	30.8
CCC	6.4	0.0	7.3	0.0
CCG	26.7	15.4	5.6	7.7
CCT	8.4	15.4	17.1	7.7
CGA	4.3	15.4	14.3	23.1
CGC	26.0	0.0	4.0	7.7
CGG	4.1	0.0	5.0	0.0
CGT	21.1	7.7	12.3	30.8
CTA	5.3	7.7	13.6	7.7
CTC	10.5	15.4	7.9	7.7
CTG	46.9	76.9	7.4	0.0
CTT	11.9	23.1	22.6	7.7
GAA	43.7	53.8	39.6	53.8
GAC	20.5	15.4	8.6	7.7

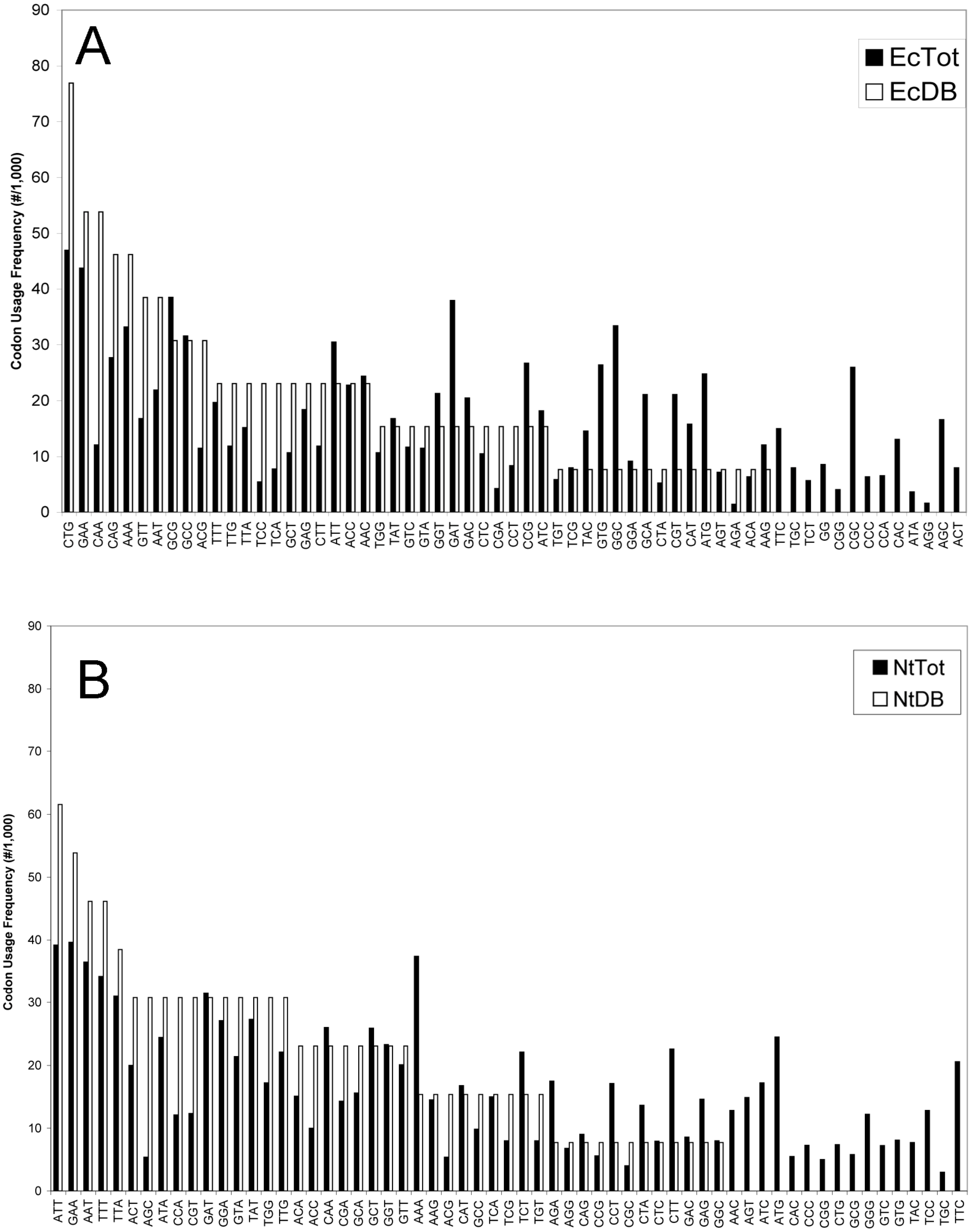
GAG	18.4	23.1	14.6	7.7
GAT	37.9	15.4	31.5	30.8
GCA	21.1	7.7	15.6	23.1
GCC	31.6	30.8	9.8	15.4
GCG	38.5	30.8	5.8	0.0
GCT	10.7	23.1	25.9	23.1
GGA	9.2	7.7	27.1	30.8
GGC	33.4	7.7	8.0	7.7
GGG	8.6	0.0	12.2	0.0
GGT	21.3	15.4	23.3	23.1
GTA	11.5	15.4	21.4	30.8
GTC	11.7	15.4	7.2	0.0
GTG	26.4	7.7	8.1	0.0
GTT	16.8	38.5	20.1	23.1
TAC	14.6	7.7	7.7	0.0
TAT	16.8	15.4	27.3	30.8
TCA	7.8	23.1	15.0	15.4
TCC	5.5	23.1	12.8	0.0
TCG	8.0	7.7	8.0	15.4
TCT	5.7	0.0	22.1	15.4
TGC	8.0	0.0	3.0	0.0
TGG	10.7	15.4	17.2	30.8
TGT	5.9	7.7	8.0	15.4
TTA	15.2	23.1	31.0	38.5
TTC	15.0	0.0	20.6	0.0
TTG	11.9	23.1	22.1	30.8
TTT	19.7	23.1	34.2	46.2

Supplementary Table 4.S2. Primers used in this study

Primer Name	Sequence
EcTotGFP-Cel6A-fwd	ATCATATGGCTAGCGGCAAAGGCGAAGAACTGTTTACCGGCGTGGTGCCGATTAATGATTCTCCGTTCTAC
EcDBGFP-Cel6A-fwd	ATCATATGGCTAGCGGTAAAGGTGAAGAACTGTTTACGGGTGTTGTTCCGATTAATGATTCTCCGTTCTAC
Cel6A-rev	ATAGACTAGGCCAGGATCGCGGCCGCTCAGCTGGCGGCGCAGGT
Prrn-fwd	ATGGCGCGCCGCTCCCCCGCCGTCGTTC
Prrn-rev	ATGGCGCGCCAAATCCCTCCCTACAACT
probe-fwd	ATAGTATCTTGTACCTGA
ptDNA-rev	TAAAGCTTTGTATCGGCTA
C6probe-fwd	GTAACGAGTGGTGCGACC
aadA-fwd	CGTGAAGCGGTTATCGCC
aadA-rev	GTCCAAGATAAGCCTGTC



Supplementary Figure 4.S1: Predicted secondary structures of GFP, HF(EcTot)GFP, and HF(EcDB)GFP/caGFP DB regions. (A) GFP DB region. (B) HF(EcTot)GFP DB region. (C) HF(EcDB)GFP/caGFP DB region. The start codon is indicated in each figure by a green box.



Supplementary Figure 4.S2: Codon usage frequencies of *E. coli* and *N. tabacum* plastid genomes. (A) EcTot (black bars) and EcDB (white bars) CUFs. (B) NtTot (black bars) and NtDB (white bars) CUFs.

REFERENCES

Bulmer M (1988) “Codon usage and intragenic position” *J Theor Biol* **133**: 67-71.

Golds T, Maliga P, Koop H-U (1993) “Stable plastid transformation in PEG-treated protoplasts of *Nicotiana tabacum*” *Bio/Technology* **11**: 95-97.

Gonzalez de Valdivia EI, Isaksson LA (2004) “A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*” *Nucleic Acids Res* **32**: 5198-5205.

Gray BN, Ahner BA, Hanson MR (2009a) “High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions,” *Biotechnology and Bioengineering* **102**: 1045-1054.

Gray BN, Ahner BA, Hanson MR (2009b) “Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants,” *Transgenic Research* in press.

Hood EE, Woodard SL (2002) “Industrial proteins produced from transgenic plants” in *Plants as Factories for Protein Production* (Hood EE and Howard JA, eds.), Kluwer Academic Publishers.

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) “Coding-sequence determinants of gene expression in *Escherichia coli*” *Science* **324**: 255-258.

Kuroda H, Maliga P (2001a) “Sequences downstream of the translation initiation codon are important determinants of translation efficiency in chloroplasts” *Plant Physiol* **125**: 430-436.

Kuroda H, Maliga P (2001b) “Complementarity of the 16S rRNA penultimate stem with sequences downstream of the AUG destabilizes the plastid mRNAs” *Nucleic Acids Res* **29**: 970-975.

La Teana A, Brandi A, O’Connor M, Freddi S, Pon CL (2000) “Translation during cold adaptation does not involve mRNA-rRNA base pairing through the downstream box” *RNA* **6**: 1393-1402.

Maliga P (2003) “Progress towards commercialization of plastid transformation technology” *Trends Biotechnol* **21**: 20-28.

Moll I, Huber M, Grill S, Sairafi P, Mueller F, Brimacombe R, Londei P, Bläsi U (2001) “Evidence against an interaction between the mRNA downstream box and 16S rRNA in translation initiation” *J Bacteriol* **183**: 3499-3505.

O’Connor M, Asai T, Squires CL, Dahlberg AE (1999) “Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA” *Proc Natl Acad Sci USA* **96**: 8973-8978.

Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G (1984) “Codon usage can affect efficiency of translation of genes in *Escherichia coli*” *Nucleic Acids Res* **12**: 6663-6671.

Sprengart ML, Fuchs E, Porter AG (1996) “The downstream box: an efficient and independent translation initiation signal in Escherichia coli” *EMBO J* **15**: 665-674.

Stenström CM, Isaksson (2002) “Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3’ side” *Gene* **288**: 1-8.

Svab Z, Maliga P (1993) “High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene” *Proc Natl Acad Sci USA* **90**: 913-917.

Tregoning JS, Nixon P, Kuroda H, Svab Z, Clare S, Bowe F, Fairweather N, Ytterberg J, van Wijk KJ, Dougan G, Maliga P (2003) “Expression of tetanus toxin Fragment C in tobacco chloroplasts” *Nucleic Acids Res* **31**: 1174-1179.

Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) “Molecular farming in plants: host systems and expression technology” *Trends in Biotechnol* **21**: 570-578.

Ye GN, Hajdukiewicz PT, Broyles D, Rodriguez D, Xu CW, Nehra N, Staub JM (2001) “Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco” *Plant J* **25**: 261-270.

Chapter 5: Design of Codon-Optimized Synthetic Downstream Boxes for Foreign Protein Production in Tobacco Chloroplasts

ABSTRACT

Plastid transformation of higher plants for high-level expression of valuable proteins *in planta* is an attractive strategy to lower protein production costs. The downstream box (DB) region, located immediately downstream of the start codon of the ORF of interest, is important for protein production in chloroplasts, but a lack of a mechanistic understanding of this region necessitates empirical, costly and time consuming trial-and-error optimization of the DB region. A less empirical method of DB optimization is desirable. One such method is described here, based on an analysis of codon usage in the DB regions of highly expressed plastid genes as compared with the codon usage of the plastid genome as a whole. Synthetic DB regions utilizing high-frequency or low-frequency DB codons were constructed. It was hypothesized that high-frequency codons would lead to high-level protein production, but surprisingly, low-frequency codons were more effective than high-frequency codons in at least one case.

INTRODUCTION

Expression of foreign proteins *in planta* is desirable due to the potential for significant projected production cost savings and ease of scale-up relative to the more well-established microbial, Chinese hamster ovary (CHO), and insect cell culture protein production systems (Hood and Woodard 2002; Twyman et al 2003). A number of proteins have been expressed *in planta* from both the nuclear and from the plastid genome, and in all cases expression from the plastid genome has resulted in significantly higher protein accumulation than expression from the nuclear genome (reviewed in Chapter 1). Plastid expression of foreign proteins has resulted in extremely high foreign protein accumulation, i.e., 30% of total soluble protein (%TSP) or higher, in several cases (De Cosa et al 2001; Oey et al 2009a, 2009b), and there are many reports of foreign protein yields of 5-25%TSP from plastid transformants (reviewed in Maliga 2003). Hence, plastid transformation is an attractive method for foreign protein production *in planta* due to the potential for high protein yields.

An important determinant of foreign protein accumulation in plastid transformants is translation efficiency, i.e., the rate of protein production from a given mRNA molecule. Translation efficiency in plastids is limited by translation initiation, which is controlled primarily by the 5' untranslated region (5'UTR) immediately upstream of the start codon and the downstream box (DB) region immediately downstream of the start codon (Chapter 1). Alterations in the DB region have resulted in order of magnitude differences in protein accumulation for several proteins expressed in transplastomic tobacco, including EPSPS (Ye et al 2001), Cel6A (Gray et al 2009a), and BglC (Chapter 3). In all of these cases, DB regions were fused to the 5' end of the ORF of interest through an empirical, trial-and-error optimization process. A recent study by Gray et al (Chapter 4) examined the differential codon usage in the DB regions of highly expressed plastid and *E. coli* genes, and described a method for optimization of DB regions by the inclusion of codons used preferentially at the 5' ends of highly expressed ORFs. By applying this DB optimization technique to the GFP DB region fused to the *cel6A* ORF, significantly increased GFP-Cel6A protein accumulation was achieved in both *E. coli* and in transplastomic tobacco. Notably, a 30-fold increase in GFP-Cel6A protein accumulation in transplastomic tobacco was achieved by Gray et al (Chapter 4) through silent mutations in the DB region. This is significantly greater than the 1.5- to 2-fold increases in foreign protein accumulation resulting from codon optimization of entire ORFs to include primarily plastid-preferred codons (e.g., Ye et al 2001; Tregoning et al 2003). Differential codon usage preferences in the DB region and in the remainder of the ORF, such that some rare codons may be preferred in the DB region, though they may be detrimental for translation when used later in the ORF, could explain the tremendous increase in GFP-Cel6A protein accumulation observed by Gray et al (Chapter 4). This type of differential codon usage has been described previously in *E. coli*, with the conclusion being reached that highly expressed genes have an especially pronounced codon usage bias at the 5' end of the ORF (e.g., Bulmer 1988). While some researchers have concluded that nucleotides, rather than codons, appear to be selected for in the DB region (e.g., Fuglsang 2004), others have concluded just the opposite (e.g., Stenström and Isaksson 2002; Gonzalez de Valdivia and Isaksson 2004). A recent study in *E. coli* found that expression levels of GFP protein depended in large part on mRNA secondary structure in the DB region (Kudla 2009). These studies suggest that the DB region may act through several different mechanisms, and that a number of parameters (e.g., codon usage and mRNA secondary structure) may be important in determining the effectiveness of a given DB region for high-level foreign protein production.

Previous studies on the use of DB fusions to affect foreign protein production in plastids have relied on empirical optimization of the DB region for the ORF of interest (e.g., Ye et al 2001; Gray et al 2009a; Chapter 3). While these studies have resulted in high-level accumulation of the foreign protein of interest, other studies of DB fusions (e.g., Lenzi et al 2008) have been less successful, resulting in only moderate accumulation of the foreign protein of interest. While trial-and-error optimization of DB regions can result in major increases in foreign protein accumulation, this process can be costly and time-consuming, and is not always successful. A less empirical method of DB optimization is therefore desirable. This report describes a method for the generation of synthetic DB regions based on an analysis of codon usage in tobacco plastids.

MATERIALS AND METHODS

Analysis of Tobacco Chloroplast Downstream Box Codon Usage and Generation of Synthetic Downstream Box Regions

Codon usage frequencies (CUFs) were analyzed as described previously (Chapter 4). Briefly, overall CUFs were taken from the online Codon Usage Database (http://www.kazusa.or.jp/codon/). The DB CUFs of the *N. tabacum* plastid *rbcL*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, and *psbF* ORFs (NtDB CUFs) were calculated by counting the occurrences of each codon in the first 14 codons of these ORFs. Overall plastid codon usage frequencies and NtDB CUFs are shown in supplementary Table 5.S1.

A synthetic DB region containing high-frequency codons (the HF[NtDB]Syn DB region) was generated by assigning a number from 1-14 to each of the 14 codons with the highest NtDB CUF. A random number generator was then used to generate a random number between 1 and 14 for each of the 14 positions in the synthetic DB. A synthetic DB region containing low-frequency codons (the LF[NtDB]Syn DB region) was generated by maintaining the amino acid sequence encoded by the HF(NtDB)Syn DB region, but substituting the codon with the lowest NtDB CUF at each position. The HF(NtDB)Syn and LF(NtDB)Syn DB sequences are shown in Table 5.1.

Plasmid Construction

The *HF(NtDB)Syn-bglC* and *LF(NtDB)Syn-bglC* ORFs were PCR amplified from the pNS6 plasmid (Spiridonov and Wilson 2001) using primer pairs HFSynBglC-fwd/BglC-rev and LFSynBglC-fwd/BglC-rev (primer sequences are shown in supplementary Table 5.S2), respectively. The *HF(NtDB)Syn-cel6A* and *LF(NtDB)Syn-cel6A* ORFs were PCR amplified from the pGFPCel6A plasmid (Gray et al 2009a)

Table 5.1. Sequences of HF(NtDB)Syn and LF(NtDB)Syn DB regions

DB	Sequence
HF(NtDB)Syn	GCT AGC ATA AAT CCA TAT GTA CGT TTT TGG GGA CCA AAT ATT TTA
LF(NtDB)Syn	GCT AGC ATC AAC CCC TAC GTC CGG TTC TGG GGG CCC AAC ATC CTG

using primer pairs HFSynCel6A-fwd/Cel6A-rev and LFSynCel6A-fwd/Cel6A-rev, respectively. These PCR products were *Nhe*I/*Not*I digested and ligated into the *Nhe*I/*Not*I backbone of the pGFPCel6A plasmid (Gray et al 2009a). The resulting plasmids were pHF(NtDB)Syn-BglC, pLF(NtDB)Syn-BglC, pHF(NtDB)Syn-Cel6A, and pLF(NtDB)Syn-Cel6A, respectively.

Chloroplast Transformation

Tobacco seedlings were bombarded as described previously (Gray et al 2009a) with the plastid transformation vectors described above. Following bombardment with the appropriate vector, leaf pieces were transferred to RMOP medium containing 500 mg/L spectinomycin. Spectinomycin-resistant shoots were screened by PCR for the presence of the appropriate *bglC* or *cel6A* ORF at the expected location of the plastid genome and subjected to several additional rounds of regeneration to generate homoplasmic plants. For regeneration of putative HF(NtDB)Syn-Cel6A shoots, 50 mg/L silver nitrate (Purnhauser et al 1987) was added to the RMOP medium along with 500 mg/L spectinomycin.

DNA Extraction and Blotting

DNA was extracted from tobacco leaves as described previously (Gray et al 2009a). Following *Xho*I/*Hind*III digestion, DNA was electrophoresed in a 1% (w/v) agarose gel, transferred to a Hybond N+ nylon membrane (Amersham Biosciences, Piscataway, NJ), and hybridized with a ^{32}P-labeled probe. The probe was synthesized using a PCR product generated from primer pair Aprobe-fwd/Aprobe-rev, using WT tobacco DNA as a template, and radiolabeled with the DecaPrime II Random Primed DNA Labeling Kit (Ambion, Austin, TX) according to the manufacturer's instructions. Following hybridization, the blot was washed and exposed to a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA) for visualization.

Protein Extraction and Immunoblotting

Protein was extracted from tobacco leaves as described previously (Gray et al 2009a). Immunoblotting of LF(NtDB)Syn-Cel6A protein was performed as described previously (Gray et al 2009a), with 1:100,000 diluted anti-Cel6A antibody (kindly donated by David Wilson, Cornell University, Ithaca, NY) as primary antibody and 1:25,000 diluted HRP-labeled anti-rabbit antibody (Sigma, St. Louis, MO) as secondary antibody. Immunoblotting of HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC proteins was performed using essentially the same protocol, except that the primary antibody was 1:250 diluted anti-BglC antibody (kindly donated by David Wilson, Cornell University, Ithaca, NY).

RNA Extraction and Blotting

Total RNA was extracted from tobacco leaves as described previously (Gray et al 2009a). RNA was electrophoresed in a 1% (w/v) agarose gel, transferred to a Hybond N+ membrane (Amersham Biosciences), and hybridized with radiolabeled PCR probes generated using primer pairs Cel6Aint-fwd/Cel6A-rev (*cel6A*) or BglCint-fwd/BglC-rev (*bglC*). Following hybridization, RNA blots were washed and exposed to a Phosphorimager screen (Molecular Dynamics) for visualization.

RNA Secondary Structure Predictions

RNA secondary structures were predicted using the m-fold online tool (Zuker 2003; http://mfold.bioinfo.rpi.edu/). RNA secondary structures were determined for the isolated HF(NtDB)Syn and LF(NtDB)Syn DB regions (the region from -13, beginning with the Shine-Dalgarno region of the T7g10 5'UTR, to +48, relative to the AUG start codon at +1), as well as for the 5' region of the likely *HF(NtDB)Syn-bglC*, *LF(NtDB)Syn-bglC*, *HF(NtDB)Syn-cel6A*, and *LF(NtDB)Syn-cel6A* monocistronic transcripts (the regions from -124 to +60, relative to the AUG start codon at +1).

RESULTS

Tobacco Chloroplast Transformation

Chloroplast transformation and regeneration of plants transformed with the HF(NtDB)Syn-BglC, LF(NtDB)Syn-BglC, and LF(NtDB)Syn-Cel6A vectors (diagrammed in Figure 5.1) was performed as described previously (Gray et al 2009a), using RMOP medium containing spectinomycin (500 mg/L) to regenerate transformed shoots. Transformation was verified by DNA blotting with a *trnA*-specific probe after *Xho*I/*Hind*III digestion (Figure 5.2a). Wild-type tobacco gave the expected 1.3 kb signal. In the *bglC*-containing plants, the expected 3.6 kb band was observed; the expected 4.0 kb band was observed in LF(NtDB)Syn-Cel6A plants.

Transformation and regeneration of HF(NtDB)Syn-Cel6A tobacco was problematic. Most putative transformants regenerated on RMOP containing spectinomycin (500 mg/L) were revealed to contain only the WT plastid genome, presumably acquiring spectinomycin resistance from a point mutation in ribosomal RNA (Svab and Maliga 1991), or to contain a DNA species formed by recombination between the native and introduced copies of P*psbA*, rather than between the *trnI* genes in the transformation vector and in the plastid genome (Gray et al 2009b). A small number of putative transformants (i.e., green shoots on RMOP medium with 500 mg/L spectinomycin) were observed that grew only for a brief period, then stalled their growth (Figure 5.3a). One of these stalled shoots (HF[NtDB]Syn-Cel6A#5) was able to be regenerated by the addition of silver nitrate (50 mg/L) to the

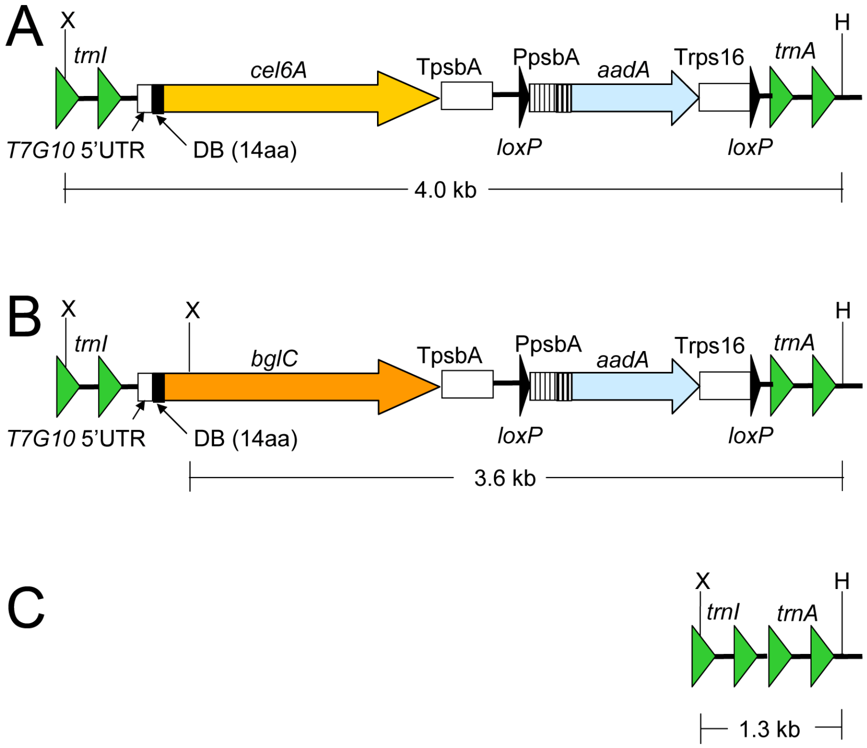


Figure 5.1: Schematic diagrams (not to scale) of plastid transformation vectors and the *trnI/trnA* region of the wild-type plastid genome. (A) Schematic diagram of HF(NtDB)Syn-Cel6A and LF(NtDB)Syn-Cel6A plastid transformation vectors. (B) Schematic diagram of HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC plastid transformation vectors. (C) Wild-type plastid genome *trnI/trnA* region. *Xho*I (X) and *Hind*III (H) sites relevant to DNA blotting experiments are shown.

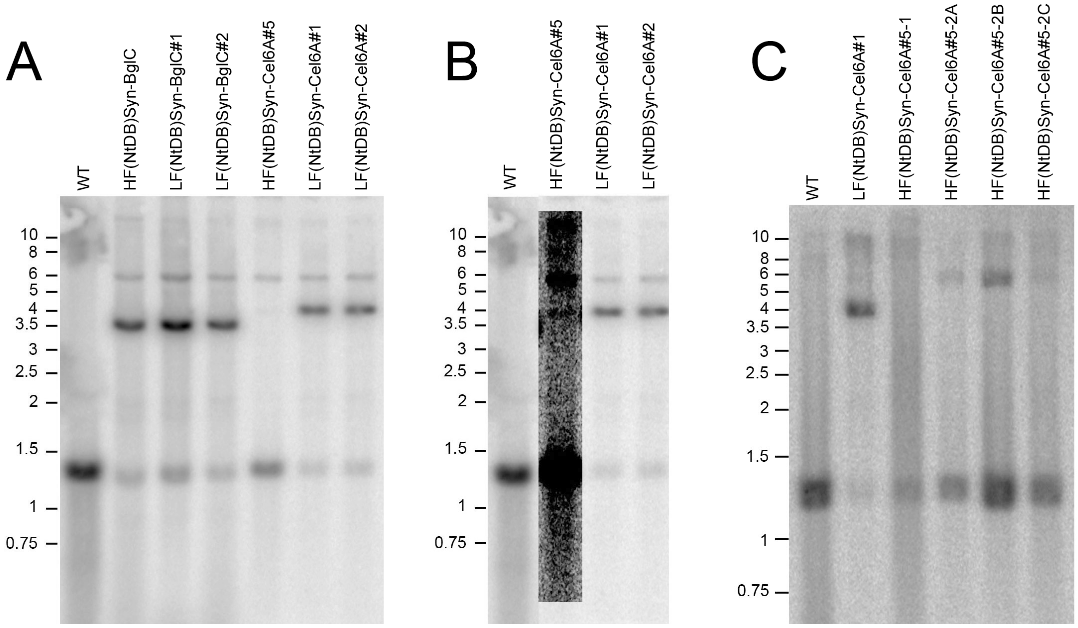


Figure 5.2: DNA blotting experiments with a *trnA*-specific probe after *Xho*I/*Hind*III digestion. (A) DNA extracted from wild-type (WT), HF(NtDB)Syn-BglC, LF(NtDB)Syn-BglC, LF(NtDB)Syn-Cel6A, and a putative HF(NtDB)Syn-Cel6A plant. #1 or #2 after the name (e.g., LF[NtDB]Syn-Cel6A#1) indicates an independent transformation event. (B) The same DNA blot shown in (A), with the contrast and brightness heavily adjusted in the HF(NtDB)Syn-Cel6A#5 lane to show the faint 4.0 kb band. (C) DNA extracted from second round regenerants of HF(NtDB)Syn-Cel6A#5 and from LF(NtDB)Syn-Cel6A tobacco. The number after HF(NtDB)Syn-Cel6A indicates the round of regeneration. Each shoot was lettered for identification purposes. HF(NtDB)Syn-Cel6A#5-2A and HF(NtDB)Syn-Cel6A#5-2B are two shoots regenerated from the HF(NtDB)Syn-Cel6A#5 shoot used for DNA blotting in (A) and (B).

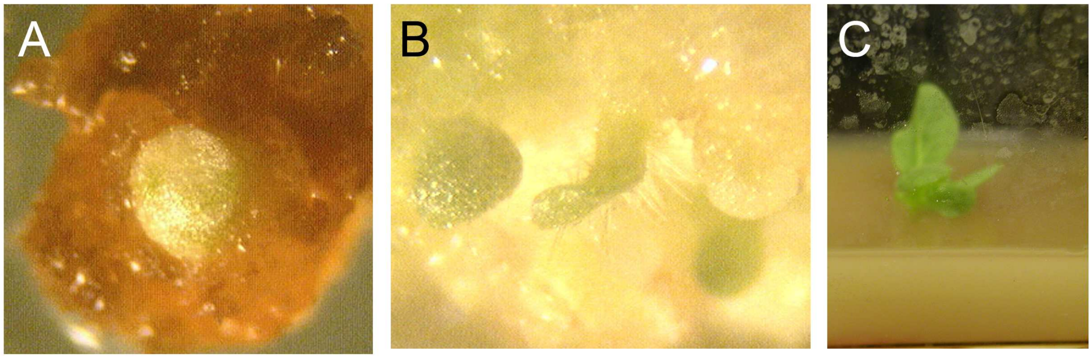


Figure 5.3: Photographs of HF(NtDB)Syn-Cel6A#5 tissue culture regeneration. (A) 15X magnification of stalled shoot on RMOP with 500 mg/L spectinomycin, approximately 10 weeks after bombardment with pHF(NtDB)Syn-Cel6A vector. (B) 15X magnification of HF(NtDB)Syn-Cel6A#5 shoot 17 days after transfer to RMOP medium containing 50 mg/L $AgNO_3$ and 500 mg/L spectinomycin. (C) HF(NtDB)Syn-Cel6A#5 shoot transferred to MS medium with 50 mg/L $AgNO_3$ and 500 mg/L spectinomycin for rooting.

RMOP/spectinomycin medium (Purnhauser et al 1987; Figures 5.3b and 5.3c). DNA blotting with this putative transformant revealed a weak *Xho*I/*Hind*III band at 4.0 kb, with relatively strong signals at 1.3 kb (deriving from both WT plastid DNA and from an extraplastidic copy of *trnA*; Chapter 3) and at 5.6 kb. The 5.6 kb band observed in this line and in all other transplastomic plants examined here derives from a recombination event between the P*psbA* sequence regulating *aadA* expression and the native plastid P*psbA* sequence (Gray et al 2009b). Closer examination of the HF(NtDB)Syn-Cel6A#5 lane, however, revealed an extremely weak signal at 4.0 kb, suggesting that this plant contained a small amount of the expected plastid DNA species with the *HF(NtDB)Syn-cel6A* and the *aadA* genes inserted between the plastid *trnI* and *trnA* genes (Figure 5.2b). The putative HF(NtDB)Syn-Cel6A shoot was subjected to a second round of regeneration on RMOP medium containing 50 mg/L silver nitrate and 500 mg/L spectinomycin. DNA blotting with DNA extracted from these plants revealed that in all cases, the expected 4.0 kb *Xho*I/*Hind*III band was lost (Figure 5.2c and data not shown).

BglC and Cel6A Protein Accumulation

Having obtained fully transformed HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC plants, the BglC protein content of these plants was analyzed by immunoblotting (Figure 5.4a). LF(NtDB)Syn-BglC protein was clearly detectable, but HF(NtDB)Syn-BglC protein could not be detected by immunoblotting. Quantification of the immunoblot in Figure 5.4a revealed that LF(NtDB)Syn-BglC protein accumulated to approximately 0.6%TSP, while HF(NtDB)Syn-BglC protein accumulation could not be quantified, though it was much less than 0.07%TSP as evidenced by the detection of 0.01 μg of purified BglC.

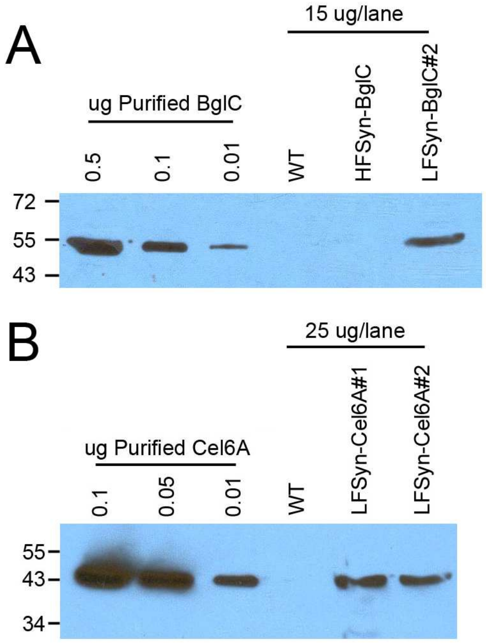


Figure 5.4: Immunoblots with protein extracted from HF(NtDB)Syn-BglC, LF(NtDB)Syn-BglC, and LF(NtDB)Syn-Cel6A tobacco lines. (A) Immunoblot with HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC tobacco. LF(NtDB)Syn-BglC accumulated to approximately 0.6%TSP, while HF(NtDB)Syn-BglC protein was undetectable (<<0.07%TSP). (B) Immunoblot with LF(NtDB)Syn-Cel6A#1 and LF(NtDB)Syn-Cel6A#2. LF(NtDB)Syn-Cel6A accumulation was approximately 0.04%TSP in both of these plants.

LF(NtDB)Syn-Cel6A protein accumulation was analyzed in fully transformed LF(NtDB)Syn-Cel6A plants by immunoblotting (Figure 5.4b). Quantification of the bands in this immunoblot revealed that LF(NtDB)Syn-Cel6A accumulated to approximately 0.04%TSP. HF(NtDB)Syn-Cel6A protein accumulation could not be assayed due to difficulties in the regeneration of HF(NtDB)Syn-Cel6A tobacco, as described above.

Transcript Accumulation

Total RNA extracted from HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC plants was hybridized with a *bglC*-specific probe (Figure 5.5a). Levels of *HF(NtDB)Syn-bglC* and *LF(NtDB)Syn-bglC* transcripts were lower than the levels of *nptII-bglC* transcripts included as a positive control (Chapter 3), but all *bglC*-containing transcripts observed in NPTII-BglC tobacco were observed in both HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC tobacco. No major differences in transcript abundance were observed between HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC tobacco after correcting for loading differences.

Probing of total LF(NtDB)Syn-Cel6A tobacco RNA revealed the accumulation of mainly polycistronic transcripts containing the *LF(NtDB)Syn-cel6A* ORF (Figure 5.5b). This is consistent with a previous report showing that GFP-Cel6A tobacco, which accumulated GFP-Cel6A protein at levels comparable to LF(NtDB)Syn-Cel6A tobacco, accumulated primarily polycistronic transcripts, with low steady-state levels of the monocistronic *cel6A* transcript (Gray et al 2009a). As expected, RNA from HF(NtDB)Syn-Cel6A#5-2B did not contain any *cel6A* transcript, confirming that this plant does not contain the *cel6A* ORF.

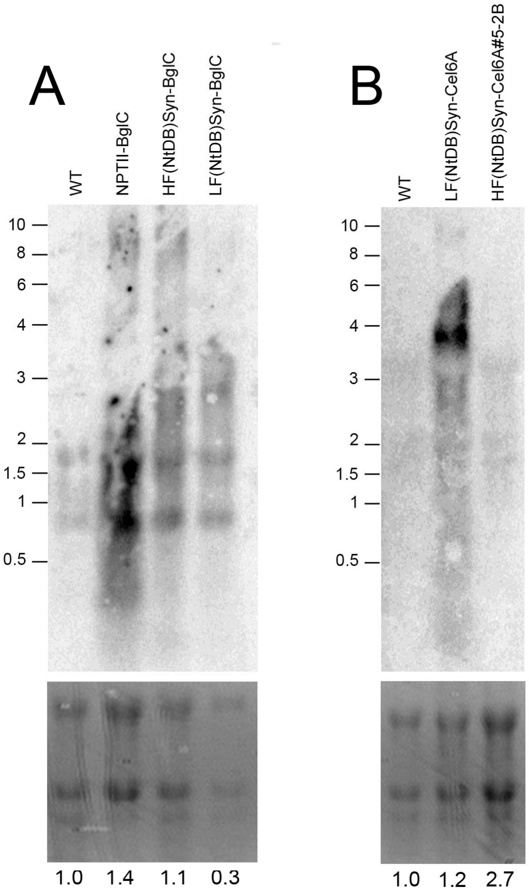


Figure 5.5: RNA blotting to detect *bglC* and *cel6A* transcripts. (A) RNA blot with RNA extracted from WT, NPTII-BglC, HF(NtDB)Syn-BglC, and LF(NtDB)Syn-BglC tobacco, hybridized with a *bglC*-specific probe. (B) RNA blot with RNA extracted from WT, LF(NtDB)Syn-Cel6A, and HF(NtDB)Syn-Cel6A#5-2B tobacco, hybridized with a *cel6A*-specific probe. Ethidium bromide-stained ribosomal RNA bands are shown below each blot. The numbers below the ribosomal RNA bands indicate relative loading, as determined from the intensity of the ethidium bromide-stained 25S rRNA band.

RNA Secondary Structure Predictions

The secondary structures of the HF(NtDB)Syn and LF(NtDB)Syn DB regions were predicted using m-fold software (Zuker 2003; Figure 5.6a, 5.6b). The HF(NtDB)Syn DB region was predicted to contain a large hairpin-loop structure extending from the Shine-Dalgarno region upstream of the AUG start codon to the +31 nucleotide, relative to A̲UG at +1. The LF(NtDB)Syn DB region was predicted to contain considerably less secondary structure than the HF(NtDB)Syn DB region (ΔG = -12.50 kcal/mol for *HF[NtDB]Syn*; ΔG = -6.93 kcal/mol for *LF[NtDB]Syn*). Monocistronic *nptII-bglC, tetC-bglC,* and *gfp-bglC* transcripts have been shown previously to have a terminus at -124 (Chapter 3). Secondary structures of the 5' ends of the mature monocistronic *HF(NtDB)Syn-bglC* and *LF(NtDB)Syn-bglC* transcripts extending from -124 to +60 were therefore predicted (Figure 5.6c, 5.6d). When the region of mRNA included for secondary structure prediction was enlarged, differences in secondary structure between the *HF(NtDB)Syn-bglC* and *LF(NtDB)Syn-bglC* transcripts were less pronounced. Strikingly, the free energy associated with these secondary structures showed a more stable secondary structure for the *LF(NtDB)Syn-bglC* transcript than for the *HF(NtDB)Syn-bglC* transcript (ΔG = -38.75 kcal/mol for *HF[NtDB]Syn-bglC*; ΔG = -43.62 kcal/mol for *LF[NtDB]Syn-bglC*). Secondary structure predictions of the 5' ends of predicted monocistronic *HF(NtDB)Syn-cel6A* and *LF(NtDB)Syn-cel6A* transcripts appeared qualitatively similar, though there was a difference in free energy (ΔG = -42.05 kcal/mol for *HF[NtDB]Syn-cel6A*; ΔG = -39.89 kcal/mol for *LF[NtDB]Syn-cel6A*; Figure 5.6e, 5.6f).

DISCUSSION

Transformation of the plastid genome to generate HF(NtDB)Syn-BglC, LF(NtDB)Syn-BglC, and LF(NtDB)Syn-Cel6A tobacco lines was performed using

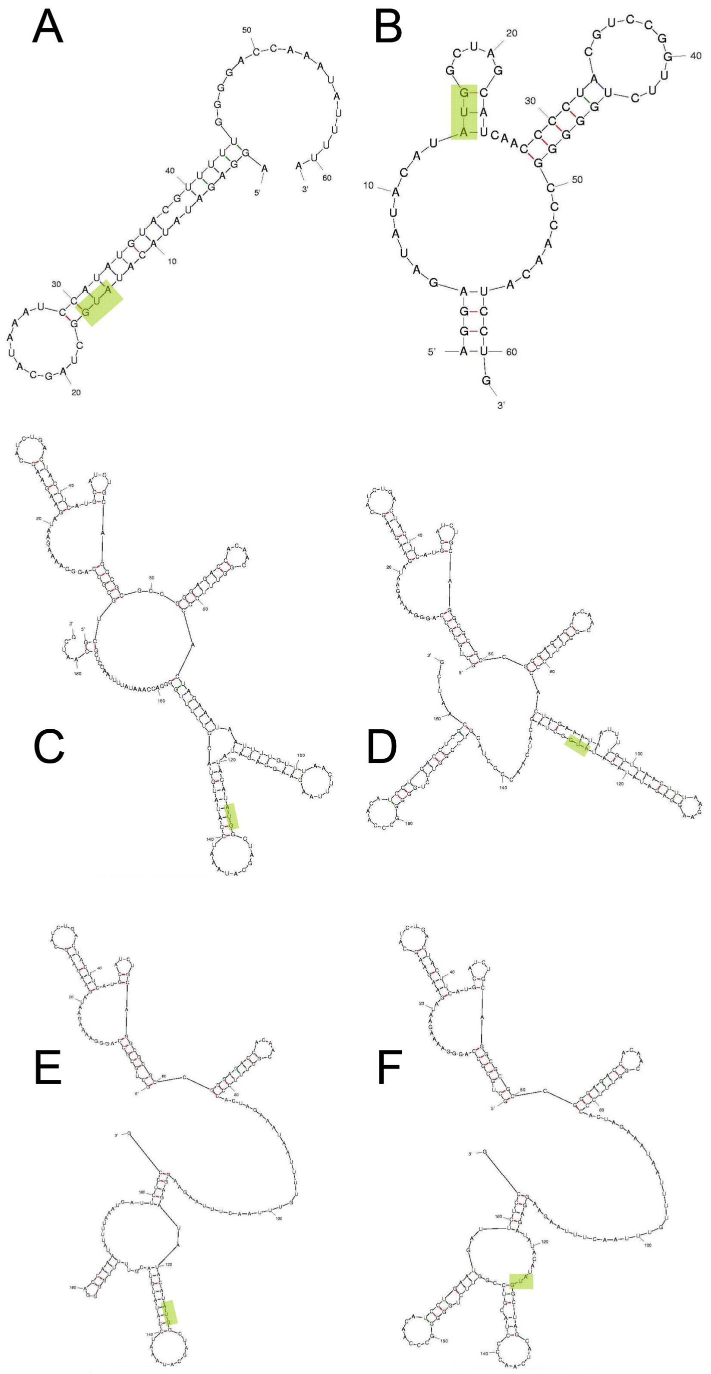


Figure 5.6: RNA secondary structure predictions. (A) HF(NtDB)Syn DB region, with nucleotides from -13 to +48, relative to the AUG start codon at +1. (B) LF(NtDB)Syn DB region. (C) 5' region of monocistronic HF(NtDB)Syn-BglC transcript, with nucleotides from -124 to +60, relative to the AUG start codon at +1. (D) 5' region of monocistronic LF(NtDB)Syn-BglC transcript. (E) 5' region of monocistronic HF(NtDB)Syn-Cel6A transcript. (F) 5' region of monocistronic LF(NtDB)Syn-Cel6A transcript. The AUG start codon in each figure is indicated by a green box.

routine procedures, and resulted in the expected transformation events. Attempts to generate a HF(NtDB)Syn-Cel6A plant line, however, resulted in unexpected transformation events (Figures 5.2 and 5.3). The majority of transplastomic plants regenerated on spectinomycin-containing tissue culture medium following bombardment with pHF(NtDB)Syn-Cel6A were transformed by a recombination event between the native plastid P*psbA* sequence and the P*psbA* sequence used to regulate the *aadA* gene in pHF(NtDB)Syn-Cel6A, resulting in plant lines like the UR-1 plant line described by Gray et al (2009b). A minority of regenerated shoots bombarded with pHF(NtDB)Syn-Cel6A, however, exhibited stalled growth on standard RMOP tissue culture medium containing spectinomycin (Figure 5.3a). The addition of 50 mg/L $AgNO_3$ to the medium (Purnhauser et al 1987) allowed for successful regeneration of one of these stalled shoots (Figures 5.3b and 5.3c). DNA blotting with DNA extracted from this shoot revealed that a small proportion of plastid DNA appeared to contain the expected HF(NtDB)Syn-Cel6A transformation product (Figure 5.2b). This shoot was subjected to a second round of regeneration on RMOP medium containing 50 mg/L $AgNO_3$ and 500 mg/L spectinomycin, and all of the second round regenerants lost the putative HF(NtDB)Syn-Cel6A plastome (Figure 5.2c and data not shown). Purnhauser et al (1987) first described the use of $AgNO_3$ to stimulate plant tissue culture growth, and attributed the growth stimulation to the ethylene inhibiting properties of $AgNO_3$. This suggests that the difficulty in regenerating a HF(NtDB)Syn-Cel6A plant line could result from changes in ethylene regulation following transformation with the pHF(NtDB)Syn-Cel6A plasmid through an unknown mechanism. Although it is possible that extremely high-level accumulation of HF(NtDB)Syn-Cel6A protein could cause changes in hormone signaling, high-level (~10%TSP) accumulation of TetC-Cel6A in transplastomic tobacco did not cause a detectable phenotype (Gray et al 2009a).

Whatever the mechanism causing difficulty in regenerating a HF(NtDB)Syn-Cel6A tobacco line, several lines of proposed research could be successful in regenerating a HF(NtDB)Syn-Cel6A plant. In the experiments described here, $AgNO_3$ was added to the shoot regeneration medium only after the identification of a stalled shoot. It is possible that the inclusion of 50 mg/L $AgNO_3$ in the RMOP/spectinomycin medium for the duration of shoot regeneration could result in the successful regeneration of a HF(NtDB)Syn-Cel6A plant. A second possibility would be to replace the P*psbA* sequence regulating *aadA* in pHF(NtDB)Syn-Cel6A with a different promoter. Ideally, this promoter would not be derived from plastid sequences. For example, bacterial, mitochondrial, or phage promoters and 5'UTRs could be used for regulating transcription and translation of *aadA*. Alternately, a different plastid promoter and 5'UTR (e.g., P*clpP*, P*rrn*, or P*rbcL*), or a hybrid promoter/5'UTR combination (e.g., P*clpP* with the *psbA* 5'UTR) could be used to regulate *aadA*. A third possibility is to generate a modified pHF(NtDB)Syn-Cel6A vector containing a second antibiotic resistance gene upstream of the *cel6A* ORF. A proposed vector is diagrammed in figure 5.7. This proposed vector contains the features of the original pHF(NtDB)Syn-Cel6A plasmid, but would also include a second antibiotic selection gene. In the vector diagram in figure 5.7, the *aphA-6* gene, conferring kanamycin resistance (Huang et al 2002), is shown. In principle, other selection genes could be used, but the *aphA-6* gene has been shown to mediate effective selection in plastid transformation experiments. The *aphA-6* gene would be regulated by a promoter and 5'UTR (P) and a 3'UTR (T). These sequences could be either plastid derived (e.g., P*clpP*, P*rrn*, P*rbcL*, as discussed above), or could be derived from non-plastid sequences (e.g., bacteria, mitochondria, or phages). The *aphA-6* gene would be flanked by *attP/attB* sequences for excision using a plastid-targeted phiC31 integrase gene (Kittiwongwattana et al 2007). Following

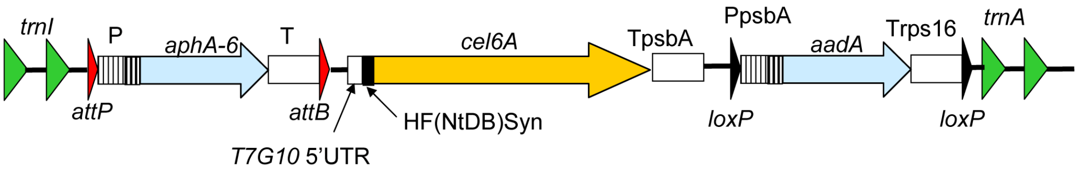


Figure 5.7: Proposed modified pHF(NtDB)Syn-Cel6A plastid transformation vector containing two selection marker genes. The *aphA-6* kanamycin-resistance gene upstream of the *HF(NtDB)Syn-cel6A* ORF and the *aadA* spectinomycin/streptomycin-resistance gene downstream of the *HF(NtDB)Syn-cel6A* ORF are expected to result in plastid transformation via the *trnI* and *trnA* flanking regions, resulting in integration of the *HF(NtDB)Syn-cel6A* ORF into the plastid genome.

bombardment with the vector diagrammed in figure 5.7, shoots would be regenerated on tissue culture medium containing both spectinomycin and kanamycin. This vector and selection scheme should force recombination via the *trnI* and *trnA* flanking regions, rather than via P*psbA*, as both marker genes would be required for shoot regeneration. The proposed second marker gene (*aphA-6* in figure 5.7) expression cassette could be inserted as an *Asc*I fragment into the unique *Asc*I site in pHF(NtDB)Syn-Cel6A, and could be removed from transformed plants by targeting phiC31 integrase to the plastid following the generation of fully transformed plants (Kittiwongwattana et al 2007).

The low-level accumulation of HF(NtDB)Syn-BglC and LF(NtDB)Syn-BglC protein is somewhat surprising and disappointing. The methods used to optimize the high-frequency codon downstream boxes were expected to generate efficiently translated DB regions, based on promising results with codon-altered GFP-Cel6A expression (Chapter 4). When the *bglC* ORF was fused to the high-frequency (HF[NtDB]Syn) and to the low-frequency (LF[NtDB]Syn) DB regions, the LF(NtDB)Syn DB resulted in higher accumulation of BglC protein (HF[NtDB]Syn-BglC and LF[NtDB]Syn-BglC accumulated to ~0.6%TSP and <<0.07%TSP, respectively). This suggests that the method of DB codon optimization used to generate the high- and low-frequency codon synthetic DB regions tested here is not effective for improving the accumulation of BglC protein in transplastomic tobacco.

Kudla et al (2009) suggested that mRNA secondary structure in the DB region is a better determinant of GFP protein accumulation than codon usage. Because LF(NtDB)Syn-BglC unexpectedly accumulated to higher levels than HF(NtDB)Syn-BglC, we hypothesized that secondary structure in the DB regions could play a role in the observed low-level accumulation of HF(NtDB)Syn-BglC. Indeed, secondary structure predictions of the isolated HF(NtDB)Syn and LF(NtDB)Syn DB regions showed greater secondary structure associated with the HF(NtDB)Syn DB region than with the LF(NtDB)Syn DB region (Figure 5.6a, 5.6b). The 5’ terminus of monocistronic *bglC* transcripts inserted at this locus in the plastid genome, however, is typically located at -124 relative to the AUG start codon at +1 (Chapter 3). We therefore determined the secondary structures of the 164 nucleotides at the 5’ ends of monocistronic *HF(NtDB)Syn-bglC and LF(NtDB)Syn-bglC* transcripts (Figures 5.6c, 5.6d). When a larger region of RNA was included in the secondary structure prediction, differences were less pronounced, and greater secondary structure stability was associated with the *LF(NtDB)Syn-bglC* transcript than with the *HF(NtDB)Syn-bglC* transcript. While RNA secondary structure may play a role in regulating the accumulation of BglC proteins, it is not clear from these secondary structure predictions what that role might be. Further research on RNA secondary structures, e.g., through X-ray crystallography experiments or RNase protection assays, could be merited in order to determine an accurate structure of the mature monocistrons of interest.

Because a HF(NtDB)Syn-Cel6A plant could not be regenerated, it is not possible at this point to compare the translation efficiency of the HF(NtDB)Syn and LF(NtDB)Syn DB fusions to the *cel6A* ORF. The relatively low accumulation of LF(NtDB)Syn-Cel6A (0.04%TSP) is expected from an inefficiently translated DB fusion to the *cel6A* ORF (Gray et al 2009a; Chapter 4). Speculatively, the difficulties in regenerating a HF(NtDB)Syn-Cel6A tobacco line could result from extremely high-level expression of HF(NtDB)Syn-Cel6A protein, causing changes in ethylene regulation through an unknown mechanism, but this hypothesis is unproven and difficult to test in the absence of a HF(NtDB)Syn-Cel6A tobacco line. For this hypothesis to be true, the accumulation of HF(NtDB)Syn-Cel6A protein would need to be extremely high in order to cause the observed low rates of shoot generation on

RMOP/spectinomycin medium. Accumulation of TetC-Cel6A was tolerated in tobacco at up to 11%TSP (Gray et al 2009a). Further, the plastid protein production machinery has been shown to tolerate foreign protein accumulation to over 30%TSP in at least two cases with no observed phenotype (De Cosa et al 2001; Oey et al 2009a), though a slow-growth phenotype was observed when a foreign protein accumulated to ~70%TSP (Oey et al 2009b). Speculatively, translation of the HF(NtDB)Syn DB region when fused to the *cel6A* ORF could be extremely rapid, causing misfolding of HF(NtDB)Syn-Cel6A protein in the plastid. These misfolded proteins could be insoluble, interfering with normal plastid function. Expression of HF(NtDB)Syn-Cel6A in *E. coli* did not cause any problems with cell growth (data not shown).

The data presented here showing low level accumulation of HF(NtDB)Syn-BglC protein (<<0.07%TSP), and moderate accumulation of LF(NtDB)Syn-BglC protein (0.6%TSP) suggest that the method of synthetic DB generation described here is not suitable for at least some ORFs. Because of our inability to regenerate a HF(NtDB)Syn-Cel6A tobacco line, it is not possible to decisively conclude whether the high- and low-frequency codon synthetic DBs are effective when fused to the *cel6A* ORF, though it is likely that some aspect of HF(NtDB)Syn-Cel6A expression is responsible for the difficulties in plant regeneration. The HF(NtDB)Syn DB region developed here may be successful for driving high-level accumulation of some as-yet-untested proteins in transplastomic plants.

It is apparent that DB function is context-specific, with the results of a given DB fusion depending on the identity of the ORF to which it is fused. This has been observed with the TetC and NPTII DB regions, where the TetC DB region mediated high-level Cel6A protein production in transplastomic tobacco, but only moderate BglC production, while the NPTII DB region mediated high-level BglC protein production, but only moderate Cel6A production (Gray et al 2009a; Chapter 3). It is likely that the DB region affects translation efficiency through several different mechanisms. Kudla et al (2009) suggest that *gfp* mRNA secondary structure in the DB region is a major determinant of GFP accumulation in *E. coli*, while codon usage is a poor predictor of GFP accumulation. It is likely that one mechanism of DB-mediated translation stimulation (e.g., codon usage or RNA secondary structure) may be of greater relative importance than another mechanism in a specific context and when expressing the protein of interest in the preferred host. For GFP expression in *E. coli*, RNA secondary structure in the DB region appears to be of greater importance than codon usage (Kudla et al 2009). RNA secondary structures in *E. coli* and in plastids should be very similar, to a first approximation. Codon usage in plastids and in *E. coli*, on the other hand, is very different. If RNA secondary structure is the major determinant of DB function for the expression of a given protein, then expression of the ORF of interest in *E. coli* should be a good prediction method for determining whether that construct will result in high foreign protein yields in tobacco plastids. If codon usage is the major determinant of DB function, then expression of that construct in *E. coli* will not be a good indicator of expression levels in tobacco plastids. Our data on TetC-Cel6A, NPTII-Cel6A, and GFP-Cel6A expression in *E. coli* suggests that *E. coli* DB-Cel6A expression is not a good predictor of expression levels in transplastomic tobacco, as GFP-Cel6A was expressed well in *E. coli*, but accumulated to low levels in transplastomic tobacco (data not shown and Gray et al 2009a). This suggests that, although RNA secondary structure may be an important factor in determining DB function, other mechanisms are likely to be important as well. Further experiments with DB regions fused to multiple ORFs and in various expression hosts to test parameters that may be important for DB function will result in a better understanding of the mechanism of translational enhancement by the DB region.

ACKNOWLEDGEMENTS

The authors wish to acknowledge David Wilson for his kind donation of anti-Cel6A and anti-BglC antibodies and of the pNS6 plasmid. BNG was supported by an NSF Graduate Research Fellowship. This work was supported by a USDA grant (USDA NRI 2007-02133) to BAA and MRH.

APPENDIX

Supplementary Table 5.S1. Overall tobacco plastid CUFs (NtTot CUF) and CUFs in the DB regions of highly expressed plastid genes (NtDB CUF), expressed as number of uses per 1,000 codons

Codon	NtTot CUF	NtDB CUF
ATT	39.2	61.5
GAA	39.6	53.8
AAT	36.5	46.2
TTT	34.2	46.2
TTA	31.0	38.5
ACT	20.0	30.8
AGC	5.4	30.8
ATA	24.4	30.8
CCA	12.1	30.8
CGT	12.3	30.8
GAT	31.5	30.8
GGA	27.1	30.8
GTA	21.4	30.8
TAT	27.3	30.8
TGG	17.2	30.8
TTG	22.1	30.8
ACA	15.1	23.1
ACC	10.0	23.1
CAA	26.0	23.1
CGA	14.3	23.1
GCA	15.6	23.1
GCT	25.9	23.1
GGT	23.3	23.1
GTT	20.1	23.1
AAA	37.4	15.4
AAG	14.5	15.4
ACG	5.4	15.4
CAT	16.8	15.4
GCC	9.8	15.4
TCA	15.0	15.4
TCG	8.0	15.4
TCT	22.1	15.4
TGT	8.0	15.4
AGA	17.5	7.7
AGG	6.8	7.7
CAG	9.0	7.7
CCG	5.6	7.7
CCT	17.1	7.7

CGC	4.0	7.7
CTA	13.6	7.7
CTC	7.9	7.7
CTT	22.6	7.7
GAC	8.6	7.7
GAG	14.6	7.7
GGC	8.0	7.7
AAC	12.8	0.0
AGT	14.9	0.0
ATC	17.2	0.0
ATG	24.5	0.0
CAC	5.5	0.0
CCC	7.3	0.0
CGG	5.0	0.0
CTG	7.4	0.0
GCG	5.8	0.0
GGG	12.2	0.0
GTC	7.2	0.0
GTG	8.1	0.0
TAC	7.7	0.0
TCC	12.8	0.0
TGC	3.0	0.0
TTC	20.6	0.0

Supplementary Table 5.S2. Primers used in this study

Primer Name	Sequence
HFSynBglC-fwd	CATATGGCTAGCATAAATCCATATGTACGTTTTTGGGGACCAAATATTTTAACCTCGCAATCGACGACT
LFSynBglC-fwd	CATATGGCTAGCATCAACCCCTACGTCCGGTTCTGGGGGCCCAACATCCTGACCTCGCAATCGACGACT
BglC-rev	ATGCGGCCGCTATTCCTGTCCGAAGAT
HFSynCel6A-fwd	CATATGGCTAGCATAAATCCATATGTACGTTTTTGGGGACCAAATATTTTAAATGATTCTCCGTTCTAC
LFSynCel6A-fwd	CATATGGCTAGCATCAACCCCTACGTCCGGTTCTGGGGGCCCAACATCCTGAATGATTCTCCGTTCTAC
Cel6A-rev	ATAGACTAGGCCAGGATCGCGGCCGCTCAGCTGGCGGCGCAGGT
Aprobe-fwd	ATAGTATCTTGTACCTGA
Aprobe-rev	TAAAGCTTTGTATCGGCTA
BglCint-fwd	TTCGTCCAGGACGGCGAC
Cel6Aint-fwd	GTAACGAGTGGTGCGACC

REFERENCES

Bulmer M (1988) “Codon usage and intragenic position” *J Theor Biol* **133**: 67-71.

De Cosa B, Moar W, Lee SB, Miller M, Daniell H (2001) “Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals” *Nat Biotechnol* **19**: 71-74.

Fuglsang A (2004) “Nucleotides downstream of start codons show marked non-randomness in *Escherichia coli* but not in *Bacillus subtilis*” *Antonie van Leeuwenhoek* **86**: 149-158.

Gonzalez de Valdivia EI, Isaksson LA (2004) “A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*” *Nucleic Acids Res* **32**: 5198-5205.

Gray BN, Ahner BA, Hanson MR (2009a) “High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions,” *Biotechnology and Bioengineering* **102**: 1045-1054.

Gray BN, Ahner BA, Hanson MR (2009b) “Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants,” *Transgenic Research* in press.

Hood EE, Woodard SL (2002) “Industrial proteins produced from transgenic plants” in *Plants as Factories for Protein Production* (Hood EE and Howard JA, eds.), Kluwer Academic Publishers.

Huang FC, Klaus SM, Herz S, Zou Z, Koop HU, Golds TJ (2002) “Efficient plastid transformation in tobacco using the *aphA-6* gene and kanamycin selection” *Mol Genet Genomics* **268**: 19-27.

Kittiwongwattana C, Lutz K, Clark M, Maliga P (2007) “Plastid marker gene excision by the phiC31 phage site-specific recombinase” *Plant Mol Biol* **64**: 137-143.

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) "Coding-sequence determinants of gene expression in *Escherichia coli*" *Science* **324**: 255-258.

Lenzi P, Scotti N, Alagna F, Tornesello ML, Pompa A, Vitale A, De Stradis A, Monti L, Grillo S, Buonaguro FM, Maliga P, Cardi T (2008) "Translational fusion of chloroplast-expressed human papillomavirus type 16 L1 capsid protein enhances antigen accumulation in transplastomic tobacco" *Transgenic Res* **17**: 1091-1102.

Maliga P (2003) "Progress towards commercialization of plastid transformation technology" *Trends Biotechnol* **21**: 20-28.

Oey M, Lohse M, Scharff LB, Kreikemeyer B, Bock R (2009a) "Plastid production of protein antibiotics against pneumonia via a new strategy for high-level expression of antimicrobial proteins" *Proc Natl Acad Sci USA* doi: 10.1073/pnas.0813146106.

Oey M, Lohse M, Kreikemeyer B, Bock R (2009b) "Exhaustion of the chloroplast protein synthesis capacity by massive expression of a highly stable protein antibiotic" *Plant J* **57**: 436-445.

Purnhauser L, Medgyesy P, Czakó M, Dix PJ, Márton L (1987) "Stimulation of shoot regeneration in *Triticum aestivum* and *Nicotiana plumbaginifolia* Viv. tissue cultures using the ethylene inhibitor $AgNO_3$" *Plant Cell Reports* **6**: 1-4.

Spiridonov NA, Wilson DB (2001) "Cloning and biochemical characterization of BglC, a beta-glucosidase from the cellulolytic actinomycete *Thermobifida fusca*" *Curr Microbiol* **42**: 295-301.

Stenström CM, Isaksson (2002) "Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side" *Gene* **288**: 1-8.

Svab Z, Maliga P (1991) “Mutation proximal to the tRNA binding region of the Nicotiana plastid 16S rRNA confers resistance to spectinomycin” *Mol Gen Genet* **228**: 316-319.

Tregoning JS, Nixon P, Kuroda H, Svab Z, Clare S, Bowe F, Fairweather N, Ytterberg J, van Wijk KJ, Dougan G, Maliga P (2003) “Expression of tetanus toxin Fragment C in tobacco chloroplasts” *Nucleic Acids Res* **31**: 1174-1179.

Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) “Molecular farming in plants: host systems and expression technology” *Trends in Biotechnol* **21**: 570-578.

Ye GN, Hajdukiewicz PT, Broyles D, Rodriguez D, Xu CW, Nehra N, Staub JM (2001) “Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco” *Plant J* **25**: 261-270.

Zuker M (2003) “Mfold web server for nucleic acid folding and hybridization prediction” *Nucleic Acids Res* **31**: 3406-3415.

Chapter 6: Extensive Homologous Recombination Between Introduced and Native Regulatory Plastid DNA Elements in Transplastomic Plants[1]

ABSTRACT

Homologous recombination within plastids directs plastid genome transformation for foreign gene expression and study of plastid gene function. Though transgenes are generally efficiently targeted to their desired insertion site, unintended homologous recombination events have been observed during plastid transformation. To understand the nature and abundance of these recombination events, we analyzed transplastomic tobacco lines derived from three different plastid transformation vectors utilizing two different loci for foreign gene insertion. Two unintended recombinant plastid DNA species were formed from each regulatory plastid DNA element included in the transformation vector. Some of these recombinant DNA species accumulated to as much as 10 to 60% of the amount of the desired integrated transgenic sequence in T0 plants. Some of the recombinant DNA species undergo further, "secondary" recombination events, resulting in an even greater number of recombinant plastid DNA species. The abundance of novel recombinant DNA species was higher in T0 plants than in T1 progeny, indicating that the ancillary recombination events described here may have the greatest impact during selection and regeneration of transformants. A line of transplastomic tobacco was identified containing an antibiotic resistance gene unlinked from the intended transgene insertion as a result of an unintended recombination event, indicating that the homologous recombination events described here may hinder efficient recovery of plastid transformants containing the desired transgene.

[1] Gray BN, Ahner BA, Hanson MR (2009) *Transgenic Res* doi:10.1007/s11248-009-9246-3

INTRODUCTION

Mapping of the plastid genome of most higher plants results in a monomeric circular structure containing two inverted repeat (IR) regions, a small single copy (SSC) region, and a large single copy (LSC) region. Higher plant plastids typically contain approximately 50-100 copies of their genome (Li et al. 2006). Although much of this DNA is found in its monomeric circular form, many other forms of plastid DNA have been observed. In tobacco chloroplasts, 45% of plastid DNA molecules were found in non-canonical conformations (i.e., linear molecules, circular multimers, and/or irregular molecules containing uneven numbers of IR regions; Lilly et al. 2001).

Plastid DNA rearrangement via intermolecular recombination was described in an early study in which two distinct arrangements of the bean plastid genome were identified with the IR regions in reverse orientation with respect to the LSC and SSC regions (Palmer 1983). Later studies identified plastid DNA deletions mediated by 7-14 bp direct (Kawata et al. 1997) or indirect (Kanno et al. 1993) repeats. These and other studies of wild-type plastid genomes have revealed multiple forms of the plastid genome resulting from intermolecular recombination among the many copies of plastid DNA in a given plastid and/or from intramolecular recombination between the IR regions of the plastid genome.

Transformation of the plastid genome relies on the plastid recombination machinery to direct homologous recombination between the plastid genome and a transformation vector, typically in the form of plasmid DNA. Most plastid transformation vectors contain an antibiotic resistance gene for selection of transformants and often a gene of interest to be inserted into the plastid genome. These genes are flanked by 0.5-1.5 kb stretches of plastid DNA to direct homologous recombination at the desired insertion site. Transgenes in transplastomic plants are usually regulated by 50-300 bp plastid DNA elements, e.g., promoters, 5’ untranslated regions (5'UTRs), and 3'UTRs. At least one study has shown that recombination efficiency is positively correlated with the length of homologous DNA sequence used to direct recombination (Blowers et al. 1989).

Observations of recombination events in transplastomic plants have advanced our understanding of plastid DNA recombination. The short plastid regulatory elements used in transformation vectors have been observed to direct unintended recombination events between duplicated DNA sequences. For example, DNA from a line of transplastomic tobacco that could not be cultured to homoplasmy exhibited a signal of unexpected size on a DNA blot hybridized with an *aadA*-specific probe (Svab and Maliga 1993). In this transplastomic tobacco line, it was shown that recombination had been mediated by the 400 bp *psbA* 3'UTR used to regulate the *aadA* transgene rather than by the downstream 1.29 kb flanking sequence from *accD* intended to direct transgene insertion. This unintended recombination was hypothesized to cause a large deletion, giving rise to an unstable transformed plastid genome that remained heteroplasmic on selective medium.

Homologous recombination events in the plastids of transplastomic plants appear to be widespread. In another transplastomic tobacco line, a small circular extrachromosomal element designated NICE1, which contained plastid DNA sequence from the *trnI* gene and resulted from recombination between imperfect repeats in direct orientation, was observed. This DNA species was found in transplastomic, but not in wild-type, tobacco plastids, despite the presence of the imperfect repeats in both transplastomic and WT tobacco (Staub and Maliga 1994). It was concluded that the process of plastid transformation somehow resulted in the formation of this extrachromosomal element via homologous recombination between these imperfect repeats. A similar small circular extrachromosomal element was observed by McCabe et al. (2008) following a homologous recombination event between the native T*rbcL* and an introduced copy of T*rbcL* in direct orientation. It was proposed that this deleted region of DNA was circularized, resulting in a 21 kb molecule.

The tendency for homologous recombination to occur between repeated DNA elements in plastids was exploited by Iamtham and Day (2000) in a novel marker gene removal technique. By introducing two copies of a promoter element in direct orientation on either side of the marker gene *aadA*, a 'loop-out' homologous recombination event resulted in deletion of the marker gene. A conceptually similar experiment exploited both the *psbA* 5'UTR and the *psbA* 3'UTR to generate heteroplasmic knockouts of the *psbA* gene following transformation of the plastid genome at the *trnI*/*trnA* locus in the IR region (Khan et al. 2007).

In a separate study, unintended recombinations between *loxP* sites and the *psbA* promoter (P*psbA*), between *loxP* sites and the *rps12* 5'UTR, and between the native ribosomal promoter (P*rrn*) and an inserted P*rrn* sequence regulating a transgene were observed when plastid-targeted CRE recombinase was stably expressed in transplastomic tobacco containing *loxP* sites (Corneille et al. 2003). The *loxP*/P*psbA* and *loxP*/*rps12* 5'UTR recombination events were hypothesized to be mediated by CRE recombinase due to the presence of imperfect CRE binding sites in P*psbA* and in the *rps12* 5'UTR, while P*rrn* recombination events were not likely to depend on CRE recombinase, instead resulting from the plastid recombination machinery but apparently up-regulated by CRE-mediated recombination events.

The recombination events in transplastomic plants described above all resulted from recombination between native and introduced regulatory plastid DNA elements in direct orientation, resulting in deletion of the intervening DNA. Homologous recombination between introduced and native copies of T*psbA* in inverse orientations with respect to each other have also been observed following transgene integration in the *rpl33*/*rpl20* (Rogalski et al. 2006), *trnG*/*trnfM*, *trnG*/*trnR* (Rogalski et al. 2008a), and *trnS*/*trnT* (Rogalski et al. 2008b) loci, all in the LSC region of the plastid genome. The recombination events observed by Rogalski et al. (2006; 2008a; 2008b) resulted in a reversed orientation for much of the LSC region in DNA molecules where this recombination occurred (i.e., between the native T*psbA* sequence and the transgene insertion site). A similar "flip-flop" recombination event between two imperfect P*rrn* copies introduced in opposite orientations with respect to each other was observed in transgenic tobacco lines engineered to express HIV antigens from the *trnG*/*trnfM* locus in the LSC region (McCabe et al. 2008; Zhou et al. 2008). In cases where recombination occurs between two plastid DNA elements in inverse orientation with respect to each other, the intervening DNA is inverted, with no DNA sequence being lost following recombination.

The prior observations of multiple forms of plastid DNA in wild-type plants formed by recombination among the multiple copies of the plastid genome present within each plastid, the ability to generate transplastomic plants *via* homologous recombination, and the finding of extrachromosomal DNA elements and unintended recombination events in transplastomic plants point to an active recombination system functioning in higher plant plastids and to a population of DNA in flux within the plastid. We report here the identification and quantification of recombinant DNA species in transplastomic tobacco resulting from the interaction of regulatory plastid DNA elements (promoters, 5'UTRs, and 3'UTRs) in transgenic plants with their native counterparts. All predicted recombinant plastid DNA species were observed in transplastomic plants created from three different plastid transformation vectors.

MATERIALS AND METHODS

Plant Material

Transplastomic tobacco seeds from plants expressing GFP (Reed et al. 2001), and GFP-Cel6A or TetC-Cel6A (Gray et al. 2008) were planted in Magenta boxes (Magenta Corporation, Chicago, IL) containing sterile MS agar medium. T0-generation 22XE2 transplastomic tobacco expressing rbcL-Cel6A (Yu et al. 2007) was maintained in a Magenta box (Magenta Corporation) containing sterile MS agar medium. Unintended Recombination-1 (UR-1) tobacco was maintained on RMOP agar medium containing 500 mg/L spectinomycin. All plants were grown under fluorescent lights with a 14 hour photoperiod.

DNA Isolation

Leaves were harvested from T0 and from seed-grown tobacco plants and immediately flash-frozen in liquid nitrogen. DNA was extracted as described previously (Gray et al. 2008). Purity of the DNA extraction was assayed by measuring the spectrophotometric absorbance at 260 nm and 280 nm.

PCR and DNA Sequencing

PCR reactions were carried out in 50 μL volumes using Taq Master Mix (Qiagen, Valencia, CA) according to the manufacturer's instructions. PCR products were visualized following electrophoresis in a 0.8% agarose gel containing ethidium bromide.

PCR products were sequenced following either gel purification of the desired band using a Qiaquick Gel Extraction Kit (Qiagen) or primer removal using a Qiaquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions.

DNA sequencing was performed by the Cornell University Life Sciences Core Laboratories Center (Cornell University, Ithaca, NY).

DNA Blots

DNA blotting was performed as described previously (Gray et al. 2008). Briefly, tobacco DNA was *Xho*I, *Hind*III/*Xho*I, or *Bam*HI/*Xho*I digested overnight, then electrophoresed in a 1% agarose gel and transferred to a Hybond N+ nylon membrane (Amersham Pharmacia, Piscataway, NJ) and probed with ^{32}P-labelled PCR products. Probes were generated by the primer pairs Iprobe-fwd/Iprobe-rev (*trnI*), Cel6Aint-fwd/Cel6A-TpsbA-rev (*cel6A*), psbAprobe-fwd/psbAprobe-rev (*psbA*), and aadAprobe-fwd/aadAprobe-rev (*aadA*; primer sequences are shown in Supplementary Table 6.S1). Between each probing, the membrane was stripped in a boiling 0.1% SDS solution to remove all isotope. DNA blots were visualized following exposure to a Phosphorimager screen (Molecular Dynamics, Sunnyvale, CA). Contrast and brightness of the scanned images were adjusted using Photoshop (Adobe, San Jose, CA). Bands were quantified using Scion Image software (Scion Corporation, Frederick, MD).

RESULTS

Detection of recombinant DNA species in GFP-Cel6A plastids

In order to detect recombinant DNA species in the plastids of transplastomic plants, DNA was extracted from WT and from T0 GFP-Cel6A tobacco plants (these plants express the *cel6A* ORF with a 14 amino acid fusion from GFP; Gray et al. 2008) and digested by *Xho*I and *Xho*I/*Hind*III for DNA blotting. In addition to the expected *trnI*-containing restriction fragments in WT and GFP-Cel6A plastid DNA, two other major unexpected *trnI*-containing *Xho*I and *Xho*I/*Hind*III fragments were observed in GFP-

Cel6A DNA (Figure 6.1a). These bands were not artifacts of incomplete digestion, as they were consistently observed even after very long digestion times with high enzyme loading to ensure complete digestion. As shown in Figures 6.1b and 6.1c, respectively, *cel6A* and *psbA* probes also hybridized with multiple bands on this DNA blot. Hybridization with a *cel6A* probe (Figure 6.1b) revealed the presence of the same unexpected *Xho*I and *Xho*I/*Hind*III fragments observed in the *trnI*-probed blot (Figure 6.1a), suggesting that these DNA species contained both the *trnI* and *cel6A* genes. A *psbA* probe hybridized with plastid DNA fragments of the expected size in both GFP-Cel6A and in WT tobacco and also hybridized with other unexpected GFP-Cel6A DNA fragments (Figure 6.1c), as with the other probes. One DNA fragment (6.9 kb and 4.3 kb *Xho*I and *Xho*I/*Hind*III fragments, respectively) hybridized with all three probes used for these DNA blots, suggesting that this DNA species contained the *trnI*, *cel6A*, and *psbA* genes. This fragment was present at a concentration approximately half that of *trnI* (compare the 6.9 kb and 5.7 kb bands in Figure 6.1a *Xho*I digests), and approximately equal to that of *psbA* (compare the 6.9 kb and 8.5 kb bands in Figure 6.1c *Xho*I digests). The *psbA* signal is expected to be half the intensity of the *trnI* and *cel6A* signals, as the *psbA* gene is located in the LSC region of the plastid genome, while the *trnI* and *cel6A* genes are located in the duplicated IR regions of the genome. DNA blots with three independently transformed GFP-Cel6A tobacco lines (Gray et al. 2008) revealed that all of these transplastomic plants contained the minor plastid DNA species described above in approximately equal amounts (Gray et al. 2008 and data not shown).

The *trnI* probe hybridized with a faint band in GFP-Cel6A plants of the same size expected from WT plants following *Xho*I digestion (Figure 6.1a); this band likely derives from a small amount of plastid DNA transferred to the nuclear and/or mitochondrial genomes and does not indicate heteroplasmy of the transformed plant

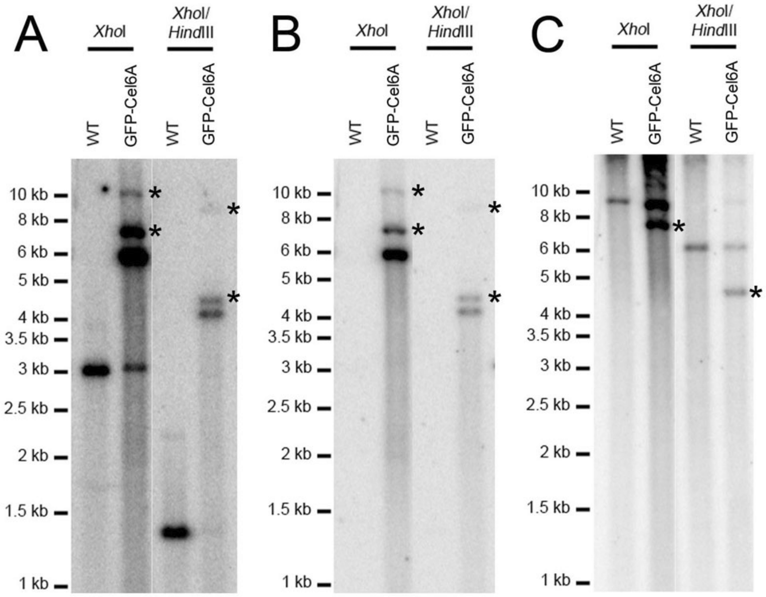


Figure 6.1 Detection of recombinant DNA species on DNA blots of *Xho*I- and *Xho*I/*Hind*III-digested T0 GFP-Cel6A tobacco DNA. (A) Blot hybridized with a *trnI* probe; (B) Blot hybridized with a *cel6A* probe; (C) Blot hybridized with a *psbA* probe. Unexpected DNA fragments are marked by asterisks (*).

(Ruf et al. 2000). Digestion of GFP-Cel6A DNA with both *Xho*I and *Hind*III resulted in the loss of the apparent WT signal in *trnI*-hybridized GFP-Cel6A DNA (Figure 6.1a). This suggests that an extraplastidic copy of *trnI* is present in the tobacco nuclear and/or mitochondrial genome that has lost the *Hind*III sites surrounding the region of DNA used to synthesize the *trnI* probe, and that GFP-Cel6A tobacco is fully transformed.

Characterization of the recombinant DNA species present in GFP-Cel6A transplastomic plants

Because it was likely that the novel bands in GFP-Cel6A tobacco lines detected in Figure 6.1 resulted from recombination events mediated by regulatory plastid DNA elements included in plastid transformation vectors, PCR was performed to amplify products predicted to result from recombination events mediated by the *psbA* promoter + 5'UTR (P*psbA*), *psbA* 3'UTR (T*psbA*), and *rps16* 3'UTR (T*rps16*) in GFP-Cel6A tobacco. Figure 6.2 shows the result of a representative PCR reaction using primers internal to *psbA* and *aadA* (psbAprobe-fwd/aadAint-rev) to amplify a 1.4 kb product between these genes. No detectable product was amplified from WT tobacco DNA. The GFP-Cel6A PCR product was sequenced to confirm that it derived from amplification of a region of DNA between the *psbA* and *aadA* genes, with *psbA* upstream of *aadA* as a result of recombination between the native *psbA* 3'UTR (T*psbA*) and an introduced copy of T*psbA* present in the transformation vectors (DNA species B in Figure 6.3, described below).

PCR reactions similar to those shown in Figure 6.2 were performed with GFP-Cel6A DNA using primer combinations to amplify products of all the predicted homologous recombinations between introduced and native regulatory plastid DNA elements (data not shown; primer combinations are shown in Supplementary Table

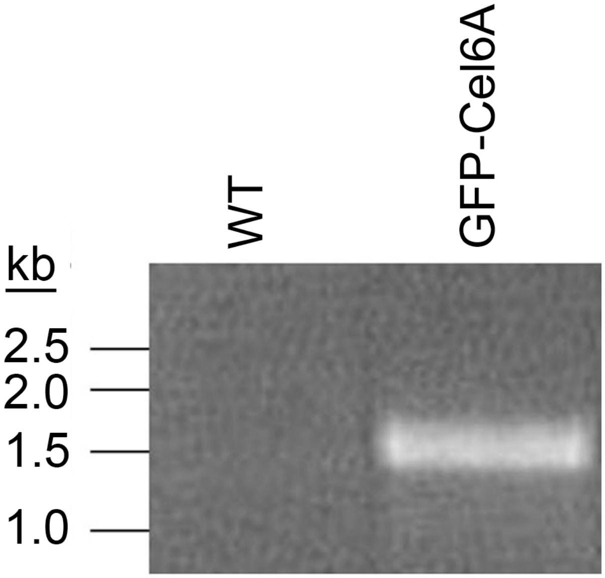


Figure 6.2 Screening GFP-Cel6A tobacco DNA for recombinant DNA species by PCR. A 1.4 kb product was PCR amplified from GFP-Cel6A tobacco that was confirmed by sequencing to result from a recombination event involving the introduced and native copies of T*psbA*. No detectable product was amplified from WT tobacco DNA. PCR reactions similar to the one shown were used to identify further recombinant plastid DNA species formed by homologous recombination mediated by introduced copies of regulatory plastid DNA elements in GFP-Cel6A tobacco.

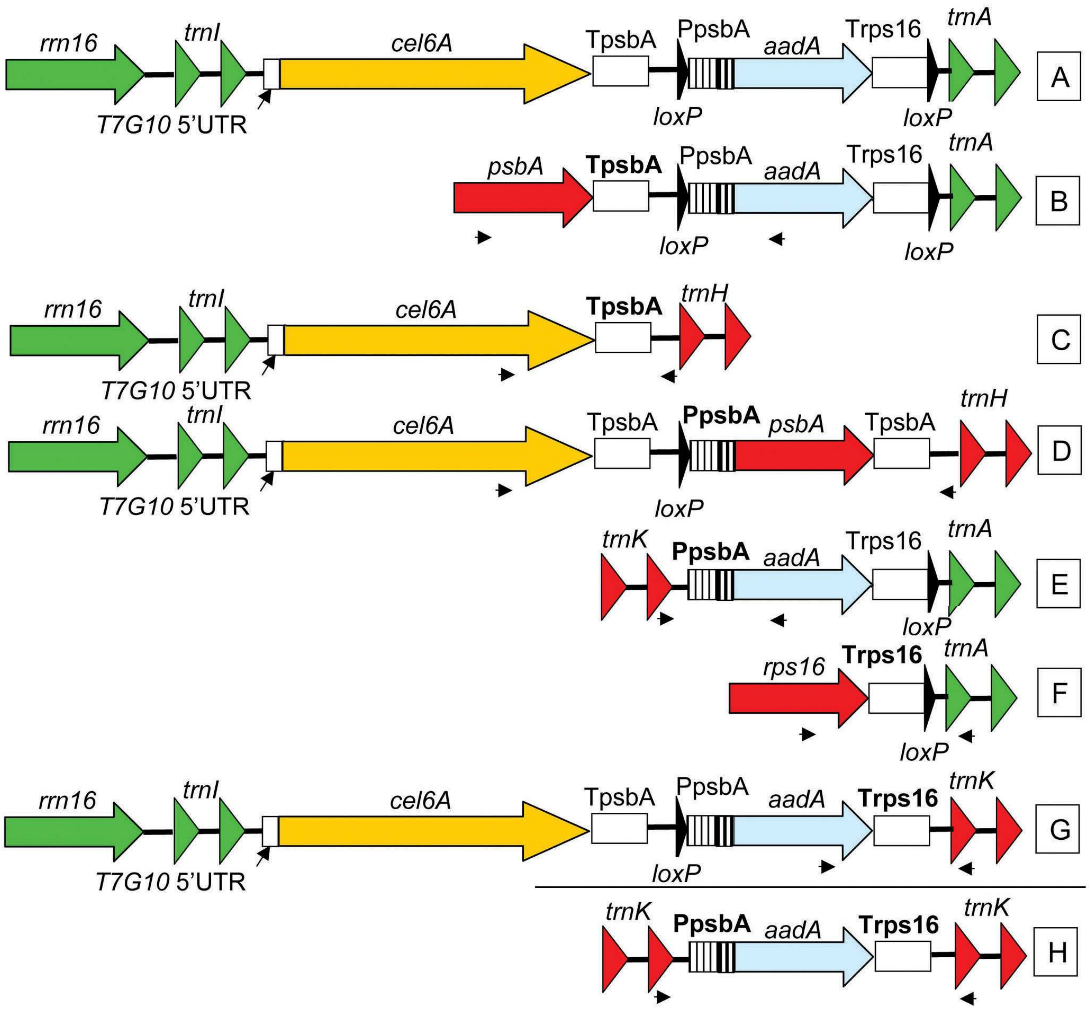


Figure 6.3 Schematic diagrams (not to scale) of the eight DNA species identified in GFP-Cel6A tobacco plastids. Species A is the main DNA species and was expected to form following plastid transformation. Species B-H resulted from recombination events mediated by the plastid DNA element(s) in bold. Small arrows underneath each schematic diagram show the locations of the primers used to amplify the given species. Species H is the result of a secondary recombination event (Figure 6.8).

6.S2). In addition to the expected DNA species resulting from transformation with the vectors described in Gray et al. (2008) with the *cel6A* and *aadA* genes inserted at the *trnI*/*trnA* intergenic region (species A in Figure 6.3), seven additional DNA species were amplified by PCR and confirmed by sequencing. Each regulatory plastid DNA element present in the transformation vector mediated the formation of two additional DNA species (i.e., species B-G in Figure 6.3) via homologous recombination with the native plastid DNA element. DNA species H in Figure 6.3 is the result of a secondary recombination between species E and the native *rps16* 3'UTR, and is described in further detail below. The sequences of the DNA species diagrammed in Figure 6.3 have been confirmed only in the region amplified by PCR (primer binding sites are shown as small arrows in Figure 6.3). The regions flanking those actually sequenced are included in Figure 6.3 as an aid to the reader, showing which DNA species are expected to hybridize with *trnI*, *cel6A*, and/or *psbA* probes. Schematic diagrams of the DNA species amplified by PCR were constructed by assuming that the DNA sequences upstream (DNA species C, D, and G) or downstream (DNA species B, E, and F) of the DNA element mediating homologous recombination remained unchanged from the transformation vector or the native plastid DNA sequence. The models of GFP-Cel6A DNA species D, E, F, G, and H diagrammed in Figure 6.3 were supported by detection of appropriately-sized restriction fragments on DNA blots (Figure 6.1, Figure 6.7, and data not shown).

To compare the relative abundance of the DNA species diagrammed in Figure 6.3 in T0 GFP-Cel6A tobacco, the bands in Figure 6.1 corresponding to DNA species D and G were quantified (Table 6.1). Species B and C were not detected by DNA blotting, presumably due to their low abundance *in vivo*; species E, F, and H are not expected to hybridize with the probes used for the blots in Figure 6.1.

Table 6.1 Quantification of the recombinant DNA species in T0 generation GFP-Cel6A tobacco

Species[a]	Detected By	Abundance (% of A)
A	*trnI, cel6A*	100
B	*psbA*	~0.0
C	*trnI, cel6A*	~0.0
D	*trnI, cel6A, psbA*	62 ± 14
E	n/d[b]	n/d[b]
F	n/d[b]	n/d[b]
G	*trnI, cel6A*	13 ± 2.7
H	n/d[b]	n/d[b]

[a] Schematic diagrams of DNA species A-H are shown in Figure 6.3

[b] n/d, this species was not detected by the probes used for these DNA blots

Detection of recombinant DNA species in 22XE2 and MR220 tobacco lines

To determine whether the novel recombinant DNA species described above are specific to GFP-Cel6A tobacco or whether they are generally formed in transplastomic plants, rbcL-Cel6A- and GFP-expressing transplastomic tobacco lines generated from plastid transformation vectors distinct from those used to generate GFP-Cel6A tobacco were analyzed. The transformation vector used to generate 22XE2 tobacco was similar to that used to generate GFP-Cel6A tobacco, with the *T. fusca cel6A* ORF fused at its 5' end to the first 14 codons from *N. tabacum rbcL* and inserted between the plastid *trnI* and *trnA* genes, but differed in the exact transgene insertion site, the order of transgenes in the transformation vector, and in some of the regulatory DNA sequences (compare transformation vectors for GFP-Cel6A in Figure 6.3 DNA species A and 22XE2 in Figure 6.4 DNA species A). The PCR amplification and sequencing strategy described above resulted in the identification of ten unexpected plastid DNA species in T0 22XE2 tobacco (Yu et al. 2007; primers used are shown in Supplementary Table 6.S2). The ten additional DNA species shown in Figure 6.4 (species B through K) were formed by recombination events mediated by the five plastid DNA elements in the 22XE2 transformation vector. As with the DNA species detected in GFP-Cel6A tobacco and diagrammed in Figure 6.3, DNA sequence data were obtained only in the regions between the primer annealing sites depicted as small arrows in Figure 6.4 and the remainder of the sequence from either the transformation vector or from native plastid DNA was assumed to remain unchanged. All the DNA species depicted in Figure 6.4 were detected by PCR; 22XE2 DNA species C was also detected by DNA blotting (Figure 6.7, described below).

In the third transgenic line analyzed, a modified GFP gene was inserted into tobacco chloroplasts at the *trnV/rps12* intergenic region instead of the *trnI*/*trnA* intergenic region (Reed et al. 2001), resulting in the MR220 line of transplastomic

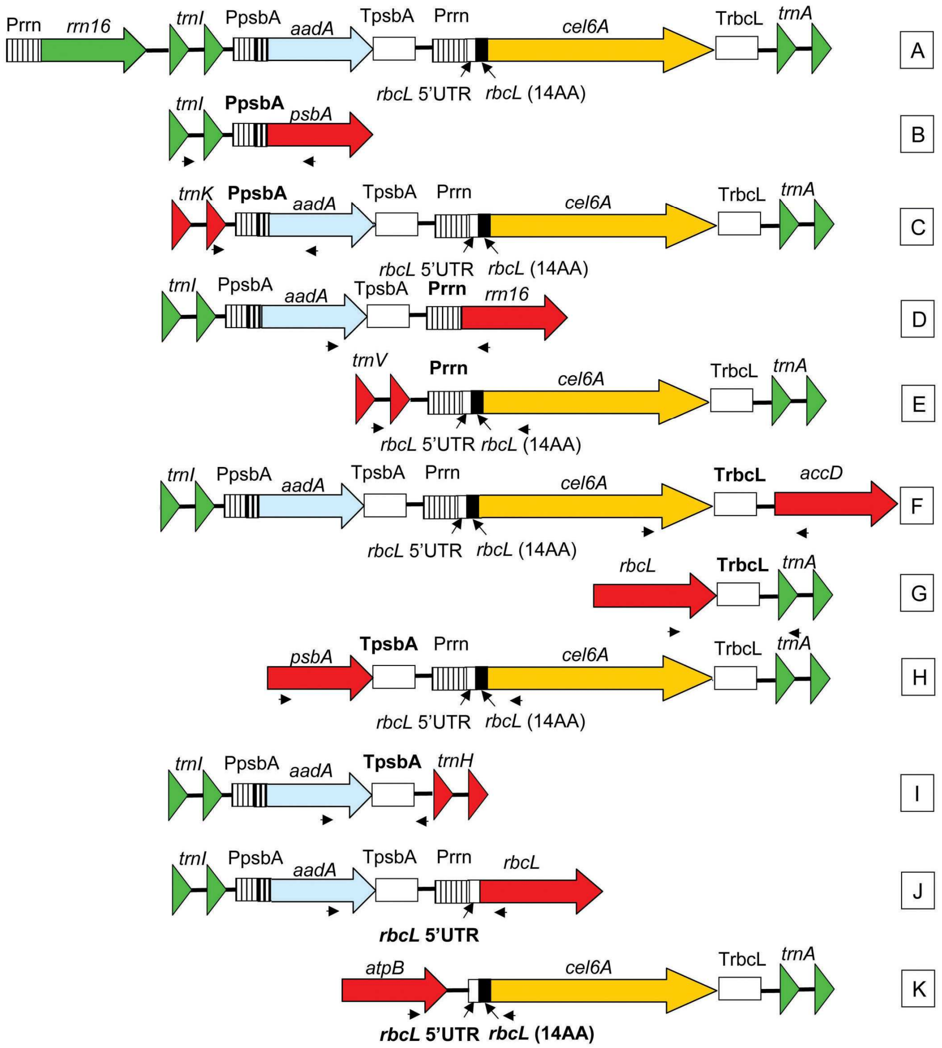


Figure 6.4 Schematic diagrams (not to scale) of the eleven DNA species identified in 22XE2 tobacco plastids. Species A is the main DNA species and was expected to form following plastid transformation. Species B-K resulted from recombination events mediated by the plastid DNA element in bold. Small arrows underneath each schematic diagram show the locations of the primers used to amplify the given species.

tobacco that differs considerably from the GFP-Cel6A and 22XE2 tobacco lines described in Figures 6.1-6.4. Figure 6.5 shows schematic diagrams of the main plastid DNA species generated by plastid transformation with the MR220 transformation vector (species A) and eight unintended plastid DNA species (species B-I) identified by PCR amplification and sequencing in seed-grown MR220 tobacco. As in GFP-Cel6A and 22XE2 tobacco, each plastid DNA element in the MR220 transformation vector mediated the formation of two additional plastid DNA species. The detection of multiple DNA species in MR220 plants indicated that the recombination events identified in GFP-Cel6A and 22XE2 tobacco were not specific to transformation of the *trnI*/*trnA* intergenic region or to transplastomic plants that accumulate Cel6A fusion proteins. Sequence data were obtained only between the primer annealing sites depicted as small arrows in Figure 6.5; MR220 DNA species G and H were also detected by DNA blotting (Figure 6.7, described below).

Unintended plastid transformation event mediated by regulatory plastid DNA elements in the transformation vector

Transplastomic plants are typically identified by PCR screening of antibiotic resistant shoots growing on selective medium following bombardment of seedlings with the plastid transformation vector. Shoots lacking the desired transgene sometimes arise on selection medium and are usually interpreted as false positive transformants resulting from mutations that give rise to antibiotic resistance (Svab and Maliga 1991). The detection of unintended recombinant DNA species in stable homoplasmic transformants (i.e., in transplastomic plants with no detectable WT plastid genomes) suggested that the regulatory elements present in plastid transformation vectors, rather than the plastid DNA sequence flanking the transgenes to be inserted into the plastid

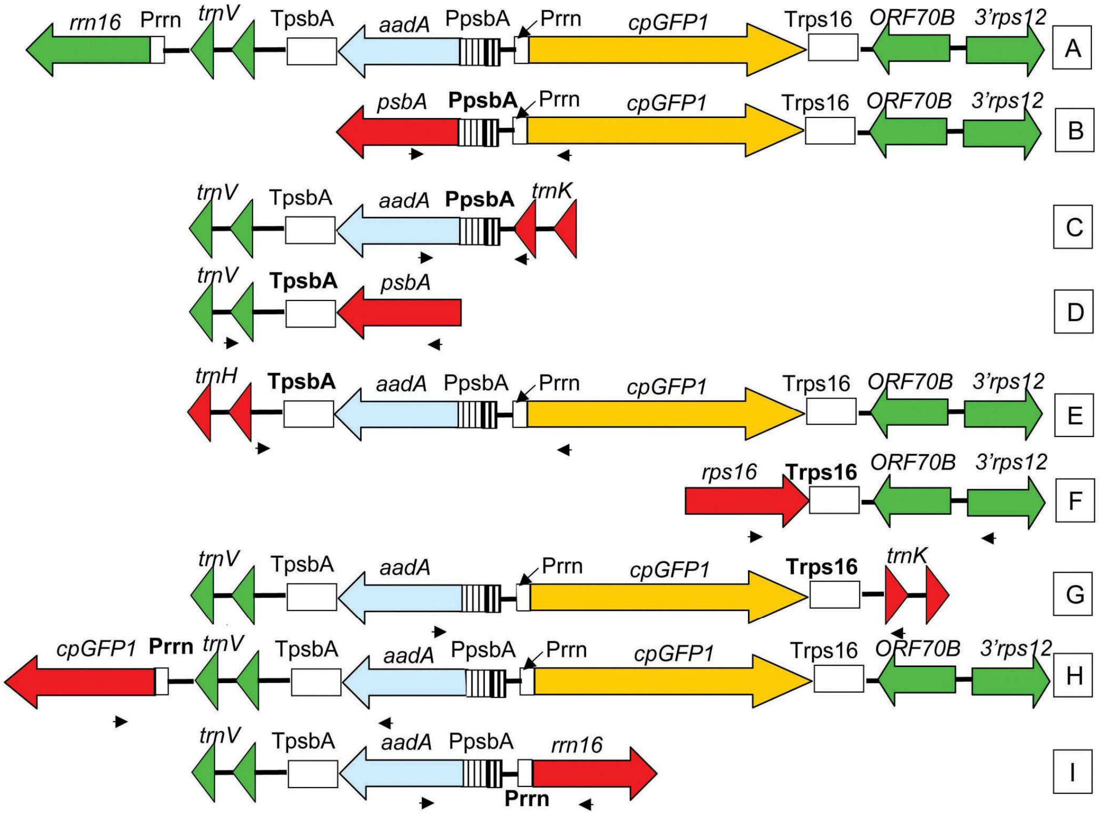


Figure 6.5 Schematic diagrams (not to scale) of the nine DNA species identified in MR220 tobacco plastids. Species A is the main DNA species and was expected to form following plastid transformation. Species B-I resulted from recombination events mediated by the plastid DNA element in bold. Small arrows underneath each schematic diagram show the locations of the primers used to amplify the given species.

genome, could mediate homologous recombination to direct the integration of an antibiotic resistance-encoding gene into an unexpected location in the plastid genome.

To test whether spectinomycin-resistant shoots shown by PCR screening to lack the *cel6A* transgene and therefore identified as false positive transformants might actually have resulted from incorporation of the antibiotic resistance gene in an unintended location of the plastid genome, further PCR analysis was performed. Three spectinomycin-resistant shoots generated from leaf pieces bombarded with the vectors described by Gray et al. (2008) were analyzed with PCR primers internal to the *aadA* gene. None of these shoots contained the *cel6A* gene, but one of these shoots contained *aadA* and was therefore named Unexpected Recombination-1 (UR-1) tobacco. As shown in Figure 6.6a, PCR indicated that *aadA* was present upstream of the *trnA* gene in UR-1, as would be expected following transformation with the vectors described by Gray et al. (2008) and demonstrated by the presence of a band amplified from TetC-Cel6A tobacco (Figure 6.6a, lane 3). The TetC-Cel6A and GFP-Cel6A tobacco lines each contained the same regulatory plastid DNA elements and were transformed at the same locus with vectors whose differences occur only in the *cel6A* ORF and do not affect homologous recombination with regulatory plastid DNA elements, as confirmed by DNA blotting and PCR analysis showing that the same recombinant DNA species were present in the same relative amounts in GFP-Cel6A and TetC-Cel6A tobacco (Gray et al. 2008 and data not shown). TetC-Cel6A DNA is therefore equivalent to GFP-Cel6A with respect to unintended homologous recombination events mediated by regulatory plastid DNA elements. A faint band in Figure 6.6a, lane 1 is likely the result of a non-specific PCR reaction in WT tobacco DNA. Conversely, Figure 6.6b shows that PCR using primers internal to the *cel6A* gene failed to amplify any product from UR-1. The PCR results depicted in Figures 6.6a and 6.6b suggested that a homologous recombination event mediated by either

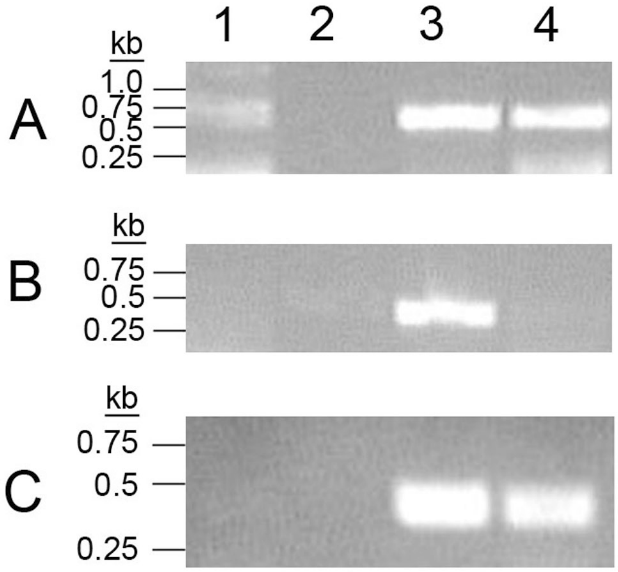


Figure 6.6 Screening for recombinant DNA species by PCR. Lane 1, WT; Lane 2, no DNA; Lane 3, TetC-Cel6A; Lane 4, UR-1. (A) PCR aadAint-fwd/trnAint-rev, amplifying a 0.6 kb region between the *aadA* and *trnA* genes; (B) PCR Cel6Aint-fwd/Cel6A-TpsbA-rev, amplifying a 0.5 kb region internal to the *cel6A* gene; (C) PCR trnKint-fwd/aadAint-rev, amplifying a 0.5 kb region between the *trnK* and *aadA* genes.

P*psbA* or T*psbA*, located between the *cel6A* and *aadA* genes in the transformation vector, may have occurred. Figure 6.6c shows PCR amplification of a 450 bp *trnK*-*aadA* fragment in TetC-Cel6A and UR-1 tobacco. No detectable PCR product was amplified from WT tobacco DNA or from a negative control lacking any DNA substrate; the positive signal in lane 3 is due to the presence of species E (Figure 6.3) in the TetC-Cel6A DNA sample. This indicated that *trnK*, normally located upstream of *psbA* in the plastid genome, was located upstream of *aadA* in UR-1 tobacco. Sequencing of the UR-1 PCR product shown in Figure 6.6c confirmed that this was species E from Figure 6.3, formed by recombination between the native P*psbA* sequence and the P*psbA* sequence present in the transformation vector. These data suggested that integration of the *aadA* gene into UR-1 tobacco was directed by homologous recombination events mediated by P*psbA* and *trnA*, located at the 5' and 3' ends of the *aadA* expression cassette, respectively, rather than by the *trnI* and *trnA* regions of DNA flanking the *cel6A* and *aadA* genes in the transformation vector plasmid.

Characterization and quantification of recombinant species in MR220, TetC-Cel6A, 22XE2, and UR-1 DNA

The strategy of PCR amplification followed by sequencing of the PCR products described above was useful for the identification, but not for the quantification, of multiple DNA species in transplastomic tobacco formed by recombination between native plastid DNA and plastid DNA elements in the transformation vector. *In vivo* abundance of some of the DNA species containing *aadA* identified via PCR in MR220, TetC-Cel6A, 22XE2 and UR-1 tobacco was determined by a DNA blot of *Xho*I/*Bam*HI-digested DNA (Figure 6.7). Equal loading of the DNA in Figure 6.7 was

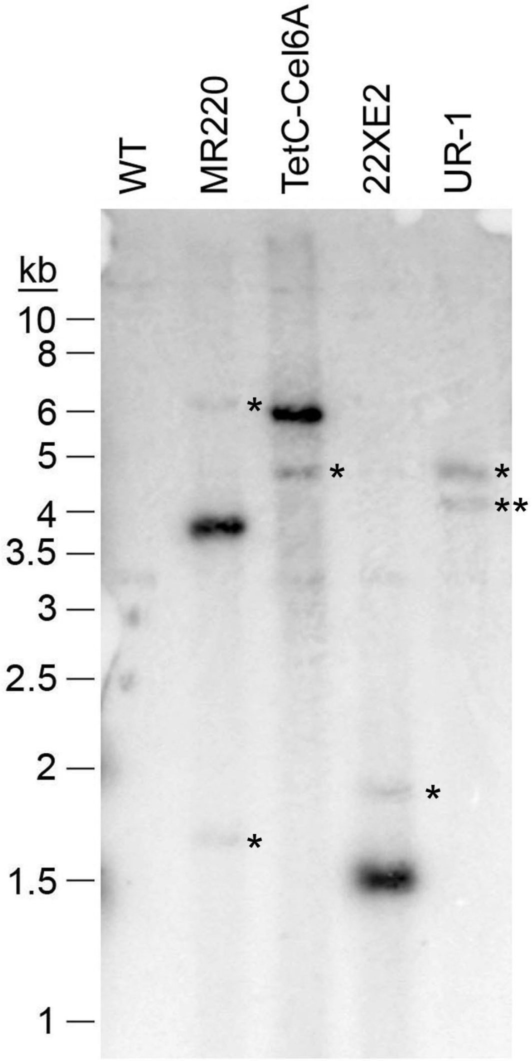


Figure 6.7 Screening for recombinant DNA species containing *aadA* on DNA blots of *Bam*HI/*Xho*I digested WT, MR220, TetC-Cel6A, 22XE2, and UR-1 tobacco DNA. The blot was hybridized with an *aadA*-specific probe. Minor DNA species formed by recombination events with plastid DNA elements are marked by asterisks (*). UR-1 Plastid DNA species H (Figure 6.3), formed from a secondary recombination event, is marked by two asterisks (**).

confirmed by visualizing the ethidium bromide-stained agarose gel used for electrophoresis (data not shown).

In addition to the expected *aadA*-containing bands found in MR220, TetC-Cel6A, and 22XE2 transplastomic tobacco lines, several less abundant bands were observed but as with GFP-Cel6A not all predicted DNA species were present at detectable levels. The sizes of observed bands were consistent with the predicted *Xho*I/*Bam*HI fragment sizes produced by the DNA species depicted in Figures 6.3-6.5. In MR220 tobacco, species G and H (6.8 and 1.7 kb, respectively) were both found at 2% of the abundance of species A (3.8 kb) whereas species C and I (4.6 and 5.9 kb, respectively) were not detected. In TetC-Cel6A tobacco, species E (4.6 kb) was detected at an abundance of 7% of species A (5.7 kb); species B, G, and H (6.0, 5.3, and 4.2 kb, respectively) were not detected. In 22XE2 tobacco, DNA species C (1.9 kb) was detected at 3% of species A (1.5 kb); species D, I, and J produced *Xho*I/*Bam*HI fragments whose sizes were equivalent to that produced by species A and therefore could not be distinguished on this blot.

Notably, species G could not be detected by DNA blotting in TetC-Cel6A tobacco grown from seed, while this species was present at 13 ± 3% of species A in T0 GFP-Cel6A tobacco (Table 6.1). This result indicates that the minor plastid DNA species formed via recombination with regulatory plastid DNA elements in the transformation vector are more abundant in the initial transformant (i.e., in the T0 plant) than in progeny plants. A drastic reduction in the concentration of several unintended DNA species in GFP-Cel6A/TetC-Cel6A tobacco between the T0 and T1 generations was observed (Supplementary Figure 6.S1a and data not shown). The *in vivo* abundance of recombinant DNA species also showed a clear positive correlation with the size of the regulatory plastid DNA elements mediating their creation by homologous recombination in both GFP-Cel6A/TetC-Cel6A and MR220 tobacco lines

(Supplementary Figure 6.S1; promoter and UTR sizes are shown in Supplementary Table 6.S3).

Figure 6.7 shows that the major *aadA*-containing band detected in *Xho*I/*Bam*HI digested T0 UR-1 DNA was found at 4.6 kb, consistent with species E in GFP-Cel6A/TetC-Cel6A tobacco (Figure 6.3). Surprisingly, a second band at approximately 4.2 kb was detected in UR-1 tobacco that was not observed in TetC-Cel6A tobacco DNA. This species was subsequently identified by PCR amplification and sequencing as described below.

Secondary recombination events

The 4.2 kb *Bam*HI/*Xho*I fragment observed in UR-1 tobacco DNA could result from recombination between species E (Figure 6.3), which appears to be the main *aadA*-containing species in UR-1 tobacco, and the native *rps16* 3'UTR (T*rps16*). Theoretically, a secondary recombination event between species G (Figure 6.3) and the native P*psbA* could also result in the formation of species H, though this is unlikely because species E is far more abundant than species G in UR-1 tobacco as determined by DNA blotting, where species G could not be detected. Figure 6.8a shows a schematic diagram of the hypothesized recombination, along with the *Bam*HI sites in the *trnK* gene that would produce a 4.2 kb fragment from this species. PCR was performed using primers trnKint-fwd/trnKint-rev to amplify a 1.9 kb fragment from both UR-1 and TetC-Cel6A tobacco that was not present in WT tobacco (Figure 6.8b). The 1.9 kb PCR products amplified from UR-1 and TetC-Cel6A tobacco were purified and sequenced, confirming the presence of the hypothesized DNA species (species H in Figure 6.3). Although this DNA species was detected by PCR and sequencing in both UR-1 and TetC-Cel6A tobacco, it could be detected by DNA

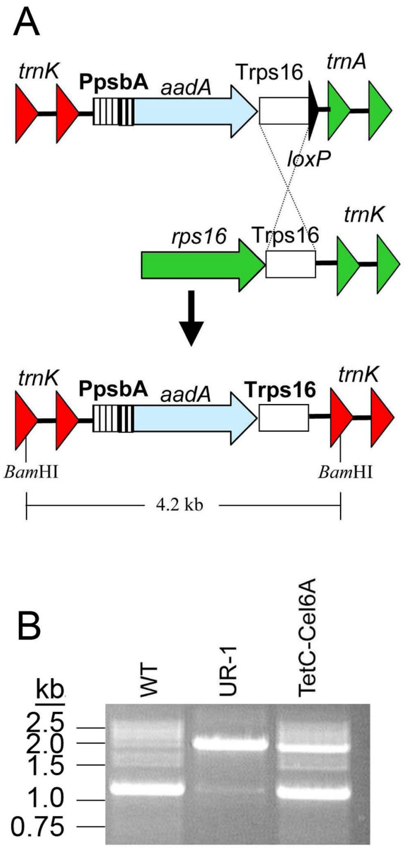


Figure 6.8 Characterization of the secondary recombination event giving rise to the 4.2 kb *aadA*-containing *Bam*HI fragment observed in UR-1 tobacco (Figure 6.7). (A) Schematic diagram of the proposed recombination event leading to the formation of this DNA species (species H in Figure 6.3), along with the location of the *Bam*HI sites in the *trnK* gene resulting in the 4.2 kb fragment observed in Figure 6.7; (B) PCR reaction with primers trnKint-fwd/trnKint-rev and WT, UR-1, and TetC-Cel6A tobacco DNA. A 1.9 kb product was PCR amplified from UR-1 and TetC-Cel6A tobacco DNA that was sequenced to confirm the presence of the DNA species depicted in Figure 6.8a.

blotting only in UR-1 tobacco, indicating a greater abundance of this DNA species in UR-1 tobacco than in TetC-Cel6A tobacco.

The ability to detect species H (Figure 6.3) in TetC-Cel6A tobacco by PCR and sequencing suggests that other secondary recombination events may occur and accumulate to very low levels in GFP-Cel6A, TetC-Cel6A, and presumably 22XE2, MR220, and other transplastomic tobacco lines. PCR amplification of these other hypothesized "secondary recombination" DNA species was not attempted.

DISCUSSION

Plastid genomes in higher plants are known to undergo *in vivo* homologous recombination in both wild-type and transplastomic plants. Reflecting their bacterial origin, plastids have maintained the RecA-mediated recombination system, with a plastid-targeted homolog to the *E. coli recA* gene identified in *Arabidopsis thaliana* (Cerutti et al. 1992) and in *Chlamydomonas reinhardtii* (Nakazato et al. 2003). Homologous recombination is responsible for plastid genome rearrangements and deletions (e.g., Kawata et al. 1997) and has been implicated in the evolution of *trnF* pseudogenes in *Arabidopsis lyrata* (Ansell et al. 2007). Intermolecular recombination in plastids can remove deleterious mutations and may be responsible for the lower mutation rate in plastid genomes than in plant nuclear genomes (Khakhlova and Bock 2006). Homologous recombination is required for plastid DNA repair (Odom et al. 2008). Intermolecular recombination events among the multiple copies of the plastid genome and/or intramolecular recombination events between inverted repeat regions may be responsible for the many forms of the plastid genome observed by fiber-FISH analysis (Lilly et al. 2001). Homologous recombination is the mechanism that allows for integration of transgenes into the plastid genome (Staub and Maliga 1992).

Plastid DNA (e.g., plastid promoters, 5'UTRs, and 3'UTRs) is often incorporated into plastid transformation vectors for gene regulation. In all lines of transplastomic tobacco tested here, each plastid DNA element incorporated into the plastid transformation vector mediated the creation of at least two unintended DNA species in transformed plastids. As the Cel6A- and GFP-expressing transplastomic tobacco lines tested here were phenotypically indistinguishable from WT tobacco, it is unclear whether these plastid DNA species play any physiological role in the plant. Corneille et al. (2003) also observed multiple unintended recombination events in transgenic plastids following CRE-*lox* recombination. In this case, the DNA species formed following these recombination events were not observed in plants grown from seed. This is consistent with the decrease in unintended DNA species in transplastomic tobacco grown from seed relative to T0 plants observed here. Intentional recombination events such as those occurring during plastid transformation or CRE-*lox* removal of a marker gene may promote further recombination events in the plastid. The decrease in the abundance of these unintended DNA species in seed-grown plants suggests that the physiological effects of the unintended recombination events described here, if any, would tend to decrease in subsequent generations.

PCR-based recombination (Bradley and Hillis 1997) cannot be completely ruled out in the amplification of some of the DNA species diagrammed in Figures 6.3-6.5. In this case, a given promoter or UTR element could act as a primer for PCR amplification as in overlap extension PCR (Ho et al. 1989). Many of the DNA species diagrammed in Figures 6.3-6.5 were also detected by DNA blotting, however, in the absence of any DNA amplification step. PCR-based recombination of DNA species cannot be ruled out but the supporting evidence provided by the DNA blots argues against this mechanism as a primary explanation for the observed results.

In addition to the potential for either ‘loop-out’ or ‘flip-flop’ recombination mediated by homologous DNA sequences in direct or inverse orientation, respectively, on a given DNA molecule, recombination can occur between homologous sequences on two separate DNA molecules, resulting in the creation of multimeric plastid genomes through a process similar to bacterial plasmid cointegrate formation (Peterson et al. 1982). Cointegrates have been observed previously in plastids (Staub and Maliga 1994, 1995; Klaus et al. 2004). Experiments using the fiber-FISH technique with isolated chloroplast DNA (Lilly et al. 2001) detected multimeric plastid genomes that may have formed via homologous recombination among the multiple copies of the plastid genome. Similar experiments using plastid DNA isolated from transplastomic plants could reveal the size and structure of transgene-containing plastid DNA, and could potentially also reveal the presence of novel recombinant DNA species not detected in this study. Further study of the structure of the DNA molecules formed from the homologous recombination events described here will help to elucidate the recombination mechanisms (i.e., ‘flip-flop’ vs. ‘loop-out’ and intermolecular vs. intramolecular recombination).

Spectinomycin-resistant shoots not containing the transgene of interest can arise from ribosomal point mutations and, as we have shown here, from unintended homologous recombination events (e.g, UR-1). Plastid transformation vectors are typically designed so that the antibiotic resistance gene and the gene of interest are linked. The UR-1 line of tobacco described here appears to have been formed from two recombination events; one mediated by the *trnA* flanking DNA, located downstream of *aadA* in the transformation vector and included for this purpose, and the other mediated by P*psbA*, located between *aadA* and the gene of interest in the transformation vector. The antibiotic resistance gene is thus decoupled from the gene of interest. Thorough characterization of transplastomic plants is clearly necessary to avoid further analysis of transgenic plants lacking chloroplast genomes with the desired transgene configuration.

The data described here suggest that the use of regulatory plastid DNA elements should be minimized in plastid transformation vectors in order to avoid this phenomenon and to promote the recovery of the intended plastid transformant. Unintentional recombination events could be reduced by using the smallest possible plastid DNA elements in transformation vectors. To completely avoid unintended recombination events mediated by regulatory plastid DNA elements, plastid vectors can be designed with regulatory DNA sequences from phages (e.g., the T7g10 5'UTR), bacteria (Newell et al. 2003), and/or mitochondria (Bohne et al. 2007) lacking significant similarity to plastid DNA sequences. The use of read-through transcription for transgenes rather than inclusion of a plastid promoter element (Gray et al. 2008; Chakrabarti et al. 2006) also decreases the number of plastid DNA elements available to mediate unintended recombination events. Alternatively, plastid transformation vectors can be designed with a T7 promoter in conjunction with a plastid-targeted, nuclear-encoded T7 RNA polymerase (McBride et al. 1994) or with a eubacterial promoter in conjunction with a plastid-targeted, nuclear-encoded sigma factor (Buhot et al. 2006).

Because all regulatory plastid DNA elements in all transplastomic lines tested mediated two homologous recombination events that were detectable by PCR, though not always by DNA blotting, it is proposed that all transplastomic plants contain a low level of unintended plastid DNA species. This is supported by the many observations of unintended DNA species in transplastomic tobacco lines transformed at multiple different loci and with multiple transgenes (e.g., Svab and Maliga 1993; Staub and Maliga 1994; Corneille et al. 2003; Rogalski et al. 2006, 2008a, 2008b; McCabe et al. 2008; Zhou et al. 2008). This suggests that transplastomic plants are more accurately described as ‘fully transformed’ rather than ‘homoplasmic’ to indicate that no WT plastome copies remain, though a non-homogenous plastid DNA population is likely to exist in all transplastomic tobacco lines containing regulatory plastid DNA elements.

The results presented here document the effects of an active plastid recombination system that serves to maintain a genome that is both relatively static at the nucleotide level (i.e., a low mutation rate relative to that of the nucleus) due to gene conversion that depends on intermolecular recombination among plastid DNA molecules (Khakhlova and Bock 2006) and relatively flexible at the macromolecular level (Lilly et al. 2001 and this report). Our data indicates that the design of plastid transformation vectors can be improved by the use of regulatory DNA elements not derived from plastid DNA sequences.

ACKNOWLEDGEMENTS

This work was supported by the United States Department of Agriculture [USDA NRI 2007-02133 to B.A.A. and M.R.H.]; and the National Science Foundation [Graduate Research Fellowship, B.N.G.].

APPENDIX

Supplementary Table 6.S1. Primers used in this study.

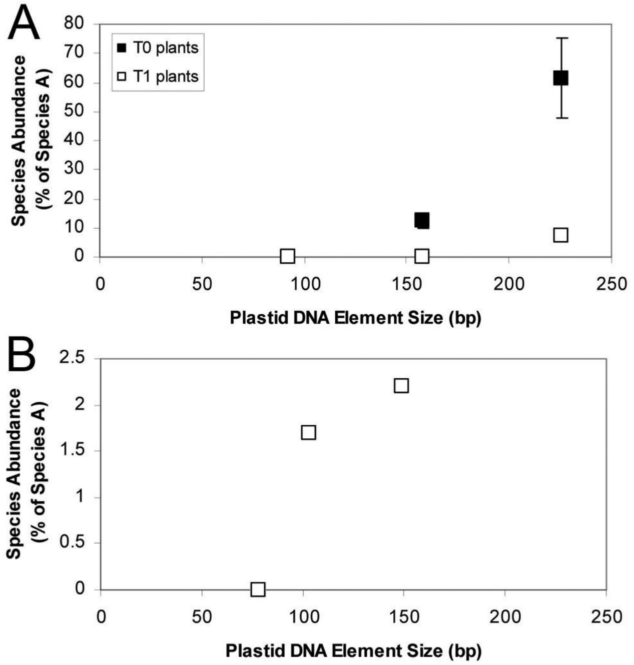
Primer Name	Primer Sequence
Iprobe-fwd	CACAGGTTTAGCAATGGG
Iprobe-rev	GAAGTAGTCAGATGCTTC
Aprobe-fwd	ATAGTATCTTGTACCTGA
Aprobe-rev	TAAAGCTTTGTATCGGCTA
psbAprobe-fwd	GAGACGCGAAAGCGAAAG
psbAprobe-rev	AGTACCAGAGATTCCTAG
Cel6Aint-fwd	GTAACGAGTGGTGCGACC
Cel6A-TpsbA-rev	ATAGACTAGGCCAGGATCGCGGCCGCTCAGCTGGCGGCGCAGGT
GFPprobe-fwd	TTCAAAAGGTGAAGAATT
GFPprobe-rev	TCTTCGATATTATGTCTG
aadAprobe-fwd	CGTGAAGCGGTTATCGCC
aadAprobe-rev	GTCCAAGATAAGCCTGTC
aadAint-rev	GCCAACTACCTCTGATAG
aadAint-fwd	ACGCTATGGAACTCGCCG
trnAint-rev	TCAGGTACAAGATACTAT
trnKint-fwd	AATCAACTGAGTATTCAA
trnKint-rev	AAAGAGACTAGCCGCACT
trnIint-fwd	CTGGGGTGACGGAGGGAT
Cel6Aint-rev	TGCTGTGGTTGCCGCAGT
Cel6Aint-fwd2	ACAACGGAACGCTCTCCC
trnHint-rev	AGTCTATGTAAGTAAAAT
trnKint-fwd	TTATCAGATTCTGATATTAT
rps16int-fwd	GAGCCGTCTATCGAATCG
16sint-rev	ATGTGTTAAGCATGCCGC
trnVint-fwd	CAGTTCGAGCCTGATTAT
accDint-rev	ATACAATAGATGAATAGT
rbcLint-fwd	CATGGTATCCACTTCCGG
rbcLint-rev	TAAGTCAATTTGTACTCT
atpBint-rev	AGAACCAGAAGTAGTAGG
GFPint-rev	TAATTTACCATATGTAGCA
trnVint-rev	ACACTCTACCGCTGAGTT
GFPint-fwd	ATCATTACTTAAGTACAC
rps12int-rev	TGGCAATGTAGTTGGACT

Supplementary Table 6.S2. Primer pairs used for PCR amplification of plastid DNA species

Species	DB-Cel6A primers	22XE2 primers	MR220 primers
A	n/a	n/a	n/a
B	psbAprobe-fwd/aadAint-rev	trnIint-fwd/ psbAprobe-rev	psbAprobe-rev/GFPint-rev
C	Cel6Aint-fwd2/trnHint-rev (small product)	trnKint-fwd/aadAint-rev	aadAint-rev/trnKint-fwd
D	Cel6Aint-fwd2/trnHint-rev (large product)	aadAint-fwd/16sint-rev	trnVint-rev/psbAprobe-fwd
E	trnKint-fwd/aadAint-rev	trnVint-fwd/Cel6Aint-rev	GFPint-fwd/trnKint-rev
F	Rps16int-fwd/trnAint-rev	Cel6Aint-fwd2/accDint-rev	rps16int-fwd/rps12int-rev
G	aadAint-fwd/trnKint-rev	rbcLint-fwd/trnAint-rev	trnHint-rev/aadAint-fwd
H	trnKint-fwd/trnKint-rev	psbAprobe-fwd/Cel6Aint-rev	GFPint-rev/aadAint-fwd
I		aadAint-fwd/trnHint-rev	aadAint-rev/16sint-rev
J		aadAint-fwd/rbcLint-rev	
K		atpBint-rev/Cel6Aint-rev	

Supplementary Table 6.S3. Sizes (in bp) of plastid DNA elements in transformation vectors

DNA element	DB-Cel6A	22XE2	MR220
P*psbA*	226	170	78
T*psbA*	92	104	189
T*rps16*	158		149
P*rrn*		97	103
rbcL 5'UTR + DB		97	
T*rbcL*		149	



Supplementary Figure 6.S1 Abundance of several plastid DNA species (relative to species A, the DNA species expected to be formed following plastid transformation) as a function of the size of the plastid DNA element giving rise to that species. (A) GFP-Cel6A/TetC-Cel6A tobacco, showing DNA species abundance in both T0 (filled squares) and T1 (open squares) generations; (B) MR220 tobacco.

REFERENCES

Ansell SW, Schneider H, Pedersen N, Grundmann M, Russell SJ, Vogel JC (2007) "Recombination diversifies chloroplast *trnF* pseudogenes in *Arabidopsis lyrata*" *J Evolution Biol* **20**:2400-2411.

Blowers AD, Bogorad L, Shark KB, Sanford JC (1989) "Studies on *Chlamydomonas* chloroplast transformation: foreign DNA can be stably maintained in the chromosome" *Plant Cell* **1**:123-132.

Bohne AV, Ruf S, Borner T, Bock R (2007) "Faithful transcription initiation from a mitochondrial promoter in transgenic plastids" *Nucleic Acids Res* **35**:7256-7266.

Bradley RD, Hillis DM (1997) "Recombinant DNA sequences generated by PCR amplification" *Mol Biol Evol* **14**:592-593.

Buhot L, Horvàth E, Medgyesy P, Lerbs-Mache S (2006) "Hybrid transcription system for controlled plastid transgene expression" *Plant J* **46**:700-707.

Cerutti H, Osman M, Grandoni P, Jagendorf AT (1992) "A homolog of *Escherichia coli* RecA protein in plastids of higher plants" *P Natl Acad Sci USA* **89**:8068-8072.

Chakrabarti SK, Lutz KA, Lertwiriyawong B, Svab Z, Maliga P (2006) "Expression of the cry9Aa2 B.t. gene in tobacco chloroplasts confers resistance to potato tuber moth" *Transgenic Res* **15**:481-488.

Corneille S, Lutz KA, Azhagiri AK, Maliga P (2003) "Identification of functional *lox* sites in the plastid genome" *Plant J* **35**:753-762.

Gray BN, Ahner BA, Hanson MR (2009) "High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions" *Biotechnol Bioeng* **102**: 1045-1054.

Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR (1989) “Site-directed mutagenesis by overlap extension using the polymerase chain reaction” *Gene* **77**:51-59.

Iamtham S, Day A (2000) “Removal of antibiotic resistance genes from transgenic tobacco plastids” *Nat Biotechnol* **18**:1172-1176.

Kanno A, Watanbe N, Nakamura I, Hirai A (1993) “Variations in chloroplast DNA from rice (*Oryza sativa*): differences between deletions mediated by short direct-repeat sequences within a single species” *Theor Appl Genet* **86**:579-584.

Kawata M, Harada T, Shimamoto Y, Oono K, Takaiwa F (1997) “Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs)” *Curr Genet* **31**:179-184.

Khakhlova O, Bock R (2006) “Elimination of deleterious mutations in plastid genomes by gene conversion” *Plant J* **46**:85-94.

Khan MS, Hameed W, Nozoe M, Shiina T (2007) “Disruption of the *psbA* gene by the copy correction mechanism reveals that the expression of plastid-encoded genes is regulated by photosynthesis activity” *J Plant Res* **120**:421-430.

Klaus SMJ, Huang F-C, Golds TJ, Koop H-U (2004) “Generation of marker-free plastid transformants using a transiently cointegrated selection gene” *Nat Biotechnol* **22**:225-229.

Li W, Ruf S, Bock R (2006) “Constancy of organellar genome copy numbers during leaf development and senescence in higher plants” *Mol Genet Genomics* **275**:185-192.

Lilly JW, Havey MJ, Jackson SA, Jiang J (2001) “Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants” *Plant Cell* **13**:245-254.

McBride KE, Schaaf DJ, Daley M, Stalker DM (1994) “Controlled expression of plastid transgenes in plants based on a nuclear DNA-encoded and plastid-targeted T7 RNA polymerase” *P Natl Acad Sci USA* **91**:7301-7305.

McCabe MS, Klaas M, Gonzalez-Rabade N, Poage M, Badillo-Corona JA, Zhou F, Karcher D, Bock R, Gray JC, Dix PH (2008) “Plastid transformation of high-biomass tobacco variety Maryland Mammoth for production of human immunodeficiency virus type 1 (HIV-1) p24 antigen” *Plant Biotechnol J* **6**:914-929.

Nakazato E, Fukuzawa H, Tabata S, Takahashi H, Tanaka K (2003) “Identification and expression analysis of cDNA encoding a chloroplast recombination protein REC1, the chloroplast RecA homologue in *Chlamydomonas reinhardtii*” *Biosci Biotech Bioch* **67**:608-2613.

Newell CA, Birch-Machin I, Hibberd JM, Gray JC (2003) “Expression of green fluorescent protein from bacterial and plastid promoters in tobacco chloroplasts” *Transgenic Res* **12**:631-634.

Odom OW, Baek K-H, Dani RN, Herrin DL (2008) “*Chlamydomonas* chloroplasts can use short dispersed repeats and multiple pathways to repair a double-strand break in the genome” *Plant J* **53**:842-853.

Palmer JD (1983) “Chloroplast DNA exists in two orientations” *Nature* **301**:92-93.

Peterson BC, Hashimoto H, Rownd RH (1982) “Cointegrate formation between homologous plasmids in *Escherichia coli*” *J Bacteriol* **151**:1086-1094.

Reed ML, Wilson SK, Sutton CA, Hanson MR (2001) “High-level expression of a synthetic red-shifted GFP coding region incorporated into transgenic chloroplasts” *Plant J* **27**:257-265.

Rogalski M, Ruf S, Bock R (2006) “Tobacco plastid ribosomal protein S18 is essential for cell survival” *Nucleic Acids Res* **34**:4537-4545.

Rogalski M, Karcher D, Bock R (2008a) “Superwobbling facilitates translation with reduced tRNA sets” *Nat Struct Mol Biol* **15**:192-198.

Rogalski M, Schöttler MA, Thiele W, Schulze WX, Bock R (2008b) “Rpl33, a nonessential plastid-encoded ribosomal protein in tobacco, is required under cold stress conditions” *Plant Cell* **20**:2221-2237.

Ruf S, Biehler K, Bock R (2000) “A small chloroplast-encoded protein as a novel architectural component of the light-harvesting antenna” *J Cell Biol* **149**:369-378.

Staub JM, Maliga P (1992) “Long regions of homologous DNA are incorporated into the tobacco plastid genome by transformation” *Plant Cell* **4**:39-45.

Staub JM, Maliga P (1994) “Extrachromosomal elements in tobacco plastids” *P Natl Acad Sci USA* **91**:7468-7472.

Staub JM, Maliga P (1995) “Marker rescue from the *Nicotiana tabacum* plastid genome using a plastid/*Escherichia coli* shuttle vector” *Mol Gen Genet* **249**:37-42.

Svab Z, Maliga P (1991) “Mutation proximal to the tRNA binding region of the *Nicotiana* plastid 16S rRNA confers resistance to spectinomycin” *Mol Gen Genet* **228**:316-319.

Svab Z, Maliga P (1993) “High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene” *P Natl Acad Sci USA* **90**:913-917.

Yu LX, Gray BN, Rutzke CJ, Walker LP, Wilson DB, Hanson MR (2007) “Expression of thermostable microbial cellulases in the chloroplasts of nicotine-free tobacco” *J Biotechnol* **131**:326-369.

Zhou F, Badillo-Corona J, Karcher D, Gonzalez-Rabade N, Piepenburg K, Borchers AM, Maloney AP, Kavanagh TA, Gray JC, Bock R (2008) “High-level expression of human immunodeficiency virus antigens from the tobacco and tomato plastid genomes" *Plant Biotechnol J* **6**:897-913.

Chapter 7: Conclusions

The goals of my thesis work were to express the *T. fusca* Cel6A and BglC proteins at a high level in tobacco chloroplasts, and to elucidate the mechanism of downstream box (DB) function. High-level production of both Cel6A and BglC proteins was achieved, and some progress was made toward understanding the mode of action of the DB region, though further questions remain as to the mechanism of DB function.

The experiments described here demonstrate that changes in the DB region of an ORF of interest can mediate order of magnitude changes in protein (i.e., Cel6A and BglC) accumulation when expressed from the tobacco plastid genome. Both non-silent (Chapters 2 and 3) and silent (Chapters 4 and 5) mutations in the DB region were found to cause large changes in foreign protein accumulation. In addition to affecting protein accumulation, non-silent changes to the DB region fused to the *cel6A* ORF were shown to affect the accumulation of monocistronic *cel6A* transcript (Chapter 2). Neither silent (Chapter 5) nor non-silent (Chapter 3) changes to the DB region fused to the *bglC* ORF affected transcript levels as judged by RNA blotting, though a close examination of the monocistronic transcripts revealed partial degradation of the RNAs (Chapter 3). It appears that the changes in transcript abundance are not the primary cause of the differences in protein accumulation among the various DB-Cel6A and DB-BglC proteins tested, but that RNA degradation is an effect of altered translation efficiency (Chapters 3 and 4).

It was hypothesized that a given DB region would be useful for enhancing foreign protein accumulation for a number of different proteins. This hypothesis was tested by fusing the TetC, NPTII, and GFP DB regions to both the *cel6A* and *bglC* ORFs (Chapters 2 and 3). Surprisingly, the TetC DB region was the most effective DB region for high-level Cel6A production from the plastid genome, while the NPTII

DB region was more effective than the TetC DB region for high-level BglC production. It is not immediately apparent why these differences exist. It was hypothesized that RNA secondary structure differences could partially explain why one DB region worked best with *cel6A*, while another DB region worked best with *bglC,* but secondary structure predictions did not show any obvious differences among the *tetC-cel6A*, *nptII-cel6A*, *tetC-bglC*, and *nptII-bglC* transcripts that could readily explain the observed differences in protein accumulation (data not shown). This remains an open question that will require further testing to explain.

The use of DB regions to enhance foreign protein production in transplastomic plants thus far has involved empirical trial-and-error testing of DB regions. One strategy for this type of testing is to choose the DB region of a gene that has been expressed effectively from the plastid genome (e.g., TetC, NPTII, or GFP; Chapters 2 and 3), and then fuse that DB region to the ORF of interest. A less empirical method of testing DB fusions is desirable. It was hypothesized that codon usage in the DB region could partially explain the differences in protein accumulation when various DB regions are fused to ORFs of interest. DB regions with silent mutations were therefore constructed and fused to the *cel6A* (Chapters 4 and 5) and *bglC* (Chapter 5) ORFs. By altering the codon usage of the GFP DB region to more closely match the codon usage in the DB regions of highly expressed plastid genes, GFP-Cel6A protein accumulation in transplastomic tobacco was increased 30-fold. This type of DB optimization (i.e., using codons preferred by the DB region as opposed to the genome as a whole) was also shown to be effective in *E. coli*, suggesting that DB codon usage could be a major determinant of DB function (Chapter 4). Synthetic DB regions were constructed from codons that are preferred (HF[NtDB]Syn DBs) or avoided (LF[NtDB]Syn DBs) in the DB regions of highly expressed plastid genes. These DB regions were fused to the *cel6A* and *bglC* ORFs. Surprisingly, the LF(NtDB)Syn DB region mediated higher accumulation of BglC protein than the HF(NtDB)Syn DB region, though BglC protein accumulation was modest (Chapter 5). The LF(NtDB)Syn DB region mediated modest accumulation of Cel6A protein, but a tobacco transformant containing the *HF(NtDB)Syn-cel6A* ORF could not be obtained (Chapter 5). It is hypothesized that this may be due to extremely high level production of HF(NtDB)Syn-Cel6A protein, hindering the recovery of a tobacco transformant.

From the data obtained, it is proposed that codon usage in the DB region of the *cel6A* ORF may be an important determinant of Cel6A protein accumulation in transplastomic tobacco. For BglC expression, DB codon usage does not appear to be as important a determinant of protein accumulation (Figure 7.1). A comparison of Cel6A and BglC protein accumulation (Chapters 2-5) as a function of the median NtDB CUF of the DB region fused to the *cel6A* or *bglC* ORF shows that the highest Cel6A accumulation (i.e., TetC-Cel6A) resulted from the DB region with the highest median NtDB CUF tested. In contrast, the highest BglC accumulation achieved (i.e., NPTII-BglC) did not result from the DB region with the highest median NtDB CUF. The two DB regions with the highest NtDB CUFs (i.e., the TetC and HF[NtDB]Syn DB regions) tested resulted in only moderate accumulation of BglC protein. While the reasons for the observed differences are unknown, it appears that increasing the NtDB CUF of the DB region fused to some ORFs (e.g., the *cel6A* ORF) can have a beneficial effect on protein accumulation. The CUF of the DB region of interest cannot fully explain the observed differences in protein accumulation, however, as demonstrated by the lack of a correlation between NtDB CUF in the DB region and BglC protein accumulation. Some other aspect of the DB region, possibly including RNA secondary structure or codon pair usage, is likely to also be important in determining the effect of a DB fusion to the ORF of interest. Speculatively, folding of the N-terminal region of the BglC protein could require a less efficiently translated DB

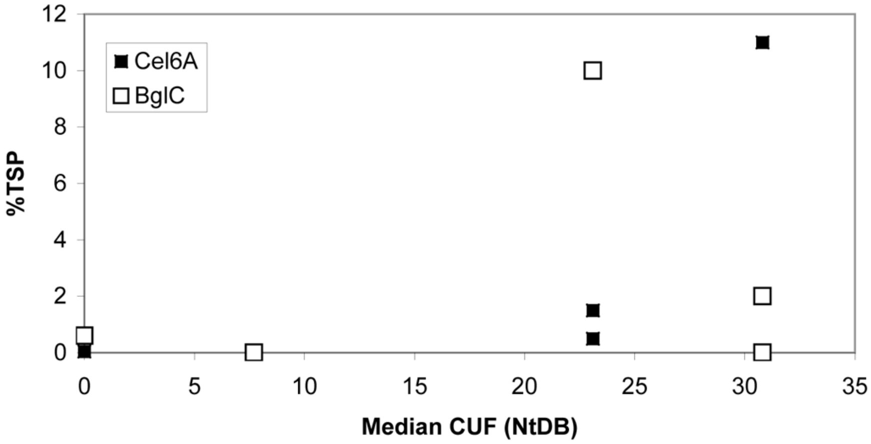


Figure 7.1: Cel6A and BglC protein accumulation in transplastomic tobacco as a function of median *N. tabacum* plastid (NtDB) CUF of the DB region fused to the *cel6A* or *bglC* ORF.

region. By including codons that are often avoided by the DB regions of highly expressed genes, translation of the DB region could be slowed, allowing for proper folding of the BglC protein. Folding of the N-terminal region of the Cel6A protein could be more efficient than folding of the BglC N-terminal region, allowing for efficient production of Cel6A protein from DB regions that are translated very efficiently.

Further experiments in transplastomic tobacco and in bacterial systems (e.g., in *E. coli*) will help to elucidate the mechanism of DB function, allowing for less empirical use of DB fusions. From the experiments described here, it can be concluded that the codon usage frequency of the DB region can partially explain the effects of DB fusions to at least some ORFs. The CUF of the DB regions of highly expressed genes in the foreign protein expression host of choice appears to be more important than the CUF of the genome as a whole (Chapter 4).

Appendix: Tobacco Chloroplast Transformation

Seed Sterilization (all procedures should be done in the sterile hood)

1. In an eppendorf tube, wash seeds in 1 mL of 100% EtOH for 2 minutes while constantly mixing.
2. Pipette off the ethanol using a P-200 pipette tip (P-1000 tips will pick up the tobacco seeds).
3. Wash seeds in 1 mL of 10% (v/v) bleach solution for 15 minutes while constantly mixing.
4. Pipette off the bleach solution with P-200 tip.
5. Wash the seeds four times in 1 mL of sterile ddH_2O.
6. Resuspend in 0.6 mL of sterile ddH_2O and dump it in the lid of a petri dish (this makes it easier to pick up the seeds with forceps).
7. Immediately plate 20-25 seeds per plate (Figure A.1) on an shallow MS-agar plate (do not store the sterilized seeds), wrap the plates in parafilm, and transfer the plates to the growth room.

Bombardment

1. Grow a 100-mL *E. coli* culture containing the DNA to be bombarded. Prepare the DNA by a 100-mL midiprep, then check the OD260/280 and dilute to 1 mg/mL and coat 0.6 μm gold beads with the DNA.
2. Once the tobacco seedlings have grown for approximately two weeks and the first true leaves are visible (approx. 2 mm), they are ready to be bombarded (Figure A.2).
3. Start bombardment with the plate in the second slot below the screen (Figure A.3) and adjust position for next shots, one slot higher or lower. If the plate is too close

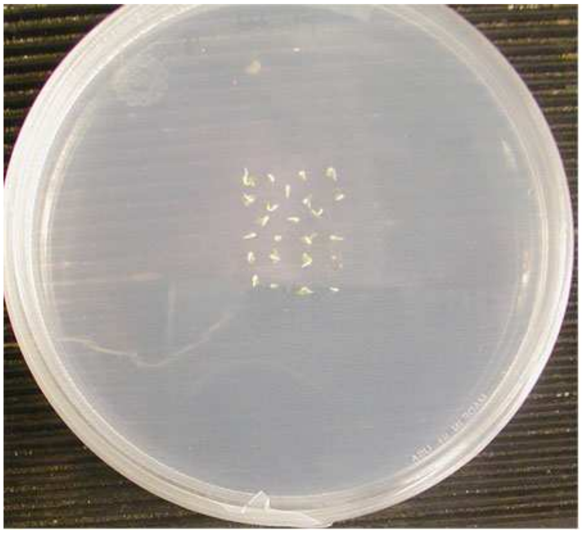


Figure A.1: Tobacco seedlings on non-selective MS agar medium, plated at a spacing and location appropriate for bombardment.

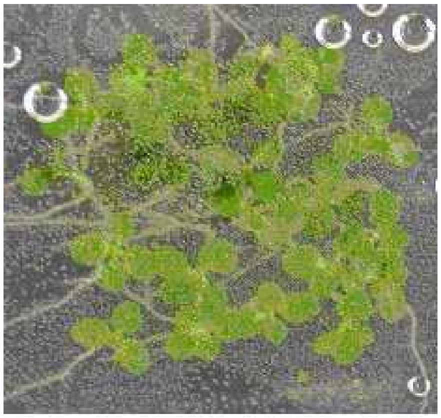


Figure A.2: Tobacco seedlings ready for bombardment, with cotyledons and the first pair of true leaves visible.

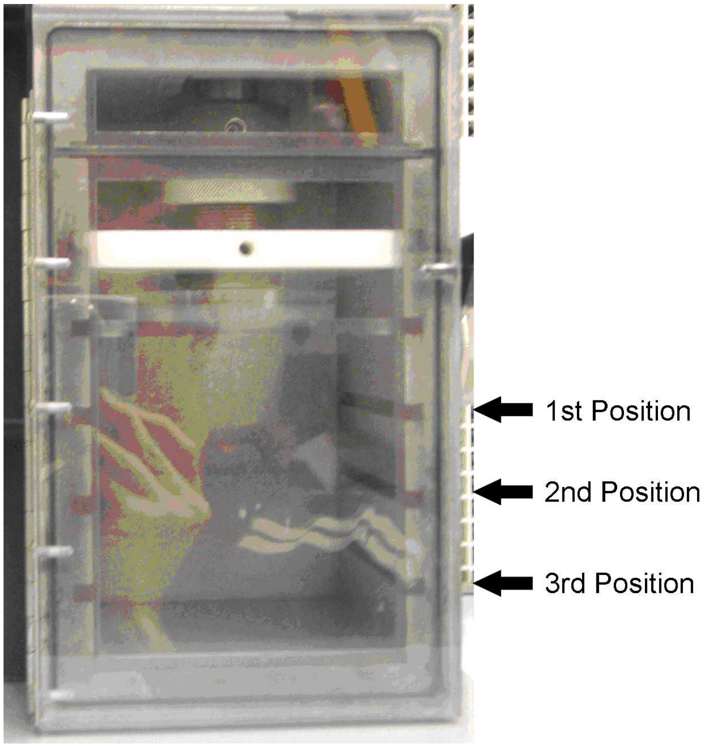


Figure A.3: Photograph of the gene gun, showing the first, second, and third positions that can be used for bombardment.

to the screen, the gold beads will only bombard within a small radius at the center of the plate and could blow the leaves off the stems. If the plate is too far from the screen, however, the beads may not be moving quickly enough to penetrate the leaves.

   a. In my hands, using the short tissue culture plates, the first slot below the screen seems to give good penetration of the leaves without destroying the plants. The plants in the very center of the plate are destroyed (leaves turn brown within a few days), but just outside from the center, the plants are not too badly injured. The tall tissue culture plates do not fit in the first position.

4. Once the plates have been bombarded, wrap the plates in parafilm and allow the plants to grow for 2-3 days in the growth room before transferring to selective medium.

Transfer of Leaves to Selective Medium

1. The leaves should be cut in half with a flame-sterilized sharp scalpel in the sterile hood. If the scalpel is too dull, the leaves will be damaged in this process. A sterile tissue culture plate can be used as a ‘cutting board’ to cut the leaves, and should be replaced occasionally to avoid contamination. The leaves need to be cut in order to take up nutrients from the RMOP medium.
2. All leaves from the bombarded plates should be transferred to RMOP medium containing spectinomycin at 500 μg/mL. It is not critical which side of the leaf is facing up (Figure A.4).
3. Wrap the RMOP plates in parafilm and transfer the plates to the growth room for 3-8 weeks until potential transformants can be identified and genotyped.

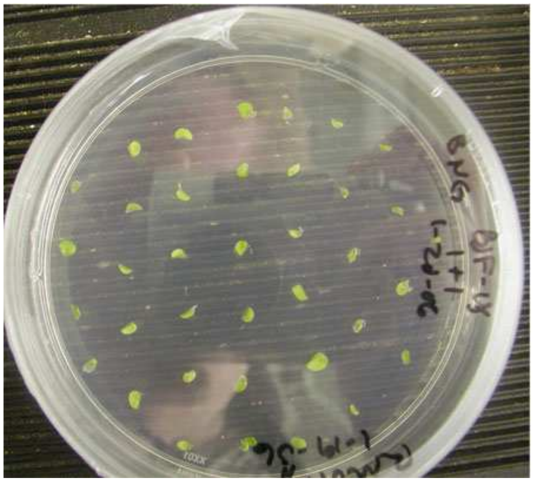


Figure A.4: Bombarded leaf pieces that have been cut and transferred to selective RMOP medium containing spectinomycin.