The origin and characterisation of new nuclear genes originating from a cytoplasmic organellar genome

Andrew Henry Mark Lloyd

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics

School of Molecular and Biomedical Science

The University of Adelaide

December 2010

Table of Contents

Abstract

Endosymbiotic transfer of DNA and functional genes from the cytoplasmic organelles (mitochondria and chloroplasts) to the nucleus has been a major factor driving the origin of new nuclear genes, a process central to eukaryote evolution. Recent developments have allowed the experimental reconstruction of DNA transfer and functional gene transfer, enabling investigation of the molecular mechanisms involved in these important evolutionary processes.

To simulate the process of functional endosymbiotic gene transfer, a chloroplast reporter gene *aadA,* which had been transferred from the chloroplast to the nucleus, was monitored for nuclear activation. In total 16 plant lines were screened, each line representing an independent nuclear insertion of the *aadA* gene. For each line ~50 million cells were screened resulting in three plants being recovered in which *aadA* showed strong nuclear activation. Activation occurred by acquisition of the CaMV 35S nuclear promoter or by nuclear activation of the native chloroplast promoter. Two fortuitous sites resident within the 3' UTR of *aadA* mRNA both promoted poly-adenylation without any sequence change. In addition, cryptic nuclear activity of the chloroplast promoter was revealed which became conspicuous when present in multiple nuclear copies.

To determine the method of chloroplast DNA insertion into the nucleus the insertion site was sequenced in line kr2.2. Complete characterisation of the nuclear sequence before and after gene transfer demonstrated simultaneous insertion of multiple chloroplast DNA fragments *via* synthesis dependent non-homologous end joining, probably at a site of double strand break (DSB) repair.

To further investigate the role of DSB repair in the nuclear insertion of organelle DNA, DSBs were induced at a specific nuclear location using the rare-cutting endonuclease I-SceI and the resulting repair events were observed. Analysis of ~300 DSB repair events indicated that most involved the loss of nucleotides from one or both ends being joined. Insertions were observed in five repair events. None of the inserted sequences were of organelle origin. Notably, the amount of nuclear sequence deleted was significantly larger in repair events involving insertion than in those without insertion, indicating that the two types of repair may be mediated by distinct pathways.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, *via* the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed ……*Date*……

Acknowledgements

A PhD is quite a journey and has taken me to many places that I would have otherwise never known, both literally and figuratively. I have learnt a great many things writing and working on this thesis, things about science, about people, about failure and success and about myself. It is hard to believe that it is now coming to an end.

First and foremost I would like to thank my supervisor Jeremy Timmis for all of his support, guidance, encouragement and supervision. Jeremy has a wonderful approach to science, and great warmth, patience and knowledge. From the time spent in his lab as an undergraduate student he has been a wonderful mentor and has instilled in me a great enthusiasm and respect for science.

I would like to thank all members of the Timmis lab, past and present, for making it such an enjoyable, engaging time. In particular I would like to thank Anna Sheppard for all of her help, especially when starting out on this journey and also for many wonderfully productive conversations that helped this work take shape. Thanks must also go to Yuan Li for all of her good cheer and technical assistance along the way. Much of this would have been impossible without her. Thanks also to Matt, Dong, Rory and Sven who as well as being a lot of fun, have been great people to work with and bounce ideas off. I would also like to thank everyone in the Genetics discipline, it has been a wonderful, supportive and enjoyable place to work.

To my family and friends your support and encouragement have been unending, you are truly wonderful people. In particular I would like to thank all past and present members of 'Palmerston'. It has been an amazing, enriching and thoroughly enjoyable time and a welcome escape from the lab.

Finally, I would like to Mairead. For all of your love, support, understanding and encouragement throughout the last three years I am truly grateful.

List of abbreviations

4-MU	4-Methylumbelliferone
A	adenine
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
bp, kb	base pairs, kilobase pairs
C	cytosine
CaMV	cauliflower mosaic virus
˚C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
cv	cultivar
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
DSB	double strand break
EDTA	ethylenediaminetetraacetic acid disodium salt
FISH	fluorescence *in situ* hybridisation
g	force of gravity
g, mg, µg, ng, pg	gram, milligram, microgram, nanogram, picogram
G	guanine
hr	hour
iPCR	inverse PCR
kr	kanamycin resistant
L, mL, µL	litre, millilitre, microlitre
M, µM	moles per litre, micromoles per litre
min	minute
mRNA	messenger RNA
cm, mm	centimetre, millimetre
MUG	4-methylumbelliferyl-beta-D-glucuronide
NHEJ	non-homologous end joining
numt	nuclear integrant of mitochondrial DNA
nupt	nuclear integrant of plastid DNA
norg	nuclear integrant of organellar DNA
nt	nucleotide
PCR	polymerase chain reaction
pmol	picomole
PVPP	polyvinylpolypyrrolidone
RACE	rapid amplification of cDNA ends
RLM-RACE	RNA ligase-mediated

RNA	ribonucleic acid
RNase A	ribonuclease A
rRNA	ribosomal RNA
RT-PCR	reverse transcription PCR
RT-QPCR	real-time quantitative PCR
sec	second
smPCR	single molecule PCR
sr	spectinomycin/streptomycin resistant
T	thymine
T-DNA	transfer DNA
TAE	Tris-acetate-EDTA
TAIL-PCR	thermal asymmetric interlaced PCR
Taq	*Thermus aquaticus*
Tris	tris(hydroxymethyl)aminomethane
tRNA	transfer RNA
U	uracil
U	Unit(s) of enzyme
UTR	untranscribed region
v/v	volume per volume
w/ v	weight per volume

Chapter 1: Introduction

1.1 Introduction

Within eukaryote cells there are several genetic compartments, the nucleus housing the vast majority of the genes, the mitochondria and, in plant, algal and some protist lineages, the plastid. The mitochondria and plastids have an endosymbiotic origin and are the extant descendants of once free-living α-proteobacteria and cyanobacteria respectively.

The primary endosymbiotic event was the uptake of an α-proteobacterium which gave rise to the mitochondria. The exact nature of the relationship that gave rise to this first mitochondriate cell is still a matter of considerable debate (Embley and Martin, 2006; Cavalier-Smith, 2009; Gross and Bhattacharya, 2010) but despite the lack of agreement on the early evolutionary history it is widely held that the earliest common ancestor of all known eukaryotes was a unicellular phagotroph that had a nucleus and mitochondria. The origin of the chloroplast is somewhat more evident.

1.2 Origin of the chloroplast

A symbiotic origin for the chloroplast was first proposed in 1905 (Mereschkowsky, 1905), although it was not until the 1970s (Margulis, 1970) that the notion of an endosymbiotic origin of both the chloroplast and the mitochondrion began to garner wider support. In the following years genetic and biochemical evidence mounted in support of their prokaryote origins and finally increasing sequence data from organelle and bacterial genomes enabled this to be conclusively proven (Gray and Doolittle, 1982). As the complete sequencing of plastid and cyanobacterial genomes continues the exact nature of the chloroplast ancestor has become clearer. The current most likely contender is a nitrogen fixing cyanobacterium (Deusch *et al.*, 2008; Falcon *et al.*, 2010), which current molecular (Yoon *et al.*, 2004; Falcon *et al.*, 2010) and fossil (Butterfield, 2000) data suggest became part of the eukaryote cell around 1.3 billion years ago. The likely limited availability of nitrogen at the time of chloroplast origin (Anbar and Knoll, 2002) and the prevalence of nitrogen fixation in many modern day cyanobacterial symbiotic associations (Kneip *et al.*, 2007) has prompted recent suggestions that nitrogen fixation may have contributed to establishing the symbiosis that gave rise to the chloroplasts (Kneip *et al.*, 2007; Deusch *et al.*, 2008).

1.3 Organelle genome reduction

1.3.1 Evolutionary gene transfer to the nucleus

Consistent with their endosymbiotic origin, chloroplasts and mitochondria retain essentially prokaryote genomes and contain all of the transcriptional and translational machinery necessary

for organellar gene expression. Their respective remnant genomes are, however, vastly reduced in size when compared with those of their free-living relatives, retaining only 1-5% of the ancestral protein coding genes. Mitochondrial genomes are the most reduced in size containing only 3-67 protein coding genes (Timmis *et al.*, 2004), while chloroplast genomes generally encode several more proteins, around 80 in land plants and over 200 in some algae (Timmis *et al.*, 2004).

The reduction in genome size has been, in part, due to loss of genes made redundant when the endosymbiont became resident within the eukaryote cell. However, most of the gene loss, is due to relocation of organelle genes to the nuclear genome. In many cases proteins encoded by these relocated genes retain their original role in organellar biogenesis. These nuclear genes that control organellar biogenesis and function are transcribed in the nucleus, their mRNAs are translated on cytoplasmic ribosomes and the proteins are then imported into the appropriate organelle.

This outsourcing of gene expression and protein synthesis required the development of sophisticated protein import machinery most notably the TOM/TIM and TOC/TIC protein import pathways of the mitochondria and chloroplast respectively (Neupert, 1997; Soll and Schleiff, 2004). Other protein import pathways also exist, such as *via* the secretory pathway (Villarejo *et al.*, 2005), but these pathways are still little understood (Millar *et al.*, 2006; Li and Chiu, 2010). The establishment of protein import mechanisms may have been the rate limiting step in the transition from an endosymbiont to an organelle (Cavalier-Smith and Lee, 1985). Once protein import was established proteins had a route back to the organelle and so genes were able to make the journey to the nucleus, cementing the genetic interdependence of the organelle and host cell. Not all genes that have relocated to the nucleus encode proteins predicted to be exported to the organelles (Martin *et al.*, 2002), suggesting the acquisition of novel non-organelle related function. The algorithms such as TargetP used for these predictions are fallible (Kleffmann *et al.*, 2004) and estimating the number of extraorganellar proteins derived from former organellar genes is problematic. However, it is clear that the relocation of genes from the cytoplasmic organelles to the nucleus has been a major contributor to the complexity of nuclear genomes and has given rise to many genes of novel function. Shorter stretches of organelle DNA, rather than whole genes, have also contributed to the complexity of nuclear genomes by contributing exonic sequence to pre-existing nuclear genes (Noutsos *et al.*, 2007).

With the current availability of the nucleotide sequence of well over 2300 mitochondrial genomes and 190 plastid genomes (NCBI, 2010), it is clear that there is great diversity in the size of organelle genomes and the number of proteins that they encode. Animal mitochondrial genomes are relatively constant at around 16 kb in length but much more diversity is seen in plants, in particular in the *Cucurbitaceae* where an early study estimated genome sizes ranging from 390 to 2,900 kb

(Ward *et al.*, 1981). The mitochondrial genomes in several of these species have now been fully sequenced confirming the earlier estimates (Alverson *et al.*, 2010). A great diversity is also seen in the mitochondrial genomes of protists (Barbrook *et al.*, 2010) whose mitochondrial protein coding capacities range from the smallest known, 3 protein coding genes on a 6 kb genome in the Apicomplexa (Wilson and Williamson, 1997), to the largest known, 67 protein coding genes on a 69 kb genome in *Reclinomonas americana* (Lang *et al.*, 1997).

The largest chloroplast genome currently known is that of *Floydiella terrestris,* a chlorophycean algae whose genome has a length of ~520 kb (Brouard *et al.*, 2010). Whilst this is the largest plastome, it encodes only 70 conserved proteins, whereas some red algae have smaller genomes that encode over 200 proteins (Reith and Munholland, 1995).

Not surprisingly the most reduced organelle genomes are generally found in organisms that have lost the major biosynthetic pathways of oxidative phosphorylation (in the case of mitochondria) and photosynthesis (in the case of plastids). Hydrogenosomes, organelles that produce molecular hydrogen and ATP in anaerobic organisms, are highly reduced mitochondria found in diverse eukaryotes (Boxma *et al.*, 2005). In most cases they appear to have completely lost their entire genome (van der Giezen *et al.*, 1997; Clemens and Johnson, 2000). Similarly the smallest plastid genomes are found in lineages that have lost the ability to photosynthesise. The parasitic plant *Epifagus virginiana* chloroplast genome is only 70 kb in size and codes for only 21 proteins (Wolfe *et al.*, 1992). The non-photosynthetic apicoplast, the vestigial plastid found in apicomplexan parasites, has an even smaller genome which in the malaria parasite *Plasmodium falciparum* is around 34 kb in size (Wilson *et al.*, 1996). Although usually photosynthetic, dinoflagellates also have a highly reduced chloroplast genome which encodes around 15 proteins (Howe *et al.*, 2008). Their chloroplast genomes contain a unique arrangement with individual (or in a few cases several) genes arranged on plasmid like 'mini-circles' (Howe *et al.*, 2008).

1.3.2 Recent gene transfer events

The number of genes found in plastid and mitochondrial genomes varies between species but in all cases there are relatively few compared with prokaryotic genomes. It is therefore thought that the majority of endosymbiotic gene transfer occurred early in the evolutionary history of the organelles (Timmis *et al.*, 2004). In some lineages, such as animals where the set of mitochondrial genes is highly conserved, gene transfer appears to have ceased completely. In the few cases where genes are missing, this is probably due to complete loss rather than transfer to the nucleus (Gissi *et al.*, 2008). In plants however, there is evidence of recent gene transfer. Adams *et al.* (2000) reported 26 independent losses of *rps10* from the mitochondrial genomes of 277 angiosperms examined. Molecular characterisation of a number of the nuclear *rps10* genes indicated that each loss of

rps10 from the mitochondrial genome was likely to represent independent transfer to the nucleus. A similar study (Millen *et al.*, 2001) looked at the loss of *InfA* from the chloroplast genome and found 24 independent losses among over 300 angiosperms. Again, molecular characterisation of nuclear *InfA* genes suggested that each loss from the chloroplast genome was due to an independent transfer of *InfA* to the nucleus. A number of other genes, mainly ribosomal protein genes, have been lost from mitochondrial and chloroplast genomes in angiosperms (reviewed by Rousseau-Gueutin *et al.*, In Press), leading to considerable diversity in gene content.

1.3.3 Why relocate?

Given the frequency of endosymbiotic gene transfer (both historic and ongoing), many have speculated on the possible advantages of organelle genes being relocated to the nucleus. Proposed explanations include the high rate of oxidative stress-induced mutation within organelles (Allen and Raven, 1996), genome streamlining (Selosse *et al.*, 2001), easier fixation of beneficial mutations (Blanchard and Lynch, 2000) and avoidance of Muller's ratchet (the accumulation of mutation in asexually reproducing organelles) due to the benefits of sexual recombination for elimination of deleterious mutation in nuclear genes (Lynch, 1996; Martin and Herrmann, 1998). These reasons, however, seem not to apply to plant organelles which have a much lower rate of mutation (Wolfe *et al.*, 1987), have larger organelle genomes with more non-coding DNA (Timmis *et al.*, 2004), and where more gene loss is observed in taxa that reproduce asexually or by self-fertilisation (Brandvain *et al.*, 2007).

Perhaps a key explanation is one of unidirectional transfer of genes to the nucleus promoted by the high frequency of organelle DNA introgression into the nucleus and the relatively rare or entirely absent transfer of DNA from the nucleus to the organelles. If transfer leads to two functional copies, one copy may then be lost. If the nuclear copy is lost the gene is free to transfer again at a later stage, but if the organelle copy is lost then the nucleus becomes the permanent location of the gene establishing a 'gene-transfer ratchet' (Doolittle, 1998). As long as there remains a polarity in DNA transfer then gene transfer to the nucleus would be an inevitable consequence of neutral drift (Berg and Kurland, 2000).

From this background of neutral gene transfer, the various mutational and/or selective pressures described above may contribute to the likelihood of gene transfer by altering the respective likelihoods of organelle or nuclear gene inactivation. These pressures may have been considerably different early in evolution when the majority of transfer is likely to have occurred. The low rate of mutation in extant plant organelle genomes (Wolfe *et al.*, 1987), presumably due to the establishment of plant specific DNA repair and/or recombination pathways (Marechal and Brisson, 2010) together with efficient gene conversion mediated by polyhaploidy, may well have led to a

slowing in the rate of gene transfer. This could explain the difference in genome size and gene content seen between plant and animal mitochondrial genomes. If this is the case it would suggest that mutational pressure rather than the energetic and replicative advantage of a small organelle genome drive gene transfer to the nucleus.

1.3.4 Why retain an organellar genome?

Thousands of genes have made the journey from the organelles to the nucleus, so why do any remain given the energy energetic outlay in maintaining all of the transcriptional and translational machinery required for the retention of alternative genetic systems? The hydrophobicity hypothesis suggests that highly hydrophobic proteins are hard to export from the cytosol to the organelles and that this precludes relocation of these genes to the nucleus (Vonheijne, 1986; Daley and Whelan, 2005). Counter to this theory, however, the chloroplast encoded hydrophobic protein D1 (encoded by *psbA*) can be imported from the cytosol to the chloroplast when experimentally equipped with a transit peptide and expressed from a nuclear gene (Cheung *et al.*, 1988). The Co-location for Redox Regulation or CoRR hypothesis (Allen, 2003) maintains that there is a key set of genes whose expression must be directly controlled by the redox state of their gene product, or interacting electron carriers. This requires separate (organellar rather than nuclear) gene expression as redox state is likely to vary between the many organelles within a single cell. Recently a sensor kinase has been identified that links the redox state of an electron carrier connecting the two photosystems, with chloroplast gene expression (Puthiyaveetil *et al.*, 2008). Neither of these hypotheses, however, appear to explain the retention of genomes in non-photosynthetic plastids such as those found in parasitic plants or the apicoplasts of the Apicomplexa. Several other hypotheses have been proposed to explain the situation in these cases (Barbrook *et al.*, 2006). The 'essential tRNA' hypothesis was proposed based on the observation that the tRNA encoded by the plastid gene *trnE* is involved in tetrapyrrole biosynthesis and so is likely to be essential even in the absence of protein biosynthesis. In addition retention of the plastid initiator tRNA (encoded by *trnfM*) has been suggested to be required for import into mitochondria in the Apicomplexa (Barbrook *et al.*, 2006). The 'limited transfer window' hypothesis posits that organisms containing a single organelle per cell have little opportunity for gene transfer as organelle breakdown, which is necessary for the release of DNA, will be lethal. It is likely that no one hypothesis will adequately explain the retention of organelle genomes in all cases and different combinations of factors may be responsible in different species.

1.4 Ongoing organelle DNA transfer to the nucleus

1.4.1 Organelle sequences in nuclear genomes

A prerequisite for the functional relocation of plastid and mitochondrial genes to the nucleus is a nucleic acid transfer mechanism(s). The first indications that transfer of organelle DNA to the nucleus continues today came to light about 30 years ago with the identification of nuclear sequences that have strong homology to organelle DNA (van den Boogaart *et al.*, 1982; Timmis and Scott, 1983). The relatively recent transfer of these sequences to the nucleus was subsequently inferred from their high sequence identity (i.e. >99%) to existing organelle genes (Lin *et al.*, 1999; The Rice Chromosome 10 Sequencing Consortium, 2003). Whole genome sequencing has since revealed extensive tracts of chloroplast or mitochondrial DNA in the nuclear genomes of almost all eukaryotes studied (Timmis *et al.*, 2004; Hazkani-Covo *et al.*, 2010). These insertions of organelle DNA are referred to as *numts* (nuclear integrants of mitochondrial DNA) and *nupts* (nuclear integrants of plastid DNA) or collectively as *norgs* (nuclear integrants of organellar DNA).

The arrangement of these sequences has been studied in detail in Arabidopsis and rice and found to be quite varied (Richly and Leister, 2004a; Richly and Leister, 2004b; Noutsos *et al.*, 2005). A large percentage of the total *norg* content is found in a relatively small number of large *norgs* that can be tens or hundreds of kb in length. The remainder is found in a large number of smaller *norgs* scattered throughout the genome (Richly and Leister, 2004b). Of the large *norg* loci, some are continuous sequences of chloroplast or mitochondrial origin and are clearly the result of the insertion of a single molecule, while others contain multiple fragments of DNA from diverse parts of the chloroplast or mitochondrial genome or both (Noutsos *et al.*, 2005). Some loci of the latter type, probably represent insertions of a single contiguous fragment of organelle DNA that has since undergone deletion and/or rearrangement (Matsuo *et al.*, 2005). However, the observation of mitochondrial and chloroplast sequence at a single locus indicates that the insertion of multiple organelle fragments in a single event or multiple sequential insertions at the one locus must also take place (Noutsos *et al.*, 2005). Several other *norg* loci are highly complex mosaics containing up to 80 distinct ~50-100 bp segments of the chloroplast and mitochondrial genome arranged end to end (Noutsos *et al.*, 2005). Exactly how these loci arise is yet to be satisfactorily explained, however, similar mosaics have been observed that are comprised of many short stretches of transposable element sequence (D. Adelson, personal communication).

Large *norg* insertions have also been observed in other species. Recently, *in situ* hybridization in the maize inbred line B73 identified a *nupt* that includes almost the entire 140 kb chloroplast genome on chromosome 5 (Roark *et al.*, 2010) and a *numt* containing the majority of the 570 kb mitochondrial genome on chromosome 9 (Lough *et al.*, 2008). These studies also showed that

numts and *nupts* varied greatly among different inbred maize lines indicating that there have been frequent recent insertions of organelle DNA into maize nuclear genomes. Current investigations such as the 1000 genomes projects in humans and Arabidopsis should contribute greatly to understanding the intra-species variation of *norgs* and highlight potential evolutionary ramifications.

The precise bp contribution of *nupt* and *numt* sequences to the nuclear genome is hard to determine. Based on current genome assemblies it is estimated that *nupt*s and *numt*s each generally make up about 0.1-0.2% of the nuclear genome in flowering plants (Richly and Leister, 2004a; Richly and Leister, 2004b; Ming *et al.*, 2008; Arthofer *et al.*, 2010; Hazkani-Covo *et al.*, 2010; Vogel *et al.*, 2010) and significantly less in algae and moss (Richly and Leister, 2004b; Hazkani-Covo *et al.*, 2010). This, however, may be the 'tip of the iceberg' as whole genome assemblies often underestimate the contribution of organelle-derived sequences to nuclear genomes. This is in large part an artefact of the elimination of 'contaminating' organelle DNA sequences - a process which must also often exclude *norgs*. An example is the honeybee genome which was initially thought to have little or no mitochondrial DNA within the nucleus (Leister, 2005) but has since been found, using a different assembly, to have one of the most extensive *numt* complements (Behura, 2007; Hazkani-Covo *et al.*, 2010).

Another problem lies in the assembly of regions of the genome that contain large duplications. Chromosome 2 in Arabidopsis was initially reported to contain a 270 kb *numt* (Lin *et al.*, 1999) but Stupar *et al.* (2001) later showed that this *numt* was in fact ~620 kb in length and contained several large internal duplications. The authors were only able to determine the *numt* size using fibre-FISH and showed that contig assembly using BACs tended to minimise clone length, missing large duplications. Despite this finding, this region is still only 270 kb in length in the current chromosome 2 assembly (Build 9.1, 14th Oct 2009) and recent studies (Richly and Leister, 2004a; Hazkani-Covo *et al.*, 2010) have therefore greatly underestimated total *numt* size in Arabidopsis. This problem presumably holds for *nupts* as well and will be compounded in genomes shotgun sequenced using high-throughput short-read platforms.

1.4.2 Experimental transfer to the nucleus

In some species it has been possible to determine experimentally the frequency with which organellar DNA can move into the nucleus. This was initially investigated in yeast by measuring the transfer of a mitochondrial plasmid to the nucleus (Thorsness and Fox, 1990) which was found to occur at high frequency (~2 x 10^{-5} per cell per generation). Although the plasmid DNA in these first experiments was not incorporated into the nuclear chromosomes, subsequent experiments, also in yeast, observed integration of mitochondrial DNA at sites of nuclear double strand break repair

(Ricchetti *et al.*, 1999). With the development of chloroplast transformation in tobacco (Svab *et al.*, 1990), similar studies were made possible in higher plants. In the first of these studies (Huang *et al.*, 2003) a selectable marker gene (*neo*) equipped exclusively for nuclear expression was introduced into the chloroplast genome of tobacco. Transplastomic pollen was used to fertilise female wild-type plants and the resultant progeny screened for kanamycin resistance (*neo* expression). In a large screen of 250,000 seedlings, 1 in 16,000 pollen grains were inferred to carry a copy of *neo* transferred from chloroplast DNA to the nucleus in the germline of the transplastomic male parent. A similar study measured the rate of transfer in somatic cells (Stegemann *et al.*, 2003) and transfer was shown to occur once in approximately 5,000,000 cells. Although still relatively frequent, this was substantially less common than the transfer observed in the male germline and prompted the suggestion that degradation of the chloroplast during pollen development (associated with uniparental inheritance) may provide more opportunity for nuclear DNA transfer by liberating organellar DNA. This hypothesis was supported by a third study that measured the rate of gene transfer in both the female and male germline (Sheppard *et al.*, 2008). An even greater frequency of gene transfer through the male germline was reported (1 in 11,000 male gametes) which far exceeded transfer in the female germline where a single transfer event was observed in a screen of over 270,000 ovules (Sheppard *et al.*, 2008).

In each of these screens the chloroplast gene must not only transfer to the nucleus but also integrate into the nuclear chromosomes. To investigate the steps in this process Sheppard *et al.* (2008) introduced a *GUS* reporter gene (again designed exclusively for nuclear expression) into the chloroplast genome and leaves of the transplastomic plant were stained for GUS activity to detect cells in which the gene had transferred to the nucleus. In this instance blue staining cells represented transient expression from the nucleus/cytoplasm as well as transfer followed by stable integration into a transcriptionally active region of the nuclear genome. Interestingly total transfer (transient and stable) was found to be 25-270 fold higher than the purely stable somatic transfer of *neo* detected by Stegemann *et al.* (2003) suggesting that most blue spots resulted from transient expression.

1.4.3 Mutational fate of norgs

Given the constant deluge of organellar DNA entering the nuclear genome, it must be expected that a counterbalancing eradication of these sequences occurs to prevent ever expanding genome size. This was first alluded to with the observation that, for *nupts* over 500 bp in length, there is an inverse relationship between their age (based on sequence identity to the chloroplast genome) and their size (Richly and Leister, 2004b). This finding has subsequently been found to hold true for *norgs* in *Brachypodium* (Vogel *et al.*, 2010) and Papaya (Ming *et al.*, 2008) and suggests that insertion of large *nupts* is followed by fragmentation and deletion. Direct experimental observation

of frequent deletion of about 50% of newly transferred chloroplast sequences has now conclusively demonstrated the extreme instability of plastid DNA in the tobacco nucleus (Sheppard and Timmis, 2009). So far it has not been possible to determine how much of the integrant is lost by recovering the sequence that remains. This *nupt* deletion happened within 1-2 generations of formation and it may be that more lines would show instability over longer, but still evolutionarily relevant, timescales.

The deletion of organelle DNA is most unlikely to be an exact excision and partial deletion would lead to novel arrangements of organelle and nuclear DNA. The deletion may also be accompanied by other rearrangements including inversions and new insertions of organellar DNA and transposable elements (Guo *et al.*, 2008). Richly and Leister (2004b) observed 'tight' and 'loose' clusters of organellar sequence in nuclear genomes of rice and Arabidopsis which they suggest represent progressive steps of degradation and rearrangement of large initial insertions. Deletions and other rearrangements may be part of a process of 'genetic tinkering' that in rare instances leads to the activation of newly transferred genes (Bock and Timmis, 2008).

Base substitution also appears to play a significant role in the evolution of *norgs.* In plants a significant bias in C→T and G→A mutations has been observed in large recent integrants of organelle DNA (Huang *et al.*, 2005). This mutational bias is consistent with spontaneous deamination of 5-methylcytosines within methylated DNA suggesting that these *norgs* are methylated. Studies linking the stability of *norg* sequences with methylation and chromatin structure have not yet been reported.

1.5 Mechanisms of gene transfer to the nucleus

1.5.1 Escape of genetic material from the organelle

The first step in transfer of a gene to the nucleus is the escape of genetic material from the organelle. In general, escape from the organelle is likely to be made possible through loss of integrity of the organelle membrane, either through various physiological stressors or programmed degradation during development. Various environmental stressors and developmental stages are known to trigger programmed organelle breakdown (Kundu and Thompson, 2005; Stettler *et al.*, 2009; Wada *et al.*, 2009) and these may lead to increased escape of organelle DNA to the nucleus. Recently cold stress (Ruf *et al.*, 2010) and heat and salt stress (D. Wang, personal communication) have been shown to increase the rate at which a chloroplast gene escapes to the nucleus in tobacco. For unicellular organisms that have only a single organelle per cell DNA transfer is likely to be limited as degradation of the organelle will lead to cell death (Barbrook *et al.*, 2006). Unsurprisingly, therefore, the *Chlamydomonas reinhardtii* nuclear genome has a low *norg* content

(Richly and Leister, 2004a) and large screens have failed to detect transfer of a chloroplast gene to the nucleus (Lister *et al.*, 2003).

Uniparental inheritance may also lead to increased escape of genetic material from the organelles. In most (but not all) sexually reproducing eukaryotes, only one sex contributes cytoplasmic organelles to the zygote. How this uni-parental inheritance is achieved varies across species but in general the cytoplasmic organelles are degraded and/or excluded from one of the gametes or, sex-specific loss of organelles occurs after fertilisation (Birky, 2001). In tobacco, chloroplasts are maternally inherited and DNA transfer from the chloroplast to the nucleus occurs more frequently in the male germ line than that of the female (Sheppard *et al.*, 2008). This difference has been attributed to release of chloroplast DNA into the cytoplasm during chloroplast degradation/exclusion in the developing male gametophyte.

Further understanding of how various stressors and modes of organelle inheritance affect chloroplast-to-nucleus DNA transfer will undoubtedly prove to be an interesting area of future research. The increasing wealth of genome sequence data will pave the way for analysis of *norgs* in different ecotypes and it will be interesting to see if any relationship exists between *norg* content and environmental conditions or geographical distribution. Further understanding of how stress and organelle inheritance affects endosymbiotic DNA transfer will also be of biotechnological importance in order to mitigate potential transfer of chloroplast transgenes to the nucleus.

1.5.2 Is there an RNA or DNA intermediate?

It is generally believed that the majority of gene transfer to the nucleus occurs *via* a DNA intermediate (Timmis *et al.*, 2004; Kleine *et al.*, 2009), although this still remains to be experimentally proven. Some early studies of the transfer of plant mitochondrial genes to the nucleus showed that nuclear copies resembled spliced, edited mRNAs and led to the suggestion that transfer was *via* a reverse transcribed RNA intermediate (Nugent and Palmer, 1991; Grohmann *et al.*, 1992; Adams *et al.*, 2000). There are, however, alternative explanations that explain these observations but do not invoke the involvement of RNA (Henze and Martin, 2001). Some further evidence also points toward DNA-mediated transfer: non-coding regions of the chloroplast genome are found as abundantly in nuclear genomes as highly transcribed genic regions of the organellar genomes (Matsuo *et al.*, 2005) and some very large nuclear insertions of organellar sequence (>100 kb) have been found (Stupar *et al.*, 2001; The Rice Chromosome 10 Sequencing Consortium, 2003) suggesting direct DNA transfer. Direct experimental evidence of RNA mediated transfer is lacking as is the determination of the relative contributions of RNA and/or DNA mediated transfer and at least one study designed to observe transfer *via* an RNA intermediate failed to detect any such transfer (A. Sheppard, unpublished results).

1.5.3 Integration into nuclear chromosomes

Once the organelle DNA (or RNA) has entered the nucleus it must be integrated into nuclear chromosomes and be included in the gametes of sexually reproducing organisms if it is to make a lasting contribution to the nuclear genome. It is thought that most integration of organelle DNA occurs *via* non-homologous end joining (NHEJ) at sites of double strand break (DSB) repair (Kleine *et al.*, 2009) and this has been shown to occur in yeast (Ricchetti *et al.*, 1999). DSBs were induced in the yeast nuclear genome through expression of the rare cutting endonuclease I-SceI and insertion of mitochondrial DNA was observed in a proportion of repair events. Interestingly, in some repair events, DNA from two disparate regions of the mitochondrial genome were inserted at a single location. Similar capture of non-mitochondrial DNA has also been observed at sites of DSB repair in yeast (Haviv-Chesner *et al.*, 2007) as well as in plant and mammalian systems (Salomon and Puchta, 1998; Lin and Waldman, 2001). In these studies DSBs were induced by transiently introducing a plasmid, or T-DNA, encoding a rare cutting endonuclease. This rare-cutting endonuclease cut at a specific target site introduced into the nuclear genome and repair events were then analysed by PCR. Insertion of the T-DNA or plasmid DNA was often observed, as were insertions of nuclear repetitive elements such as retrotransposons and micro-satellites. While insertion of organellar DNA has so far only been observed in yeast the fact that extra-chromosomal DNA can be captured at sites of DSB repair in plants and animals suggests this process applies more widely.

The insertion of *norgs* has been also investigated in several bioinformatic analyses and these suggest more than one pathway for integration (Leister, 2005). Some integrants show a very simple arrangement likely originating when a single organellar DNA fragment inserted at a single location. Others are much more complex and are the result of multiple fragments being inserted in a single event or multiple insertions at a single location. Organelle sequences may also insert into areas of the genome that already contain *norgs* or other repetitive sequences which also adds to the complexity of these loci. There is some evidence that organelle DNA integrates more frequently into intergenic regions (Richly and Leister, 2004b), in particular those containing mobile elements (Mishmar *et al.*, 2004). Large *nupts* have also been shown to preferentially locate to pericentromeric regions in rice (Matsuo *et al.*, 2005) which are known to be DSB hotspots (Blitzblau *et al.*, 2007) and to contain a high density of transposable elements (Hall *et al.*, 2006). A recent study has also linked *numt* insertion sites in yeast to origins of replication (Lenglez *et al.*, 2010) which led the authors to suggest these sites may be prone to DSBs resulting in high levels of insertion. These findings point toward DSB repair, possibly at sites of transposon excision (Leister, 2005), being a pathway for the nuclear insertion of organellar sequences. The presence of such a DNA repair/integration mechanism would contribute significantly to the complex arrangement of organellar sequences integrating into the nuclear genome. This would be important from an

evolutionary perspective as it would lead to the creation of novel sequence arrangements which, in some instances, may result in nuclear activation of the transferred organelle genes.

The cross-over in the insertion pathway and chromosomal location of *norgs* and repetitive elements shows that these sequences can be dealt with in very similar ways by the nuclear DNA repair/maintenance machinery. To date, studies have focussed exclusively on either organellar DNA or other repetitive sequence. There may be significant advantage to both fields if a more unified approach is taken to investigating these areas.

Although bioinformatic analyses of *norgs* due to evolutionary transfer have added considerably to our understanding of these sequences they are limited in that a *norg* sequence cannot usually be compared with that of the nuclear sequence prior to insertion. This makes identification of micro-homology and other indicators of NHEJ difficult to assess. Also it is impossible to determine how much of the observed complexity of *norg* sequence is due to the primary insertion event and how much is due to subsequent fragmentation or insertion at this locus. Some partial characterisation of experimentally transferred *norgs* has been undertaken and suggests that micro-homology is involved in the insertion of these sequences (Huang *et al.*, 2004). A fuller understanding will come with complete characterisation of *de novo norgs* and comparison with their pre-insertion sequences. This remains an important but technically challenging future step.

1.5.4 Proteins involved

While the NHEJ pathway of DSB repair is considered the major pathway for nuclear integration of organelle DNA there is no direct evidence for the involvement of any specific proteins in this process. A very limited number of *de novo* organelle DNA transfers have been experimentally observed to date and their generation and characterisation pose significant technical challenges. This, coupled with the lack of mutants in species amenable to gene transfer experimentation, has made direct investigation of the proteins involved impractical. There are significantly more studies which have investigated the role of particular proteins in DSB repair and transgene integration and some of these findings may, with some caution, inform our understanding of the integration of organellar DNA.

In yeast, the insertion of extra-chromosomal DNA at sites of DSB repair is Ku80 dependent (Haviv-Chesner *et al.*, 2007) indicating the involvement of the NHEJ pathway. Similarly, transformation in yeast also occurs mainly *via* the NHEJ pathway. However, when this pathway is disrupted (in Ku70, Ku80 and Lig4 mutant lines) and sequences flanking the transgene are homologous to the desired insertion locus, then transformation occurs *via* homologous recombination allowing efficient gene replacement (Ninomiya *et al.*, 2004). A recent study suggests the same is true, at least for Lig4 null lines, in the transformation of Arabidopsis calli by particle bombardment (Tanaka *et al.*, 2010). The

moss *Physcomitrella patens* appears to be an exception where transformation occurs preferentially *via* homologous recombination (reviewed by Schaefer, 2002). Although most Arabidopsis transformation is *via* the NHEJ pathway, the involvement of the Ku heterodimer seems to be tissue dependent, with Ku playing a considerable role in somatic cells (Li *et al.*, 2005) but not in the germline (Friesner and Britt, 2003; Gallego *et al.*, 2003). Interestingly, germline transformation usually results in higher copy-number insertion than transformation of root calli (De Buck *et al.*, 2009), suggesting that the Ku-independent pathway active in the germline may be responsible for more complex insertion events. This may well be of significance to organellar DNA insertion in higher plants as the majority of inherited plastid DNA transfer occurs in the male germline (Sheppard *et al.*, 2008).

For most species, the classical NHEJ pathway (Weterings and Chen, 2008) appears to be the main route for insertion of foreign DNA into the nucleus, however, there are some differences between species (Schaefer and Zryd, 1997) and between tissues (Friesner and Britt, 2003; Li *et al.*, 2005) within a species which may account for differences in organelle sequences observed in their nuclear genomes.

1.6 Activation of newly transferred organelle genes

1.6.1 Examples of organellar gene activation in the nucleus

Only in a very few instances will transfer of organelle DNA to the nucleus lead to the functional relocation of an organelle gene. In most cases organelle sequences transferred to the nucleus have the same fate as other non-coding DNA - freely accumulating mutations and degrading over time. The low mutation rate in plant organelle genomes means that the extant organelle genomes provide a historic reference for the sequence at the time of insertion, from which it is possible to derive many insights into the various ways in which *norg* sequences evolve. In a few rare cases these sequence rearrangements and changes in base composition lead to activation of newly transferred genes. Activation, in the majority of cases, must be a multistep process and requires the acquisition of a nuclear promoter, a polyadenylation signal and, if the protein is to be targeted back to the organelle, a transit peptide or an alternative mechanism for protein targeting.

Several bioinformatic studies have highlighted various means by which organellar genes have become activated in nuclear genomes. One such study investigated transfer of the maize gene encoding the mitochondrial protein RPS14 to the nucleus (Figueroa *et al.*, 1999). The mitochondrial gene had inserted into an intron of the nuclear gene encoding the iron-sulphur subunit of succinate dehydrogenase (*sdh2*) and was expressed by differential splicing of the mRNA with both proteins using the SDH2 target peptide for targeting to the mitochondria. In a similar case the chloroplast *rpl32* gene was transferred to the nucleus in an ancestor of mangrove and poplar (Cusack and

Wolfe, 2007) where it inserted into an intron of the gene encoding the chloroplast superoxide dismutase (SODcp) to form the chimeric *SODcp-rpl32* gene. In mangrove the SODcp protein and a SODcp amino terminus/RPL32 fusion protein are expressed from the single promoter through differential splicing. Both proteins are then targeted to the mitochondrion using the SODcp transit peptide. In poplar, evolutionary tinkering has taken the process one step further with the duplication and subfunctionalization of the *SODcp-rpl32* gene. One copy has lost the RPL32 coding sequence and now solely encodes SODcp, the other now exclusively expresses the SODcp amino terminus/RPL32 fusion protein. There are numerous other examples of genes that have recently transferred to the nucleus in angiosperms, many of which have also hijacked transit peptides from existing nuclear encoded chloroplast proteins (Liu *et al.*, 2009).

1.6.2 Experimental activation of a chloroplast gene transferred to the nucleus

Experimental attempts have been made to reconstruct functional gene transfer to gain a better understanding of the diverse processes involved and the frequency with which newly transferred genes become activated in the nucleus. Stegemann and Bock (2006) showed functional activation of a chloroplast marker gene *aadA* that had been recently transferred to the nucleus in tobacco. In each case *aadA* was activated by a short simple deletion that recruited the nearby CaMV 35S strong nuclear promoter that was integral to their experimental cassette and present in the same transcriptional polarity. In no case was activation achieved by acquisition of a native nuclear promoter and so the frequency of a 'natural' gene transfer event remains unclear.

Interestingly they found that the *aadA* transcripts were polyadenylated despite the lack of any changes in the *psbA* 3' UTR found downstream of the *aadA* open reading frame. Examination of the *psbA* terminator revealed a sequence that matched the rather flexible AT-rich plant polyadenylation consensus sequence and this was the site of *aadA* polyadenylation. This led the authors to suggest that the AT-rich nature of plant non-coding sequences may provide many fortuitous polyadenylation sites greatly aiding the process of functional gene transfer. This could possibly be extended to other AT-rich regulatory motifs such as a TATA box. Indeed, the tobacco chloroplast *psbA* promoter has been shown to have weak nuclear activity that is dependent upon TATA and CAAT boxes fortuitously present (Cornelissen and Vandewiele, 1989) but cryptic nuclear activity of any other chloroplast promoters remains unknown.

1.7 Summary

The constant integration of organellar DNA has had profound consequences in the evolution of eukaryote nuclear genomes (Timmis *et al.*, 2004; Kleine *et al.*, 2009). The ingress of DNA is followed by decay, deletion and rearrangement of these sequences which leads to novel sequence combinations. Ultimately this provides new genetic material for the action of natural selection and

in rare instances, can lead to the functional relocation of organelle genes to the nucleus or the generation of genes with novel function. This process is of great evolutionary interest as it has been a major pathway for the generation of new genes in eukaryote nuclear genomes. It is also of great interest both to plant biotechnologists and the wider public in assessing the level of transgene containment provided by chloroplast transformation.

1.8 Project Aims

Broadly, this project aims to investigate the mechanisms through which chloroplast transgenes can transfer to the nuclear genome and become activated in this new genetic environment.

The initial aim of this project, presented in Chapter 2, is to investigate the frequency of activation of a chloroplast gene newly transferred to the nucleus. This will extend work undertaken by Stegemann and Bock by screening over 5 times the number of lines corresponding to 16 genomic locations of *aadA* and by using an experimental arrangement of the transgenes that precludes recruitment of the 35S promoter by simple deletion enabling novel means of activation to be uncovered.

The second objective, described in Chapter 3, is the characterisation of a *de novo* chloroplast DNA integrant. In investigating the frequency of DNA transfer from the chloroplast to the nucleus in tobacco a number of lines have been generated that contain independent *de novo* transfers of DNA from the chloroplast to the nucleus (Huang *et al.*, 2003; Sheppard *et al.*, 2008). While some partial characterisation has been undertaken in some of these lines, none has been fully described (Huang *et al.*, 2005). This chapter describes the complete sequencing of the chloroplast DNA integrant, and some flanking nuclear DNA, in line kr2.2. This sequence information is then used to derive a model for the insertion of the chloroplast DNA in this line.

The third project aim, described in Chapter 4, is to design and evaluate an experimental system for the detection of organelle sequence insertion at sites of DNA DSB repair. DSBs are initiated through inducible expression of the rare cutting endonuclease I-SceI which can then be followed by PCR and sequencing to analyse the molecular nature of the repair. Finally, in chapter 5, the experimental system described in chapter 4 is used in concert with single molecule PCR to directly observe sites of DSB repair in tobacco. This procedure is used to determine the proportion of repair events that incorporate cytoplasmic organellar DNA and to gain insights into the molecular mechanisms involved.

Chapter 2: Functional transfer of a Chloroplast Transgene to the Nucleus in Tobacco

2.1 Introduction

Throughout eukaryote evolution many genes have relocated from the cytoplasmic organelles to the nucleus. While in some lineages, such as multi-cellular animals, functional gene transfer appears to have ceased, in angiosperms the process continues today (Timmis *et al.*, 2004). Several studies have investigated a number of recent evolutionary gene transfers in some detail (Adams *et al.*, 2000; Millen *et al.*, 2001; Cusack and Wolfe, 2007; Liu *et al.*, 2009) and have revealed different evolutionary events that have resulted in the activation of transferred mitochondrial or plastid genes and import of the encoded proteins into the organelle. Analyses of evolutionary gene transfers to the nucleus, by their nature, reveal only fully functional genes whose activity has replaced that of the organellar copy. It is less clear what are (if any) the intermediate steps in the process of functional transfer which lead to full activation. Also these studies give no indication of the frequency with which a gene can be functionally relocated to the nucleus. This is something of keen interest to plant biotechnologists for whom it is essential to fully understand the level of transgene containment provided by maternal inheritance (Maliga, 2003). Investigation of these questions can only be satisfactorily achieved experimentally.

Recently, new techniques have enabled the experimental recapitulation and dissection of molecular events involved in endosymbiotic gene transfer. The frequency of DNA transfer from the chloroplast to the nucleus in tobacco (*Nicotiana tabacum*) has been measured (Huang *et al.*, 2003; Stegemann *et al.*, 2003; Sheppard *et al.*, 2008) and was found to be unexpectedly high. These studies used transplastomic plants containing, within the chloroplast genome, a chloroplast-specific aminoglycoside resistance gene (*aadA*) used to select the transplastomic lines, and a kanamycin resistance gene (*neo*) designed for exclusive nuclear expression. Though inactive in the chloroplast, migration of *neo* to the nucleus could be monitored by screening for kanamycin resistance in transplastomic cells in culture or the seedling progeny of transplastomic plants. In the initial experiments (Huang *et al.*, 2003) pollen from transplastomic flowers was back-crossed to wild-type female and the resulting progeny screened for kanamycin resistance. Sixteen plants among ~250,000 progeny (1 in 16,000) showed heritable kanamycin resistance, indicating transfer of the *neo* gene from the chloroplast to the nucleus. Southern analysis of the 16 gene transfer lines revealed different lengths of flanking chloroplast DNA transferred in each case indicating that these were all independent events. In a subsequent study (Sheppard *et al.*, 2008), an identical screen was undertaken and an even greater frequency of one gene transfer event in 11,000 seedlings was

observed. In this study, progeny from the reciprocal cross (where wild-type pollen was used to fertilise transplastomic ovules) were also screened. Interestingly only one atypical kanamycin resistant seedling was observed in this screen of over 270,000 seedlings, indicating that the frequency of DNA transfer in the female germline is much lower than in the male germline. Similar experiments were also undertaken to investigate the frequency of chloroplast DNA transfer in somatic cells (Stegemann *et al.*, 2003). A frequency of one transfer in ~5,000,000 cells was observed, which is also much lower than the frequency observed in the male germline. The elevated frequency of transfer through the male germline is likely due to the release of chloroplast DNA into the cytoplasm during plastid breakdown in the developing pollen grain which accompanies (and probably causes) maternal plastid gene inheritance (Sheppard *et al.*, 2008).

In each of these studies the gene transferred (*neo*) was engineered for nuclear expression and needed no molecular changes to allow its expression, other than to be integrated into a domain of the nuclear genome that was amenable to transcription. The frequencies observed, therefore, indicate purely the minimum rate at which DNA transfers from the chloroplast to the nucleus. Perhaps the more interesting question both from evolutionary and biotechnological perspectives relates to the likelihood of functional gene transfer. Functional gene transfer is a more involved process and requires not only the transfer of DNA encoding the gene to the nucleus but also the acquisition of a eukaryote-type promoter, a polyadenylation signal and, if the protein is to be targeted back to the organelle, a transit peptide. This chapter describes a secondary screen designed to recover activation of a chloroplast gene previously transferred to the nucleus in the original screens outlined above. The DNA fragments transferred in the original experiments were large (Huang *et al.*, 2003; Sheppard *et al.*, 2008) and so, in the majority of lines generated, *neo* transfer was accompanied by several kilobases of flanking chloroplast DNA that almost always included the closely linked chloroplast selectable marker gene *aadA*. As *aadA* in these experiments was driven by the *psbA* chloroplast promoter (Huang *et al.*, 2003) this gene was expected to be inactive in the nucleus/cytoplasm genetic compartment (Huang *et al.*, 2003). The *aadA* gene would be expressed only if it acquired a nuclear promoter and the transcript was translated on 80S ribosomes - paralleling the pathway of genes functionally transferred during endosymbiotic evolution.

A previous study (Stegemann and Bock, 2006) used this pair of reporter genes arranged in the nucleus in a head-to-tail transcriptional polarity with 35S-driven *neo* upstream of *aadA* (with *aadA* driven by the rRNA operon promoter). This set of experiments monitored nuclear activation of the *aadA* gene and uncovered a high frequency of acquisition of aminoglycoside resistance. In every case activation was due to simple small deletions which brought the upstream 35S promoter driving *neo* (arranged in the same transcriptional polarity) to the 5' end of *aadA* and no activation

by a native nuclear promoter or other evolutionarily relevant process was observed. Interestingly, *aadA* transcripts were found to be polyadenylated, despite the lack of any sequence change in the *psbA* 3' UTR found downstream of the *aadA* coding sequence. 3' RACE showed that mRNA cleavage and polyadenylation occurred at a site within the *psbA* 3' UTR. This region fortuitously matched the flexible AT-rich plant polyadenylation site consensus sequence.

To simulate more closely the natural process of functional endosymbiotic gene transfer, we used an arrangement of the reporter gene cassette that precludes activation through recruitment of the 35S promoter by simple deletion, thereby enabling more complex and evolutionarily authentic mechanisms of nuclear activation to be revealed. In addition, activation of *aadA* in many different genomic contexts was assessed by screening 16 independent lines each representing a different nuclear location of the transgene, thereby allowing a greater chance of revealing different sorts of activation events.

2.2 Results

2.2.1 Screen for *aadA* activation

To identify events leading to nuclear expression of *aadA,* leaf explants from 16 kanamycin resistant tobacco lines (Appendix 1) in which *aadA* and *neo* had been co-transferred to independent nuclear genomic loci (Huang *et al.*, 2003; Sheppard *et al.*, 2008) were screened for aminoglycoside resistance. 1000 explants (each explant ~34 mm^2) from each line were placed on plant regeneration medium containing spectinomycin (400 mg L^{-1}) and streptomycin (200 mg L^{-1}) to select for nuclear activation of *aadA*. Spontaneous mutations in the ribosomal RNA genes can give rise to either spectinomycin (Svab and Maliga, 1991) or streptomycin (Etzold *et al.*, 1987) resistance. The spontaneous mutants are resistant to only one drug and so inclusion of both antibiotics prevented their growth, ensuring that shoots only grew as a result of *aadA* activation.

Two distinct types of resistance were observed: several lines displayed immediate resistance to aminoglycosides in all cells of all explants screened; two other lines showed very rare growth foci where *aadA* was activated as a result of an acquired mutation in a single cell. All the remaining lines were entirely susceptible to the antibiotics (Appendix 1).

2.2.2 Multiple copy insertion of *aadA* leads immediately to aminoglycoside resistance

The *psbA* promoter shows weak nuclear activity thought to be dependent upon the presence of fortuitous TATA and CAAT boxes (Cornelissen and Vandewiele, 1989). For this reason, the *psbA* promoter used in this study was truncated at the 5' end to remove the CAAT box. Despite this modification, all cells of all explants from two of the 16 lines screened (kr2.7 and kr2.9) displayed strong resistance to aminoglycosides (Figure 2.1A) and three others showed various degrees of

partial resistance (e.g. kr2.10, Figure 2.1A; Appendix 1). Most lines were completely sensitive (e.g. kr2.2, kr2.3, Figure 2.1A). Southern blots (Sheppard *et al.*, 2008) indicate that the two highly resistant lines contain multiple insertions of chloroplast DNA in the nucleus, whereas, lines that are susceptible have Southern signals consistent with single or low copy numbers of the chloroplast integrant (Sheppard *et al.*, 2008). To determine if the copy number of *aadA* was correlated with *aadA* expression and aminoglycoside resistance, the relative copy number was ascertained using quantitative slot-blot hybridisation and this was compared with *aadA* transcript accumulation which was assessed by RT-PCR. Relative *aadA* gene copy numbers were determined for all lines in this study that arose from the screen undertaken by Sheppard *et al.*, (2008): kr2.2, kr2.3, kr2.7, kr2.9 and kr2.10 (Figure 2.1B). Assuming that kr2.2, the line giving the lowest *aadA* hybridisation signal, harbours a single copy of the gene, kr2.3, kr2.7, kr2.9 and kr2.10 contain 3, 11, 17 and 5 copies respectively (these are approximate figures given the compound variance associated with the calculations). Transcript accumulation of *aadA,* assessed by RT-PCR relative to *RPL25* in these lines (Figure 2.1C) was found to be significantly correlated ($r = 0.93$, $p < 0.05$) with the number of nuclear copies of the gene.

2.2.3 Frequency of multiple copy *aadA* insertion

Of the 16 gene transfer lines tested, 2 (12.5%) contained multiple, probably unchanged, copies of *aadA* in the nucleus sufficient to confer strong aminoglycoside resistance (kr2.7 and kr2.9). Given that gene transfer from the chloroplast to the nucleus occurs once in every 11,000-16,000 pollen grains (Huang *et al.*, 2003; Sheppard *et al.*, 2008) this corresponds to a frequency of 1 active gene transfer due to multiple copy insertion in ~88,000-128,000 pollen grains when the modified *psbA* promoter is used.

2.2.4 Mutational activation of *aadA*

Lines kr2.7, kr2.9 and kr2.10 showed growth without any secondary change to *aadA* sequence due to multiple copy insertion of *aadA.* The remaining 13 kr lines showed sufficient aminoglycoside sensitivity to be effectively screened for mutational activation of *aadA*. Figure 2.2A shows one of three resistant shoots that emerged from such a screen. Expression of *aadA* was examined by RT-PCR of leaf RNA samples and demonstrated elevated *aadA* transcripts in the three resistant shoots (Figure 2.2B). The shoots were placed on 0.5 × MS agar medium to generate roots, then transferred to soil and grown into mature plants designated sr1, sr2 (both from kr2.2) and sr3 (from kr18). After plant regeneration, *aadA* activation in sr1 and sr2 was confirmed by real-time quantitative PCR [RT-QPCR](Figure 2.2C) and progeny were tested for the inheritance of aminoglycoside resistance (Figure 2.3). Sr1 showed a significant deviation from the expected 3:1 (R:S) ratio, with a large excess of resistant progeny ($R = 116$, $S = 3$, $p = 1.49^{-08}$). Sr2 segregated normally ($R = 212$, $S = 74$, $p = 0.73$). The sr1 and sr2 parent line, kr2.2, showed a small number (< 1%) of apparently

resistant progeny, always at the plate periphery (Figure 2.3), which became fully bleached with time. This is likely attributable to a very low level of *aadA* expression (Figure 2.2C) combined with a local, perhaps condensation induced, fluctuation in antibiotic concentration. In line sr3 increased *aadA* transcript accumulation, though clear in the first assay (Figure 2.2B), could not be detected by RT-QPCR at plant maturity (Figure 2.2C) and the spectinomycin resistance phenotype was not inherited by progeny (Figure 2.4), demonstrating that the resistance phenotype was unstable in sr3.

Possible explanations for a high number (~97%) of resistant progeny observed in sr1 are either multiple copies of the activated *aadA* gene segregating independently, or that sr1 is homozygous at this locus and that the few sensitive progeny are due to silencing or *nupt* instability (Sheppard and Timmis, 2009). In an attempt to distinguish between these two possibilities quantitative slot-blot hybridisation was performed to determine *aadA* copy number and also a PCR was performed using primers kr2.2NJR and kr2.2AJR2 which flank the kr2.2 chloroplast insertion site (see chapter 3). Amplification of a 399 bp product in the pre-insertion site PCR (Figure 2.5) indicated that sr1 was hemizygous for the chloroplast DNA insertion locus. Slot blot hybridisation was inconclusive and so the presence of a second copy of *aadA* could not be excluded (Figure 2.6).

2.2.5 Activation of *aadA* in sr1 occurred by acquisition of the 35S nuclear promoter

To determine the mechanism of gene activation in the two lines showing activation of *aadA*, the sequence 5' of this gene was recovered (Figure 2.7) using TAIL-PCR and genome walking. The transcription start site was determined using 5' RLM-RACE (Figure 2.7). In sr1 *aadA* was activated by acquisition of the nearby 35S promoter which, together with the first 47 bp of *neo,* was duplicated, inverted and inserted in reverse polarity upstream of *aadA* (Figure 2.7B).

2.2.6 Activation of *aadA* in sr2 occurred by nuclear activation of the native chloroplast promoter

In sr2 activation of *aadA* was achieved by the recruitment of nuclear gene regulatory elements within the 35S promoter to enhance nuclear expression from the *psbA* promoter (Figure 2.7C). Regulatory elements were recruited through a 98 bp deletion which brought the 35S promoter closer to the *psbA* promoter. Two adjacent single nucleotide substitutions 25 bp upstream from the site of the deletion were also observed that did not appear to have any functional significance (see below).

To verify that the relocated 35S promoter was the cause of activation an *in vivo* transient expression assay was performed. This assay quantified the expression of the GUS reporter gene from a series of promoters involving variations of the initial *psbA* promoter and additional

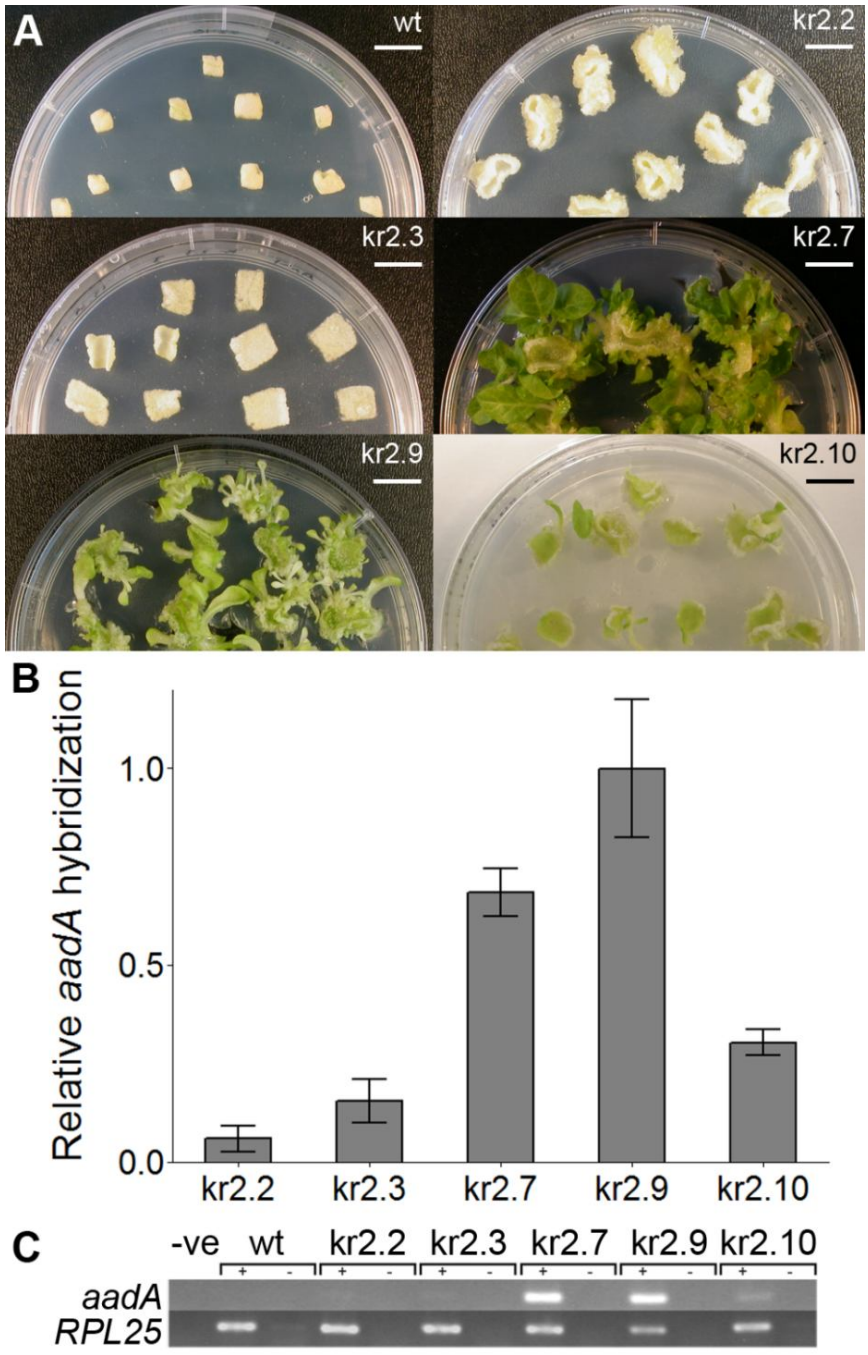


Figure 2.1 Comparison of aminoglycoside resistance, *aadA* copy number and *aadA* transcript accumulation. (**A**) Explants grown on regeneration medium containing 400 mg L^{-1} spectinomycin and 200 mg L^{-1} streptomycin. Scale bars = 10 mm. (**B**) DNA slot blot was performed using DNA from known hemizygous plants. Triplicates of each sample were probed with an *aadA* probe followed by stripping the filter and probing with ribosomal DNA as a loading control. The graphs show average *aadA* hybridisation, relative to kr2.9, after normalisation to ribosomal DNA hybridisation. Error bars show standard deviation. (**C**) RT-PCR showing *aadA* transcript accumulation with (+) and without (-) reverse transcriptase. A no template control is included (-ve). Control RT-PCRs with *RPL25* primers are also shown.

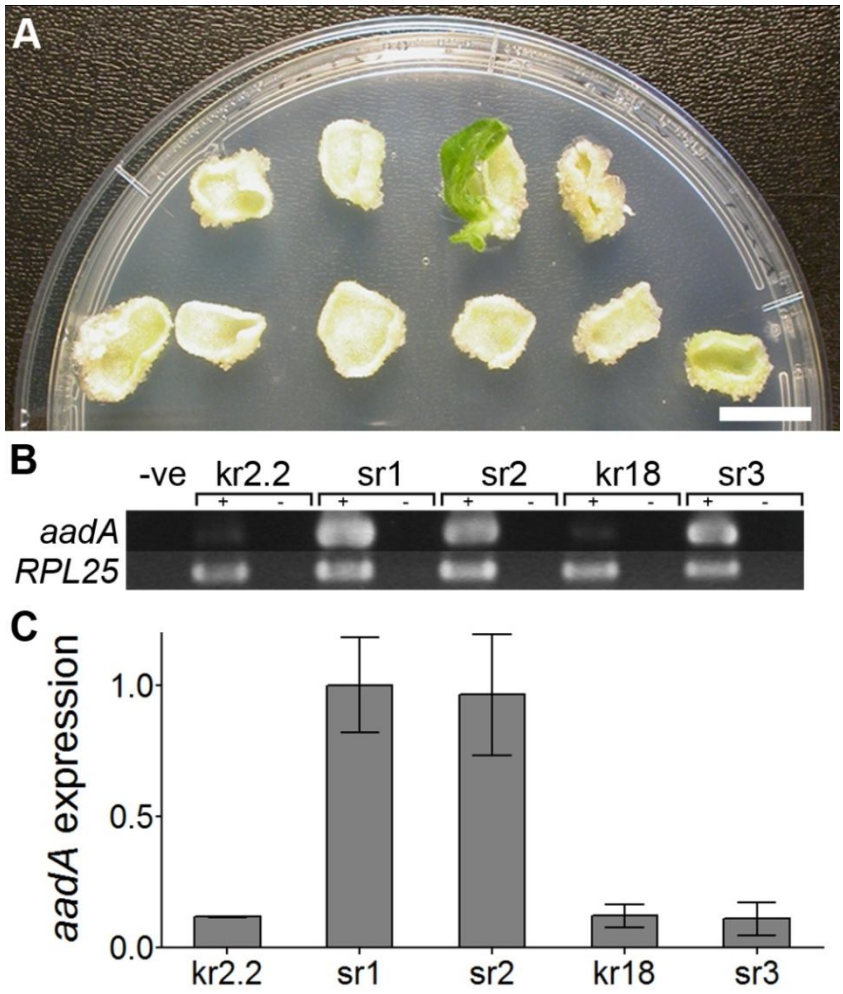


Figure 2.2 Nuclear activation of *aadA* in lines sr1 and sr2. (**A**) Selection for *aadA* activation in line kr2.2. Explants were grown on plant regeneration medium with 400 mg L^{-1} specinomycin and 200 mg L^{-1} streptomycin. A single resistant shoot (sr1) can be seen. Scale bar = 10 mm. (**B**) RT-PCR demonstrating *aadA* transcript accumulation (+), no RT (-). A no template control is included (-ve). Template control RT-PCRs with *RPL25* mRNA primers are also shown. (**C**) RT-QPCR shows *aadA* mRNA accumulation in leaves. *RPL25* normalised data are shown relative to sr1 mRNA levels. Error bars show standard deviation. Sr1 and sr2 show a significant increase in expression compared with the parent line kr2.2 ($p < 0.05$, Student's t-test). Sr3 showed no increase in *aadA* expression.

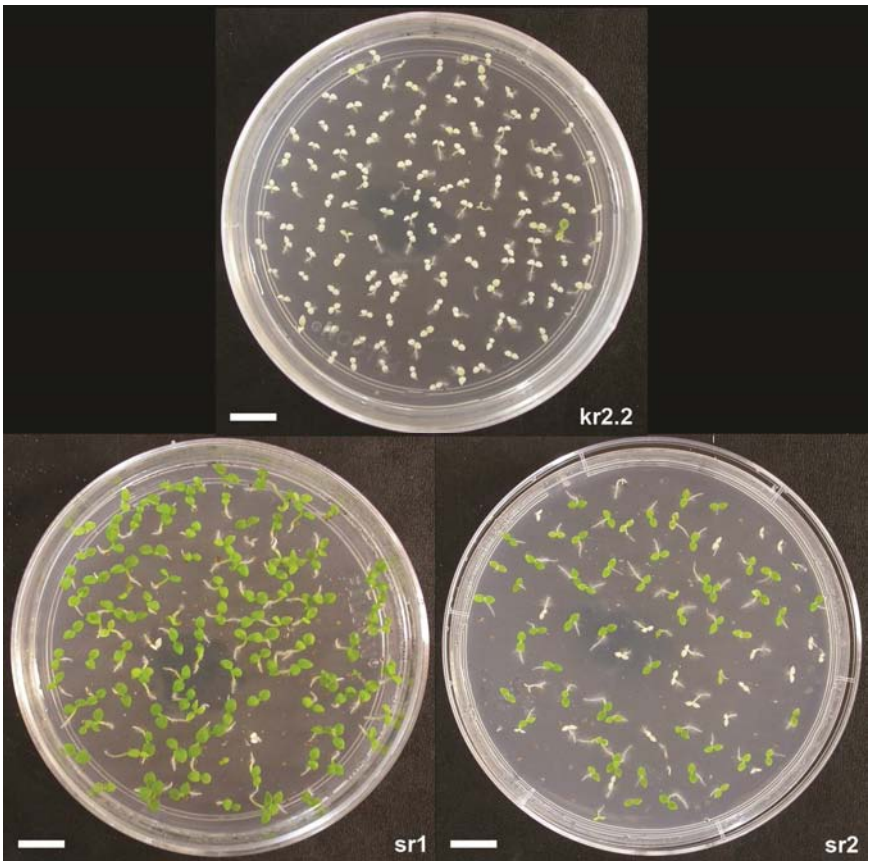


Figure 2.3 Inheritance of spectinomycin resistance in sr1 and sr2. Seedlings from self-fertilised capsules were grown on 0.5 × MS medium containing 200 mg L^{-1} spectinomycin. Sr1 shows a significant deviation from the expected 3:1, resistant:sensitive ratio (R = 116, S = 3, p = 1.49^{-8}). Sr2 segregates according to expectation (R = 60, S = 23 p = 0.57). Kr2.2 showed a small number (< 1%) of apparently resistant progeny, always at the plate periphery, which became fully bleached with time. Scale bars = 10 mm.

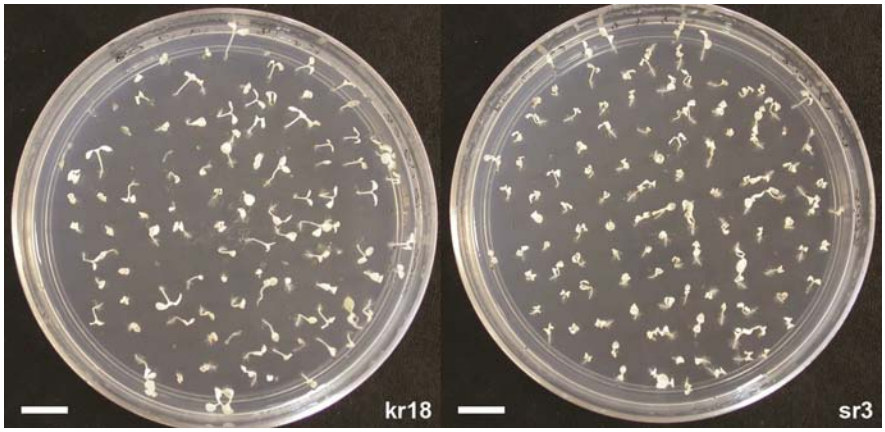


Figure 2.4 Inheritance of spectinomycin resistance in sr3. Kr18 and sr3 seedlings grown on plates containing 0.5 × MS medium with 200 mg L^{-1} spectinomycin. No spectinomycin resistance was seen in sr3 seedlings. Scale bar = 10mm.

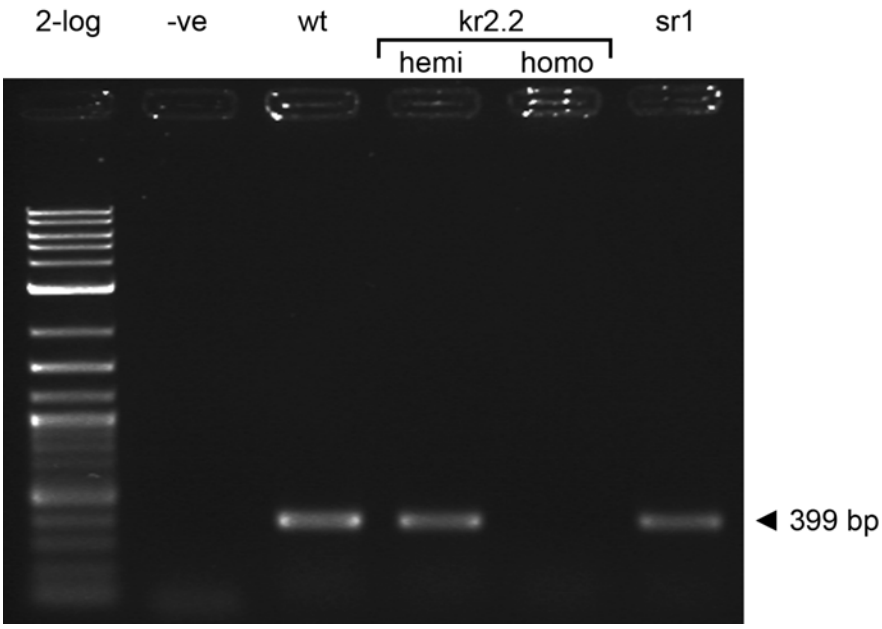


Figure 2.5 Sr1 is hemizygous for the kr2.2 chloroplast insertion locus. PCR used primers flanking the line kr2.2 chloroplast insertion locus (Chapter 3) and wild-type (wt), kr2.2 hemizygous (hemi), kr2.2 homozygous (homo) or sr1 DNA as template. A no template control was also included (-ve). A product (399 bp) is expected only from wild-type chromosomes (insertion of the ~17 kb chloroplast integrant between the primer binding sites prevents amplification).

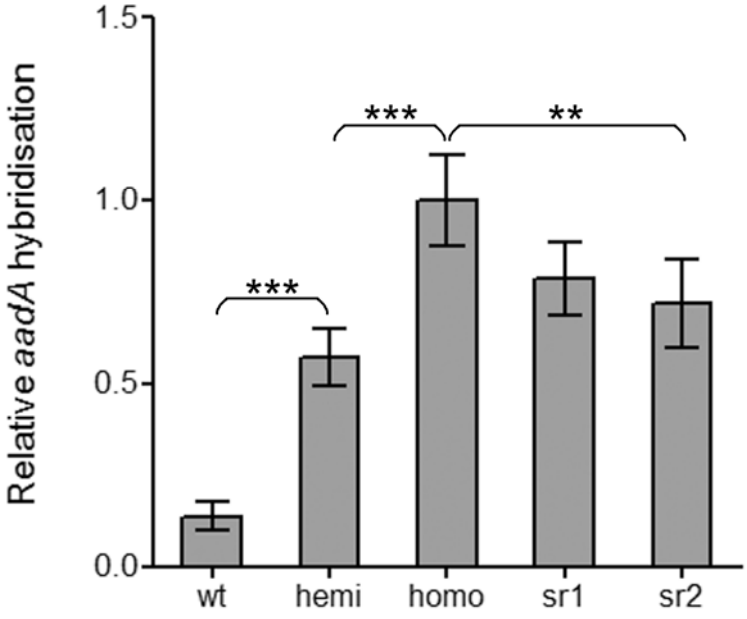


Figure 2.6 sr2 has a single copy of *aadA*; sr1 copy number is unclear. DNA slot blot was performed using DNA from the following; wild-type (wt), hemizygous kr2.2 (hemi) containing a single copy of *aadA*, homozygous kr2.2 (homo) containing two copies of *aadA* and sr1 and sr2 both hemizygous for the kr2.2 insertion locus. Quadruplicates of each sample were probed with an *aadA* probe followed by probing with ribosomal DNA as a loading control. The graphs show average *aadA* hybridisation, relative to homozygous kr2.2, after normalisation to ribosomal DNA hybridisation.

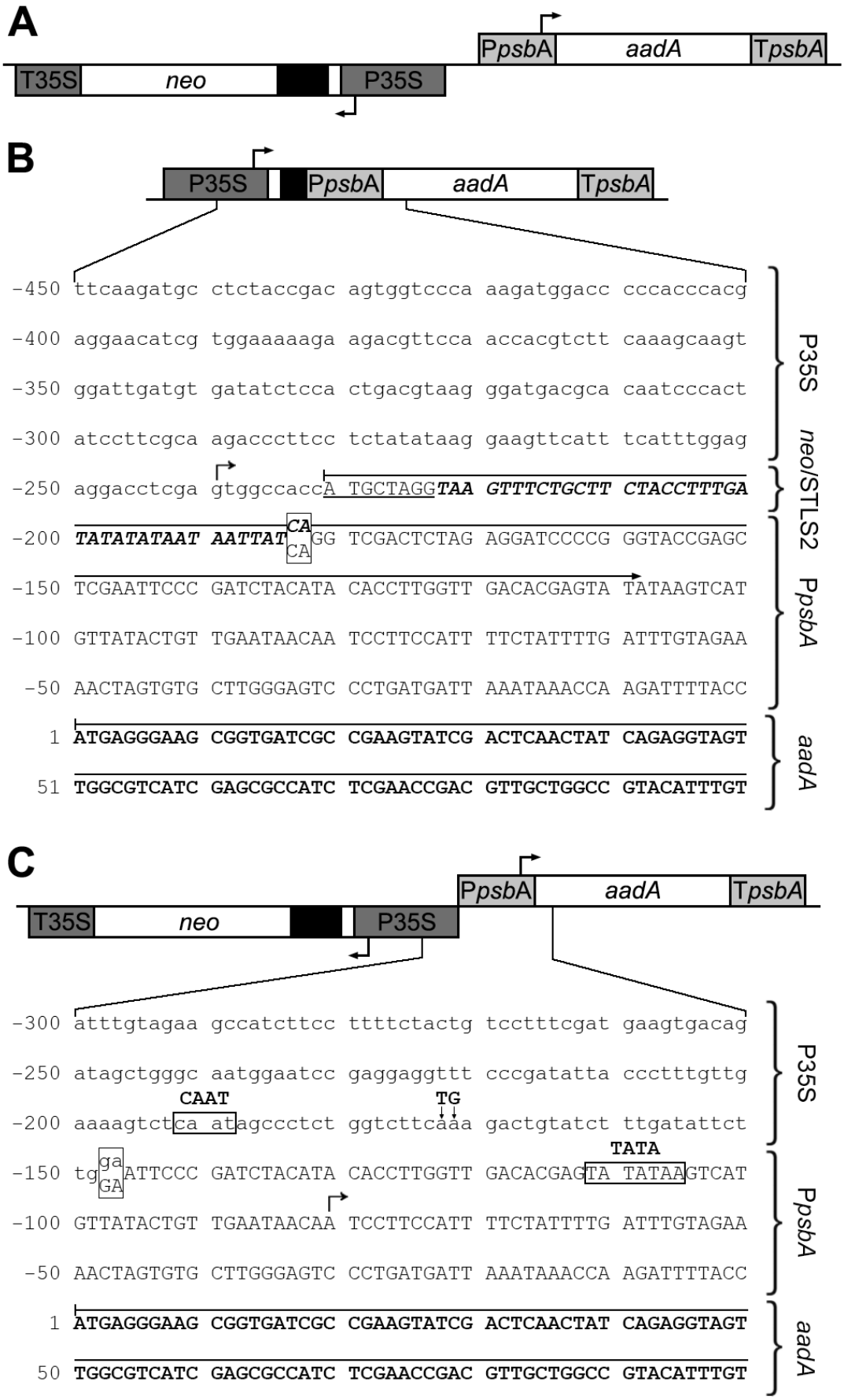

A
PpsbA
aadA
TpsbA
T35S
neo
P35S
B
P35S
PpsbA
aadA
TpsbA
-450 ttcaagatgc ctctaccgac agtggtccca aagatggacc cccacccacg
-400 aggaacatcg tggaaaaaga agacgttcca accacgtctt caaagcaagt
-350 ggattgatgt gatatctcca ctgacgtaag ggatgacgca caatcccact
-300 atccttcgca agacccttcc tctatataag gaagttcatt tcatttggag
-250 aggacctcga gtggccaccA TGCTAGGTAA GTTTCTGCTT CTACCTTTGA
-200 TATATATAAT AATTATCA GG TCGACTCTAG AGGATCCCCG GGTACCGAGC
CA
-150 TCGAATTCCC GATCTACATA CACCTTGGTT GACACGAGTA TATAAGTCAT
-100 GTTATACTGT TGAATAACAA TCCTTCCATT TTCTATTTTG ATTTGTAGAA
-50 AACTAGTGTG CTTGGGAGTC CCTGATGATT AAATAAACCA AGATTTTACC
1 ATGAGGGAAG CGGTGATCGC CGAAGTATCG ACTCAACTAT CAGAGGTAGT
51 TGGCGTCATC GAGCGCCATC TCGAACCGAC GTTGCTGGCC GTACATTTGT
P35S
neo/STLS2
PpsbA
aadA
C
PpsbA
aadA
TpsbA
T35S
neo
P35S
-300 atttgtagaa gccatcttcc ttttctactg tcctttcgat gaagtgacag
-250 atagctgggc aatggaatcc gaggaggttt cccgatatta ccctttgttg
CAAT
TG
-200 aaaagtctca atagccctct ggtcttcaaa gactgtatct ttgatattct
TATA
-150 tgga ATTCCC GATCTACATA CACCTTGGTT GACACGAGTA TATAAGTCAT
GA
-100 GTTATACTGT TGAATAACAA TCCTTCCATT TTCTATTTTG ATTTGTAGAA
-50 AACTAGTGTG CTTGGGAGTC CCTGATGATT AAATAAACCA AGATTTTACC
1 ATGAGGGAAG CGGTGATCGC CGAAGTATCG ACTCAACTAT CAGAGGTAGT
50 TGGCGTCATC GAGCGCCATC TCGAACCGAC GTTGCTGGCC GTACATTTGT
P35S
PpsbA
aadA

Figure 2.7 Sequence rearrangements leading to *aadA* activation in sr1 and sr2. (**A**) Arrangement of experimental cassette in parent line kr2.2. (**B**) Sequence of part of the experimental cassette in sr1, including: 35S promoter (lowercase), *neo* first exon (underlined), *STLS2* intron (bold italics), *psbA* promoter (uppercase) and *aadA* ORF (bold). Transcription of *aadA* (black arrow) starts at the expected 35S transcription site. An upstream 41 aa ORF exists from -241 → -89. The *aadA* ORF begins at +1 (ORFs are indicated by black bars above the sequence). A black box marks the junction of *STLS2* intron with the *psbA* promoter. The CA dinucleotide is shared by both the *STLS2* intron and the *psbA* promoter representing microhomology that is commonly found at sites of double strand break repair. (**C**) Sequence of part of the experimental cassette in sr2, including: 35S promoter (lowercase), *psbA* promoter (uppercase) and *aadA* (bold). Transcription of *aadA* is from the *psbA* promoter at position -81 relative to the start of translation (black arrow), 4 nt downstream of the plastid transcription start site. A TATA box is present within the *psbA* promoter and a putative CAAT box is present within the 35S promoter. A 98 bp deletion occurred in line sr2 between the 35S promoter and the *psbA* promoter. The junction sequence on either side of the deletion is marked by a black box. The GA dinucleotide present at this site is found in both the 35S promoter and the *psbA* promoter. Two adjacent single nucleotide substitutions are also found in sr2 at positions -172 and -173. The bases in bold shown above the substitutions show the original sequence of kr2.2 at each position. Arrows indicate the direction of transcription.

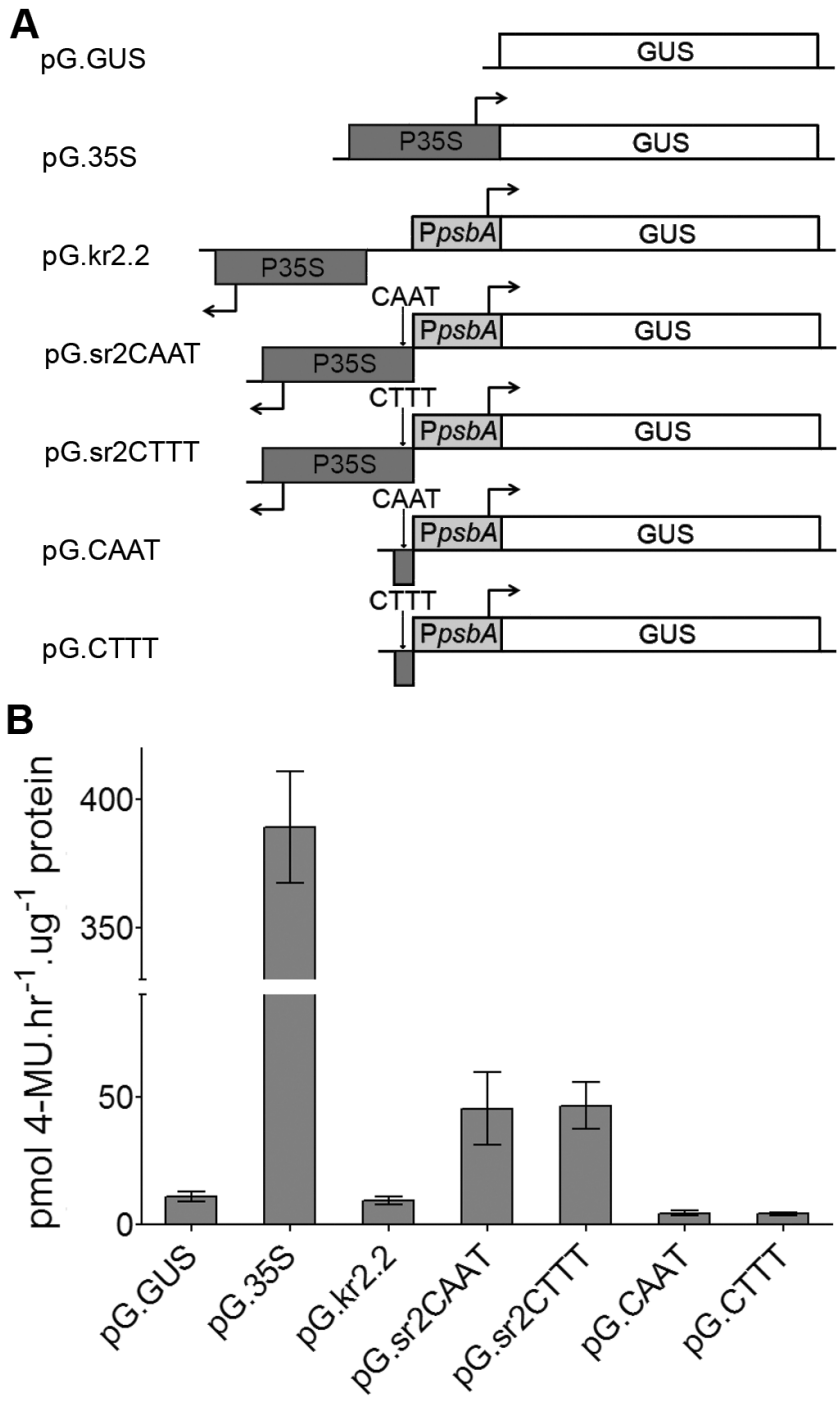


Figure 2.8 Transient expression analysis of the novel sequence motifs that activate *aadA*. (**A**) GUS expression constructs. Promoters as follows: pG.GUS – negative control, no promoter; pG.35S – positive control, 35S promoter; pG.kr2.2 – kr2.2 promoter (sequence 5' of *aadA* in line kr2.2); pG.sr2CAAT – sr2 promoter (sequence 5' of *aadA* in line sr2); pG.sr2CTTT – sr2 promoter with putative CAAT box sequence altered to the sequence CTTT; pG.CAAT - truncated sr2 promoter that included the putative CAAT box but no other known regulatory elements further 5'; pG.CTTT – similar to pG.CAAT but with putative CAAT box altered to the sequence CTTT. (**B**) Quantitative measurement of GUS activity. Total protein was prepared from leaf segments three days after infiltration with *Agrobacterium* strains containing the various expression constructs. GUS activity is shown as picomoles of product (4-methylumbelliferone; 4-MU) generated per hour per μg of total leaf protein. Values shown are the average activity of triplicate samples of protein extract with each sample comprising material from three independent infiltrations. Error bars show the standard deviation. Arrows below the line in (**A**) indicate 35S promoter sequence in reverse orientation.

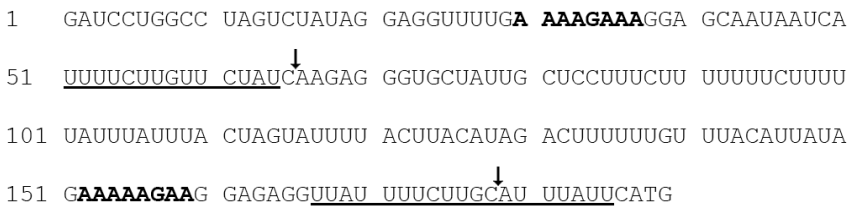


Figure 2.9 Polyadenylation sites within the *psbA* 3' UTR. There are two sites of mRNA cleavage and polyadenylation found within the *psbA* 3' UTR (↓). In each case the polyadenylation site matches the consensus sequence for plant polyadenylation which consists of a 6-10 nt A rich region (in bold) found 10-40 nt upstream of the cleavage site, with cleavage occurring at a di-nucleotide consisting of a pyrimidine followed by an A or C within a U rich region (underlined).

upstream nuclear sequences from sr2 and kr2.2 (Fig. 2.8). Kr2.2 is the parent line from which sr2 was derived (Sheppard *et al.*, 2008). Constructs pG.kr2.2 and pG.sr2CAAT contained the *psbA* promoter sequence and upstream nuclear sequences from kr2.2 and sr2 respectively (Figure 2.8A). An increase in GUS expression from construct pG.sr2CAAT relative to pG.kr2.2 ($p < 0.01$, 1 way ANOVA, Bonferroni corrected) confirmed that the closer proximity of the 35S promoter, even though it was in reverse orientation, was responsible for increased *aadA* expression *in vivo* (Figure 2.8B).

The deletion that re-positioned the 35S promoter also introduced a CAAT motif at approximately the correct position for activity relative to the TATA motif in the native *psbA* promoter. The effect on expression of this putative CAAT box within the *de novo* sr2 promoter was investigated by mutation of the CAAT sequence to CTTT in two constructs (Fig 2.8, pG.sr2CTTT, pG.CTTT). GUS expression from these 'CTTT' constructs showed no reduction when compared constructs that contained the CAAT motif but were in all other ways identical (Fig 2.8, pG.sr2CAAT, pG.CAAT). Deletion of 35S promoter sequence further 5' of *aadA* (constructs pG.CAAT and pG.CTTT), however, did lead to a reduction in GUS expression ($p < 0.001$, 1 way ANOVA, Bonferroni corrected) when compared with that observed for the full length constructs (pG.sr2CAAT, pG.sr2CTTT). These results indicate that the CAAT box had little influence on expression and that enhancer elements further 5' of the CAAT sequence were responsible for activation. In addition to the 98 bp deletion, there was a di-nucleotide substitution at a site 25 bp upstream from the site of the deletion (Figure 2.7C). The importance of these substitutions with regard to activation was able to be determined in the transient expression assay by including them in the constructs pG.sr2CAAT, pG.sr2CTTT, pG.CAAT, and pG.CTTT (Figure 2.8). Neither the presence of the CAAT box and the substitutions (pG.CAAT), nor the substitutions alone (pG.CTTT), lead to higher expression than that observed when both were absent (pG.kr2.2). These results do not rule out the possibility that the substitutions work in conjunction with upstream enhancers to increase expression, but it can be concluded that they do not contribute independently to *aadA* activation.

2.2.7 Maturation of *aadA* transcripts

Polyadenylation was determined by 3' RACE and found to occur at two separate sites within the 189 bp *psb*A UTR, 3' of the *aadA* coding sequence (Figure 2.9). Both sites matched the rather flexible AT-rich plant polyadenylation consensus sequence (Li and Hunt, 1997) and one was described previously (Stegemann and Bock, 2006). 5' RLM-RACE, which amplifies only cDNA from full length capped mRNA, was used to determine the *aadA* transcription start site and to confirm the correct 5' maturation of transcripts.

2.2.8 Frequency of gene activation

The frequency of gene activation, which is of consequence both from evolutionary and biotechnological perspectives, was determined. To calculate the frequency of gene activation a measure of cell density was needed. The total cell density was determined for leaves equivalent to those used in the *aadA* activation screen as well as for leaves of a similar age and size grown in soil. As in the *aadA* activation screen, vascular tissue was minimised by using leaf tissue taken from between leaf veins. Approximately 8 cm long leaves from plants grown in tissue culture jars were found to have a cell density of 1509 ± 113 (n=3) cells mm^2. Interestingly leaves taken from plants grown in soil were found to have a cell density 10 times higher at 14952 ± 1170 (n=3) cells mm^2. Although precise percentages of various cell types were not determined, the majority of cells were mesophyll (spongy and palisade) with a smaller percentage of epidermal cells and a minority of vascular cells.

Given an average explant area of ~34mm^2, and 1000 explants for each line, a total of 650 million cells were screened with a resulting estimated frequency of one activation in ~2 × 10^8 cells (kr2.7 and kr2.9 were excluded as they displayed homogeneous aminoglycoside resistance; kr2.10 was also excluded as it displayed sufficient basal levels of aminoglycoside resistance to prevent efficient screening for mutational *aadA* activation).

2.3 Discussion

Although it has become increasingly clear that organelle genomes have contributed large numbers of genes to the nuclear genomes of all eukaryotes, this pathway is usually not appropriately recognised in discussions of the possible origins of new nuclear genes (Kaessmann, 2010). In higher plants 14% of nuclear genes are thought to be derived from the chloroplast and its ancestor (Deusch *et al.*, 2008) and studies in yeast suggest an even greater number of nuclear genes may be derived from the mitochondrial ancestor (Esser *et al.*, 2004). While the processes involved in endosymbiotic gene transfer are beginning to be understood, a number of key questions still remain unanswered. This chapter goes some way to revealing the complex pathways by which organelle genes transferred to the nucleus can become activated in this very different genetic environment.

2.3.1 Multiple copy insertion of *aadA* leads to nuclear expression

A surprising finding of this research was that explants from two lines in which *aadA* had been transferred to the nucleus, immediately displayed strong homogeneous resistance to aminoglycosides indicating significant *aadA* expression without sequence change. In this study the *aadA* gene was driven by the *psbA* promoter. The *psbA* promoter shows weak nuclear activity which is thought to be dependent upon the presence of fortuitous TATA and CAAT boxes

(Cornelissen and Vandewiele, 1989). For this reason, the *psbA* promoter used in this study was truncated at the 5' end to remove the CAAT box. Despite removal of this motif *aadA* showed a low level of nuclear transcription, indicating that while the CAAT box may contribute to nuclear expression it is not vital.

The two lines which showed high levels of aminoglycoside resistance (kr2.7 and kr2.9) both show strong Southern signals suggesting high copy number insertion, whilst lines that were sensitive to aminoglycosides had Southern signals consistent with low copy insertion. Further investigation showed that the level of *aadA* transcription was correlated with gene copy number, which, given that the chloroplast DNA insertions are sometimes very complex (up to 10-17 copies), can lead to significant levels of expression. The correlation of gene copy number and expression could either be due to additive low levels of expression from the *psbA* promoter, or it could be that multiple copies of the *psbA* promoter provide multiple chances for activation by the recruitment of nuclear enhancers. All lines, however, display Mendelian segregation ratios of the closely linked *neo* gene (except for kr2.9 which shows fewer than expected resistant progeny) suggesting that in probably all cases the multiple copies of *aadA* have been inserted at a single locus (Sheppard *et al.*, 2008). There is, therefore, unlikely to be a direct correlation between *aadA* copy number and the potential for activation by existing nuclear enhancers. Most likely the increase in expression is due to additive low level transcription from the multiple copies of the gene, with some variation depending on the local chromatin structure.

Functional gene transfer from the chloroplast to the nucleus through multi-copy nuclear insertion of *aadA* driven by the modified *psbA* promoter occurred at a frequency of one in ~88,000-128,000 pollen grains. This frequency is up to 3 times higher that of paternal plastid transmission (Ruf *et al.*, 2007; Svab and Maliga, 2007) and is more relevant to transgene containment as it results in regular nuclear inheritance, rather than the patchy inheritance associated with the heteroplasmic plants that result from paternal plastid transmission. These findings warn that potential nuclear activity of a chloroplast promoter should be considered when designing chloroplast transgenes, particularly in cases where complete containment is required. It also highlights that removal of the CAAT box from the *psbA* promoter is not sufficient to prevent nuclear expression from this promoter.

The nuclear activity observed is obviously specific to the *psbA* promoter used in this study and whether this is likely to be the case with other chloroplast promoters has yet to be determined. Future investigation into the nuclear activity of other chloroplast promoters will therefore be important for furthering our understanding of transgene containment.

2.3.2 Multi-copy insertion must lead to very large chloroplast DNA inserts

The copy number analysis also highlights that these chloroplast DNA insertions must, in some cases, be very large. For example, kr2.9 is estimated to have 17 copies of *aadA*. Based on Southern analysis undertaken by Sheppard *et al.* (2008) *neo* is present in similar copy number to *aadA.* The Southern also shows that approximately half of the copies of *aadA* originated from inverted repeat A and half from inverted repeat B of the chloroplast genome (Sheppard *et al.*, 2008) and that the chloroplast derived sequence extends at least 10.9 kb downstream of *neo* and at least 11.4 kb or 18.5 kb (depending on which inverted repeat) downstream of *aadA*. This equates to a minimum (on average) of 25.9 kb of chloroplast DNA per copy of the experimental cassette which, given an estimate of ~17 copies of the experimental cassette, amounts to a minimum of ~440 kb (almost 3 chloroplast genomes worth) of chloroplast DNA inserted at a single nuclear locus.

This is a very significant contribution to the nuclear genome but it is not unprecedented. Several large >100 kb insertions of organelle DNA have been identified in the nuclear genomes of Arabidopsis and rice (Stupar *et al.*, 2001; The Rice Chromosome 10 Sequencing Consortium, 2003) and recently hybridisation to maize chromosomes has revealed extensive chloroplast (Roark *et al.*, 2010) and mitochondrial (Lough *et al.*, 2008) DNA insertions, some of which appear to entail almost entire organelle genomes. These insertion loci vary amongst different inbred lines indicating ongoing nuclear transfer and deletion of organelle sequences.

2.3.3 Secondary rearrangements lead to *aadA* activation

Only one previous experimental study aimed to determine the rate of gene activation within the nucleus (Stegemann and Bock, 2006). In that study a chloroplast marker *aadA* which had been transferred to the nucleus showed activation by short simple deletions bringing the closely linked 35S nuclear promoter to the 5' end of *aadA* to drive its expression. The frequency at which *aadA* becomes activated in the nucleus must be artificially high as the 35S strong nuclear promoter is always present within 1 kb of the chloroplast gene, in the same transcriptional polarity. This issue was addressed using an alternative arrangement of the *neo* and *aadA* genes that precluded activation by simple deletion. Three activation events were observed, two of which showed heritable activation of the *aadA* gene (sr1 and sr2).

Activation in sr1 occurred through duplication, inversion and insertion of the 35S promoter. At the novel sequence junction created by the rearrangement, micro-homology was observed between the two sequences joined (Fig 2.7B) suggesting that it may have been mediated by the non-homologous end joining (NHEJ) pathway of double strand break (DSB) repair.

In sr2 activation occurred by recruitment of nuclear enhancer(s) within the 35S promoter to drive expression from the erstwhile chloroplast *psbA* promoter. It is known that a number of enhancer

elements within the 35S promoter are able to act in either orientation (Fang *et al.*, 1989) and presumably one of these was responsible. This activation by the recruitment of nuclear enhancers represents a novel pathway for nuclear activation of a chloroplast gene and may have been made possible through the pre-existing low nuclear activity of the *psbA* promoter. It is as yet unknown whether other chloroplast promoters harbour any nuclear activity and if the ability to activate a chloroplast promoter *via* nuclear enhancers applies more widely.

Micro-homology was also observed between the two sequences which were joined as a result of the deletion in sr2 (Fig 2.7C). Interestingly, micro-homology was also detected by Stegemann and Bock (2006) in most of the deletions that they observed, which together with the findings presented here indicate that NHEJ is probably the major pathway of *nupt* sequence shuffling. The sequence rearrangements generated can be relatively simple, as in the deletion observed in sr2, or more complex involving duplication and inversion of the chloroplast sequence, as in sr1. These findings indicate that both simple and complex changes, which probably arise from NHEJ mediated repair of DSBs, can rearrange nearby gene elements to create new sequence compositions that in some instances lead to newly active genes. This then raises the question as to whether areas of the genome or environmental conditions conducive to the formation of DSBs would lead to an increase in the rate of gene activation.

To calculate the frequency at which a gene becomes activated in the nucleus, a measure of cell density is needed. The value of leaf cell density that has been used previously in cell number based calculations (Stegemann *et al.*, 2003; Stegemann and Bock, 2006; Sheppard *et al.*, 2008) was determined for tobacco plants grown in soil (Hannam, 1968). It is known, however, that growing plants *in vitro*, as was the case in this study, can have profound morphological and physiological effects. These effects can lead to reduced palisade tissue and larger intercellular spaces (Gaspar *et al.*, 1987) both of which would affect the number of cells in a given area of leaf tissue. For this reason we determined the tobacco leaf cell density in the *in vitro* grown plants used in this study as well as in soil grown plants. The cell density determined for leaves taken from soil grown plants was comparable to those previously reported (Hannam, 1968) but there was a 10 fold reduction in cell density in *in vitro* grown leaves.

Using this value of leaf cell density for all calculations and comparisons, a frequency of one activation per 2×10^8 cells was estimated which is an order of magnitude lower than that ascertained by Stegemann and Bock (2006) reflecting the very different events needed to activate *aadA* in the two studies. Nonetheless, despite our changes to the experimental design, which precluded activation by recruitment of the 35S promoter through simple deletion, the 35S promoter was involved in all of the events that we observed, suggesting that the majority of

sequence rearrangements which take place are local in nature. This would suggest that the chromosomal location into which the gene inserts is likely to have a significant impact on the likelihood of gene activation. In this study the inclusion of 16 different novel integrant loci enabled the investigation of *aadA* activation in a large number of genomic contexts and yet still no activation involved native nuclear DNA clearly demonstrating that this is a very rare event. Indeed, given the large size of the screen undertaken, activation due to native nuclear sequence is probably too rare for experimental simulation.

2.3.4 Maturation of *aadA* transcripts

Not only was *aadA* transcribed in the nucleus, through multiple-copy insertion and mutational activation, but *aadA* transcripts also underwent correct mRNA maturation. Correct 5' maturation was determined using RLM-RACE which only amplifies a product from full-length, capped mRNA and 3' RACE was used to identify polyadenylated *aadA* transcripts. Remarkably, despite any sequence change, mRNA cleavage and polyadenylation occurred in different mRNAs at two separate sites within the *psbA* 3' UTR which supports earlier suggestions that the AT-rich nature of chloroplast non-coding regions may provide abundant chance polyadenylation sites (Stegemann and Bock, 2006) for newly transferred genes. Cryptic control elements, such as TATA and CAAT boxes, may also be present in other AT-rich DNA sequences, as evidenced by their presence in the *psbA* promoter.

2.4 Conclusion

Local sequence rearrangements, possibly induced by DSBs, can cause activation of a chloroplast gene newly relocated to the nucleus by creating novel sequence compositions that give rise to a functioning gene. The *psbA* promoter and terminator used in this study had fortuitous elements that also aided activation and lead to significant expression when the gene was present in multiple copies. Indeed it seems possible that some organelle genes may arrive in the nucleus complete with all of the requirements for nuclear expression: a promoter that has nuclear activity, a 3' UTR containing cryptic polyadenylation signals and, according to a recent report (Ueda *et al.*, 2008), may even encode a transit peptide. While such transfer is unlikely to give a high level of expression, low level expression could provide a “starting point” for gene transfer enabling selective maintenance of the gene while enhancement of transcription, polyadenylation and protein targeting is established by post-insertional mutation. However, provided chloroplast transgenes do not contain or are not placed adjacent to, sequences that can promote nuclear activation, containment by maternal inheritance is extremely effective with plastid leakage through the male parent representing the major threat to escape.

2.5 Materials and methods

2.5.1 Plant growth conditions

Nicotiana tabacum plants were grown either in soil (in pots) or in tissue culture jars containing 0.5 × MS salt medium (Murashige and Skoog, 1962) and 0.8% agar (0.5 × MS agar). Soil grown plants were grown in a controlled environment chamber with a 14 hr light/10 hr dark and 25˚C day/18˚C night growth regime. *In vitro* grown plants were grown in a controlled temperature room with a 16 hr light/8 hr dark cycle at 25˚C.

2.5.2 Starting material

Kr plant lines used in this study were generated in two previously published screens with the kr series from Huang *et al.* (2003) and the kr2 series from Sheppard *et al.* (2008). These lines were produced by crossing pollen from a transplastomic parent to female wild-type to remove the transplastome. Plants used in the aminoglycoside resistance explant screen were kanamycin resistant progeny from self fertilised capsules of the original kr plants. Known hemizygous plants were kanamycin resistant progeny obtained by crossing the original kr plants to wild-type.

2.5.3 Selection for spectinomycin resistance

Leaf explants with an average area of ~34 mm^2 were taken from 4-6 week old *in vitro* grown plants and placed on culture plates containing regeneration medium MS104 (Mathis and Hinchee, 1994) supplemented with 400 mg L^{-1} spectinomycin and 200 mg L^{-1} streptomycin. Resistant shoots were transferred to 0.5 × MS agar medium to root and were subsequently transferred to soil. ImageJ software (US National Institutes of Health, Bethesda, MD) was used to determine the average area of explants taken from two representative plates.

2.5.4 Analysis of antibiotic resistance in seedlings

Surface-sterilised seeds were grown on 0.5 × MS agar medium supplemented with 150 mg L^{-1} kanamycin or 200 mg L^{-1} spectinomycin. Plates were grown in a controlled temperature room at 25^{o}C with a 16 hr light/8 hr dark cycle.

2.5.5 Nucleic acid isolation

DNA was prepared either using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions, or by phenol/chloroform extraction. For phenol/chloroform extraction, 100 mg frozen leaf tissue was ground to a fine powder for 1 min at 30 htz in a Retsch MM300 grinder (F-Kurt Retsch GmbH & Co., Haan, Germany). The frozen ground tissue was suspended in 700 µL of DNA extraction buffer (1% [w/v] sarkosyl, 100 mM Tris-HCl, 100 mM NaCl, 10 mM EDTA, 2% [w/v] PVPP, pH 8.5) to which 700 µL of phenol/chloroform/isoamyl alcohol (25:24:1, v/v) was then added and tube mixed for 20 min. Samples were centrifuged at 1,500 × g

for 10 min, the upper aqueous layer transferred to a fresh tube and another 600 µL of phenol/chloroform/isoamyl alcohol (25:24:1, v/v) added. Samples were mixed again for 10 mins and centrifuged at 1,500 × g for 5 min. The upper aqueous layer was transferred to a fresh tube to which 60 µL of 3 M sodium acetate (pH 4.8) and 600 µL of isopropanol was added. DNA was allowed to precipitate for 5 min at room temperature with constant gentle mixing and then centrifuged at 16,000 × g for 2 mins. The supernatant was removed and the pellet washed in 70% ethanol before air drying. DNA was resuspended in 50-100 µL of H_2O containing 40 µg mL^{-1} RNase A.

RNA was prepared using an RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions.

2.5.6 PCR and sequencing

PCR for vector construction and sequencing was performed with *PfuTurbo* DNA polymerase (Stratagene, La Jolla, Ca) according to the manufacturer's instructions. All other PCRs were performed using *Taq* DNA polymerase (New England Biolabs, Ipswich, MA; or Roche, Basel, Switzerland) according to the manufacturers' instructions. Gel electrophoresis of PCR products was performed using 1% agarose gels in 1x TAE buffer and 2-Log DNA Ladder (New England Biolabs) was used for size comparison.

Prior to sequencing, PCR products were purified using either a PCR Purification Kit (Qiagen) or a Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. Products were either directly sequenced after purification using product specific primers or cloned into pGEM-T-Easy (Promega, Madison, WI) according to manufacturer's instructions and sequenced using primers T7 and SP6 as well as product specific primers where applicable. Details of these and all subsequent primers can be found in Appendix 2. Sequencing was performed using BigDye Terminator v3.1 (Applied Biosystems, Foster City, CA). Each 20 µL reaction contained 1 µL Ready Reaction Mix, 3 µL Sequencing Buffer, 3.2 pmol primer and template DNA. Thermal cycling was performed with an initial denaturation step at 96˚C for 2 min followed by 26 cycles of 96˚C for 30 sec, 50˚C for 15 sec and 60˚C for 4 min. Extension products were purified by isopropanol precipitation. Each 20 µL reaction was mixed with 80 µL 75% isopropanol, incubated for 15 min and centrifuged at 16,000 x g for 20 min. The supernatant was removed and a further 250 µL 75% isopropanol added. Centrifugation was then performed at 16,000 x g for 5 min, the supernatant removed and the pellet dried. Purified extension products were analysed using a 3730 DNA Analyzer (Applied Biosystems) by the Institute of Medical and Veterinary Science (Adelaide, South Australia).

2.5.7 RT-PCR

For kr line RT-PCR, RNA was extracted from leaf tissue taken from 3 week-old, *in vitro* grown, wild-type and hemizygous kr plants. For sr line RT-PCR, RNA was extracted from part of the resistant shoot that was transferred to rooting medium. DNA was removed from RNA samples using a TURBO DNA-free kit (Ambion, Austin, TX). Reverse transcription was performed using an Advantage RT-for-PCR kit (Clontech, Mountain View, CA) with an oligo(dT) primer in accordance with the manufacturer's instructions. Samples were also prepared without reverse transcription. For amplification of *aadA* cDNA, primers aadAF5 and aadAR3 were used. *RPL25* cDNA was amplified using primers L25F and L25R. ImageQuant TL software (GE Healthcare, Buckinghamshire, UK) was used to quantify bands in RT-PCR for correlation with slot blot hybridisation.

2.5.8 Real Time Quantitative PCR

Sr line RNA was prepared from plants two months after transfer to soil. Real Time Quantitative PCR was performed as described (Schmidt and Delaney, 2010). For amplification of *aadA* cDNA, primers QaadAF and QaadAR were used with. For relative quantification *RPL25* mRNA, amplified using primers QL25F and QL25F, was used as an internal standard.

2.5.9 TAIL-PCR

Template DNA was extracted from mature kr2.2 and sr line plants. TAIL-PCR was performed as described (Liu *et al.*, 1995) using degenerate primer AD3 (Sessions *et al.*, 2002) and gene specific primers aadAT1, aadAT2 and aadAT3 (Stegemann and Bock, 2006).

2.5.10 Genome walking

Template DNA was extracted from mature kr2.2 and sr line plants. Genome walking utilised a Universal GenomeWalker Kit (Clontech) according to manufacturer's instructions. To isolate sequence upstream of *aadA* in sr lines, primers aadAT1 and aadAT2 were used.

2.5.11 RACE

Total leaf RNA was extracted from 3-4 week old *in vitro* grown sr line and kr line plants. 5' and 3' RACE were performed using Marathon cDNA Amplification Kit (Roche) or FirstChoice RLM-RACE Kit (Ambion) according to manufacturers' instructions. For 5' RACE PCR of *neo,* primers neoR1 and neoR2 respectively were used in two successive rounds of PCR; for *aadA*, primers aadAT1, aadAT2 and aadAT3 respectively were used in three successive rounds of PCR. For 3' RACE PCR, primers aadAF1 and aadAF2 (Sheppard and Timmis, 2009) respectively were used in two successive rounds of PCR.

2.5.12 Cell counts

Cell counts were performed essentially as described (Humphries and Wheeler, 1960). Leaves from plants grown both in soil and *in vitro* on 0.5 × MS agar medium were tested. For both soil and *in*

vitro grown plants, four leaf pieces from each of three leaves were tested. Each piece of leaf tissue (50 – 150 mm^2) was individually digested with pectinase (Sigma, St Louis, MO), macerated to single cells and counted using a haemocytometer with a minimum of 10 technical replicates. The area of the leaf piece, the volume of the cell suspension and the cell count values were then used to determine the average cell density of each leaf. Leaf pieces from plants grown in tissue culture were digested for one hour at 60˚C, leaf pieces from plants grown in soil were vacuum infiltrated and digested for 14 hours at 37˚C. Vascular tissue was minimised by using leaf tissue taken from between leaf veins. ImageJ software (US National Institutes of Health) was used to determine area of leaf pieces. 95% confidence intervals were calculated using Prism 5 (GraphPad Software, Inc, CA, USA).

2.5.13 Construct design of transient expression vectors

The GUS.ocs expression cassette was excised as a SacI/HindIII fragment from pJKKmf(–):GUS.ocs (Kirschman and Cramer, 1988) and cloned into pGreen0029 (Hellens *et al.*, 2000) to create pG.GUS. Kr2.2 and sr2 promoters (regions upstream of *aadA*) were amplified using psbaR_NcoI and 35SR_BglII and cloned into NcoI/BamHI digested pG.GUS to create pG.kr2.2 and pG.sr2CAAT respectively. The 35S promoter was isolated from the kr2.2 PCR product by restriction with SacI and BglII and cloned into SacI/BamHI-digested pG.GUS to create pG.35S. The CAAT-containing promoter was amplified using psbAR_NcoI and 35SR2_BglII and cloned into NcoI/BamHI digested pG.GUS to create pG.CAAT. PCR mutagenesis of pG.sr2CAAT and pG.CAAT was performed using primers CTTTF and CTTTR to alter the CAAT box sequence to CTTT. The PCR products, encompassing the entire vectors, were circularised using blunt end ligation to form pG.sr2CTTT and pG.CTTT respectively. Expression cassettes were all sequenced prior to use.

2.5.14 Transient Expression Analysis

The pGreen system of binary vectors (Hellens *et al.*, 2000) was used in the transient expression assay. The transient expression assay was performed using an adapted *Agrobacterium* infiltration method (Sparkes *et al.*, 2006). Triplicate samples were prepared for each expression construct with each sample comprising material from three infiltrations. All samples were measured with three technical replicates. Each replicate (40 µL) contained 3 µg total protein. GUS activity was determined using a MUG fluorescent assay (Delaney *et al.*, 2007). The plate was incubated for 40 minutes at 37˚C prior to detection of 4-MU fluorescence. Fluorescence was measured and 4-MU levels were calculated by comparison with a standard curve generated using known concentrations of 4-MU. For the concentration of total protein used (75 µg/mL), linearity of fluorescence with respect to GUS concentration was determined ($R^2 > 0.99$). Data were compared using one-way ANOVA with Bonferroni post-hoc comparison. ANOVA was performed using Prism 5 (GraphPad Software, Inc).

2.5.15 DNA Blot Analysis

For slot blot analysis DNA was extracted from leaves of 4-6 week old *in vitro* grown wild-type and hemizygous kr and sr line plants. DNA slot blot was performed and image quantified as described (Sheppard and Timmis, 2009). Levels of *aadA* hybridisation and expression (as shown by RT-PCR) were compared. Correlation coefficient and p-value were calculated using Prism 5 (GraphPad Software, Inc).

Chapter 3: Characterisation of a *de novo* nuclear insertion of chloroplast DNA

3.1 Introduction

Within the nuclear genomes of all plants that have been sequenced to date are found tracts of DNA of chloroplast and mitochondrial origin that range from a few tens of base pairs up to many kilobases in size (Arthofer *et al.*, 2010; Hazkani-Covo *et al.*, 2010), the result of a constant deluge of organelle DNA entering the nucleus and integrating into nuclear chromosomes. A number of detailed analyses have highlighted interesting aspects of the arrangement of these sequences and provided insights into their sequence evolution.

Insertions of organelle DNA are often very large and in some cases encompass essentially the entire mitochondrial or chloroplast genome. Chromosome 2 in Arabidopsis contains a large (~620 kb) insertion of mitochondrial DNA that covers the entire mitochondrial genome and has several large internal duplications (Stupar *et al.*, 2001). This sequence shares >99% identity with the mitochondrial genome indicating that it is a recent insert (Lin *et al.*, 1999). Similarly, chromosome 10 of rice contains a 131 kb insertion of chloroplast DNA which covers almost the entire chloroplast genome and it also shares >99% identity with the organelle sequence from which it was derived (The Rice Chromosome 10 Sequencing Consortium, 2003). Recently hybridisation to maize chromosomes has revealed extensive chloroplast (Roark *et al.*, 2010) and mitochondrial (Lough *et al.*, 2008) DNA insertions, some of which appear to entail almost entire organelle genomes. These insertion loci vary amongst different inbred lines indicating ongoing nuclear transfer and deletion of organelle sequences. This supports the intraspecific variation in *norg* content reported previously (Ayliffe *et al.*, 1998).

Many insertions are much smaller than the examples above and appear to be the result of individual integrations of small organelle DNA fragments. Other loci show complex arrangements comprising fragments of chloroplast and mitochondrial DNA from quite distant parts of the organelle genomes which are often arranged in different polarity with respect to the organelle genomes. These may represent either individual insertions followed by subsequent rearrangements and/or deletions, sequential insertions at a single locus or concomitant insertions of multiple fragments in a single event. While the exact nature of the insertion events is not known, there have been several recent advances in our understanding of what happens to *norgs* after insertion.

Richly and Leister (2004b) recognised that, for *nupts* over 500 bp in length, there is an inverse relationship between their age (based on sequence identity to the chloroplast genome) and their size. This suggests that insertion of large *nupts* is followed by fragmentation and deletion. These authors also observed 'tight' and 'loose' clusters of organellar sequence in nuclear genomes of rice and Arabidopsis. Organelle sequences in tight clusters had higher sequence identity to the extant organelle genomes indicating recent insertion whilst loose clusters appeared to be the result of older insertion events. Tight and loose clusters were therefore suggested to represent progressive stages of degradation and rearrangement of large initial insertions. Consistent with this finding a number of complex chloroplast sequence arrangements in rice have been explained by large initial insertions followed by a series of sequential rearrangements (Matsuo *et al.*, 2005; Guo *et al.*, 2008), though hard experimental evidence for this inference is still lacking. Analysis of chloroplast insertions in tobacco has also revealed a high degree of instability of *nupts*, with transferred chloroplast DNA being deleted within a generation in approximately 50% of integrant loci (Sheppard and Timmis, 2009).

While these findings suggest that shuffling and deletion may play a large role in the evolution of *norgs*, the observation that some loci contain both mitochondrial and chloroplast sequences indicates that in at least some instances sequential insertion at a single locus or the insertion of multiple fragments in a single event can take place. This last possibility is supported by studies of DNA transfer from the chloroplast to nucleus which show that multiple copies (up to 10-15 copies, see Chapter 2) of a chloroplast marker gene can be inserted at a single nuclear locus (Huang *et al.*, 2003; Sheppard *et al.*, 2008). Whether these loci represent the insertion of multiple fragments or duplications introduced close by, during insertion is, however, unclear.

The majority of studies investigating the insertion of chloroplast DNA have focused on evolutionary transfer events by looking for regions of chloroplast and mitochondrial origin in sequenced nuclear genomes. Despite the wealth of information these studies have provided there are a number of drawbacks. One of these is that they rely on the correct assembly of the genomic sequencing data. Nuclear sequence data are particularly prone to miss-assembly when it comes to organelle DNA insertions as nuclear organelle sequences (*norgs*) are often mistakenly presumed to be contaminating organelle DNA and discarded (Rousseau-Gueutin *et al.*, In Press). In addition, *norgs* often contain large internal duplications which can be missed during assembly (Stupar *et al.*, 2001). Another downside to analysis of evolutionary insertions is that it is impossible to determine the sequence of the region prior to the insertion. Without the complete before and after picture it is impossible to determine exactly what changes have taken place during the insertion. For example, in many cases chloroplast DNA in the nucleus is found adjacent to other regions of chloroplast DNA or other repetitive sequences, without the pre-insertion site sequence it is not possible to tell if

these fragments were inserted in a single event or are the result of chloroplast DNA insertion into a pre-existing *nupt*. Similarly, filler DNA or base substitutions introduced at the junction of the nuclear and chloroplast DNA cannot be determined without knowledge of the original nuclear pre-insertion sequence and the sequence of the organelle genome at the time of insertion.

Although, for the reasons given above, sequencing of *de novo* chloroplast DNA insertions and their pre-insertion sites is preferential for understanding the mechanisms of integration, obtaining such data poses significant technical difficulties and, as yet, none have been completely sequenced. The technical difficulties are due to the large size of the chloroplast DNA insertions, which can be well over 100 kb in length (see Chapter 2) and the complex arrangements of the chloroplast DNA inserted - up to 10-15 copies of a particular region of the chloroplast genome at a single locus (see Chapter 2). In addition, the presence of thousands of chloroplast genomes per cell, and the presence of multiple nuclear copies of chloroplast sequences that result from ancestral transfer events, make specific isolation of newly transferred *nupts* very difficult. Finally the nuclear sequences into which organelle DNA fragments integrate are themselves often repetitive, posing further problems for specific amplification and sequencing of these regions.

Despite these issues partial sequencing of *de novo* chloroplast insertions has been possible (Huang *et al.*, 2004) and has revealed complex arrangements at these loci. Several junctions between two distinct regions of chloroplast DNA show micro-homology, suggesting the involvement of the NHEJ repair pathway, whilst other junctions show the introduction of short stretches of filler DNA which is either of chloroplast or undetermined (presumably nuclear) origin (Huang *et al.*, 2004).

This chapter describes the complete cloning and sequencing of the chloroplast DNA insertion and its pre-insertion site in line kr2.2 which was one of the gene transfer lines arising from the screen undertaken by Sheppard *et al.* (Sheppard *et al.*, 2008). Kr2.2 is of particular interest in that it gave rise to the two spectinomycin resistant lines characterised in chapter 2 suggesting that this line may be particularly susceptible to change. Southern analysis of kr2.2 revealed a single insertion of the experimental cassette containing the *neo* and *aadA* genes and indicated that the length of chloroplast DNA inserted was smaller than in most other lines (Sheppard *et al.*, 2008), making the *nupt* and the nuclear flanking DNA in this line amenable to amplification by inverse PCR. The resulting sequence of the chloroplast DNA insertion and its pre-insertion site, obtained from wild-type tobacco, reveal insertion of multiple chloroplast fragments in a single event and the involvement of a synthesis dependent DSB repair pathway.

3.2 Results

3.2.1 Cloning and confirmation of the kr2.2 integrant and pre-insertion site

Both *aadA* activations reported in chapter 2 (sr1 and sr2) were due to rearrangements of the chloroplast sequence transferred to the nucleus in line kr2.2, which suggested that this nuclear locus may be particularly susceptible to change. The kr2.2 insertion locus was therefore sequenced to determine its arrangement and identify the mechanism of chloroplast DNA integration.

A combination of conventional PCR, inverse PCR (iPCR) and genome walking was used to amplify the chloroplast DNA insertion in kr2.2 and its pre-insertion site in wild-type tobacco. The left (*neo* side) nuclear flanking sequence was obtained by iPCR, using kr2.2 DNA digested with XbaI and subsequently circularised as template. Primers used in iPCR were located in the CaMV 35S promoter / *neo* region of the experimental cassette so as to be unique within the tobacco genome. This approach was only possible due to the (relatively) small size of the integrant in line kr2.2. Using primers designed within the left flanking nuclear sequence obtained by the iPCR, genome walking was then used to amplify the pre-insertion side in wild-type tobacco. The whole region including the full chloroplast insert and the flanking nuclear DNA was then amplified in three overlapping PCR products. The two largest products used primers within *neo* and the left (*neo* side) nuclear flanking sequence and primers within *aadA* and the right (*aadA* side) nuclear flanking sequence. A shorter region overlapping these first two PCR products was amplified using one primer in *aadA* and one in *neo* and sequenced to obtain the entire sequence of the insert.

The pre-insertion site was confirmed by PCR (Figure 3.1) using a primer (kr2.2NJR1) within the left flanking nuclear sequence and a primer (kr2.2NJR2) within the right flanking nuclear sequence (Figure 3.2A). A 399 bp product was expected from wild-type chromosomes but no product was expected in chromosomes containing the chloroplast insertion due to the ~16.8 kb insertion between the two primer binding sites. As expected a product was obtained when wild-type or hemizygous kr2.2 DNA was used as a template but not when homozygous kr2.2 DNA was used (Figure 3.1).

3.2.2 The kr2.2 integrant and pre-insertion site sequence

The kr2.2 integrant was 16,757 bp in length and comprised three fragments of chloroplast DNA (cp1-3) from disparate parts of the chloroplast genome (Figure 3.2A). Fragment cp1 corresponded to nucleotides 95667 – 107107 in inverted repeat B (IRB) (or 135523 – 146963 in IRA) of the tobacco plastid genome (Yukawa *et al.*, 2005) and included the transgenes *aadA* and *neo* from the chloroplast transformation cassette (Figure 3.2A). Fragment cp2 and cp3 corresponded to nucleotides 64850 – 64880 and 48075 – 50450 respectively in the large single copy region. Four XbaI sites were present within the region sequenced, giving an 8,741 bp *neo* containing XbaI

fragment and a 6,626 bp *aadA* containing XbaI fragment (Figure 3.2A). These XbaI fragment sizes correspond to those observed hybridising to *aadA* and *neo* probes in the kr2.2 Southerns (Sheppard *et al.*, 2008), further confirming the arrangement of the sequence. Among the 16,757 bp of the *de novo* integrant in kr2.2 a single nucleotide deletion (nucleotide 101286 in IRB or 141344 in IRA), within a homopolymeric stretch, was the only change compared with the original plastid sequence. To rule out a PCR error, this single nucleotide deletion was confirmed by sequencing a second clone obtained from an independent PCR.

In total 1,265 bp of flanking nuclear DNA was sequenced and this region was analysed for similarity to known plant nuclear sequences. The nuclear sequence has no significant ORFs but contains three stretches of sequence that each share significant sequence identity with multiple unclassified tobacco EST accessions (Table 3.1), as well as EST accessions from a wide range of species within the Solanaceae. Blast analysis against a plant repeat database indicated that one of these regions (2) has moderate identity (75%, e = 0.081) with the solSINE2 transposon from *Solanum lycopersicum*.

Table 3.1 Blastn matches to pre-insertion site sequence

	co-ordinates[a]	length	indentity[b]	E-value[b]	Accession #[b]
1	-441 / -152	290	91%	3×10^{-103}	AM809504
2	-93 / 145	238	97%	3×10^{-103}	EB449160
3	378 / 474	97	81%	2×10^{-14}	AM831992

[a] nucleotides -1 and +1 are those immediately upstream and downstream respectively of the chloroplast DNA insertion site
[b] the accession number, identity and E-value of the closest match is shown

To investigate the molecular mechanisms of insertion, the sequences of the nuclear/chloroplast junctions, the chloroplast/chloroplast junctions and the pre-insertion site were analysed. No deletion of nuclear DNA occurred during chloroplast DNA integration but filler DNA (5-15 bp) was added at each of the four junctions (Figure 3.2B-C). In one case (junction a-b) filler DNA (α) was due to a duplication of the 12 bp of chloroplast sequence adjacent to the junction [α*](Figure 3.2C). In the other three cases filler sequence was derived from an ectopic location (Figure 3.2C). In one case [γ*] this was over 14.3 kb from the junction containing the filler DNA [γ](Figure 3.2C). The filler template sequence was always no more than 15 bp from the end of one of the chloroplast DNA molecules being inserted (Figure 3.2C). Sequences adjacent to the filler DNA were usually homologous with the template that promoted the formation of the junction such that, β and β*, γ and γ* and δ and δ* are complementary (Figure 3.2C). Figure 3.3 presents a model of how this

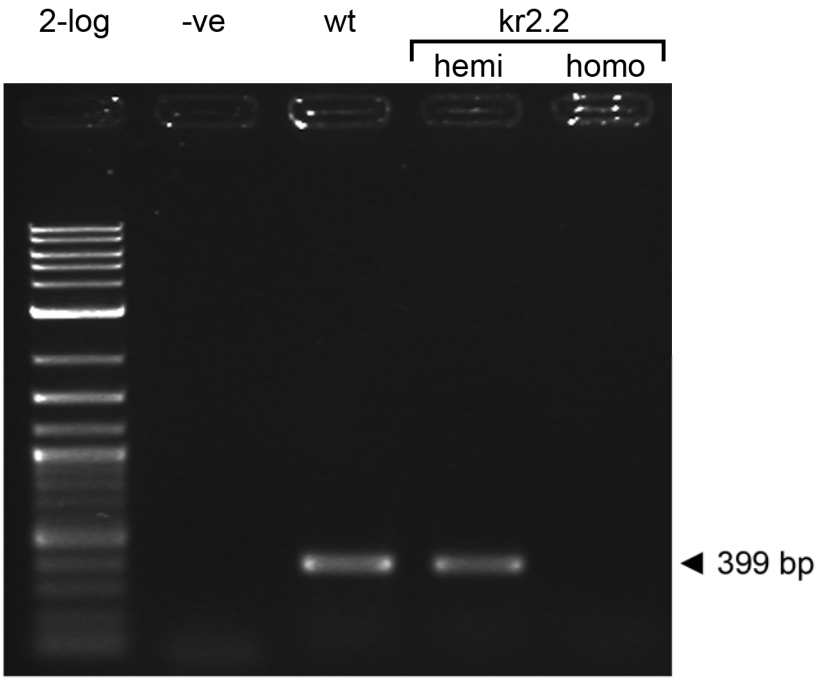


Figure 3.1 The kr2.2 pre-insertion site is only present in wild-type plants and plants hemizygous for the insertion. Primers flanking the kr2.2 pre-insertion site were used to amplify a 399 bp product. Products were obtained from wild-type and hemizygous kr2.2 DNAs as both contain at least one chromosome lacking the insertion. The ~17 kb insertion between the two primer binding sites prevented amplification from homozygous kr2.2 DNA. A no template control (-ve) is also shown.

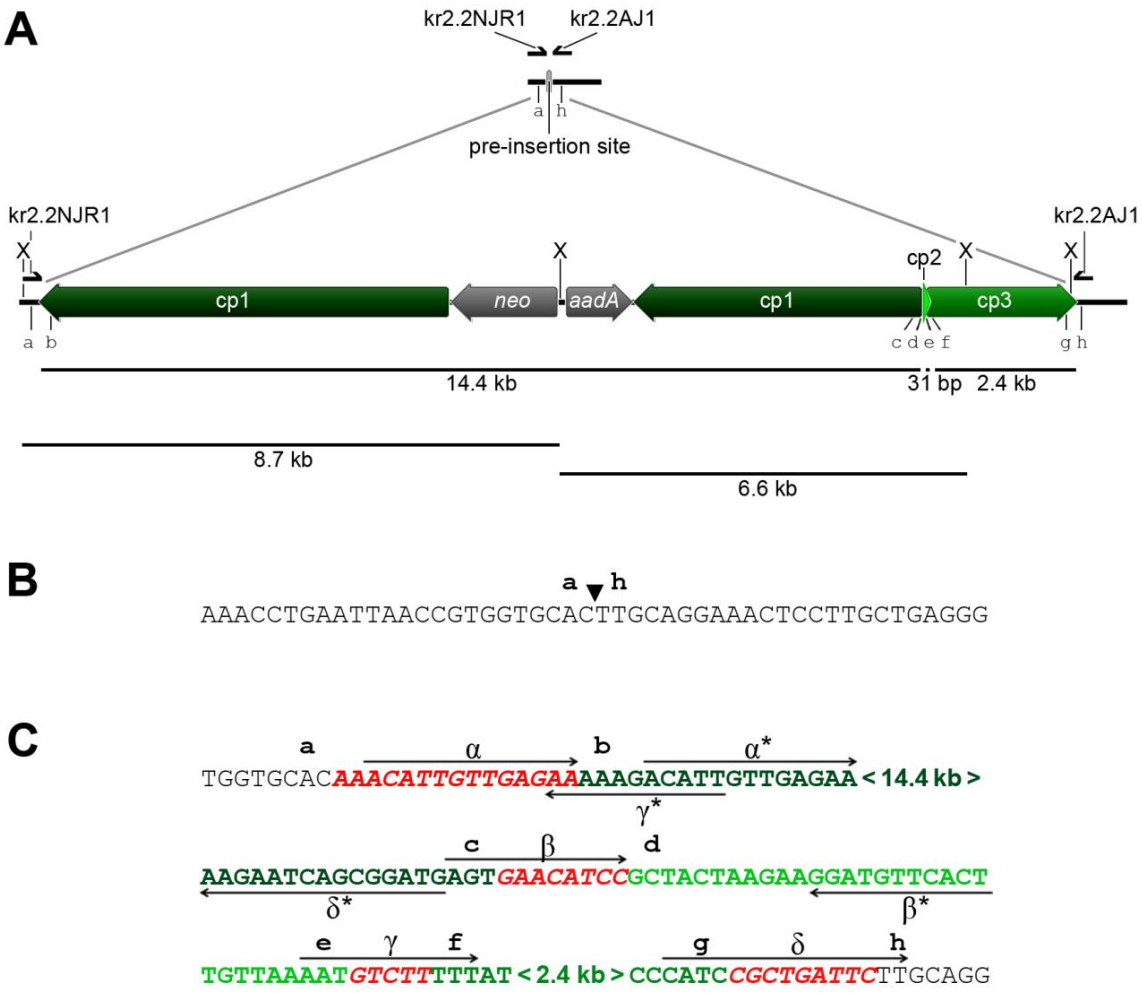


Figure 3.2 Characterization of the kr2.2 integrant. Three disparate fragments of chloroplast DNA (cp1-3, green) with sizes of 14.4 kb, 31 bp and 2.4 kb (upper black bars) comprise the 16.8 kb integrant in kr2.2 (**A**). XbaI sites are shown (X); the XbaI fragment containing neo is 8.7 kb in length, the XbaI fragment containing *aadA* is 6.6 kb in length (lower lack bars). The nuclear genomic sequence before (**B**) and the nuclear/chloroplast sequence after integration (**C**) is shown. The site of chloroplast DNA integration is marked in **B** (▼). Filler DNA sequences (red) are found at each of the junctions (a-b, c-d, e-f and g-h) and are derived from sequences near the end of chloroplast DNA fragments (**α***, **β***, **γ*** and **δ***). In junctions c-d, e-f and g-h sequences adjacent to the filler DNA are homologous with the template that promoted the formation of the junction such that **β** and **β***, **γ** and **γ***, and **δ** and **δ*** are complementary.

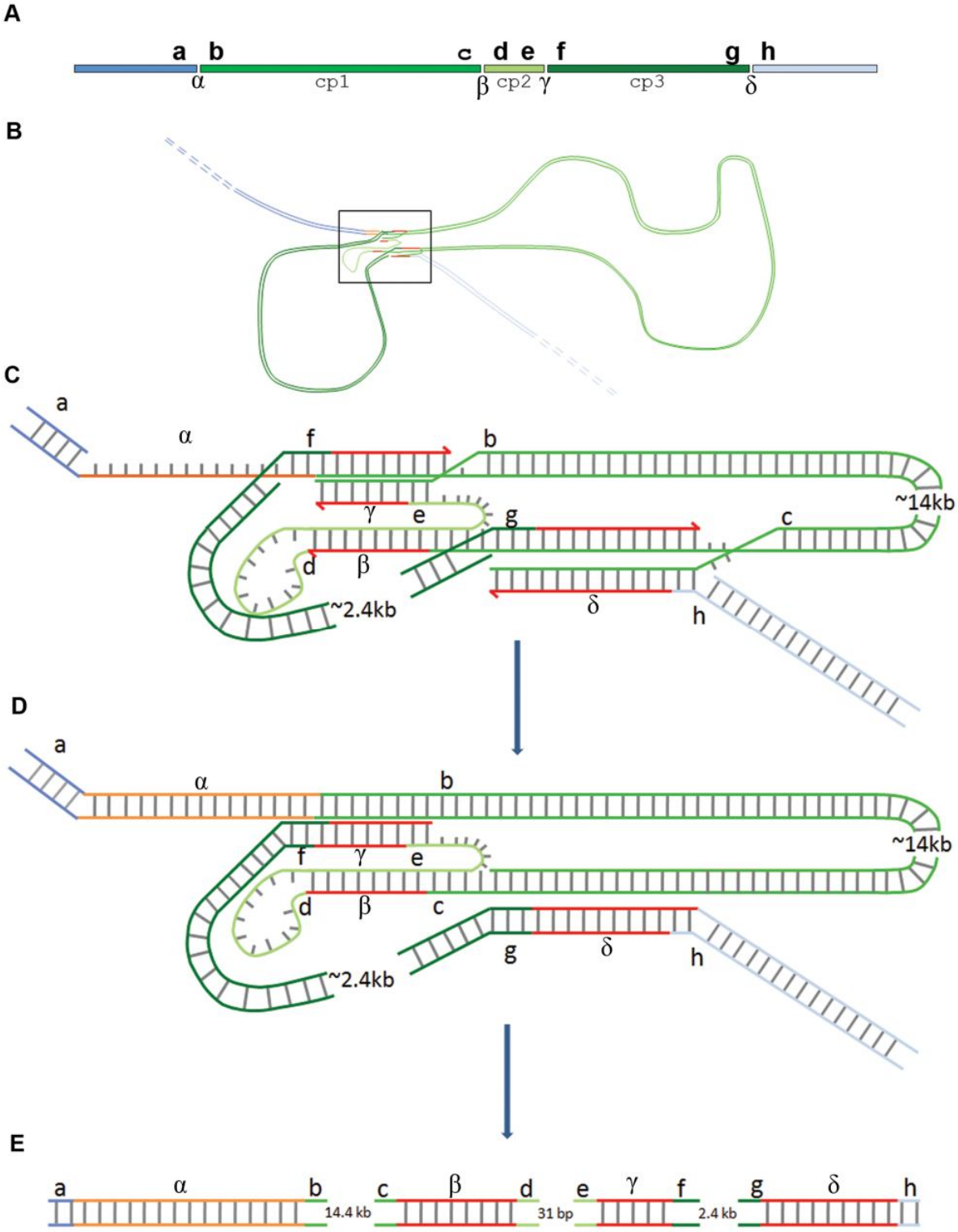

A
a b c d e f g h
α cp1 β cp2 γ cp3 δ
B
C
a α f b ~14kb γ e g c d β ~2.4kb δ h
D
a α b ~14kb f γ e d β c g δ h ~2.4kb
E
a α b 14.4 kb c β d 31 bp e γ f 2.4 kb g δ h

Figure 3.3 A model of chloroplast DNA insertion in line kr2.2. (**A**) Insertion of three chloroplast fragments (green) at a single nuclear location (blue)[not to scale]. Lower case letters indicate DNA ends involved. (a) and (h) indicate nuclear DNA ends. (b) and (c), (d) and (e), and (f) and (g) represent the ends of the three chloroplast DNA fragments inserted (**A, C**-**E**). (**B**) The ends of chloroplast and nuclear DNA are brought together to facilitate joining at a central repair node. At this site (**C**), using primarily polymerase-dependent non-homologous end joining, 3' overhangs are extended (red lines) to generate regions of complementarity enabling subsequent joining of two DNA molecules by single strand annealing. 3' overhangs at (f) and (e) are extended using opposite DNA strands of (b) as a template (**C**). This generates complementary filler sequence (**γ**) which can be used to facilitate their joining, generating junction f-e (**D**). Similarly, 3' overhangs at (**g**) and (**h**) are extended using opposite strands of (c) as template (**C**). Again this generates complementary filler sequence (**δ**) which can be used to join these two DNA ends. Filler sequence (**β**) at junction c-d is synthesized using template sequence at (e). Sequence at junction a-b (**α**) is introduced through a local duplication.

insertion event may have occurred. In this model, sites of filler template and sites of filler synthesis co-localise during insertion necessitating the co-location of the two ends of each of the chloroplast DNA molecules and the two nuclear DNA ends (Figure 3.3C-D). Short regions of homology are used to prime the synthesis of filler DNA which mediates joining (Figure 3.3C-D).

3.3 Discussion

Despite the high frequency with which nuclear integration of chloroplast DNA occurs (Huang *et al.*, 2003; Stegemann *et al.*, 2003), complete characterisation of any *de novo* chloroplast DNA insertions has so far proven elusive due to the large size of the inserts, the high copy number of chloroplast genomes present in each cell and the presence of multiple ancestral copies of chloroplast DNA fragments within the nuclear genome (Huang *et al.*, 2004). Therefore, this complete characterisation of the kr2.2 integrant and flanking nuclear DNA is the first such characterisation and it has shed light on a number of previously unanswered questions.

Naturally-occurring regions of chloroplast DNA within the nucleus often contain contiguous fragments of chloroplast DNA from disjunct parts of the chloroplast genome which may show different polarities (Leister, 2005). It has not hereto been clear whether this is due to multiple sequential insertions at one location, concomitant insertions of multiple fragments or insertion of a single fragment, followed by fragmentation, deletion, inversion and shuffling (Richly and Leister, 2004b; Matsuo *et al.*, 2005). Our results demonstrate concomitant insertion of three chloroplast DNA fragments and provide the first evidence that much of the observed complexity may be generated during the primary integration events.

Adding to the complexity of this locus the pre-insertion site contained a SINE retrotransposon-like sequence and flanking this SINE-like region, several other stretches of sequence which showed high identity to multiple uncharacterised EST accessions. This is perhaps not surprising as the tobacco nuclear genome contains abundant repetitive sequences. One of the potential mechanisms proposed for organelle DNA integration is that DSBs caused by transposon excision, form sites into which organelle DNA can be inserted (Leister, 2005). However, as the chloroplast DNA inserted into the middle of the SINE-like sequence (SINEs do not undergo excision) and no deletions were associated with the insertion, it does not appear that this was a factor in this particular event. No sequence similarity was observed between the insertion site and the chloroplast DNA suggesting also that no homology based recombination or repair mechanisms were responsible. Short stretches (5-10 bp) of filler DNA, however, were observed at each of the nuclear/chloroplast junctions and also at the junctions of two chloroplast fragments.

The term ‘filler DNA’ has been applied to a number of different classes of DNA insertion. In studies of DSB repair the term is generally applied to any DNA that is introduced at the site of the DSB. In

some cases these are fragments of nuclear DNA, usually repetitive sequence (Gorbunova and Levy, 1997; Salomon and Puchta, 1998), or may be fragments of introduced T-DNA or plasmid DNA that encode the enzyme responsible for the double strand breaks (Gorbunova and Levy, 1997; Salomon and Puchta, 1998). This filler DNA may be hundreds or thousands of base pairs in length. A more restricted use of the term filler DNA is often found in studies interested in the insertion of specific sequences into the nuclear genome such as T-DNA integration or the insertion of organellar DNA. In these instances the term is applied to short sequences that are introduced at the junctions of nuclear sequence and the sequence of interest (Kumar and Fladung, 2002; Windels *et al.*, 2003; Huang *et al.*, 2004; Zhu *et al.*, 2006). This filler DNA is much shorter, usually 5-20 bp, in length and often shares similarity with sequence near one of the junctions, suggesting that it may be introduced through *de novo* synthesis rather than insertion of a pre-existing DNA fragment. This type of filler has also be observed at sites of DSB repair (Salomon and Puchta, 1998). Filler DNA referred to in this chapter is of this latter class.

In kr2.2 all filler DNA was of chloroplast origin and derived from one of the ends of the three fragments of chloroplast DNA inserted. For three out of four junctions the filler DNA was derived from an ectopic location (i.e. from a DNA end involved in the insertion but not being joined at that junction). Thus the formation of these junctions must have required the co-location of three DNA ends, one containing the filler DNA template sequence and the two DNA ends being joined (Figure 3.3C). Together this would have required all eight DNA ends (two ends for each of the three chloroplast fragments and the two nuclear DNA ends) to be in close proximity during the insertion process (Figure 3.3B-C). It is possible that this was a chance association attributable to the potentially large amount of chloroplast DNA present in the nucleus during male gametogenesis – when this insertion occurred and, significantly, when paternal plastids are eliminated (Sheppard *et al.*, 2008). Another possibility is that free DNA ends of multiple chloroplast DNA fragments are brought together through active recruitment to a single repair node, possibly a site of DSB repair, where they are processed and joined to mediate nuclear integration. An insertion mechanism such as this would contribute to the highly complex nature of chloroplast DNA insertions observed in sequenced nuclear genomes and may also explain the observation that many new nuclear integrations of chloroplast DNA contain multiple copies of a chloroplast gene at a single locus (Huang *et al.*, 2003; Sheppard *et al.*, 2008).

The lack of any homology at the insertion site and the presence of filler DNA at the junctions indicates that NHEJ is involved in the insertion. Both classical Ku dependent and Ku independent forms of NHEJ are responsible for the insertion of T-DNA in Arabidopsis (Gallego *et al.*, 2003; Li *et al.*, 2005) but it is not clear which (or if both) of these pathways is involved in the integration of chloroplast DNA. In this instance it is likely that the end-joining was synthesis dependent given the

template mediated insertion of filler DNA. Recently a polymerase theta dependent, Ku independent pathway of DSB repair that introduces filler DNA has been described (Chan *et al.*, 2010) and is a possible candidate pathway for insertion. However, many more experiments are needed before this question will be answered.

3.4 Conclusions

The complete sequencing of the chloroplast DNA integrant and its pre-insertion site reveals several factors that contribute to the complexity of *norg* loci. Firstly, organelle DNA may be inserted into regions of the nuclear genome that already contain repetitive elements and secondly, multiple fragments of organelle DNA in differing polarities can be inserted in a single event. The observed pattern of filler DNA insertion gives some indication that the insertion of multiple fragments of organelle DNA may be mediated through active recruitment to a single repair locus. A DNA repair/integration mechanism such as this would significantly contribute to the complex arrangement of organellar sequences in the nuclear genome. This is likely to be important from an evolutionary perspective as it will lead to the creation of novel sequence arrangements which in some instances may result in nuclear activation of the transferred organelle genes.

3.5 Materials and methods

3.5.1 Plant growth

Plant growth was as described in section 2.5.1.

3.5.2 DNA extraction

DNA extraction was undertaken as described in section 2.5.5.

3.5.3 Inverse PCR

Inverse PCR was adapted from Ochman *et al.* (1988). 1 µg DNA was digested for 2 hours in a 15 µl reaction with 30 U XbaI (New England Biolabs, Ipswich, MA), diluted to 500 µl and ligated with 100 U T4 DNA ligase (New England Biolabs) for 16 h at 4°C. Ligation products were ethanol precipitated and resuspended in 15 µl of water of which 1 µl was amplified using Expand Long Range dNTPack (Roche) according to manufacturer's instructions. Forward and reverse primers within *neo* but facing away from each other were used to amplify the left (*neo* side) nuclear flanking region, *neo* and the intervening chloroplast DNA from the circularised template. Three rounds of PCR were necessary using sequentially, primer pairs neoR3/neoF1, 35SR1/neoF2 and 35SR2/neoF3.

3.5.4 Genome walking

Genome Walking utilised a Universal GenomeWalker Kit (Clontech, Palo Alto, CA) according to manufacturer's instructions. The left (*neo* side) nuclear flanking sequence was determined using iPCR (above). In wild-type tobacco, primers kr2.2NJR2 and kr2.2NJR1 (within the left nuclear

flanking sequence) were used amplify the pre-insertion site. In kr2.2, primers kr2.2AJR1 and kr2.2AJR2 (within the right nuclear flanking sequence) were used to walk back from the right hand side nuclear sequence (determined by sequencing the pre-insertion site) into the integrant.

3.5.5 PCR

The chloroplast insertion was amplified in two PCRs using Expand Long Range dNTPack (Roche) according to manufacturer's instructions. The left (*neo* side) was amplified using, successively, primer pairs kr2.2NJR2/neoF4 and kr2.2NJR1/neoF1. The right (*aadA* side) was amplified using, successively, primer pairs kr2.2AJR1/aadAF3 and kr2.2AJR2/aadAF4. To complete the sequence of the insert a shorter region overlapping the first two PCR products was amplified with primers neoR4 and aadAR3, using *PfuTurbo* DNA polymerase (Stratagene) according to the manufacturer's instructions.

3.5.6 Sequencing

Sequencing was undertaken as described in section 2.5.6, using primers in Appendix 3.

3.5.7 Sequence analysis

All sequences were initially analysed using Geneious (ver.Pro.5.0; Biomatters, Auckland, New Zealand). The pre-insertion site was analysed for the presence of repetitive elements using the Plant Repeat Databases website (http://plantrepeats.plantbiology.msu.edu) and analysed for matches to ESTs using the blastn algorithm through the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Junction sequences were also analysed using the blastn algorithm through the NCBI website.

Chapter 4: Design and evaluation of an experimental system for the detection of organelle sequence insertion at sites of DNA double strand break repair

4.1 Introduction

Insertion of DNA at sites of double strand break (DSB) repair was first observed in yeast, where several studies (Moore and Haber, 1996; Teng *et al.*, 1996) noted the insertion of retro-transposon DNA in a proportion of the DSBs repaired by non-homologous end joining (NHEJ). Several years later mitochondrial DNA was also observed at sites of DSB repair, indicating that nuclear insertions of organelle DNA could occur *via* the same process (Ricchetti *et al.*, 1999). In this study DSBs were induced in the yeast nuclear genome through expression of the rare cutting endonuclease I-SceI and insertion of mitochondrial DNA was regularly observed at the sites of repair.

DNA DSBs are repaired using two main pathways, homologous recombination (HR) and NHEJ. HR uses the sister chromatid or homologous chromosome as a template for repair of the broken chromosome. Any nucleotides deleted as a consequence of the DSB are copied from the template molecule which results in repair without loss of any genetic information. While studies of DSB repair by HR in plants lag behind those in yeast and mammals (Chittela and Sainis, 2010), there is a high degree of conservation of the HR pathway, and much of what has been observed in other systems is broadly applicable. The details of HR in DSB repair have been comprehensively reviewed recently (Heyer *et al.*, 2010; Mazon *et al.*, 2010).

In flowering plants most somatic repair of DSBs occurs *via* the NHEJ pathway (Puchta, 2005). NHEJ involves ligation of the two DNA ends formed by the DSB without the use of a homologous template and consequently the process is inherently error-prone (reviewed by Lieber, 2010). NHEJ falls into two main categories: classical-NHEJ and alternative-NHEJ (alt-NHEJ). In classical-NHEJ the ku70/ku80 heterodimer binds to DNA ends (Downs and Jackson, 2004) and recruits a number of proteins including the DNA ligase IV(Lig4)/XRCC4 complex which ligates the two strands (Grawunder *et al.*, 1997). The term alt-NHEJ is generally used to describe any NHEJ lacking in one or more of the core classical NHEJ proteins e.g. ku70, ku80, Lig4, XRCC4 (Lieber, 2010). Alt-NHEJ, sometimes referred to as backup-NHEJ (B-NHEJ; Wang *et al.*, 2003) or micro-homology-mediated end joining (MMEJ; McVey and Lee, 2008), is not so well characterised and may well encompass

several distinct repair pathways (Lieber, 2010). It has been suggested that this pathway is inhibited by classical-NHEJ (Fattah *et al.*, 2010; Simsek and Jasin, 2010).

While insertion of organellar DNA *via* the NHEJ repair pathway has so far only been observed in yeast, the fact that extra-chromosomal DNA can be captured at sites of DSB repair in plants (Salomon and Puchta, 1998) and animals (Lin and Waldman, 2001) suggests this process applies more widely. For this reason NHEJ is thought to be the main pathway for the nuclear insertion of organelle sequences in plants (Kleine *et al.*, 2009).

Both nuclear DNA and T-DNA have been experimentally observed at sites of DSB repair in plants (Salomon and Puchta, 1998; Kirik *et al.*, 2000). It was likely that these sequences would have been inserted more frequently than organelle DNA as both nuclear DNA and T-DNA must have been in the nucleus of every cell undergoing a DSB in these experiments. Nuclear DNA is obviously always present in the nucleus and the T-DNA must have also been present as *Agrobacterium* transformation was used to transiently introduce the enzyme causing the DSBs (Salomon and Puchta, 1998; Kirik *et al.*, 2000). Organelle DNA very rarely makes it into the nucleus during normal somatic cell growth (Sheppard *et al.*, 2008) and therefore, entry into the nucleus is probably the limiting step in insertion of organelle DNA at sites of DSB repair. One possible route to detect the insertion of organelle DNA may be to induce DSBs in cell types or growth conditions more likely to lead to entry of organelle DNA into the nucleus.

DNA insertion at sites of DSB repair has so far only been investigated in mitotically dividing plant cells. In such cells, a chloroplast gene transfers to the nucleus at a frequency of 2×10^{-7} (Stegemann *et al.*, 2003). In the plant male germline however, chloroplast gene transfer occurs ~300 times more often, at a frequency of $\sim 6 \times 10^{-5}$ (Huang *et al.*, 2003). In addition, environmental stresses such as heat and salt are known to increase the frequency with which organelle DNA enters the nucleus in somatic cells (D. Wang, personal communication). Investigating DSB repair in the male germline or in somatic cells that have undergone environmental stress is therefore more likely to lead to observation of organelle DNA insertion.

This chapter describes the design and evaluation of an experimental system intended to allow efficient observation of organelle-to-nucleus DNA transfer by enabling the induction of DSBs in specific tissues and at specific stages of development or after stress treatment. The DSBs are introduced at a specific genomic location allowing characterisation of repair events through standard PCR based approaches.

4.2 Outline of the experimental system

The experimental system involves the inducible expression of the rare cutting endonuclease I-SceI which, in the current design, cleaves at two I-SceI target sites flanking a dual selectable marker gene *dao1* (Figure 4.1). Some DSB repair events will lead to *dao1* excision (Figure 4.2) and these can be identified by selecting for the absence of the *dao1* gene. Primers flanking the I-SceI sites can then be used to PCR amplify and sequence the repair junction (Figure 4.2).

Two *Agrobacterium* transformation vectors were generated. One vector, pAlcR:ISceI, contains the requirements for inducible expression of the rare-cutting endonuclease I-SceI (Figure 4.1A). Inducible expression of *I-SceI* was enabled using the AlcR ethanol inducible promoter system, initially isolated from *Aspergillus nidulans* (Lockington *et al.*, 1985) and adapted for use in plants (Caddick *et al.*, 1998). This system involves the constitutively expressed transcriptional activator AlcR which, in the presence of ethanol, binds to a synthetic promoter comprising the activator region of the *Aspergillus alcA* promoter fused to a minimal 35S promoter (Caddick *et al.*, 1998). Upon binding, the AlcR protein transcriptionally activates the *alcA*:35S promoter. This enables expression of the experimental gene, *I-SceI* in this instance, to be induced through the application of ethanol either by spraying or root drenching (Salter *et al.*, 1998). The *I-SceI* gene, driven by the *alcA:35S* promoter, encodes a rare-cutting endonuclease which has an 18 bp recognition site. By chance there are expected to be < 0.1 I-SceI target sites in the tobacco genome and therefore expression of this gene is unlikely to cause DSBs in wild-type tobacco.

The second construct, pdao1, contains the *dao1* gene flanked by I-SceI target sites (Figure 4.1B). The *dao1* gene encodes a D-amino acid oxidase (DAAO) which metabolises D-amino acids to α-keto acids and ammonia (Pollegioni *et al.*, 2007). D-amino acids vary in their phytotoxicity as do their corresponding α-keto acids: D-alanine and D-Serine are both toxic to plants but they are catabolised by DAAO into non-toxic products. In contrast, D-valine and D-isoleucine have low phytotoxicity but are catabolised by DAAO into their highly toxic α-keto acids (Erikson *et al.*, 2004). This allows *dao1* to be used as both a negative and positive selectable marker: growing plants on media containing D-alanine or D-serine will select for the presence of *dao1*, whilst media containing D-valine or D-isoleucine will select for the absence of *dao1* (Erikson *et al.*, 2004).

After ethanol induction of I-SceI expression, the two I-SceI target sites flanking *dao1* will be cleaved (Figures 4.1B and 4.2). The resulting DSBs must be repaired before cell division can continue and, in some cases, this will lead to the excision of the *dao1* gene (Figure 4.2). Repair events leading to the excision of *dao1* can be selected using D-valine or D-isoleucine and the DSB repair site can then be amplified by PCR using primers within the T-DNA that flank the I-SceI sites (Figure 4.2). A larger

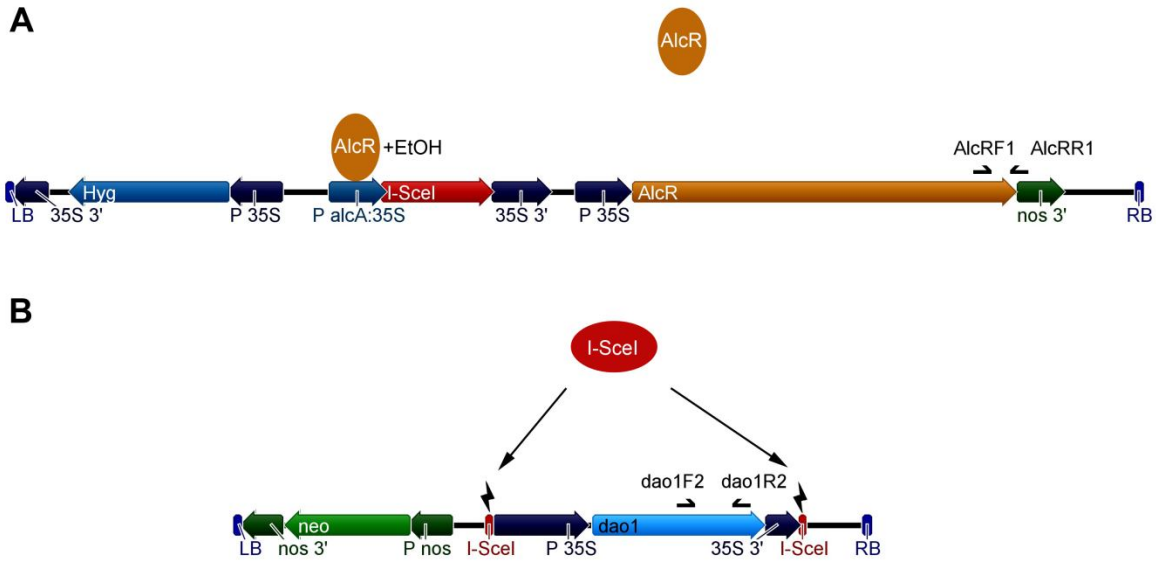


Figure 4.1 Overview of EtOH induced induction of DSBs. The T-DNA of vector pAlcR:ISceI (**A**) contains a hygromycin selectable marker gene (*hyg*); the *AlcR* gene constitutively expressed from the 35S promoter; and the *I-SceI* gene driven by the *alcA*:35S promoter. In the presence of EtOH, AlcR binds to and transcriptionally activates the *alcA*:35S promoter, driving expression of *I-SceI*. The T-DNA of vector pdao1 (**B**) contains a kanamycin selectable marker gene (*neo*); and the *dao1* gene driven by the 35S promoter and flanked by I-SceI target sites. Upon EtOH induction of *I-SceI* (**A**) the I-SceI protein cleaves the two I-SceI target sites flanking the dao1 gene (**B**).

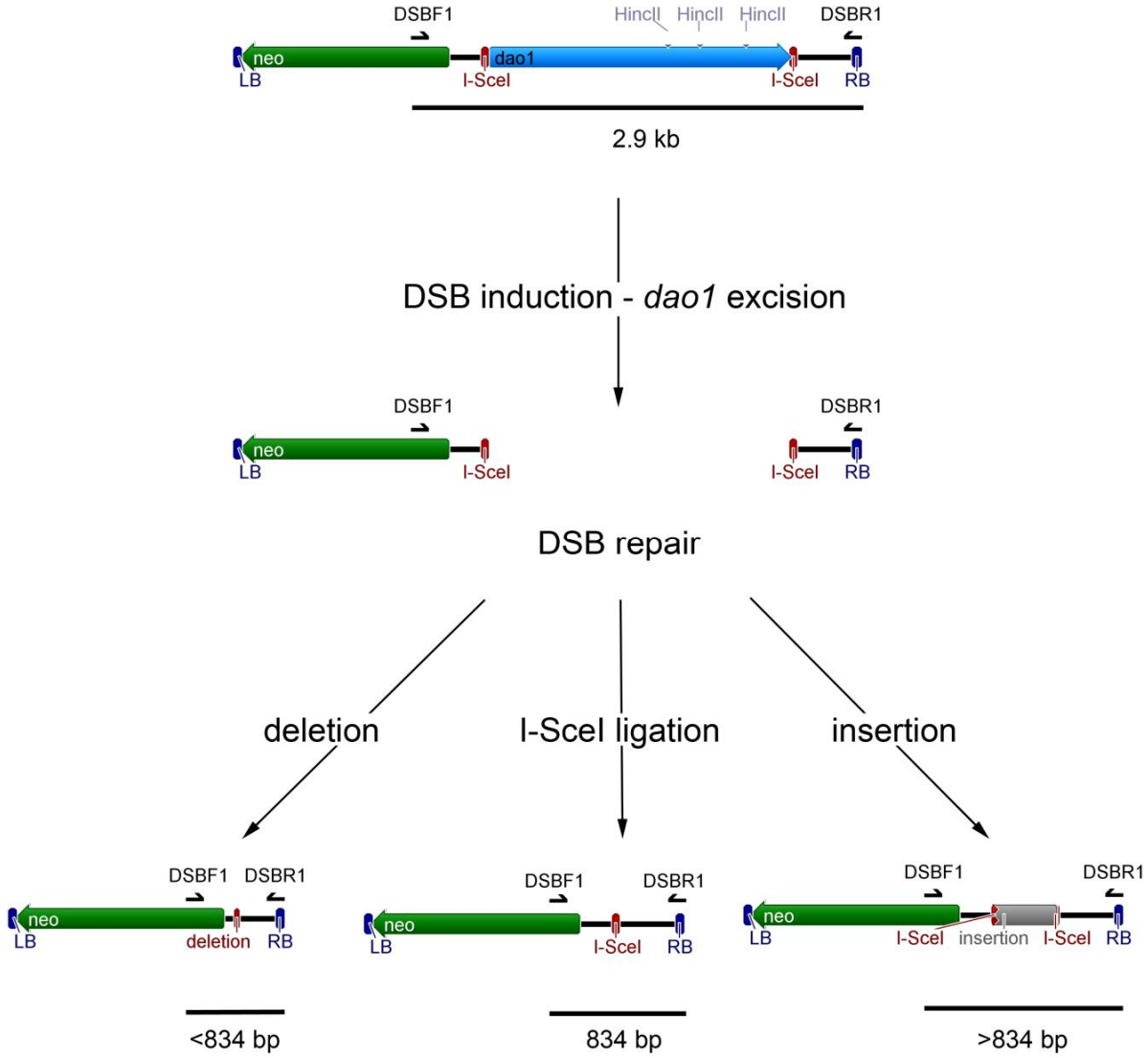


Figure 4.2 Overview of DSB repair. The *dao1* gene is flanked by I-SceI target sites. Upon I-SceI expression these sites are cleaved leading to the excision of *dao1*. DSB repair will then result in the joining of the cleaved sequences. This may result in direct joining of the I-SceI restriction sites, deletion of sequence on either side of the DSB or insertion of sequence at the site of DSB repair. These three types of repair can be distinguished by PCR using primers DSBF1 and DSBR1 that flank the site of DSB repair. Direct joining will result in an 834 bp product whilst deletion will result in a smaller product and insertion in a larger product. In the absence of DSB induction or where DSB repair events do not lead to *dao1* excision a PCR product of 2.9 kb will be obtained. The *dao1* gene contains three HincII sites and so digestion of template DNA with HincII will prevent amplification.

than expected PCR product would indicate insertion at the site of DSB repair, while a smaller than expected PCR product would indicate deletion of nucleotides at the site of the DSB (Figure 4.2).

Negative *dao1* selection was found to be unsuccessful in tissue culture regeneration and so no plants lacking the *dao1* gene were able to be regenerated using this method. Negative *dao1* selection in seedling germination proved possible but no cases of *dao1* excision in the male germline were identified.

4.3 Results

4.3.1 Transformation with vectors pdao1 and pAlcR:ISceI

To establish the experimental system in tobacco, two binary *Agrobacterium* transformation vectors, pdao1 and pAlcR:ISceI, were generated. The pdao1 T-DNA contains the *neo* gene for kanamycin selection and the 35S promoter-driven *dao1* gene flanked by two I-SceI target sites (Figure 4.1B). The pAlcR:ISceI T-DNA contains the *hyg* gene for hygromycin selection of transformants, *AlcR* driven by a 35S promoter and *I-SceI* driven by the *alcA*:35S promoter (Figure 4.1A). The pdao1 and pAlcR:ISceI constructs were independently transformed into *Nicotiana tabacum* cv. Wisconsin 38 *via Agrobacterium* transformation to generate D and A lines respectively. D line transformants were selected on regeneration medium containing 150 mg L^{-1} kanamycin and A line transformants were selected on medium containing 15 mg L^{-1} hygromycin. Resistant shoots were transferred to 0.5 × MS agar medium to generate roots and later transferred to soil.

Resistant shoots were assayed by PCR to confirm the presence of the *dao1* and *AlcR* genes. Ten D line primary transformants (D4, D5, D10, D11, D12, D13, D14, D15, D16 and D17) and four A line primary transformants (A1, A2, A3 and A4) were confirmed by PCR. For both constructs, T_1 progeny from self fertilised T_0 plants were grown on the relevant antibiotic to determine segregation ratios. Four of the six D lines tested (D4, D11, D13 and D14) and two of the three A lines tested (A2 and A3) were found to have 3:1 segregation ratios indicating T-DNA insertion at a single locus (Appendix 4).

4.3.2 Transformation with vector pGU.D.US

Additional transformed lines (G lines) were generated *via Agrobacterium* transformation using the transformation vector pGU.D.US. The pGU.D.US vector contains two overlapping sections of the GUS reporter gene (GU and US), between which is located the *dao1* gene flanked by I-SceI sites. In lines transformed with this vector, *dao1* excision can be repaired by homologous recombination between the U regions of GU and US to reconstitute the GUS reporter gene. G lines were used in the initial experiments which ascertain the effectiveness of *dao1* selection in tobacco but were not

used in any of the downstream double strand break experiments. For the purposes of this chapter D and G lines can be considered equivalent, i.e. both contain the *dao1* gene.

The pGU.D.US transformation vector was tansformed into *Nicotiana tabacum* cv. Wisconsin 38 *via Agrobacterium* transformation to generate G lines. Transformants were selected on regeneration medium containing 150 mg L^{-1} kanamycin and resistant shoots were transferred to 0.5 × MS agar medium to generate roots and later transferred to soil.

Resistant shoots were assayed by PCR to confirm the presence of *dao1*. Six G line primary transformants (G1, G4, G8, G9, G10 and G11) were confirmed by PCR. T_1 progeny from self fertilised T_0 plants were grown on 150 mg L^{-1} kanamycin to determine segregation ratios. Five of the six lines tested (G1, G4, G8, G9 and G10) were found to have 3:1 segregation ratios indicating T-DNA insertion at a single locus (Appendix 4).

4.3.3 Evaluation of the use of *dao1* as a selectable marker gene in tobacco

The *dao1* dual selectable marker gene has been used successfully in both Arabidopsis and maize (Erikson *et al.*, 2004; Lai *et al.*, 2007), but there are no reports of its use in tobacco. To determine the effectiveness of *dao1* as a dual selectable marker gene in tobacco, seedling selection and tissue culture regeneration were undertaken on media containing either D-alanine or D-valine.

4.3.3.1 dao1 seedling selection

T_1 seedlings of two single locus lines G10 and D11 and one multiple locus line D10 as well as wild-type seedlings were grown on a range of D-alanine concentrations (1 mM, 3 mM, 5 mM and 10 mM), a range of D-valine concentrations (5 mM, 15 mM, 30mM and 50 mM) and on media containing neither D-amino acid. After two weeks seedlings were weighed and assessed for growth. All transgenic lines showed strong growth on all concentrations of D-alanine whilst wild-type showed significantly stunted growth at 5 mM D-alanine and did not grow past germination on 10 mM D-alanine (Figure 4.3A). Growth at 10 mM D-alanine showed the greatest distinction between transgenic and wild-type seedlings. At this concentration wild-type and transgenic seedlings were easily distinguishable by sight (Figure 4.4A-D). Transgenic lines had significantly reduced growth at all concentrations of D-valine tested and were unambiguously distinguishable visually from wild-type at concentrations of 15 mM and 30 mM (Figure 4.3B). At the highest concentration (50 mM), D-valine became toxic to wild-type seedlings and they were no-longer easily distinguishable from seedlings containing *dao1* (Figure 4.3B). 15-30 mM D-valine was therefore determined to be the optimum concentration range for *dao1* negative selection in tobacco seedlings. At this concentration wild-type and transgenic seedlings were easily

distinguishable by sight (Figure 4.4E-H). These findings indicate that *dao1* is suitable for use as both a positive and negative selectable marker gene in tobacco seedlings.

4.3.3.2 dao1 explant selection

The suitability of *dao1* as a marker gene in tissue culture regeneration experiments was also assessed. D line plants and wild-type plants were grown in tissue culture jars. After 4 weeks, leaf explants taken from both D line and wild-type plants were transferred to regeneration medium containing either 10 mM D-alanine or 30 mM D-valine - these being the concentrations identified as optimum for seedling selection. As expected, in the presence of D-alanine (positive selection) shoots were generated from leaf explants containing *dao1* whilst all wild-type explants died (Figure 4.5A-B). However, in the presence of 30 mM D-valine (negative selection) both wild-type and transgenic explants failed to grow (Figure 4.5C-D). Lower concentrations of D-valine were then tested. At concentrations of 15 mM D-valine and 5 mM D-valine both transgenic and wild-type explants failed to grow (Figure 4.6A-D), at 2 mM D-valine shoots were generated from both transgenic and wild-type explants (Figure 4.6E). Although it is possible that an intermediate D-valine concentration (2 mM < X < 5 mM) might allow wild-type growth but prevent transgenic growth, this would most likely lead to a high rate of false positives and false negatives due to the minimal window of suitable D-valine concentration (Hattasch *et al.*, 2009). D-valine was therefore found to be unsuitable for negative *dao1* selection in explant regeneration.

4.3.4 Evaluation of *I-SceI* ethanol induction

The other component of the experimental system that required assessment prior to screening was the ethanol induction of *I-SceI*. Leaf tissue and anthers were taken from the T_0 A2 plant immediately prior to, and three days after, induction with 0.7 M ethanol. From this tissue RNA was prepared and cDNA synthesised for use in RT-PCR. A very faint band was observed for *I-SceI* mRNA prior to induction in leaf tissue (Figure 4.7A) indicating a very low level of leaky transcription of *I-SceI* in the absence of ethanol. A strong band was observed post induction (Figure 4.7A), indicating a marked increase in *I-SceI* transcription in the presence of ethanol. A greater level of leaky *I-SceI* expression was observed in anthers (Figure 4.7B) although ethanol induction still led to a considerable increase in expression (Figure 4.7B). This I-SceI expression may cause premature induction of DSBs in this tissue.

4.3.5 Generation of experimental lines

Five T_1 plants were grown for all D and A lines that showed single locus inserts. The T_1 plants were allowed to self fertilise and the resulting T_2 progeny were screened for either kanamycin (D lines) or hygromycin (A lines) resistance to determine segregation ratios and thus establish the zygosity of the respective T_1 parents. Homozygous T_1 plants were identified for lines D4, D13, D14 and A2.

Homozygous D line T_1 plants were crossed to the homozygous A2 line T_1 plant to generate double hemizygous T_2 progeny (i.e. all progeny contain a single copy of both transformation constructs) which were the starting material for the DSB induction. The double hemizygous lines generated were designated D4A2, D13A2 and D14A2.

4.3.6 Induction of DSBs and selection for *dao1* excision

Initially it had been intended to induce DSBs in both somatic cells (leaf) and germline cells (meiotic cells or pollen). DSB repair events resulting in the loss of *dao1* were then to be selected using D-valine either in leaf explant tissue culture (somatic cells) or as seedlings (germline cells). As *dao1* selection in tissue culture was shown to be ineffective, it was not possible to investigate DSB repair in somatic cells using this method. As a result only seedling selection was performed.

Flowers of double hemizygous lines were sprayed with 0.7 M ethanol to induce *I-SceI* expression. The induction was undertaken at a flower bud size of 15-20 mm when it is estimated that microspores enter pollen mitosis I (Villanueva *et al.*, 1985). This is the stage at which organelle DNA is most likely to enter the generative cell nucleus (Yu and Russell, 1994; Sheppard *et al.*, 2008) and therefore DSB repair most likely to lead to insertion of organelle DNA. After induction, flowers were allowed to self-fertilise and the resulting progeny were screened for DSB repair events that had excised *dao1*.

For a DSB to be induced in either male or female gametes both the *dao1* T-DNA and the *AlcR:I-SceI* T-DNA must be present. This allows both production of the I-SceI enzyme (from the *AlcR:I-SceI* T-DNA) and cleavage of the I-SceI sites flanking *dao1* (within the *dao1* T-DNA). Both T-DNAs will be present in a quarter of all gametes (Figure 4.8). The allele resulting from *dao1* excision will be referred to as 'DΔ', the allele without *dao1* deletion 'D' and absence of the *dao1* T-DNA as 'd'. Of the 16 possible combinations of gametes giving rise to seedlings from self fertilised capsules, 7 have the potential to have had *dao1* excision and inherit a DΔ allele (Figure 4.8). In 2 of these cases even if *dao1* is excised another *dao1* will be inherited from the other gamete (DΔD; K^r, D-v^s) making these seedlings D-valine sensitive and so unable to survive selection; one will have the potential to inherit a second DΔ allele from the other parent (DΔDΔ; K^r, D-v^r) making these seedlings D-valine resistant if there have been two *dao1* excisions; and the remaining 4 will be D-valine resistant if the single *dao1* excision occurs [DΔd; K^r, D-v^r](Figure 4.8).

Seedlings that inherit at least one functioning copy of *dao1* (DD, K^r, D-v^s; Dd, K^r, D-v^s; DDΔ, K^r, D-v^s) will be D-valine sensitive and those that do not inherit D or DΔ (dd; K^s, D-v^r) will be kanamycin sensitive (Figure 4.8).

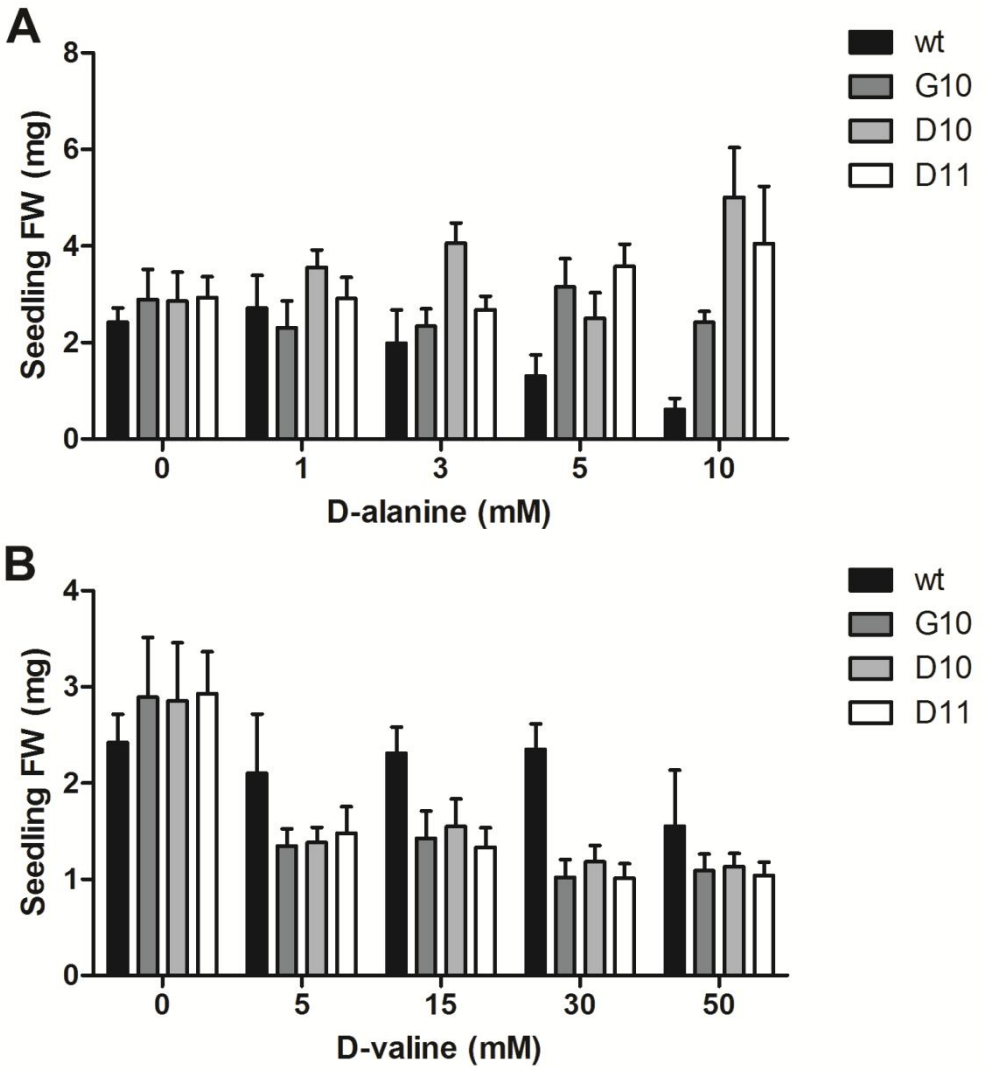


Figure 4.3 D-alanine and D-valine are suitable for positive and negative selection of dao1 respectively in tobacco. Seedlings of transgenic lines containing *dao1* and wild-type (wt) seedlings were grown on various concentrations of D-alanine (**A**) and D-valine (**B**) or media containing neither amino acid (**A-B**). D-alanine was most effective at a concentration of 10mM leading to a strong reduction in the growth of wt seedlings while not affecting the growth of transgenic seedlings (**A**). D-valine was most effective at a concentration of 30 mM leading to a marked reduction in the growth of transgenic seedlings while not affecting the growth of wt seedlings. 50 mM D-valine was toxic to both transgenic and wt seedlings and wt seedlings grown at this concentration were unable to be distinguished from transgenic seedlings. Error bars for both A and B show SD.

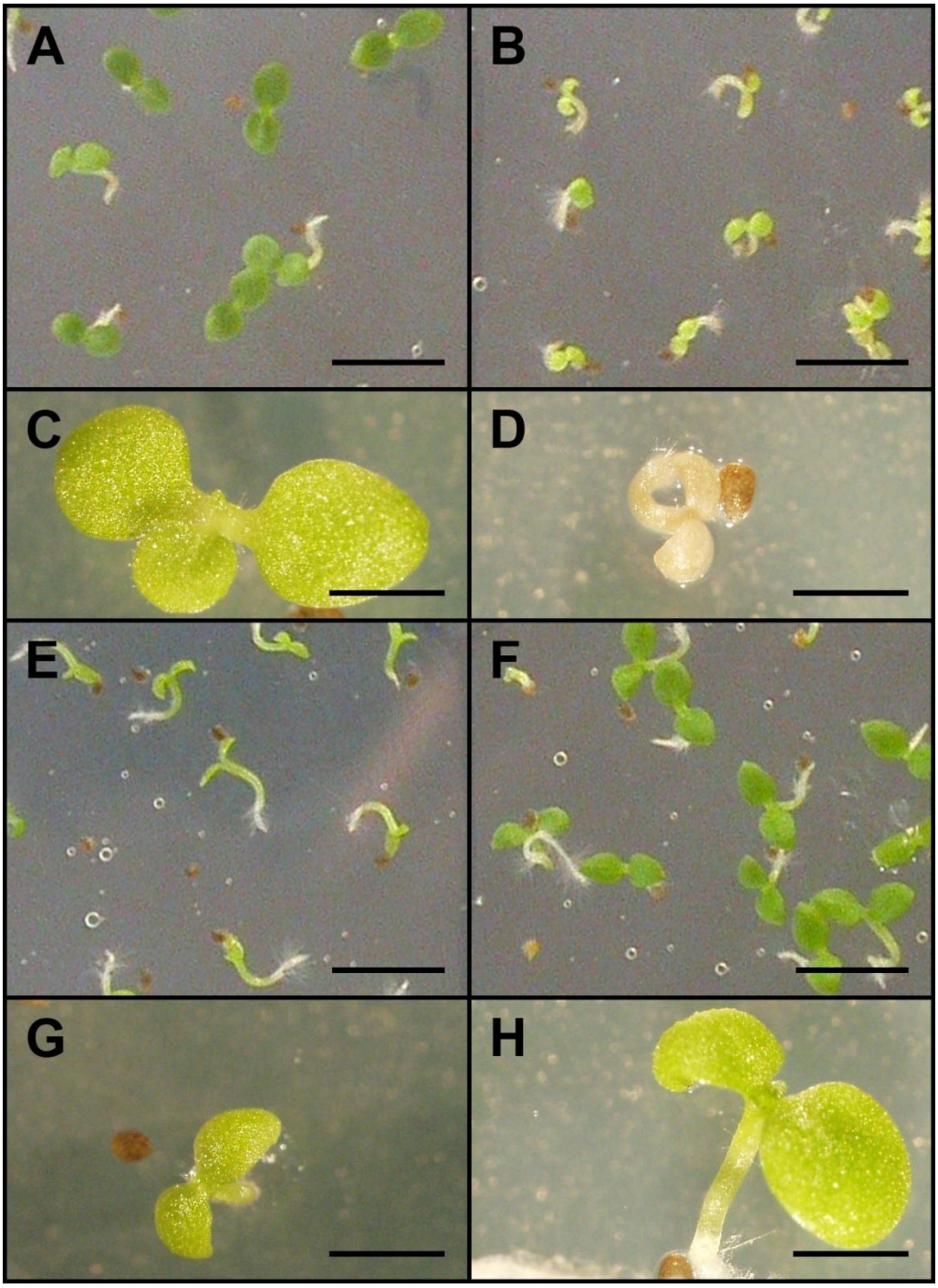


Figure 4.4 *dao1* transgenic and wild-type seedlings were easily distinguishable by sight when grown on both 10 mM D-alanine and 30 mM D-valine. *dao1* transgenic seedlings grown on 10 mM D-alanine showed strong growth (**A,C**), wild-type (wt) seedlings grown on the same medium bleached soon after germination (**B,D**). *dao1* transgenic seedlings grown on 30 mM D-valine had reduced growth (**E,G**) although seedlings did not bleach, cotyledons failed to fully expand and there was no growth of the first true leaf, wt seedlings grown on the same medium showed strong growth (**F,H**). Scale bars for A, B, E and F = 5 mm, scale bars for C, D, G and H = 2mm.

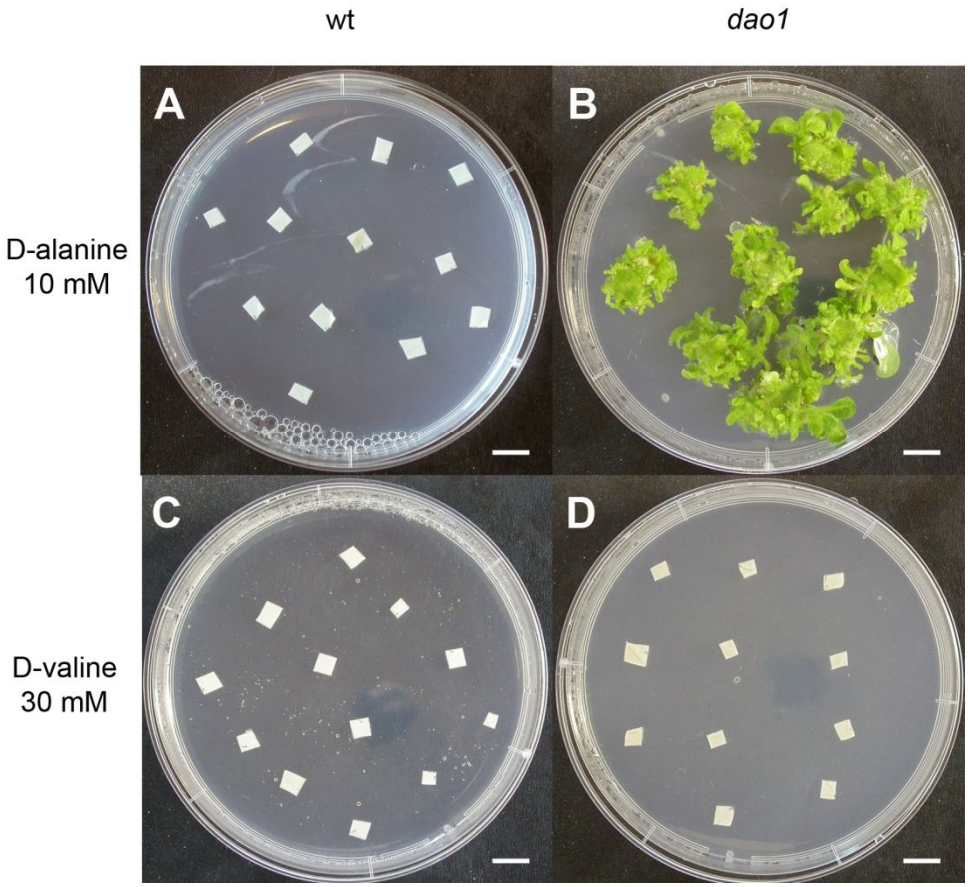


Figure 4.5 10 mM D-alanine is suitable for positive selection of tobacco leaf tissue explants but 30mM D-valine is not suitable for negative selection. Leaf explants taken from wild-type plants (wt) were killed when grown on regeneration medium containing 10 mM D-alanine (**A**). Resistant shoots were generated from *dao1* positive leaf explants grown on same media (**B**). Leaf explants from both wt and *dao1* positive plants were killed when grown on regeneration medium containing 10 mM D-valine (**E-F**). Scale bar = 10 mm.

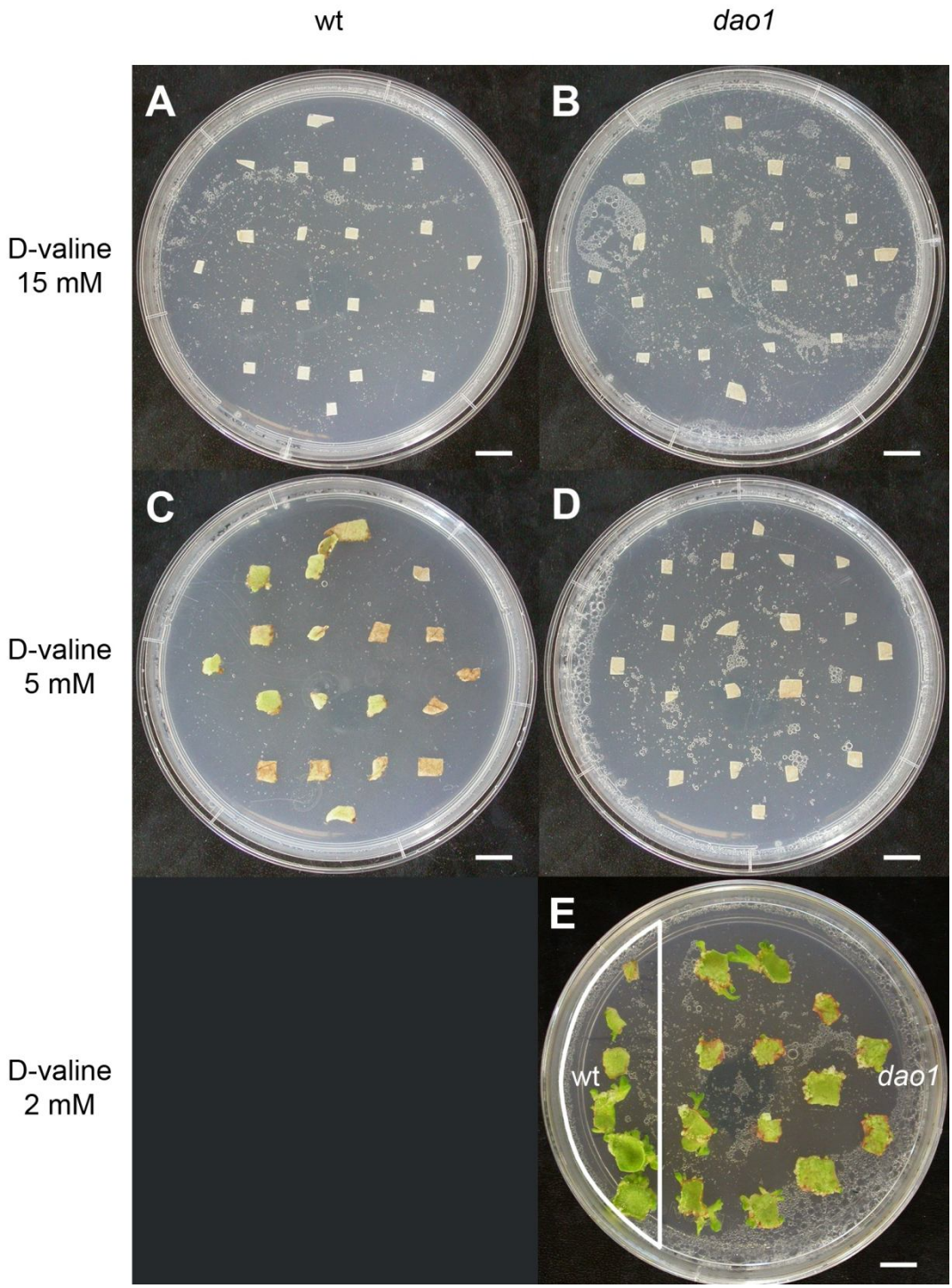


Figure 4.6 D-valine is not suitable for negative selection of tobacco leaf tissue explants. At concentrations of both 15 mM and 5 mM D-valine both *dao1* positive and wild-type (wt) explants failed to generate resistant shoots (**A-D**). At a concentration of 2 mM D-valine both *dao1* positive and wt explants (white boxed area) generated shoots (**E**). Scale bar = 10 mm.

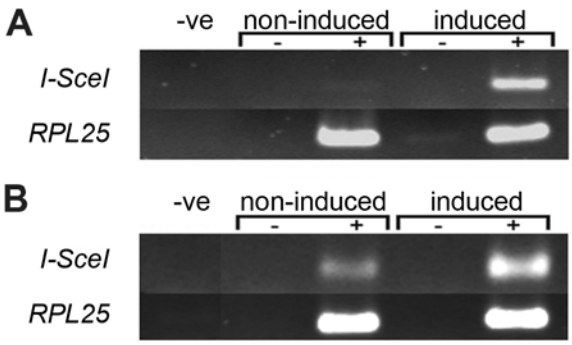


Figure 4.7 I-SceI expression is induced by ethanol. RT-PCR (+) demonstrates increased *I-SceI* mRNA accumulation after induction with 0.7 M ethanol in both leaf tissue (**A**) and anthers (**B**). Considerable I-SceI mRNA accumulation is observed in non-induced anthers (**B**) and low levels of I-SceI mRNA accumulate in non-induced leaf tissue (**A**). No reverse transcriptase (-) and no template (-ve) controls are shown. Template control RT-PCRs used *RPL25* mRNA primers.

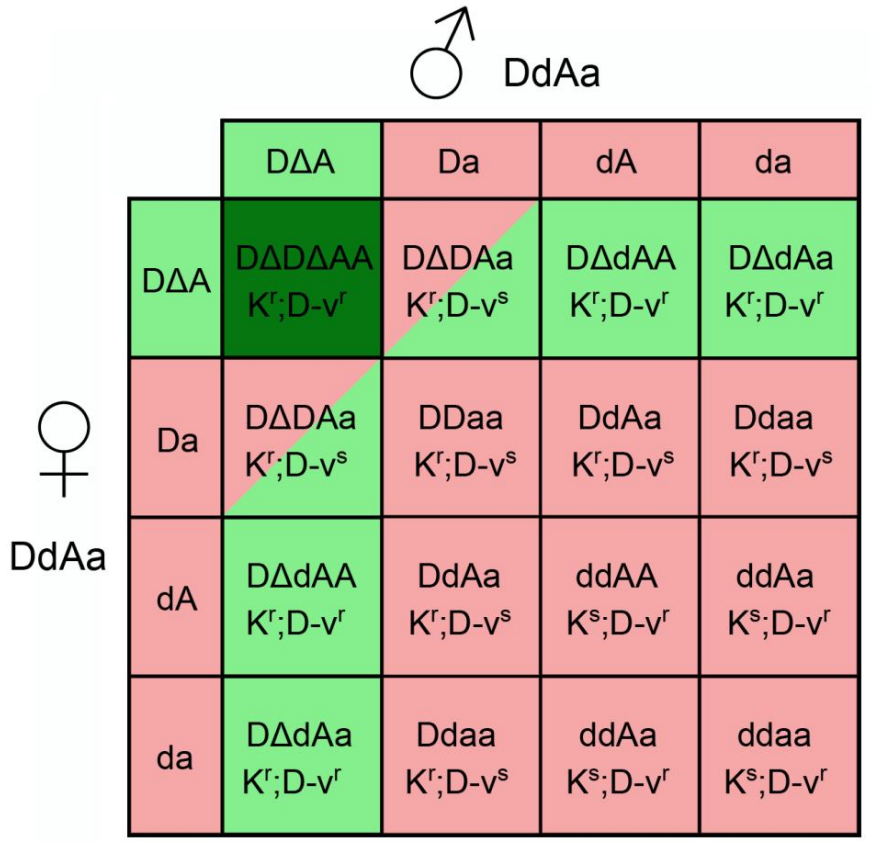


Figure 4.8 Punnett square: pdao1 and pAlcR:ISceI inheritance in progeny of double hemizygous lines. The seedling selection is designed to identify seedlings that have undergone *dao1* excision (DΔ). This is only possible if both the pdao1 T-DNA (D) and pAlcR:ISceI T-DNAs (A) are present in the one gamete (DA). For the purpose of explanation let us assume that all DA gametes undergo *dao1* excision (DΔA, light green) [i.e. 100% efficiency]. Progeny that are not derived from at least one DΔA gamete (pink) will be kanamycin sensitive (K^s), D-valine sensitive (D-v^s), or both. Those derived from one DΔA gamete and one Da gamete will be D-valine sensitive (pink/light green). The remaining progeny will be kanamycin resistant (Kr) and D-valine resistant (D-v^r) [provided they have under gone one (DΔdAA, DΔdAa, light green) or two (DΔDΔAA, dark green) *dao1* excision events].

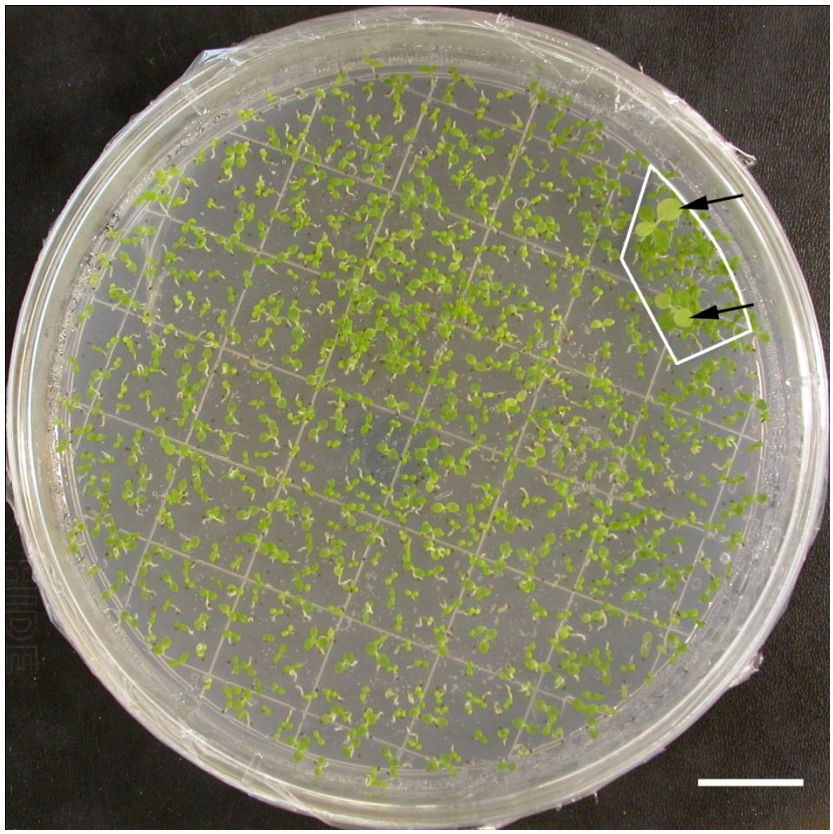


Figure 4.9 No seedlings with *dao1* excision were recovered. Selfed progeny from D4A2 flowers induced with ethanol were grown on media containing 150 mg L^{-1} kanamycin and 30 mM D-valine, none were resistant to both selective agents. Controls consisted of wild-type seedlings (kanamycin sensitive, white box) and kr2.2 seedlings (kanamycin and D-valine resistant, black arrows). This is a representative plate from a larger screen. Scale bar = 20 mm.

Seedlings that contain DΔ (but not D) were selected by growing on medium containing both kanamycin (to select against null segregants) and D-valine (to select against plants retaining *dao1*). In a screen of over 23,000 seedlings no kanamycin and D-valine resistant seedlings were identified (Figure 4.9, Table 4.1), indicating that none had had the *dao1* gene excised. As only 1/8 seedlings had the potential to be both kanamycin and D-valine resistant due to a single *dao1* excision in the male germline, the number of male germline cells screened was effectively ~2875.

To achieve any meaningful results this screen would have had to recover several 100s of DSB repair events. The absence of any recoverable cases of *dao1* excision in 23,000 seedlings makes these events prohibitively rare and, as a result, this screen was not continued.

Table 4.1 Seedling selection for *dao1* excision

Plant	Seedlings screened	Resistant progeny
D4A2 #1	9500	0
D4A2 #4	8000	0
D13A2 #1	6000	0

4.4 Discussion

4.4.1 Evaluation of *dao1* as a selectable marker gene in tobacco

A number of selectable marker genes are routinely used in plants (Miki and McHugh, 2004). These can be used for either positive selection, as is the case with antibiotic resistance genes such as *neo*, *aadA* and *hyg*, or negative selection as with *codA*. Recently a selectable marker gene for use in plants has been described that is capable of both negative and positive selection (Erikson *et al.*, 2004). The gene *dao1* encodes a D-amino acid oxidase (DAAO). Growth of plants on media containing D-alanine and D-serine will select for the presence of *dao1*, whilst growth on media containing D-valine and D-isoleucine will select for the absence of *dao1* (Erikson *et al.*, 2004). Whilst this selectable marker gene has been shown to work in Arabidopsis (Erikson *et al.*, 2004) and Maize (Lai *et al.*, 2007) it was not known if this marker gene would function in tobacco. Analysis showed that *dao1* was effective for use both as a positive and a negative selectable marker gene for *in vitro* selection of germinating seedlings using concentrations of 10 mM D-alanine and 15-30 mM D-valine respectively as the selective agents. In tissue culture, positive selection but not negative selection was able to clearly distinguish *dao1* transgenic and wild-type explants. D-valine, used in negative selection, is converted to the toxic 3-methyl-2-oxo butanoic acid by *dao1,* however, D-valine is itself toxic when used at high concentrations. There is therefore a "concentration window" at which D-valine concentration is sufficiently low to allow growth of cells lacking *dao1* but sufficiently high to allow the production of phytotoxic levels of 3-methyl-2-oxo-

butanoic acid in cells containing *dao1*. In seedlings this "concentration window" was determined to be 15-30 mM D-valine, in explant regeneration there does not appear to be a suitable concentration of D-valine that will result in growth of wild-type cells and death of those containing the *dao1* gene. Recently, D-valine was also shown to be ineffective in tissue culture selection of apple calli, being both toxic to wild-type as well as *dao1* transgenic plants (Hattasch *et al.*, 2009). This same study found that wild-type and *dao1* transgenic calli could be distinguished if D-isoleucine was used as the negative selective agent. However, the use of D-isoleucine in regenerating rare cells lacking *dao1* amongst a majority of cells containing *dao1* was not investigated. Based on the findings of Hattasch *et al.* (2009) negative selection may be possible in tobacco using D-isoleucine, however, the cost of D-isoleucine (over 100 times the price of D-valine) may make large scale screens prohibitive. While D-alanine was suitable for use in distinguishing *dao1* transgenic and wild-type explants this does not ensure that it will be suitable for use in the initial selection of transformants during leaf disc transformation. When used to distinguish transgenic and wild-type explants all cells of the explants tested either contain the *dao1* gene (transgenic explants) or do not (wild-type explants). In transformation, only rare transformed cells contain the *dao1* gene and these must be able to grow surrounded by dying wild-type cells. This difference was highlighted in the study by Hattasch *et al.* (2009) which showed that D-serine was effective in distinguishing *dao1* transgenic and wild-type apple calli but was not suitable for use in selecting apple calli transformed with *dao1* as a selectable marker in *Agrobacterium* transformation.

In relation to the DSB repair experiment, *dao1* was determined to only be suitable for use in the *in vitro* selection of germinating seedlings for identification of *dao1* excision events taking place in the germline. It could not be used in leaf disc tissue culture regeneration to identify cells that had undergone somatic excision of *dao1*.

4.4.2 Seedling screen for *dao1* excision

In a screen of 23,000 seedlings (effectively a screen of ~2800 male gametes, see results), no cases of *dao1* excision were identified. There are several possible reasons for the lack of *dao1* excision, the most likely being either that DSBs were not induced in the gametes or that DSBs were effectively induced but no DSB repair events lead to excision of *dao1*. Very little is known about the specifics of DSB repair in the gametes of plants but observation of transfer of chloroplast DNA to the nucleus in pollen (Huang *et al.*, 2003; Sheppard *et al.*, 2008) suggests that a non-homologous DNA repair mechanism is active. An alternative explanation is that the *dao1* gene is excised from between the I-SceI sites but that the DSB repair machinery re-inserts this gene elsewhere in the nuclear genome. In this case *dao1* excision would not lead to the absence of *dao1* and D-valine

resistance. While it cannot be ruled out, the presence DSB repair that leads in every case to retention of *dao1* excision seems unlikely.

The other possible explanation is that DSBs are not induced in the germline. The induction of DSBs requires I-SceI expression which is dependent upon binding of the AlcR protein to the *alcA*:35S promoter. *AlcR* is itself driven by the 35S promoter. There are some conflicting reports on the activity of the 35S promoter in pollen; while 35S activity has been reported in tobacco pollen (Wilkinson *et al.*, 1997) this may have been an artifact of 35S promoter activity in the anther wall rather than in the pollen itself (Mascarenhas and Hamilton, 1992). More recently GFP expression was found to be absent in pollen of transgenic plants containing a 35S driven *GFP* gene (Hudson and Stewart, 2004). It seems, therefore, that 35S promoter activity in pollen is at most very low. While this appears to conflict with the I-SceI mRNA accumulation observed in RT-PCR, 35S promoter activity in the anther wall would also explain these observations, as the RT-PCR used mRNA derived from whole anthers. If the 35S promoter is not active in pollen then AlcR will not be expressed and will not be able to activate expression of I-SceI. It is also conceivable that the *alcA*:35S fusion promoter itself may not be able to be activated in pollen, given that is partly composed of a minimal 35S promoter.

It seems likely then that the failure to detect *dao1* excision was due to the lack of activity of the 35S promoter resulting in an absence of the AlcR transcriptional activator and therefore an inability to induce I-SceI expression. This could be overcome in future experiments by placing *AlcR* under the control of a promoter known to be active in pollen such as the *LAT52* promoter (Twell *et al.*, 1990).

4.5 Conclusion

The dual selectable marker gene *dao1* can be effectively used for both positive and negative selection of tobacco seedlings but is not suitable for use in tissue culture regeneration of cells lacking *dao1* and thus prevented the selection of cells undergoing *dao1* excision during DSB repair. The absence of *dao1* excision in the germline prevented investigation of DSB repair in the germline also.

4.6 Methods

4.6.1 Plant Growth

Plant growth was as described in section 2.5.1.

4.6.2 Nucleic Acid Extraction

DNA and RNA extraction was undertaken as described in section 2.5.5.

4.6.3 PCR and Sequencing

PCR and sequencing were undertaken as described in section 2.5.6.

4.6.4 Construct Design

pAlcR:ISceI

The *AlcR* expression cassette containing the 35S promoter, *AlcR* ORF, and *nos* terminator was isolated as a NcoI/HindIII fragment from pbinSRN (Caddick *et al.*, 1998). This cassette was blunt ended using the Klenow fragment of DNA pol I and cloned into SmaI cut pGreen0179 to generate pG.AlcR. The *I-SceI* coding region was excised from pCISceI (Puchta *et al.*, 1996) and inserted between the alcA:35S promoter and nos terminator in Alc-pUC (kindly provided by Dr V. Radchuk), using BamHI. Primers AlcF_NcoI and AlcR_NcoI were then used to amplify the *I-SceI* expression cassette and the product was ligated into pG.AlcR using NcoI to generate pAlcR:ISceI (for vector maps see Appendix 5).

pdao1

The 35S terminator from pPRVIIIA::neoSTLS2 (Huang *et al.*, 2003) and 35S promoter from p35S (kindly provided by Dr S. Delaney) were cloned into pGreen0029 (Hellens *et al.*, 2000) using HindIII/BamHI and NotI/XbaI respectively. The *dao1* coding sequence was amplified from pVC_RLM_1qcz (kindly provided by Dr. A. Renz, BASF Plant Science) using primers dao1F and dao1R (dao1R contains an XbaI site at its 5' end), the PCR product was then digested with XbaI and cloned into the pGreen0029 vector containing the 35S promoter and terminator, thus generating pG.dao1.

A multiple cloning site containing two I-SceI restriction sites flanking HindIII and NotI sites was generated by annealing two complementary oligonucleotides I-SceIMCS1 and I-SceIMCS2. This double stranded MCS had 4 bp overhangs at each end allowing ligation into SacI and XhoI cut pGreen0029, generating pG.MCS. The *dao1* expression cassette was excised from pG.dao1 with HindIII and NotI and cloned into HindIII/NotI digested pG.MCS to generate pdao1.

pGU.D.US

The GU.C.US region from pGU.C.USB (Siebert and Puchta, 2002), containing two overlapping regions of the *GUS* gene flanked by a 35S promoter and terminator and separated by a *codA* gene flanked by I-SceI target sites, was excised and cloned into pGreen0029 (Hellens *et al.*, 2000) using BamHI and HindIII to create pGU.C.US. The *dao1* expression cassette was then excised from pdao1 and cloned into pGU.C.US using I-SceI to create pGU.D.US.

4.6.5 Transformation

Transformation of tobacco was performed using the pGreen system of binary transformation vectors (Hellens *et al.*, 2000). The transformation vectors pdao1 and pAlcR:ISceI were independently transformed into an *Agrobacterium* strain containing pSoup (Hellens *et al.*, 2000) using the 'freeze-thaw' method (An *et al.*, 1988) and used in subsequent tobacco transformation. Transgenic lines were generated using a standard leaf disc method (Mathis and Hinchee, 1994) using either 150 mg L^{-1} kanamycin (pdao1, pGU.D.US) or 15 mg L^{-1} hygromycin (pAlcR:ISceI) for selection. Putative transformants were confirmed by PCR using primer pairs dao1F2/dao1R2 and AlcRF1/AlcRR1 respectively.

4.6.6 Analysis of Antibiotic Resistance in Seedlings

Antibiotic resistance in seedlings was assessed as described in section 2.5.4. Hygromycin selection was performed using 15 mg L^{-1} hygromycin.

4.6.7 *dao1* seedling selection

Positive and negative *dao1* selection was evaluated by growing T_1 seedlings from two single locus lines (G10 and D11), one multiple locus line (D10) and wild-type on 0.5 × MS agar medium containing a range of D-alanine concentrations (1 mM, 3 mM, 5 mM and 10 mM), a range of D-valine concentrations (5 mM, 15 mM, 30mM and 50 mM) and on media containing neither D-amino acid. Two weeks after germination the average seedling weight (average fresh weight of 10 seedlings) was determined for each line at each concentration of D-amino acid. At this stage the ability to visually distinguish transgenic and non-transgenic seedlings was also assessed.

4.6.8 *dao1* tissue culture selection

T_1 *dao1* line seedlings were grown on media containing kanamycin to select positive segregants and these were transferred to 0.5 × MS agar medium in tissue culture jars for further growth. Wild-type seedlings were sown directly onto 0.5 × MS agar medium in tissue culture jars. After four weeks growth in jars, leaf explants from both wild-type and *dao1*-transgenic plants were placed adaxial side down on regeneration MS104 (Mathis and Hinchee, 1994) medium containing D-alanine (10 mM) or various concentrations of D-valine (30 mM, 15 mM, 5 mM and 2 mM).

4.6.9 Ethanol induction of *I-SceI* for RT-PCR

I-SceI induction was assessed in a T_0 plant transformed with pAlcR:ISceI. For leaf induction, tissue samples were obtained from the same leaf tissue just prior to induction and again 72 hours after induction. Anthers were obtained 72 hours after induction from induced and non-induced flowers on the same plant. Induction was performed by spraying the leaf or inflorescence with ~1-2 mL 0.7 M ethanol and subsequently enclosing the leaf or inflorescence in a plastic bag to maintain the

presence of ethanol vapour. Induced and non-induced tissue samples were used in subsequent RT-PCR.

4.6.10 RT-PCR

RT-PCR was performed as described in section 2.5.7. Primers I-SceIF1 and I-SceIR1 were used in the amplification of *I-SceI* cDNA and primers L25F and L25R used in amplification of *RPL25* cDNA.

4.6.11 Experimental induction of DSBs

I-SceI expression in leaves was induced in the T_0 A2 plant grown in soil. Mature leaves were sprayed with 1-2 mL of 0.7 M ethanol and enclosed in a plastic bag to maintain the presence of ethanol vapour. The plant was then left for 4 days to allow time for I-SceI expression, generation of DSBs and subsequent DSB repair. After 4 days leaf tissue was sampled.

I-SceI expression in floral tissue was induced by spraying 15-20 mm flower buds with 1-2 mL 0.7 M ethanol and subsequently enclosing the inflorescence in a plastic bag to maintain the presence of ethanol vapour and hence I-SceI expression. After 4 days the plastic bag was removed and the flowers allowed to self fertilise.

4.6.12 Seedling screen for *dao1* excision

Seeds were harvested from capsules produced by induced flowers on double hemizygous plants and grown on selective media containing 150 mg L^{-1} kanamycin and either 15 or 30 mM D-valine. Negative control (D4; K^r, D-val^s) and positive control seeds (kr2.2; K^r, D-val^r) were included on each plate.

Chapter 5: Investigating DSB repair by single molecule PCR

5.1 Introduction

To investigate the role of DSB repair in the nuclear insertion of organelle DNA, an experimental system was designed which aimed to enable the characterisation of DSBs in specific tissues and under particular growth conditions (see Chapter 4). Briefly, DSBs were generated in nuclear DNA by expression of the rare-cutting endonuclease I-SceI which was present as a nuclear transgene driven by an ethanol inducible promoter (Figure 4.1). The DSBs were induced at two I-SceI target sites that flanked the *dao1* dual selectable marker gene which was present at a second locus (Figure 4.1). Repair of these DSBs could then lead to *dao1* excision (Figure 4.2).

Recovery of repair events resulting in the excision of *dao1* was attempted by selecting for the absence of *dao1* both in seedlings and during explant regeneration in tissue culture. In tobacco explants, *dao1* was found to be unsuitable as a negative selectable marker gene. In seedling germination screens *dao1* negative selection was shown to be possible but no *dao1* excision events could be recovered as I-SceI expression was unable to be induced in the germline (see Chapter 4).

The inability to induce *dao1* excision in the germline and the inability to use *dao1* negative selection in tissue culture explant regeneration necessitated an alternative approach to identify and characterise DSB repair events. The method chosen was single molecule (sm)PCR, which has been used previously in wide ranging applications (Ben Yehezkel *et al.*, 2008; Chhibber and Schroeder, 2008; Kraytsberg *et al.*, 2009) and is ideally suited to the analysis of somatic mutation allowing amplification of a desired locus from individual template DNAs (Kraytsberg and Khrapko, 2005).

A standard PCR, using 50 ng of tobacco genomic DNA as template, contains DNA from ~4300 nuclei. If the sequence of the region to be amplified differs between nuclei (as is expected here as a result of independent DSB repair events in each cell), the reaction will contain amplicons derived from a large number of different template molecules. As the aim is to amplify template molecules corresponding to rare insertion events, standard PCR is not appropriate on two counts; firstly, PCR will preferentially amplify those templates present in high copy number (those arising from common repair events such as repair without insertion or deletion) and secondly PCR preferentially amplifies smaller products and therefore amplification of a template molecule made larger due to an insertion between the primer binding sites will be less efficient.

In smPCR, template DNA is diluted so that each product is amplified from a single template molecule. Each product will therefore represent the sequence of an individual DSB repair event and PCR artefacts such as template jumping and allelic preference will be avoided (Kraytsberg and Khrapko, 2005). Avoiding allelic preference was particularly important in this application as the aim was to amplify rare large templates amongst a population of smaller template DNAs.

In practice DNA must be diluted to less than 1 template molecule per reaction to ensure that a significant number of reactions contain only a single template molecule. Generally DNA is diluted until over 50% of reactions receive no template DNAs at all (Kraytsberg and Khrapko, 2005). Table 5.1 shows the expected proportion of reactions containing a single template molecule for several template concentrations. The number of template molecules in each reaction follows a Poisson Distribution; the formula describing the general case is given below.

Table 5.1 Number of template molecules in smPCR reactions

Average templates per reaction	% reactions containing n templates[a]			Products amplified from a single template
	n = 0	n = 1	n ≥ 2	
10	0%	0%	100%	0%
1	37%	23%	50%	37%
0.8	45%	36%	19%	65%
0.43[b]	65%	28%	7%	80%
0.1	90%	9%	1%	92%

[a] Proportion as given by Poisson Distribution

[b] This value represents that used in experimentation and was calculated using the formula given below

$$f\ (k\ ;\ \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$f\ (k\ ;\ \lambda)$ = the proportion of reactions that contain k template molecules given an average of λ template molecules per reaction

smPCR proved to be a high-throughput method for the amplification of products representing individual DSB repair events. Analysis of ~300 DSB repair events indicated that most involve the loss of nucleotides from one or both ends being joined. Insertions were observed in five repair events, although none of the inserted sequences were of organelle origin. Notably, the amount of

nuclear sequence deleted was significantly larger in repair events involving insertion than in those without insertion, indicating that the two types of repair may be mediated by distinct pathways.

5.2 Results

5.2.1 PCR detection of DSB repair events

Standard PCR was used initially to confirm that *I-SceI* expression resulted in the induction of DSBs and *dao1* excision. From 4-5 week old D4A2 plants (chapter 4) grown in tissue culture jars, leaf tissue was taken immediately prior to and four days after ethanol induction. DNA was extracted from leaf samples and the *dao1* locus then amplified using primers flanking the two I-SceI sites. A 2.9 kb band was expected from template molecules that had not undergone *dao1* excision (Figure 4.2). An 834 bp band was expected from template molecules that had undergone *dao1* excision without any associated insertion or deletion (Figure 4.2).

A 2.9 kb band, corresponding to amplification of the locus without *dao1* excision, was amplified from all DNA templates (Figure 5.1A). In addition to the 2.9 kb product a ~834 bp product was amplified when using template DNA extracted from tissue that had undergone DSB induction (Figure 5.1A), indicating that DSB induction was resulting in *dao1* excision. As both 2.9 kb and ~834 bp bands are present it is evident that some template molecules originate from cells that have undergone DSBs leading to *dao1* excision and some from cells that have not.

Several weak unexpected high molecular weight bands were also observed in amplification of DNA from induced tissue (Figure5.1A). This may have been due to insertions at the site of DSB repair. To prevent amplification of those template molecules containing the unchanged T-DNA and favour amplification of any insertion events, samples were digested with HincII which cuts three times within the *dao1* gene but not in the sequence between the I-SceI sites and primer binding sites (Figure 4.2). Unavoidably this would also prevent amplification of template molecules arising from DSB repair events involving insertion of DNA containing a HincII site. After HincII digest only the ~834 bp product was amplified (Figure 5.1B).

5.2.2 Single molecule PCR

To amplify PCR products representing individual DSB repair events smPCR was used. In the experimental plants, the DSB locus is hemizygous and therefore present in a single copy per cell. There is ~11.5 pg of DNA in a diploid tobacco cell and so each 11.5 pg of total cellular DNA will contain one potential target molecule. In order to amplify only those molecules resulting from a DSB repair event which lead to the excision of *dao1,* the template DNA was digested with HincII. HincII cuts three times within the *dao1* gene (Figure 4.2) and so prevents amplification from potential templates that have not undergone *dao1* excision. This reduces the number of template

molecules per unit weight DNA in a manner dependent upon the efficiency of the *dao1* excision. As the efficiency of *dao1* excision was not known it was not possible to determine the average template molecules in a given weight of DNA. Instead a series of DNA concentrations were tested to arrive at a concentration resulting in approximately one third of reactions amplifying a product.

Two independent plants D4A2#2I and D4A2#6I were tested and a DNA concentration of 110-130 pg/reaction was chosen. This resulted in a product being amplified in 33-38% of reactions. For both plants this corresponds to one template molecule in ~275 pg, (or one template molecule in every 24 diploid genomes) indicating efficient DSB induction and *dao1* excision in leaf cells.

In 840 PCRs, products were amplified in 296 reactions (Table 5.2). The majority of PCR products were ~834 bp in length (Table 5.2, Figure 5.2A), the size expected with direct joining of the two I-SceI sites (Figure 4.2). Seven (2.4%) PCR products were significantly smaller, corresponding to large (>50 bp) deletions (Table 5.2, one example is shown in Figure 5.2B) and five (1.7%) were significantly larger indicating insertions (Table 5.2, one example is shown in Figure 5.2C).

Table 5.2 Overview of smPCR results

Line	reactions	products	~834 bp	Deletion (>50 bp)	Insertion (>50 bp)
D4A2#2I	424	159 (38%)	156 (96.9%)	2 (1.3%)	1 (0.6%)
D4A2#6I	416	137 (33%)	128 (93.4)	5 (3.6%)	4 (2.9%)
Total	840	296 (35%)	284 (95.3%)	7 (1.7%)	5 (2.4%)

All five insertion products and 26 other randomly chosen products were sequenced. Of the 26 random products, 6 (23%) were clearly the result of reactions containing more than one template molecule and discarded (20% of products were expected to result from reactions containing more than one template molecules given Poisson Distribution, Table 5.1). The remaining 20 represented sequences of individual DSB repair events. About half resulted in no loss of sequence, re-forming the I-SceI restriction site (Figure 5.3A). 12 had small deletions (1-62 bp) and 3 had single nucleotide insertions (Figure 5.3A). In each case the inserted base was a T (or an A on the opposite strand). In some instances microhomology was observed between the terminal bases of the fragments being joined (Figure 5.3A) although more junctions would need to be sequenced to determine if the levels were greater than those expected by chance.

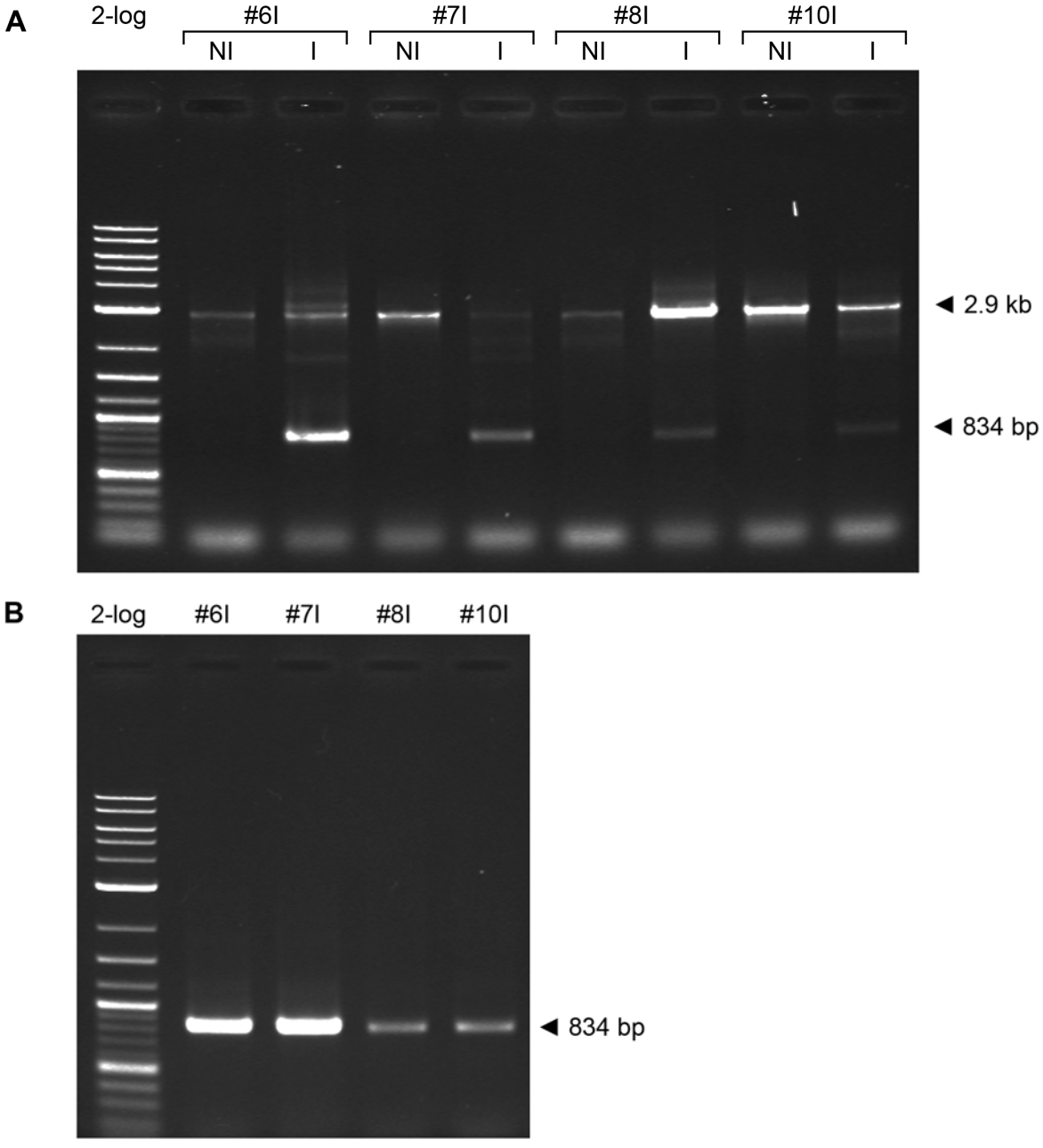


Figure 5.1 Ethanol induction causes *dao1* excision. (**A**) The DSB region was PCR amplified from 4 D4A2 plants using primers DSBF1 and DSBR1 which flank the two I-SceI sites. Only the full length 2.9 kb band is amplified from template DNA extracted prior to DSB induction (NI). An additional ~834 bp band is amplified from template DNA extracted after DSB induction (I). 834 bp is the expected size of the DSB region after excision of the *dao1* gene. (**B**) After HincII digestion of induced template DNA only the 834 bp band is amplified. No amplification is observed from those molecules which have not undergone *dao1* excision.

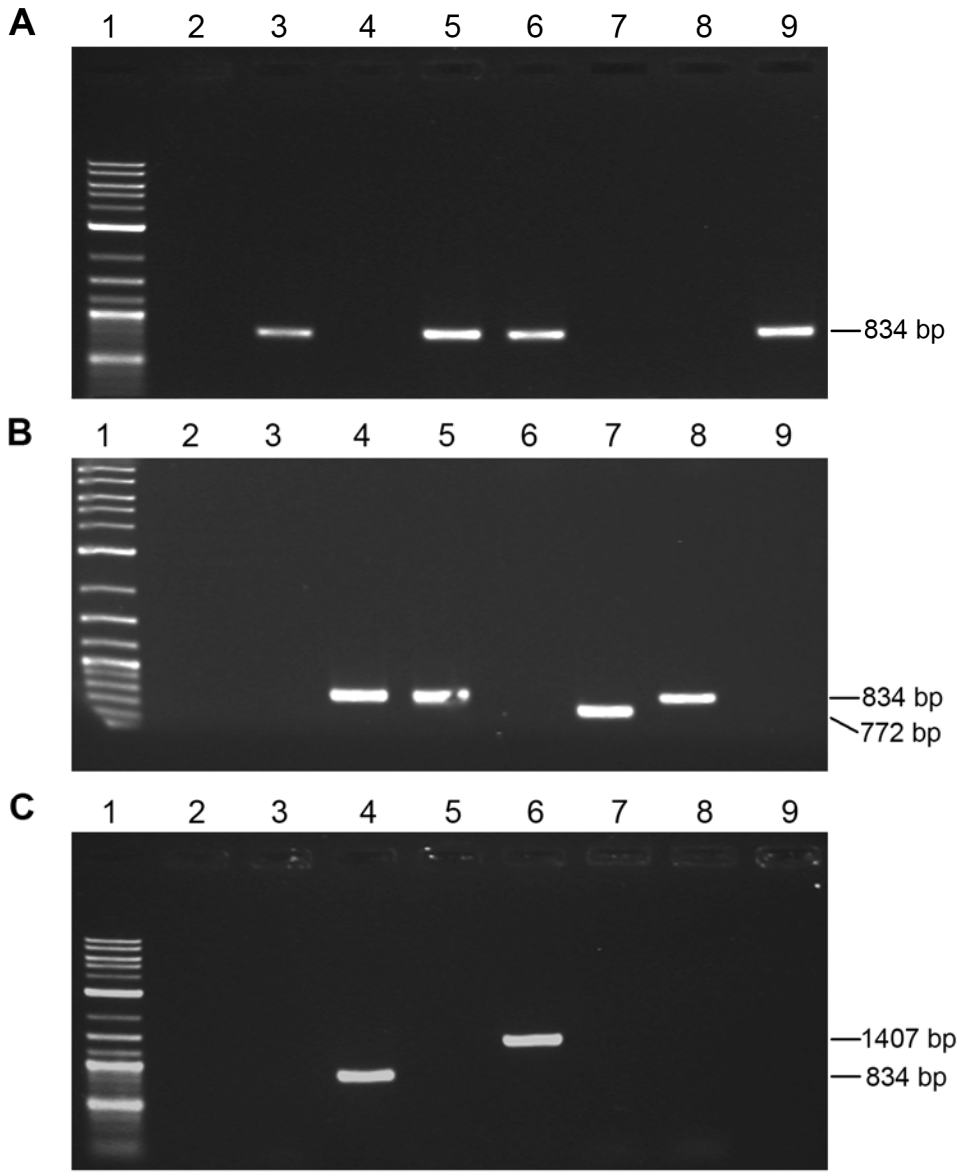


Figure 5.2 DSB events resulting in deletion, insertion or direct joining of the I-SceI sites were observed by smPCR. The majority of products amplified using smPCR were ~834 bp in size (**A-C**). Some repair events resulted in deletions leading to products < 834 bp (**B**) while others resulted in insertion leading to products > 834 bp (**C**). Examples shown are amplified from D4A2#6I template DNA.

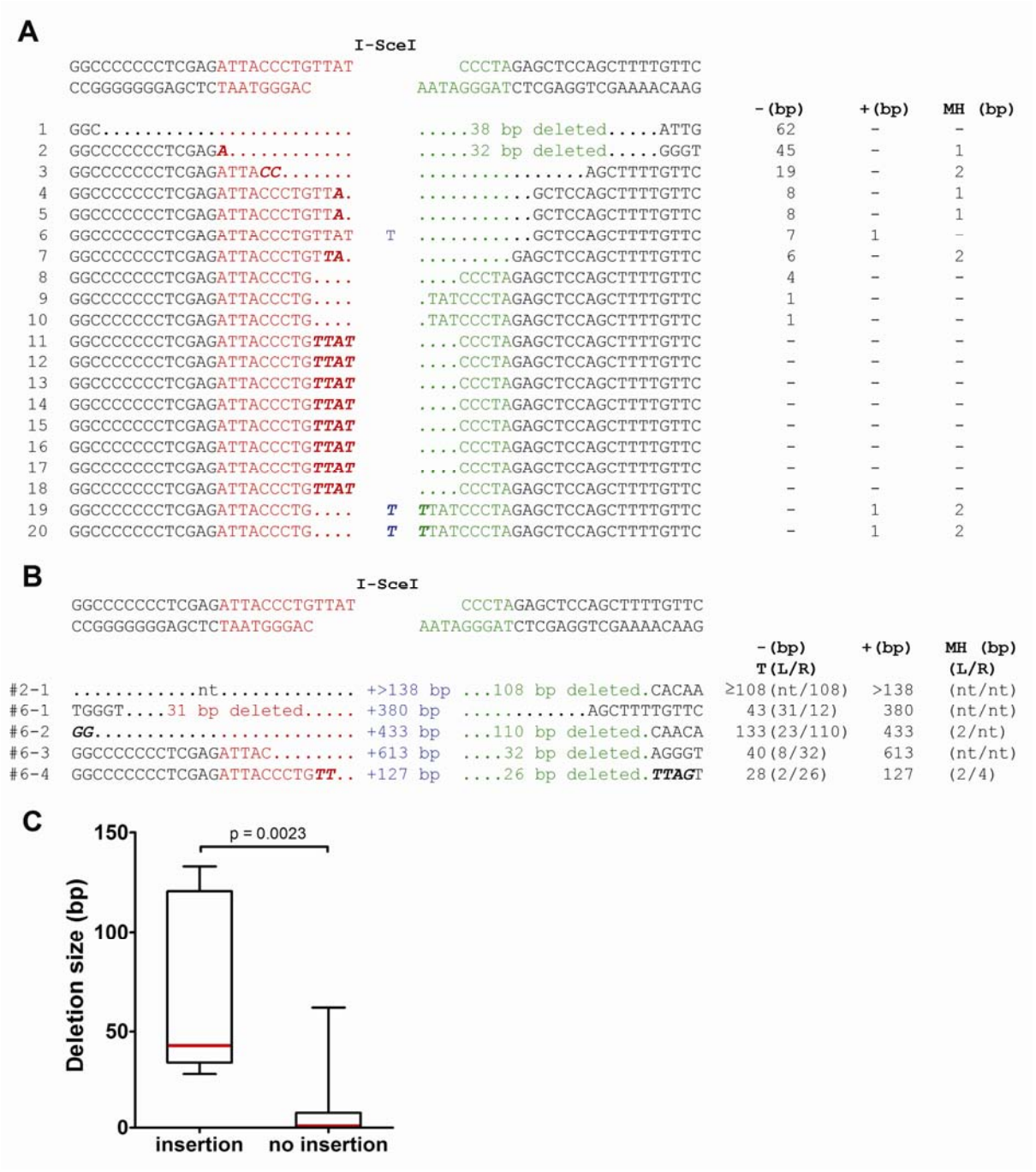


Figure 5.3 Sequence of double strand break repair events. The sequence surrounding the junction sites is shown for 20 randomly chosen repair products (**A**) and the 5 repair products that harboured insertions (**B**). The original sequence generated by I-SceI cleavage is shown at the top of both **A** and **B** (the sequence of both strands is shown). Bases from the I-SceI site upstream of *dao1* are shown in red, bases from the I-SceI site downstream of *dao1* are shown in green. Inserted bases are shown in blue. In some instances microhomology was observed between the terminal bases of the fragments being joined (bold italics). Columns to the right of **A** and **B** indicate the total length of deletion (**-**), insertion (**+**) and microhomology (**MH**, not including I-SceI site overlap), nt signifies not testable. Numbers in brackets indicate length of deletion or microhomology observed at the junction either upstream (**L**) or downstream (**R**) of the insertion. The median deletion size was considerably larger in DSB repair events involving insertion than in repair events not involving insertion (**C**). The box-and-whisker plot shows the median (red line), the first and third quartiles, and the upper and lower limits of the length of deletions (two-tailed Mann-Whitney U test).

The five insertions ranged from 127-613 bp in length (Figure 5.3B) and in all cases insertion was accompanied by deletion (Figure 5.3B). The median size of deletion was found to be significantly larger in DSB repair events involving insertion than that in repair events that did not involve insertion (Figure 5.3C, $p < 0.01$, two-tailed Mann-Whitney U test). The difference in median deletion size was still significant when DSBs repaired by direct joining of the two I-SceI sites were excluded from the analysis ($p < 0.05$, two-tailed Mann-Whitney U test). For the deletion analysis, only those DSB repair events harbouring insertions >1 bp were considered true insertions. Investigation of the presence of micro-homology at repair junctions was limited to three junctions. At these three junctions the bases that flanked the insert sequence in its original context could be inferred from the EST sequence to which the insert matched. For the other junctions, BLAST searches only indentified accessions with limited identity to the insert sequence which prevented unequivocal assessment of microhomology. This is a limitation of analyses such as this where the sequence from which the insert originates is unknown. All three junctions that could be assessed showed micro-homology (2-4 bp), although more data need to be generated to determine if this is meaningful.

None of the inserts were found to be of organelle origin. Part of insertion #2-1 was identical to an isoleucine tRNA. Insertion #6-3 contained a region, flanked by 37 bp inverted repeats, which showed 70% identity (e value, 8.2×10^{-8}) to an unclassified transposable element. The rest showed complete or partial identity to uncharacterised tobacco genomic or EST clones (Table 5.3) indicating that all insertions are probably of nuclear origin.

Table 5.3 Size and origin of insertions

Insertion	Length	Origin
#2-1	>138	Partial match to tRNA-ILE, NR_023306
#6-1	380	Complete match to uncharacterised tobacco EST, AM843263
#6-2	433	Partial match to uncharacterised tobacco ESTs, FS418974, EB695504
#6-3	613	Similar to unclassified transposable element, Plant Repeat Database
#6-4	127	Complete match to uncharacterised tobacco EST, FS392274

5.3 Discussion

5.3.1 Single molecule PCR

The inability to use *dao1* selection to identify *dao1* excision events either in seedling selection or in tissue culture regeneration necessitated an alternative approach. smPCR proved to be an effective method for generating the sequence of individual DSB repair events. In this study all repair events observed must be mediated by the NHEJ pathway of DSB repair as the sequence generated by *dao1* excision is novel and therefore not due to DSB repair by homologous recombination.

DNA templates representing DSB repair events involving both insertion and deletion were effectively amplified using smPCR. Deletion of nucleotides at the end of the two DNA fragments occurred in about 50% of repair events, these were mainly small (1-50 bp) with larger deletions (>50 bp) observed in only 2.4% of repair events. Deletions that resulted in the loss of one or both primer binding sites would not have been observed in this analysis and therefore a maximum deletion size of ~750 bp could be amplified by PCR using these primers. As a result, 2.4% is a conservative estimate of the proportion of repair events that involve large deletions.

Insertions were observed in 1.7% of repair junctions and in each case insertion was also accompanied by the deletion of flanking nuclear sequence. The amount of DNA deleted was significantly larger in these cases than in repair events not including insertion suggesting that DSB repair involving insertion may be mechanistically different from DSB repair that does not involve insertion. Larger deletions at sites of DSB repair involving insertion have also been observed in mammalian cells (Simsek and Jasin, 2010) and this difference may therefore be a general feature of DSB repair by NHEJ. A more in-depth analysis of the repair junctions was not possible as in most cases the sequence from which the insert originated was unknown, preventing the detection of filler DNA or micro-homology.

Of the five insertions recorded, none involved organelle DNA. In all cases the closest match in BLAST analysis was to tobacco genomic or tobacco EST sequence (Table 5.3). Insertion #6-2 appeared to be composite sequence with different regions of the insert matching separate EST clones (Table 5.3). Part of insertion #2-1 closely matched the sequence of tRNA-ILE (Table 5.3) suggesting that it may be SINE derived sequence (Ohshima and Okada, 1994). Two others were identical to uncharacterised ESTs and the fifth showed similarity to an unclassified transposon. These findings confirm earlier reports that DNA insertions at sites of DSB repair in tobacco are generally repetitive genomic sequences (Salomon and Puchta, 1998; Kirik *et al.*, 2000). Recently the NHEJ pathway has been shown to be involved in the retrotransposition of LINEs (Suzuki *et al.*, 2009), further highlighting the role of DSB repair in the insertion of transposable elements.

Previous studies that have investigated DSB repair in tobacco have reported a much higher proportion of repair events that involve insertion; 15-50% (Salomon and Puchta, 1998; Kirik *et al.*, 2000) compared to 1.7% reported here. There are several reasons for this. Firstly, in previous studies DSBs have been induced between the promoter and coding sequence of a negative selectable marker gene (Salomon and Puchta, 1998; Kirik *et al.*, 2000). Repair events could therefore only be recovered if relatively large deletions involving either the promoter or coding sequence of the marker gene occurred. As larger deletions are more likely to involve insertion, these studies would have been biased toward observing those repair events involving insertion. Conversely, the study described in this chapter was biased toward observation of smaller deletions as large deletions could remove one or both primer binding sites preventing amplification. In addition, HincII digestion would have prevented amplification of any inserts that contain a HincII site. The true proportion of repair events involving insertion presumably lies somewhere between that observed in this study and those observed in the studies by Salomon and Puchta (1998) and Kirik *et al.* (2000).

Although no insertions of organelle DNA were observed, this study has demonstrated that smPCR is an efficient, high-throughput method for screening and sequencing a large number of individual DSB repair events. In future this method could be used in larger screens to identify and characterise more insertion events. This may be particularly informative in conjunction with growth conditions and/or stress treatments which are more likely to result in organelle breakdown and therefore allow entry of organelle DNA into the nucleus.

5.4 Conclusion

This study has shown smPCR to be an effective high-throughput method for screening large numbers of DSB repair events and has the potential to be used in wide ranging investigations of DSB repair. Analysis of ~300 DSB repair events indicated that most involve the loss of nucleotides from one or both ends being joined. Notably, the amount of nuclear sequence deleted was significantly larger in repair events involving insertion than in those without insertion, indicating that the two types of repair may be mediated by distinct pathways.

5.5 Methods

5.5.1 Experimental induction of DSBs

I-SceI expression was induced in the leaves of D4A2 plants (see chapter 4) grown in tissue culture jars. Leaves were sprayed with 1-2 mL of 0.7 M ethanol and the jar lids replaced. The plants were then left for 4 days to allow time for I-SceI expression and generation of DSBs and their subsequent repair. After 4 days leaf tissue was sampled.

5.5.2 DSB PCR

Products were amplified with LongAmp *Taq* DNA polymerase (New England Biolabs) using suggested PCR conditions, primers DSBF1 and DSBR1, an annealing temperature of 59°C and 40 ng template DNA. Non-induced, induced/undigested and induced/digested DNA was used as template. For digested template, 2 µg genomic DNA was digested overnight at 37°C using 20 U HincII (New England Biolabs) in a 20 µL reaction and purified using a PCR purification kit (QIAGEN) according to manufacturer's instructions.

5.5.3 smPCR

Single molecule PCR was performed using LongAmp *Taq* DNA polymerase. Reactions were 2 µL in volume and contained 0.3 mM dNTPs, 0.4 µM primers (DSBF1 DSBR1), 0.2 U LongAmp *Taq* DNA polymerase, 1 × LongAmp buffer and 110-130 pg template DNA. Reactions were overlayed with mineral oil to prevent evaporation. Cycle conditions were as follows: Initial denaturation 95°C 30 sec then 45 cycles of 95°C 20 sec; 59°C 20 sec; and 65°C 4 min followed by a final extension at 65°C for 10 min. After PCR, 18 µL of H_2O was added to each reaction to give a total volume of 20 µL. 5 µL was analysed by standard agarose gel electrophoresis and the remainder used in subsequent sequencing.

5.5.4 Statistical analysis

Statistical analysis of deletion size (two-tailed Mann-Whitney U test) was performed using Prism 5 (GraphPad Software).

5.5.5 Sequence analysis

BLAST analysis was performed on several databases, including NCBI's non-redundant nucleotide collection (nr/nt) and non-human non- mouse ESTs (est_others) [http://blast.ncbi.nlm.nih.gov] and the PRD's Solanaceae repeats database (http://plantrepeats.plantbiology.msu.edu) using blastn.

Chapter 6: Discussion and Conclusions

6.1 Introduction

The cells of plant, algal and some protist lineages contain three genetic compartments. These are the nucleus, which houses the majority of the genes, and two cytoplasmic organelles - mitochondria and plastids. The mitochondria and plastids have an endosymbiotic origin and are the extant descendants of once free-living α-proteobacteria and cyanobacteria respectively. Following their incorporation into the eukaryote cell the organelle ancestors underwent massive genome reduction so that the current organelles contain only 1-5% of the genes found in their free-living prokaryotic relatives. This genome reduction has been primarily due to relocation of organelle genes to the nucleus. In flowering plants, transfer of non-functional DNA as well as functional relocation of genes from the organelles to the nucleus continues today. As a result, large tracts of DNA essentially identical to regions of the extant plastid and mitochondrial genomes are found within all angiosperm nuclear genomes.

The wealth of sequencing data now available has revealed the large extent to which the two cytoplasmic organelles have contributed to plant nuclear genomes. It is estimated that ~14% of nuclear genes are derived from the ancestral plastid (Deusch *et al.*, 2008) and potentially many more than this are derived from the ancestral mitochondrion (Esser *et al.*, 2004). Smaller fragments of organelle DNA have also contributed to nuclear genomes by inserting into pre-existing nuclear genes to which they now contribute coding sequence (Noutsos *et al.*, 2007). In this way the movement of organelle DNA to the nucleus has greatly added to the complexity of nuclear genomes, providing raw genetic material for the creation of new nuclear exons and whole genes (Timmis *et al.*, 2004). While a large number of the transferred genes retain their original organellar function, many others have acquired novel non-organelle based roles, adding new prokaryote biochemistry to various parts of the eukaryote cell (Martin *et al.*, 2002).

Recent developments have allowed the experimental reconstruction of both endosymbiotic DNA transfer and functional activation of the transferred genes, enabling investigation of the molecular mechanisms involved. This is essential for developing our understanding of this important evolutionary process and is also of great biotechnological significance in view of the desire to minimise transfer of chloroplast transgenes to the nucleus.

6.2 Insertion

This project aimed to investigate the molecular mechanisms by which chloroplast sequences integrate into nuclear chromosomes and then become functionally activated within this very different genetic environment. The insertion of chloroplast DNA into nuclear chromosomes was investigated using two methods, firstly a *de novo* chloroplast DNA integrant and its nuclear pre-insertion site were sequenced and analysed and secondly, the role of DNA double strand break (DSB) repair was assessed directly through observation of DSB repair events involving DNA insertion.

Chapter 3 reported the full sequence of the chloroplast DNA insertion in line kr2.2 and is the first full characterisation of a *de novo* chloroplast integrant and its pre-insertion site. The integrant is composed of three fragments of chloroplast DNA from separate parts of the chloroplast genome and these fragments inserted into a region of the nuclear genome containing transposable element (TE) related sequence. This region is now in effect a sequence mosaic, containing a complex arrangement of plastid and transposable element-derived sequences. From what is known as a result of the limited characterisation of other *de novo nupts* (Huang *et al.*, 2004) and also extensive bioinformatic analysis of evolutionarily transferred *nupts* (Noutsos *et al.*, 2005), the kr2.2 integrant appears to be typical of organelle DNA insertions.

The complexity of the sequence arrangement in kr2.2 is almost certainly an immediate product of the insertion mechanism. Filler DNA observed at the junctions between chloroplast and nuclear sequence and between chloroplast fragments in this line suggest that the multiple ends of chloroplast DNA may have been recruited to a single repair node where the loose DNA ends were processed and joined by synthesis dependent non-homologous end joining (NHEJ), probably at a site of DSB repair. It also suggests that at the time this insertion event occurred, during male gametogenesis, there must have been a large amount of chloroplast DNA invading the nucleus. This conclusion is based on the finding that three sequences that are widely separated in the plastome, were all available to be integrated into nuclear DNA within a very brief time frame.

Recently, filler DNA insertion very similar to that observed in the kr2.2 integration has been observed during DSB repair in Drosophila male gametes (Chan *et al.*, 2010). As was found in the kr2.2 integration, the filler DNA introduced was derived from sequence close to the end of one of the fragments being joined. The 'alternative' NHEJ pathway responsible was shown to be Ku and Lig4 independent and was reliant upon DNA polymerase theta. Interestingly, the Arabidopsis homologue of polymerase theta, TEBICHI, is also involved in DSB repair (Inagaki *et al.*, 2006) and is specifically up-regulated in sperm cells (Borges *et al.*, 2008). Given the high TEBICHI activity in the

male gamete, where the kr2.2 integration took place (Sheppard *et al.*, 2008), and the similarity in filler DNA observed between polymerase theta mediated DSB repair and the kr2.2 integration event, it would be interesting to investigate whether TEBIHCHI mediated DSB repair in the male germline is involved in organelle DNA insertion. This could be achieved by RNAi mediated knockdown of the tobacco homologue(s).

Sequence analysis of the kr2.2 integrant and numerous other studies (reviewed by Kleine *et al.*, 2009) suggest that DSB repair is involved in the nuclear insertion of organelle sequences. To further investigate the role of DSB repair in organelle DNA insertion, DNA DSBs were induced at a specific nuclear location using the I-SceI homing endonuclease and the subsequent repair events were analysed by smPCR. Insertions were observed in five of the repair events. None of the inserted sequences were of organelle origin suggesting that there was relatively little chloroplast or mitochondrial DNA, compared with TE DNA for example, available in the cells analysed. Those insertions whose origin could be traced appeared to be derived from TE sequence. Interestingly, the amount of nuclear sequence deleted was significantly larger in repair events involving insertion than in those without insertion, suggesting that the two types of repair may be mechanistically different. It was intended to induce DSBs in meiotic and postmeiotic cells where the natural insertion of ptDNA is already elevated but the 35S promoter was unsuitable. There are several promoters that would be suitable for this experiment but the length of time required to generate new transgenic lines precluded further studies.

In general it seems likely that organelle DNA is inserted into the nuclear genome through NHEJ mediated repair of DSBs. Recruitment of multiple loose DNA molecules to a single repair node may contribute to the complex sequence arrangements often seen in organelle DNA insertions, thereby resulting in the creation of new sequence arrangements within the nuclear genome.

6.3 Activation

Functional activation of a chloroplast gene transferred to the nucleus was also investigated. This involved screening for activation of the gene *aadA* which had been transferred from the chloroplast to a number of independent nuclear loci in a previous screen (Sheppard *et al.*, 2008).

Several activation events were observed, although none involved the acquisition of a native nuclear promoter. In both cases activation involved the nearby CaMV 35S promoter which was part of the experimental cassette. As this screen assessed activation of *aadA* in 16 independent genomic locations and screened a total of over 600 million cells, it is clear that activation involving native tobacco nuclear DNA is very rare – probably too rare for experimental simulation. The two activation events that did occur, sr1 and sr2, were the result of local sequence rearrangements

which, due to micro-homology observed at the new sequence junctions, were most likely caused by non-homologous repair of DSBs. DSB repair may therefore play an important role not only in the initial integration of organelle sequences into the nucleus but in subsequent re-arrangement of these sequences to bring about functional activation.

This screen also revealed that chloroplast sequences can themselves contain motifs that promote nuclear expression. The *psbA* promoter and terminator flanking the *aadA* coding sequence were found to promote nuclear transcription and mRNA cleavage and polyadenylation respectively. These sequences worked at low efficiency within the nucleus. However, in two lines where *aadA* was present in multiple copies, the combined *aadA* activity was sufficient to confer strong aminoglycoside resistance. As the plant polyadenylation consensus sequence and several motifs that regulate transcription, such as TATA and CAAT boxes are AT-rich, it may be that the AT-rich nature of plant non-coding regions contributes to the likelihood of finding such motifs by chance (Stegemann and Bock, 2006). Such chance motifs within chloroplast sequences are unlikely to lead to high levels of expression. However, weak nuclear activity, particularly if the gene is transferred in high copy number, may be enough to provide a "starting point" for gene transfer, enabling selective maintenance of the gene while transcription, polyadenylation and protein targeting are gradually strengthened under selection by post-insertional mutation.

This finding is also relevant to our understanding of the level of transgene containment provided by maternal inheritance of chloroplast located transgenes. Chloroplast transgenes have generally been considered inactive if transposed to the nucleus (Daniell and Parkinson, 2003). The observed nuclear activity of two commonly used chloroplast regulatory sequences highlights a previously unconsidered route for the escape of chloroplast transgenes.

6.4 The role of double strand break repair

DSB repair appears to play an important role both in the nuclear insertion of organelle sequences and the subsequent rearrangement of these transferred sequences to bring about functional activation. Given these dual roles, it is possible that DSB hot-spots are important areas for the integration and evolution of organelle sequences within the nuclear genome. Although by no means conclusive, several studies indicate that DSB hotspots may be associated with higher levels of organelle DNA insertion. In yeast *numts* have been found to preferentially locate to origins of replication, which the authors suggest is due to increased DSB formation in these areas (Lenglez *et al.*, 2010) in rice large *nupts* have been shown to preferentially locate to pericentromeres (Matsuo *et al.*, 2005) which contain a high density of TEs (Hall *et al.*, 2006) and are known to be DSB hotspots (Blitzblau *et al.*, 2007). However, it is possible this last point is due to the capacity for gene

poor pericentromeres to accommodate large sequence insertions and rearrangements with little fitness cost. More generally, organelle sequences throughout the genome are commonly associated with TEs (Mishmar *et al.*, 2004; Noutsos *et al.*, 2005) which, as a result of transposon excision, will likely result in relatively high rates of DSB formation and repair by NHEJ (Puchta, 2005).

A high rate of DSB formation and NHEJ repair may also lead to further organelle DNA insertions (Ricchetti *et al.*, 1999), deletions (chapter 5; Salomon and Puchta, 1998; Simsek and Jasin, 2010) and intra-chromosomal recombination (Xiao and Peterson, 2000). Accordingly, regions of the genome containing high TE density evolve rapidly (Hall *et al.*, 2006). This heightened rate of sequence shuffling may create novel sequence combinations which in some instances may lead to functional activation of a newly transferred chloroplast gene. The insertion and rearrangement of TEs containing RNA polymerase II promoters could potentially be important in this process (Kaessmann, 2010). DSBs can also lead to gene amplifications (Pipiras *et al.*, 1998) which, as observed in chapter 2, could be a potential stepping stone to nuclear activation of chloroplast genes. Non-homologous repair of DNA DSBs therefore appears to be a major generator of novel sequence diversity and central to endosymbiotic gene transfer.

6.5 A role for the male germline?

It seems that the male germline may be of particular importance in the process of endosymbiotic gene transfer. Transfer of chloroplast DNA to the nuclear genome is known to occur at a heightened rate in the male germline which has been attributed to the breakdown of the plastids as a result of their maternal inheritance (Sheppard *et al.*, 2008). In addition, pollen are exposed to a unique array of environmental factors likely to cause DNA damage (Jackson, 1987). As a result of their haploid nature, DNA DSBs in sperm cells must be repaired *via* the NHEJ pathway, leading to higher levels of intra-chromosomal homologous recombination and insertions or deletions. The male germline therefore may represent a major generator of *de novo* organelle DNA integrants and of sequence variation within these organelle sequences.

It has been proposed that in angiosperms, the male germline presents a unique opportunity for natural selection (Mulcahy, 1979; Lankinen and Armbruster, 2007). This is a result of competition between the growing pollen tubes to deliver the sperm cell, potentially allowing the elimination of metabolically weak gametophytes so that only relatively healthy genotypes progress to the next generation (Mulcahy, 1979). As pollen grains are haploid, recessive mutations would also be subject to this selection. At a population level this allows vast numbers of gametes to be screened and poorly functioning haploid genotypes to be eliminated from the population with very little

resource cost. In support of this, pollen competition has been shown to reduce inbreeding depression in *Collinsia heterophylla* (Lankinen and Armbruster, 2007), presumably by selecting against recessive mutations exposed due to the haploid nature of the pollen. Such a process is unlikely to positively select for the relocation of chloroplast genes. However, selection against deleterious mutation in pollen may permit a heightened rate of sequence insertion, deletion and other rearrangement without the danger of mutational overload. This may, indirectly, contribute to functional transfer of chloroplast genes to the nucleus in angiosperms.

6.6 Conclusion

Functional transfer of organelle genes to the nucleus is an important evolutionary process which has contributed greatly to the complexity of nuclear genomes. The non-homologous repair of DSBs in DNA appears to play a central role both in insertion of organelle genes and their subsequent nuclear activation. The complexities of the various non-homologous DNA repair pathways active in plant cells are, however, still largely unknown and therefore continued investigation, particularly with a focus on those pathways active in the germline, will be invaluable for developing an understanding of the forces shaping eukaryote nuclear genomes.

Unravelling this process will also greatly contribute in understanding gene transfer from more recently acquired endosymbionts such as bacterial endosymbionts of insects (Nikoh *et al.*, 2008) and the chromatophore of *Paulinella* which appears to be in the early stages of endosymbiotic gene transfer (Nowack *et al.*, 2008; Nowack *et al.*, 2010).

Appendix 1

Lines explant screened for spectinomycin and streptomycin resistance.

Line	explants screened	positive shoots
kr4[1]	1000	0
kr8[1]	1000	0
kr9[1]	1000	0
kr10[1]	1000	0
kr11[1]	1000	0
kr12[1]	1000	0
kr13[1]	1000	0
kr15[1]	1000	0
kr17[1, a]	1000	0
kr18[1]	1000	1
kr19[1, a]	1000	0
kr2.2[2]	1000	2
kr2.3[2]	1000	0
kr2.7[2, c]	1000	N/A
kr2.9[2, c]	1000	N/A
kr2.10[2, b]	1000	N/A

[1] Huang *et al.,* 2003

[2] Sheppard *et al.,* 2008

[a] Partial resistance low; many explants showed low levels of growth, although plants were sufficiently sensitive for further screening.

[b] Partial resistance mid; many explants showed low levels of growth. Partial resistance prevented further screening.

[c] All explants grew resistant shoots. Resistance prevented further screening.

Appendix 2

General Primers

Primer Name	Primer sequence (5´ → 3´)
35SR1	CCACTGACGTAAGGGATGAC
35SR1	CCACTGACGTAAGGGATGAC
35SR1_BglII	ATAAGATCTGTCCTCTCCAAATGAAATGAAC
35SR2	CGAGAGTGTCGTGCTCCACCATGTTGACCT
35SR2_BglII	ATAAGATCTATGGAATCCGAGGAGGTTTCCCGAT
aadAF1	CCAAGATTTTACCATGAGGGAAGCGGTG
aadAF2	GCCGAAGTATCGACTCAACTATCAGAGG
aadAF3	CAAGAGAACATAGCGTTGCCTTGG
aadAF4	GCCCGTCATACTTGAAGCTAGACAG
aadAF5	AGTATCGACTCAACTATCAGAGG
aadAR1	CGATGACGCCAACTACCTCTG
aadAR2	CTTCCCTCATGGTAAAATCTTGG
aadAR3	GACTACCTTGGTGATCTCGCCTTTC
aadAT1	TCCAAAAGGTCGTTGATCAAAGCTCGCC
aadAT2	GCGTTGTTTCATCAAGCCTTACGGTCAC
aadAT3	ACCAGCAAATCAATATCACTGTGTGGCTTCA
AD3	AGWGNAGWANCAWAGG
AlcF_NcoI	TTCCATGGGATAGTTCCGACCTAGGATGG
AlcR_NcoI	TTCCATGGGGCGATTAAGTTGGGTAACG
AlcRF1	CGTCGTTCTTATTCACTCGTTTGC
AlcRR1	TTGGAGGATGGGAAATGCGTTAG
cp140409R	CGTGGAAACGCATAAGAAC
CTTTF	GTCTCTTTAGCCCTCTGGTCTTCAAAG
CTTTR	TTTTCACACAAGGGTAATATCGG
dao1F	GAGAAAGGAAGGGAAGAAAGC
dao1F2	GGCAAACCGTCCTCGTCAAG
dao1R	ACTCTAGACCTACAACTTCGACTCCCG
dao1R2	TGACCTCCTTCTCCTTCGCC
DSBF1	GATAGTGACCTTAGGCGACTTTTGAACG
DSBR1	TCCCCTGATTCTGTGGATAACCGT
I-SceIF1	ACAAACTGGCTAACCTGTTCATCGT
I-SceIMCS1	CTAGGGATAACAGGGTAATAAGCTTGCGGCCGCTAGGGATAACAGGGTAATC
I-SceIMCS2	TCGAGATTACCCTGTTATCCCTAGCGGCCGCAAGCTTATTACCCTGTTATCCCTAGAGCT
I-SceIR1	TTCGGAGGAGATAGTGTTCGGCA
kr2.2AJR1	CTTGTGCTTGTTGGAGATAACTGGTG
kr2.2AJR2	CGGTGGGTTGGTTGAAATCA
kr2.2NJR1	TAGCCACCTTTCCTGTGTTATG
kr2.2NJR2	GCTCTGTTTTATGATGGCTGATGTTTC
L25F	AAAATCTGACCCCAAGGCAC
L25R	GCTTTCTTCGTCCCATCAGG
neoF1	GCGTTGGCTACCCGTGATATTGCTG

General Primers (cont.)

neoF2	CCTTCTATCGCCTTCTTGAC
neoF3	GGGACCCGCAAAAATCACCAGTCTCTCTC
neoR1	CGCCAAGCTCTTCAGCAATATCACG
neoR2	CCAATAGCAGCCAGTCCCTTC
neoR3	TCATAGCCGAATAGCCTCTCCACCCAAG
psbAR_NcoI	ATACCATGGTAAAATCTTGGTTTATTTAATC
QaadAF	TAACGCTATGGAACTCGCCG
QaadAR	CGGCGAGTTCCATAGCGTTA
QL25F	CCCCTCACCACAGAGTCTGC
QL25R	AAGGGTGTTGTTGTCCTCAATCTT
SP6	ATTTAGGTGACACTATA
T7	TAATACGACTCACTATAG

Restriction sites included in primers are underlined

Appendix 3

kr2.2 sequencing primers

Primer Name	Primer sequence (5´ → 3´)
kr2.2NJR1	TAGCCACCTTTCCTGTGTTATG
cp146805F	GTTTCCACGACTCTACCACCCAG
cp146406F	CAACAGGAATCACACAAGGGTAG
cp145892F	GGGACCCTATTCACCTCTTT
cp145777F	GTCATTCGGGCCTATCTAAA
cp145418F	GACTGGACGAAACCAAGAAA
cp145068F	GTTGGAGACTCAAATGATGGA
cp1443885F	AAAGGATCGGACTAATGACG
cp143890F	ACGGTCTAATGAGGCTACTATG
cp143691F	CAACTCTATGTATTCCTCTATCC
cp143216F	CTCTCGGGGACTTTCTTAAA
cp143049F	GAATGGCAGAGGCAAATAGAGC
cp142466	GAATGGCAGAGGCAAATAGAGC
cp141966F	AGATGCTGTCGGAGTAAAGG
cp141383F	GAAGTAGAAGAACCCAGATTCC
cp140986F	GACCCACCAAAGTACGAAAG
ORF131F	CCCCTATCGGAAATAGGATT
neoF1	GCGTTGGCTACCCGTGATATTGCTG
neoF4	ATTCGGCTATGACTGGGCACAACAG
neoR2	CCAATAGCAGCCAGTCCCTTC
35SF1	CCACTGACGTAAGGGATGAC
35SF2	AACATGGTGGAGCACGACACTC
aadAR1	CGATGACGCCAACTACCTCTG
aadAF1	CCAAGATTTTACCATGAGGGAAGCGGTG
aadAF3	CAAGAGAACATAGCGTTGCCTTGG
aadAF4	GCCCGTCATACTTGAAGCTAGACAG
cp140202R	GGATTAGTCAGTTCTATTTCTCG
cp140111F	TCGAACTGATGACTTCCACCAC
cp139586R	CAGCCACACTGGGACTGAGACACG
cp139023R	CGCAAGAATGAAACTCAAAGG
cp138431R	GTGAAGTCGTAACAAGGTAGC
cp137795R	ACGGCTCCTCTCTTCTCG
cp137188R	TTGCGATTACGGGTTGGATG
cp136688R	TCGGAGAAGGGCAATGACT
cp136231R	GAAATGCTTCGGGGAGTTGA
cp135661R	AGCGAAAGCGAGTCTTCATAGG
cp48556F	ACAAATGCGATGCTCTAACCTCTG
cp49175F	ATCAAAGAGGAGAAAACACG
cp49774R	CGAGAATAAAGATAGAGTCCCGT
kr2.2AJR2	CGGTGGGTTGGTTGAAATCA

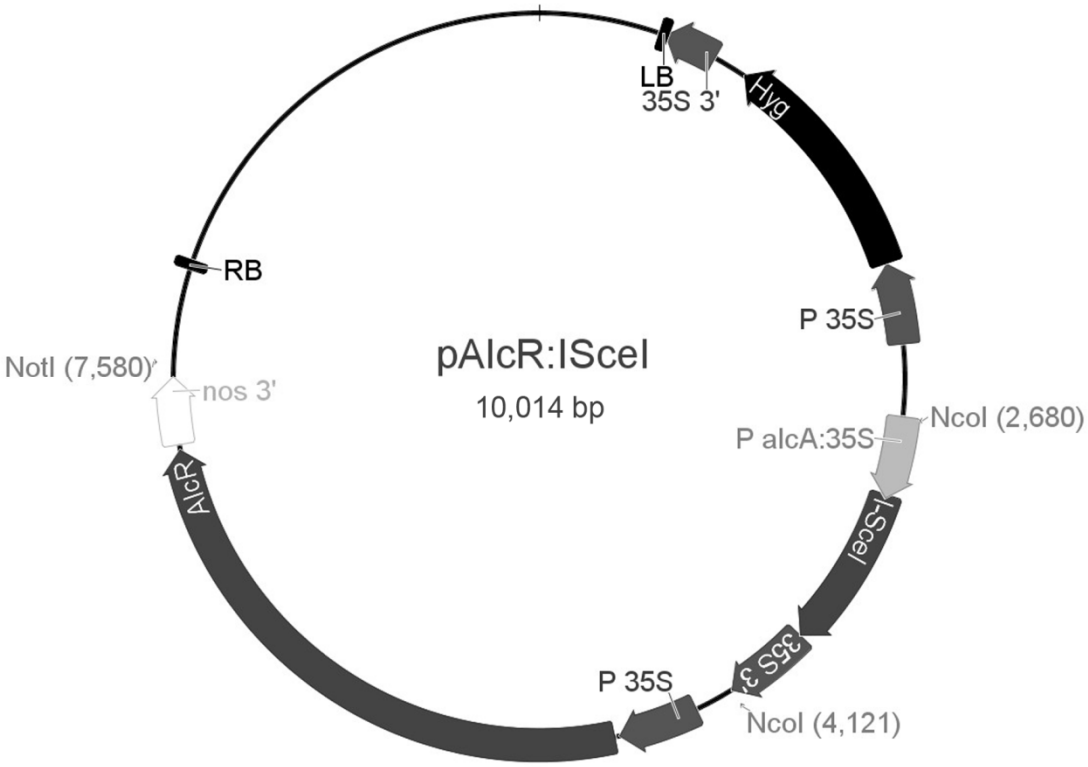
Appendix 4

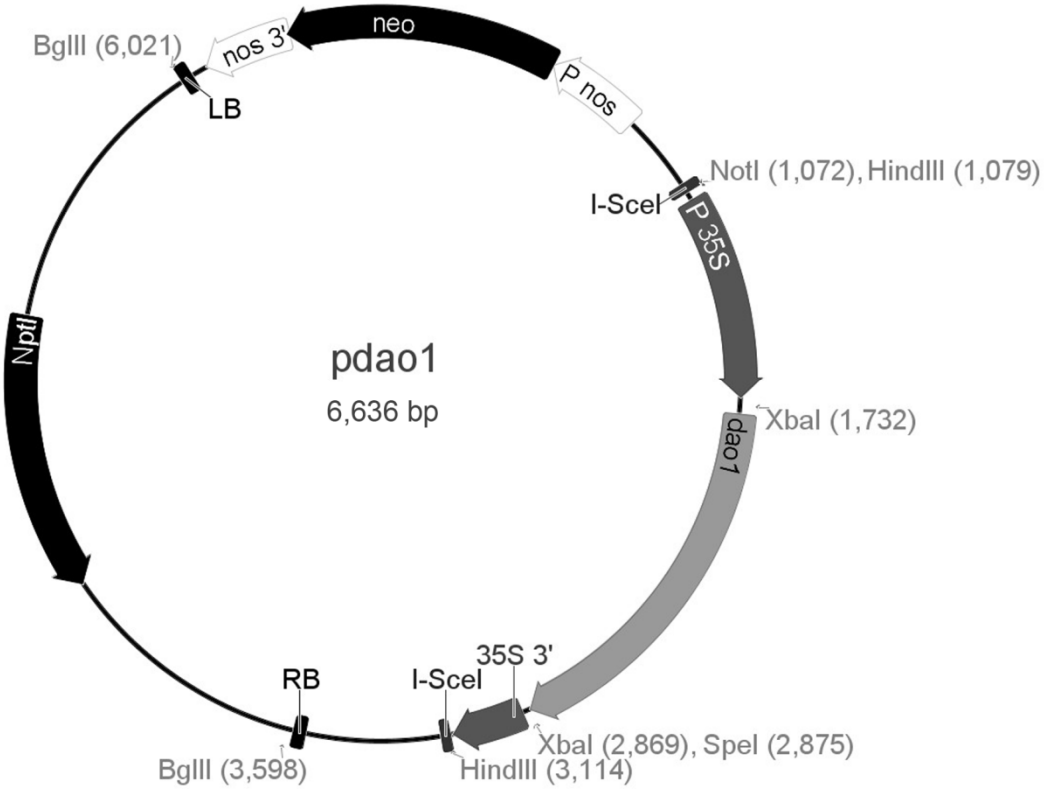
Segregation ratio in T_1 progeny of primary transformants

Line	Resistant	Sensitive	% Resistant	P value
A1[1]	45	3	94%[1]	2.70E-03
A2	42	18	70%	0.37
A3	51	26	66%	0.08
D4	73	28	72%	0.53
D10[1]	111	1	99%[1]	3.82E-09
D11	53	18	75%	0.95
D12[1]	113	8	93%[1]	2.99E-06
D13	56	24	70%	0.30
D14	73	23	76%	0.81
G1	79	32	71%	0.35
G4	88	29	75%	0.96
G8	72	26	73%	0.73
G9	80	30	73%	0.58
G10	80	24	77%	0.65
G11[1]	79	36	58%[1]	2.2 E-4

[1] Indicates a significant deviation from the expected 3:1 segregation ratio

Appendix 5


LB
35S 3'
Hyg
P 35S
NcoI (2,680)
P alcA:35S
I-SceI
35S 3'
P 35S
NcoI (4,121)
AlcR
NotI (7,580)
nos 3'
RB
pAlcR:ISceI
10,014 bp


BglII (6,021)
nos 3'
neo
P nos
NotI (1,072), HindIII (1,079)
I-SceI
P 35S
XbaI (1,732)
dao1
35S 3'
I-SceI
RB
XbaI (2,869), SpeI (2,875)
HindIII (3,114)
BglII (3,598)
LB
NptII
pdao1
6,636 bp

References

Adams, K. L., D. O. Daley, Y. L. Qiu, J. Whelan and J. D. Palmer (2000). "Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants." Nature **408**(6810): 354-357.

Allen, J. F. (2003). "The function of genomes in bioenergetic organelles." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **358**(1429): 19-37.

Allen, J. F. and J. A. Raven (1996). "Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles." Journal of Molecular Evolution **42**(5): 482-492.

Alverson, A. J., X. X. Wei, D. W. Rice, D. B. Stern, K. Barry and J. D. Palmer (2010). "Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae)." Molecular Biology and Evolution **27**(6): 1436-1448.

An, G., P. Ebert, A. Mitra and S. Ha (1988). Binary vectors. Plant molecular biology manual. S. Gelvin and R. Schilperoort, Kluwer Academic Publishers**:** A3:1-19.

Anbar, A. D. and A. H. Knoll (2002). "Proterozoic ocean chemistry and evolution: A bioinorganic bridge?" Science **297**(5584): 1137-1142.

Arthofer, W., S. Schuler, F. M. Steiner and B. C. Schlick-Steiner (2010). "Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence." Molecular Ecology **19**(18): 3853-3856.

Ayliffe, M. A., N. S. Scott and J. N. Timmis (1998). "Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants." Molecular Biology and Evolution **15**(6): 738-745.

Barbrook, A. C., C. J. Howe, D. P. Kurniawan and S. J. Tarr (2010). "Organization and expression of organellar genomes." Philosophical Transactions of the Royal Society B-Biological Sciences **365**(1541): 785-797.

Barbrook, A. C., C. J. Howe and S. Purton (2006). "Why are plastid genomes retained in non-photosynthetic organisms?" Trends in Plant Science **11**(2): 101-108.

Behura, S. K. (2007). "Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome." Molecular Biology and Evolution **24**(7): 1492-1505.

Ben Yehezkel, T., G. Linshiz, H. Buaron, S. Kaplan, U. Shabi and E. Shapiro (2008). "De novo DNA synthesis using single molecule PCR." Nucleic Acids Research **36**(17): Article No.: e107.

Berg, O. G. and C. G. Kurland (2000). "Why mitochondrial genes are most often found in nuclei." Mol Biol Evol **17**(6): 951-961.

Birky, C. W. (2001). "The inheritance of genes in mitochondria and chloroplasts: Laws, mechanisms, and models." Annual Review of Genetics **35**: 125-148.

Blanchard, J. L. and M. Lynch (2000). "Organellar genes - why do they end up in the nucleus?" Trends in Genetics **16**(7): 315-320.

Blitzblau, H. G., G. W. Bell, J. Rodriguez, S. P. Bell and A. Hochwagen (2007). "Mapping of meiotic single-stranded DNA reveals double-strand-break hotspots near centromeres and telomeres." Current Biology **17**(23): 2003-2012.

Bock, R. and J. N. Timmis (2008). "Reconstructing evolution: gene transfer from plastids to the nucleus." Bioessays **30**(6): 556-566.

Borges, F., G. Gomes, R. Gardner, N. Moreno, S. McCormick, J. A. Feijo and J. D. Becker (2008). "Comparative transcriptomics of Arabidopsis sperm cells." Plant Physiology **148**(2): 1168-1181.

Boxma, B., R. M. de Graaf, G. W. van der Staay, T. A. van Alen, G. Ricard, T. Gabaldon, A. H. van Hoek, S. Y. Moon-van der Staay, W. J. Koopman, J. J. van Hellemond, *et al.* (2005). "An anaerobic mitochondrion that produces hydrogen." Nature **434**(7029): 74-79.

Brandvain, Y., M. S. Barker and M. J. Wade (2007). "Gene co-inheritance and gene transfer." Science **315**(5819): 1685-1685.

Brouard, J. S., C. Otis, C. Lemieux and M. Turmel (2010). "The Exceptionally Large Chloroplast Genome of the Green Alga *Floydiella terrestris* Illuminates the Evolutionary History of the Chlorophyceae." Genome Biology and Evolution **2**: 240-256.

Butterfield, N. J. (2000). "*Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes." Paleobiology **26**(3): 386-404.

Caddick, M. X., A. J. Greenland, I. Jepson, K. P. Krause, N. Qu, K. V. Riddell, M. G. Salter, W. Schuch, U. Sonnewald and A. B. Tomsett (1998). "An ethanol inducible gene switch for plants used to manipulate carbon metabolism." Nature Biotechnology **16**(2): 177-180.

Cavalier-Smith, T. (2009). "Predation and eukaryote cell origins: a coevolutionary perspective." Int J Biochem Cell Biol **41**(2): 307-322.

Cavalier-Smith, T. and J. J. Lee (1985). "Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles." Journal of Protozoology **32**(3): 376-379.

Chan, S. H., A. M. Yu and M. McVey (2010). "Dual Roles for DNA Polymerase Theta in Alternative End-Joining Repair of Double-Strand Breaks in Drosophila." Plos Genetics **6**(7).

Cheung, A. Y., L. Bogorad, M. Vanmontagu and J. Schell (1988). "Relocating a gene for herbicide tolerance - a chloroplast gene is converted into a nuclear gene." Proceedings of the National Academy of Sciences of the United States of America **85**(2): 391-395.

Chhibber, A. and B. G. Schroeder (2008). "Single-molecule polymerase chain reaction reduces bias: Application to DNA methylation analysis by bisulfite sequencing." Analytical Biochemistry **377**(1): 46-54.

Chittela, R. K. and J. K. Sainis (2010). "Plant DNA recombinases: a long way to go." J Nucleic Acids **2010**.

Clemens, D. L. and P. J. Johnson (2000). "Failure to detect DNA in hydrogenosomes of *Trichomonas vaginalis* by nick translation and immunomicroscopy." Molecular and Biochemical Parasitology **106**(2): 307-313.

Cornelissen, M. and M. Vandewiele (1989). "Nuclear transcriptional activity of the plastid *psbA* promoter." Nucleic Acids Res. **17**(1): 19-29.

Cusack, B. P. and K. H. Wolfe (2007). "When gene marriages don't work out: divorce by subfunctionalization." Trends in Genetics **23**(6): 270-272.

Daley, D. O. and J. Whelan (2005). "Why genes persist in organelle genomes." Genome Biology **6**(5).

Daniell, H. and C. L. Parkinson (2003). "Jumping genes and containment." Nature Biotechnology **21**(4): 374-375.

De Buck, S., N. Podevin, J. Nolf, A. Jacobs and A. Depicker (2009). "The T-DNA integration pattern in Arabidopsis transformants is highly determined by the transformed target cell." Plant Journal **60**(1): 134-145.

Delaney, S. K., S. J. Orford, M. Martin-Harris and J. N. Timmis (2007). "The fiber specificity of the cotton *FSltp4* gene promoter is regulated by an AT-rich promoter region and the AT-hook transcription factor GhAT1." Plant Cell Physiol. **48**(10): 1426-1437.

Deusch, O., G. Landan, M. Roettger, N. Gruenheit, K. V. Kowallik, J. F. Allen, W. Martin and T. Dagan (2008). "Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor." Molecular Biology and Evolution **25**(4): 748-761.

Doolittle, W. E. (1998). "You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes." Trends in Genetics **14**(8): 307-311.

Downs, J. A. and S. P. Jackson (2004). "A means to a DNA end: the many roles of Ku." Nature Reviews Molecular Cell Biology **5**(5): 367-378.

Embley, T. M. and W. Martin (2006). "Eukaryotic evolution, changes and challenges." Nature **440**(7084): 623-630.

Erikson, O., M. Hertzberg and T. Nasholm (2004). "A conditional marker gene allowing both positive and negative selection in plants." Nature Biotechnology **22**(4): 455-458.

Esser, C., N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, *et al.* (2004). "A genome phylogeny for mitochondria

among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes." Molecular Biology and Evolution **21**(9): 1643-1660.

Etzold, T., C. C. Fritz, J. Schell and P. H. Schreier (1987). "A point mutation in the chloroplast 16 S rRNA gene of a streptomycin resistant *Nicotiana tabacum*." Febs Letters **219**(2): 343-346.

Falcon, L. I., S. Magallon and A. Castillo (2010). "Dating the cyanobacterial ancestor of the chloroplast." Isme Journal **4**(6): 777-783.

Fang, R. X., F. Nagy, S. Sivasubramaniam and N. H. Chua (1989). "Multiple *cis* Regulatory Elements for Maximal Expression of the Cauliflower Mosaic Virus 35S Promoter in Transgenic Plants." Plant Cell **1**(1): 141-150.

Fattah, F., E. H. Lee, N. Weisensel, Y. B. Wang, N. Lichter and E. A. Hendrickson (2010). "Ku Regulates the Non-Homologous End Joining Pathway Choice of DNA Double-Strand Break Repair in Human Somatic Cells." Plos Genetics **6**(2): 14.

Figueroa, P., I. Gomez, L. Holuigue, A. Araya and X. Jordana (1999). "Transfer of *rps14* from the mitochondrion to the nucleus in maize implied integration within a gene encoding the iran-sulphur subunit of succinate dehydrogenase and expression by alternative splicing." Plant Journal **18**(6): 601-609.

Friesner, J. and A. B. Britt (2003). "Ku80- and DNA ligase IV-deficient plants are sensitive to ionizing radiation and defective in T-DNA integration." Plant Journal **34**(4): 427-440.

Gallego, M. E., J. Y. Bleuyard, S. Daoudal-Cotterell, N. Jallut and C. I. White (2003). "Ku80 plays a role in non-homologous recombination but is not required for T-DNA integration in *Arabidopsis*." Plant Journal **35**(5): 557-565.

Gaspar, T., C. Kevers, P. Debergh, L. Maene, M. Paques and P. Boxus (1987). Vitrification: morphological, physiological and ecological aspects. Cell and Tissue Culture in Forestry. J. M. Bonga and D. J. Durzan. Dordrecht, Kluwer Academic Press. **1:** 152-166.

Gissi, C., F. Iannelli and G. Pesole (2008). "Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species." Heredity **101**(4): 301-320.

Gorbunova, V. and A. A. Levy (1997). "Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions." Nucleic Acids Research **25**(22): 4650-4657.

Grawunder, U., M. Wilm, X. T. Wu, P. Kulesza, T. E. Wilson, M. Mann and M. R. Lieber (1997). "Activity of DNA ligase IV stimulated by complex formation with XRCC4 protein in mammalian cells." Nature **388**(6641): 492-495.

Gray, M. W. and W. F. Doolittle (1982). "Has the Endosymbiont Hypothesis Been Proven." Microbiological Reviews **46**(1): 1-42.

Grohmann, L., A. Brennicke and W. Schuster (1992). "The mitochondrial gene encoding ribosomal protein S12 has been translocated to the nuclear genome in *Oenothera*." Nucleic Acids Research **20**(21): 5641-5646.

Gross, J. and D. Bhattacharya (2010). "Uniting sex and eukaryote origin in an emerging oxygenic world." Biol Direct **5**(1): 53.

Guo, X. Y., S. L. Ruan, W. M. Hu, D. G. Ca and L. J. Fan (2008). "Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved." Functional & Integrative Genomics **8**(2): 101-108.

Hall, A. E., G. C. Kettler and D. Preuss (2006). "Dynamic evolution at pericentromeres." Genome Research **16**(3): 355-364.

Hannam, R. V. (1968). "Leaf growth and development in young tobacco plant." Aust. J Biol. Sci. **21**(5): 855-870.

Hattasch, C., H. Flachowsky and M. V. Hanke (2009). "Evaluation of an alternative D-amino acid/DAAO selection system for transformation in apple (*Malus* X *domestica* Borkh.)." Journal of Horticultural Science & Biotechnology: 188-194.

Haviv-Chesner, A., Y. Kobayashi, A. Gabriel and M. Kupiec (2007). "Capture of linear fragments at a double-strand break in yeast." Nucleic Acids Research **35**(15): 5192-5202.

Hazkani-Covo, E., R. M. Zeller and W. Martin (2010). "Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes." Plos Genetics **6**(2).

Hellens, R. P., E. A. Edwards, N. R. Leyland, S. Bean and P. M. Mullineaux (2000). "pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation." Plant Mol. Biol. **42**(6): 819-832.

Henze, K. and W. Martin (2001). "How do mitochondrial genes get into the nucleus?" Trends in Genetics **17**(7): 383-387.

Heyer, W.-D., K. T. Ehmsen and J. Liu (2010). "Regulation of homologous recombination in eukaryotes." Annu Rev Genet **44**: 113-139.

Howe, C. J., R. E. R. Nisbet and A. C. Barbrook (2008). "The remarkable chloroplast genome of dinoflagellates." Journal of Experimental Botany **59**(5): 1035-1045.

Huang, C. Y., M. A. Ayliffe and J. N. Timmis (2003). "Direct measurement of the transfer rate of chloroplast DNA into the nucleus." Nature **422**(6927): 72-76.

Huang, C. Y., M. A. Ayliffe and J. N. Timmis (2004). "Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco." Proceedings of the National Academy of Sciences of the United States of America **101**(26): 9710-9715.

Huang, C. Y., N. Grunheit, N. Ahmadinejad, J. N. Timmis and W. Martin (2005). "Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes." Plant Physiol. **138**(3): 1723-1733.

Hudson, L. C. and C. N. Stewart (2004). "Effects of pollen-synthesized green fluorescent protein on pollen grain fitness." Sexual Plant Reproduction **17**(1): 49-53.

Humphries, E. C. and E. C. Wheeler (1960). "The Effects of Kinetin, Gibberellic Acid, and Light on Expansion and Cell Division in Leaf Disks of Dwarf Bean (*Phaseolus vulgaris*)." J. Exp. Bot. **11**(1): 81-85.

Inagaki, S., T. Suzuki, M. Ohto, H. Urawa, T. Horiuchi, K. Nakamura and A. Morikami (2006). "*Arabidopsis* TEBICHI, with helicase and DNA polymerase domains, is required for regulated cell division and differentiation in meristems." Plant Cell **18**(4): 879-892.

Jackson, J. F. (1987). "DNA repair in pollen. A review." Mutation Research **181**(1): 17-29.

Kaessmann, H. (2010). "Origins, evolution, and phenotypic impact of new genes." Genome Res **20**(10): 1313-1326.

Kirik, A., S. Salomon and H. Puchta (2000). "Species-specific double-strand break repair and genome evolution in plants." Embo Journal **19**(20): 5562-5566.

Kirschman, J. A. and J. H. Cramer (1988). "Two new tools: multi-purpose cloning vectors that carry kanamycin or spectinomycin/streptomycin resistance markers." Gene **68**(1): 163-165.

Kleffmann, T., D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem and S. Baginsky (2004). "The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions." Curr Biol **14**(5): 354-362.

Kleine, T., U. G. Maier and D. Leister (2009). "DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis." Annu. Rev. Plant Biol. **60**: 115-138.

Kneip, C., P. Lockhart, C. Voss and U. G. Maier (2007). "Nitrogen fixation in eukaryotes - New models for symbiosis." Bmc Evolutionary Biology **7**.

Kraytsberg, Y., N. Bodyak, S. Myerow, A. Nicholas, K. Ebralidze and K. Khrapko (2009). "Quantitative Analysis of Somatic Mitochondrial DNA Mutations by Single-Cell Single-Molecule PCR." Mitochondrial DNA: Methods and Protocols: 329-369.

Kraytsberg, Y. and K. Khrapko (2005). "Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations." Expert Review of Molecular Diagnostics **5**(5): 809-815.

Kumar, S. and M. Fladung (2002). "Transgene integration in aspen: structures of integration sites and mechanism of T-DNA integration." Plant J **31**(4): 543-551.

Kundu, M. and C. B. Thompson (2005). "Macroautophagy versus mitochondrial autophagy: a question of fate?" Cell Death and Differentiation **12**: 1484-1489.

Lai, F. M., K. F. Mei, L. Mankin and T. Jones (2007). Application of two new selectable marker genes, *dsdA* and *dao1* in maize transformation. Biotechnology and Sustainable Agriculture 2006 and Beyond: Proceedings of the 11th IAPTC&B Congress. Z. Xu, J. Li, Y. Xue and W. Yang. Dordrecht, Springer**:** 141-142.

Lang, B. F., G. Burger, C. J. Okelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel and M. W. Gray (1997). "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature." Nature **387**(6632): 493-497.

Lankinen, A. and W. S. Armbruster (2007). "Pollen competition reduces inbreeding depression in *Collinsia heterophylla* (Plantaginaceae)." Journal of Evolutionary Biology **20**(2): 737-749.

Leister, D. (2005). "Origin, evolution and genetic effects of nuclear insertions of organelle DNA." Trends in Genetics **21**(12): 655-663.

Lenglez, S., D. Hermand and A. Decottignies (2010). "Genome-wide mapping of nuclear mitochondrial DNA sequences links DNA replication origins to chromosomal double-strand break formation in *Schizosaccharomyces pombe*." Genome Research **20**(9): 1250-1261.

Li, H. M. and C. C. Chiu (2010). "Protein transport into chloroplasts." Annu Rev Plant Biol **61**: 157-180.

Li, J. X., M. Vaidya, C. White, A. Vainstein, V. Citovsky and T. Tzfira (2005). "Involvement of KU80 in T-DNA integration in plant cells." Proceedings of the National Academy of Sciences of the United States of America **102**(52): 19231-19236.

Li, Q. S. and A. G. Hunt (1997). "The polyadenylation of RNA in plants." Plant Physiol. **115**(2): 321-325.

Lieber, M. R. (2010). "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway." Annual Review of Biochemistry, Vol 79 **79**: 181-211.

Lin, X. Y., S. S. Kaul, S. Rounsley, T. P. Shea, M. I. Benito, C. D. Town, C. Y. Fujii, T. Mason, C. L. Bowman, M. Barnstead*, et al.* (1999). "Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*." Nature **402**(6763): 761-+.

Lin, Y. F. and A. S. Waldman (2001). "Capture of DNA sequences at double-strand breaks in mammalian chromosomes." Genetics **158**(4): 1665-1674.

Lister, D. L., J. M. Bateman, S. Purton and C. J. Howe (2003). "DNA transfer from chloroplast to nucleus is much rarer in *Chlamydomonas* than in tobacco." Gene **316**: 33-38.

Liu, S. L., Y. Zhuang, P. Zhang and K. L. Adams (2009). "Comparative Analysis of Structural Diversity and Sequence Evolution in Plant Mitochondrial Genes Transferred to the Nucleus." Molecular Biology and Evolution **26**(4): 875-891.

Liu, Y. G., N. Mitsukawa, T. Oosumi and R. F. Whittier (1995). "Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR." Plant J. **8**(3): 457-463.

Lockington, R. A., H. M. Sealylewis, C. Scazzocchio and R. W. Davies (1985). "Cloning and characterization of the ethanol utilization regulon in *Aspergillus nudulans*." Gene **33**(2): 137-149.

Lough, A. N., L. M. Roark, A. Kato, T. S. Ream, J. C. Lamb, J. A. Birchler and K. J. Newton (2008). "Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize." Genetics **178**(1): 47-55.

Lynch, M. (1996). "Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes." Molecular Biology and Evolution **13**(1): 209-220.

Maliga, P. (2003). "Plant biology - Mobile plastid genes." Nature **422**(6927): 31-32.

Marechal, A. and N. Brisson (2010). "Recombination and the maintenance of plant organelle genome stability." New Phytologist **186**(2): 299-317.

Margulis, L. (1970). Origin of Eukaryotic Cells. New Haven and London, Yale University Press.

Martin, W. and R. G. Herrmann (1998). "Gene transfer from organelles to the nucleus: How much, what happens, and why?" Plant Physiology **118**(1): 9-17.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa and D. Penny (2002). "Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus." Proc. Natl. Acad. Sci. U.S.A. **99**(19): 12246-12251.

Mascarenhas, J. P. and D. A. Hamilton (1992). "Artifacts in the localization of GUS activity in anthers of petunia transformed with a CaMV-35S GUS construct." Plant Journal **2**(3): 405-408.

Mathis, N. L. and A. W. Hinchee (1994). *Agrobacterium* inoculation techniques for plant tissues. Plant Molecular Biology Manual. S. B. Gelvin and R. A. Schilperoort. Dordrecht, The Netherlands, Kluwer Academic Publishers**:** 1-9.

Matsuo, M., Y. Ito, R. Yamauchi and J. Obokata (2005). "The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux." Plant Cell **17**(3): 665-675.

Mazon, G., E. P. Mimitou and L. S. Symington (2010). "SnapShot: Homologous Recombination in DNA Double-Strand Break Repair." Cell **142**(4).

McVey, M. and S. E. Lee (2008). "MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings." Trends in Genetics **24**(11): 529-538.

Mereschkowsky, C. (1905). "Uber Natur und Ursprung der Chromatophoren im Pflanzenreiche." Biol Centrabl. **25**: 593-604.

Miki, B. and S. McHugh (2004). "Selectable marker genes in transgenic plants: applications, alternatives and biosafety." Journal of Biotechnology **107**(3): 193-232.

Millar, A. H., J. Whelan and I. Small (2006). "Recent surprises in protein targeting to mitochondria and plastids." Current Opinion in Plant Biology **9**(6): 610-615.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Giray, C. W. Morden, P. J. Calie, L. S. Jermiin and K. H. Wolfe (2001). "Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus." Plant Cell **13**(3): 645-658.

Ming, R., S. B. Hou, Y. Feng, Q. Y. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, W. Wang, B. V. Ly, K. L. T. Lewis*, et al.* (2008). "The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)." Nature **452**(7190): 991-U997.

Mishmar, D., E. Ruiz-pesini, M. Brandon and D. C. Wallace (2004). "Mitochondrial DNA-like sequences in the nucleus (NUMTs): Insights into our African origins and the mechanism of foreign DNA integration." Human Mutation **23**(2): 125-133.

Moore, J. K. and J. E. Haber (1996). "Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks." Nature **383**(6601): 644-646.

Mulcahy, D. L. (1979). "Rise of the Angiosperms: A Genecological Factor." Science **206**(4414): 20-23.

Murashige, T. and F. Skoog (1962). "A revised medium for rapid growth and bio assays with tobacco tissue cultures." Physiol. Plantarum **15**(3): 473-&.

NCBI. (2010, September 2010). "NCBI Eukaryotae Organelles List." Retrieved 13th December, 2010, from http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=2759&hopt=html.

Neupert, W. (1997). "Protein import into mitochondria." Annual Review of Biochemistry **66**: 863-917.

Nikoh, N., K. Tanaka, F. Shibata, N. Kondo, M. Hizume, M. Shimada and T. Fukatsu (2008). "*Wolbachia* genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes." Genome Research **18**(2): 272-280.

Ninomiya, Y., K. Suzuki, C. Ishii and H. Inoue (2004). "Highly efficient gene replacements in *Neurospora* strains deficient for nonhomologous end-joining." Proceedings of the National Academy of Sciences of the United States of America **101**(33): 12248-12253.

Noutsos, C., T. Kleine, U. Armbruster, G. DalCorso and D. Leister (2007). "Nuclear insertions of organellar DNA can create novel patches of functional exon sequences." Trends Genet. **23**(12): 597-601.

Noutsos, C., E. Richly and D. Leister (2005). "Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants." Genome Research **15**(5): 616-628.

Nowack, E. C. M., M. Melkonian and G. Glockner (2008). "Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes." Current Biology **18**(6): 410-418.

Nowack, E. C. M., H. Vogel, M. Groth, A. R. Grossman, M. Melkonian and G. Glöckner (2010). "Endosymbiotic Gene Transfer and Transcriptional Regulation of Transferred Genes in *Paulinella chromatophora*." Molecular Biology and Evolution.

Nugent, J. M. and J. D. Palmer (1991). "Rna-Mediated Transfer of the Gene CoxII from the Mitochondrion to the Nucleus During Flowering Plant Evolution." Cell **66**(3): 473-481.

Ochman, H., A. S. Gerber and D. L. Hartl (1988). "Genetic Applications of an Inverse Polymerase Chain Reaction." Genetics **120**(3): 621-623.

Ohshima, K. and N. Okada (1994). "Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of 3 different tRNA-derived retrotransposons in the octopus." Journal of Molecular Biology **243**(1): 25-37.

Pipiras, E., A. Coquelle, A. Bieth and M. Debatisse (1998). "Interstitial deletions and intrachromosomal amplification initiated from a double-strand break targeted to a mammalian chromosome." Embo Journal **17**(1): 325-333.

Pollegioni, L., L. Piubelli, S. Sacchi, M. S. Pilone and G. Molla (2007). "Physiological functions of D-amino acid oxidases: from yeast to humans." Cellular and Molecular Life Sciences **64**(11): 1373-1394.

Puchta, H. (2005). "The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution." J. Exp. Biol. **56**(409): 1-14.

Puchta, H., B. Dujon and B. Hohn (1996). "Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination." Proceedings of the National Academy of Sciences of the United States of America **93**(10): 5055-5060.

Puthiyaveetil, S., T. A. Kavanagh, P. Cain, J. A. Sullivan, C. A. Newell, J. C. Gray, C. Robinson, M. van der Giezen, M. B. Rogers and J. F. Allen (2008). "The ancestral symbiont sensor kinase CSK links photosynthesis with gene expression in chloroplasts." Proceedings of the National Academy of Sciences of the United States of America **105**(29): 10061-10066.

Reith, M. and J. Munholland (1995). "Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome." Plant Molecular Biology Reporter **13**(4): 333-335.

Ricchetti, M., C. Fairhead and B. Dujon (1999). "Mitochondrial DNA repairs double-strand breaks in yeast chromosomes." Nature **402**(6757): 96-100.

Richly, E. and D. Leister (2004a). "NUMTs in sequenced eukaryotic genomes." Molecular Biology and Evolution **21**(6): 1081-1084.

Richly, E. and D. Leister (2004b). "NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs." Mol. Biol. Evol. **21**(10): 1972-1980.

Roark, L. M., Y. Hui, L. Donnelly, J. A. Birchler and K. J. Newton (2010). "Recent and Frequent Insertions of Chloroplast DNA into Maize Nuclear Chromosomes." Cytogenetic and Genome Research **129**(1-3): 17-23.

Rousseau-Gueutin, M., A. H. Lloyd, A. E. Sheppard and J. N. Timmis (In Press). Gene transfer to the nucleus. Organelle Genetics: Evolution of Organelle Genomes and Gene Expression. C. Bullerwell.

Ruf, S., S. Braune, P. Endries, C. Hasse, S. Stegemann and R. Bock (2010). Plastid transmission, gene transfer and the impact of the environment. ISCGGE. Maynooth.

Ruf, S., D. Karcher and R. Bock (2007). "Determining the transgene containment level provided by chloroplast transformation." Proc. Natl. Acad. Sci. U.S.A. **104**(17): 6998-7002.

Salomon, S. and H. Puchta (1998). "Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells." EMBO J. **17**(20): 6086-6095.

Salter, M. G., J. A. Paine, K. V. Riddell, I. Jepson, A. J. Greenland, M. X. Caddick and A. B. Tomsett (1998). "Characterisation of the ethanol-inducible alc gene expression system for transgenic plants." Plant Journal **16**(1): 127-132.

Schaefer, D. G. (2002). "A new moss genetics: targeted mutagenesis in *Physcomitrella patens*." Annual Review of Plant Biology **53**: 477-501.

Schaefer, D. G. and J. P. Zryd (1997). "Efficient gene targeting in the moss *Physcomitrella patens*." Plant Journal **11**(6): 1195-1206.

Schmidt, G. W. and S. K. Delaney (2010). "Stable internal reference genes for normalization of real-time RT-PCR in tobacco (*Nicotiana tabacum*) during development and abiotic stress." Mol Genet Genomics **283**(3): 233-241.

Selosse, M., B. Albert, B. Godelle, O. G. Berg and C. G. Kurland (2001). "Reducing the genome size of organelles favours gene transfer to the nucleus

Why mitochondrial genes are most often found in nuclei." Trends Ecol Evol **16**(3): 135-141.

Sessions, A., E. Burke, G. Presting, G. Aux, J. McElver, D. Patton, B. Dietrich, P. Ho, J. Bacwaden, C. Ko*, et al.* (2002). "A high-throughput Arabidopsis reverse genetics system." Plant Cell **14**(12): 2985-2994.

Sheppard, A. E., M. A. Ayliffe, L. Blatch, A. Day, S. K. Delaney, N. Khairul-Fahmy, Y. Li, P. Madesis, A. J. Pryor and J. N. Timmis (2008). "Transfer of plastid DNA to the nucleus is elevated during male gametogenesis in tobacco." Plant Physiol. **148**(1): 328-336.

Sheppard, A. E. and J. N. Timmis (2009). "Instability of Plastid DNA in the Nuclear Genome." Plos Genet. **5**(1).

Siebert, R. and H. Puchta (2002). "Efficient repair of genomic double-strand breaks by homologous recombination between directly repeated sequences in the plant genome." Plant Cell **14**(5): 1121-1131.

Simsek, D. and M. Jasin (2010). "Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4-ligase IV during chromosomal translocation formation." Nature Structural & Molecular Biology **17**(4): 410-U443.

Soll, J. and E. Schleiff (2004). "Protein import into chloroplasts." Nature Reviews Molecular Cell Biology **5**(3): 198-208.

Sparkes, I. A., J. Runions, A. Kearns and C. Hawes (2006). "Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants." Nature Protoc. **1**(4): 2019-2025.

Stegemann, S. and R. Bock (2006). "Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus." Plant Cell **18**(11): 2869-2878.

Stegemann, S., S. Hartmann, S. Ruf and R. Bock (2003). "High-frequency gene transfer from the chloroplast genome to the nucleus." Proc. Natl. Acad. Sci. U.S.A. **100**(15): 8828-8833.

Stettler, M., S. Eicke, T. Mettler, G. Messerli, S. Hortensteiner and S. C. Zeeman (2009). "Blocking the metabolism of starch breakdown products in *Arabidopsis* leaves triggers chloroplast degradation." Mol Plant **2**(6): 1233-1246.

Stupar, R. M., J. W. Lilly, C. D. Town, Z. Cheng, S. Kaul, C. R. Buell and J. M. Jiang (2001). "Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats." Proceedings of the National Academy of Sciences of the United States of America **98**(9): 5099-5103.

Suzuki, J., K. Yamaguchi, M. Kajikawa, K. Ichiyanagi, N. Adachi, H. Koyama, S. Takeda and N. Okada (2009). "Genetic Evidence That the Non-Homologous End-Joining Repair Pathway Is Involved in LINE Retrotransposition." Plos Genetics **5**(4).

Svab, Z., P. Hajdukiewicz and P. Maliga (1990). "Stable transformation of plastids in higher plants." Proceedings of the National Academy of Sciences of the United States of America **87**(21): 8526-8530.

Svab, Z. and P. Maliga (1991). "Mutation proximal to the tRNA binding region of the *Nicotiana* plastid 16S rRNA confers resistance to spectinomycin." Molecular & General Genetics **228**(1-2): 316-319.

Svab, Z. and P. Maliga (2007). "Exceptional transmission of plastids and mitochondria from the transplastomic pollen parent and its impact on transgene containment." Proceedings of the National Academy of Sciences of the United States of America **104**(17): 7003-7008.

Tanaka, S., C. Ishii, S. Hatakeyama and H. Inoue (2010). "High efficient gene targeting on the *AGAMOUS* gene in an *Arabidopsis AtLIG4* mutant." Biochemical and Biophysical Research Communications **396**(2): 289-293.

Teng, S. C., B. Kim and A. Gabriel (1996). "Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks." Nature **383**(6601): 641-644.

The Rice Chromosome 10 Sequencing Consortium (2003). "In-depth view of structure, activity, and evolution of rice chromosome 10." Science **300**(5625): 1566-1569.

Thorsness, P. E. and T. D. Fox (1990). "Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*." Nature **346**(6282): 376-379.

Timmis, J. N., M. A. Ayliffe, C. Y. Huang and W. Martin (2004). "Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes." Nature Rev. Genet. **5**(2): 123-U116.

Timmis, J. N. and N. S. Scott (1983). "Sequence Homology between Spinach Nuclear and Chloroplast Genomes." Nature **305**(5929): 65-67.

Twell, D., J. Yamaguchi and S. McCormick (1990). "Pollen-specific gene expression in transgenic plants: coordinate regulation of two different tomato gene promoters during microsporogenesis." Development **109**(3): 705-&.

Ueda, M., M. Fujimoto, S. I. Arimura, N. Tsutsumi and K. I. Kadowaki (2008). "Presence of a latent mitochondrial targeting signal in gene on mitochondrial genome." Molecular Biology and Evolution **25**(9): 1791-1793.

van den Boogaart, P., J. Samallo and E. Agsteribbe (1982). "Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa*." Nature **298**(5870): 187-189.

van der Giezen, M., K. A. Sjollema, R. R. Artz, W. Alkema and R. A. Prins (1997). "Hydrogenosomes in the anaerobic fungus *Neocallimastix frontalis* have a double membrane but lack an associated organelle genome." FEBS Lett **408**(2): 147-150.

Villanueva, V. R., V. Mathivet and R. S. Sangwan (1985). "RNA, proteins and polyamines during gemetophytic and androgenetic development of pollen in *Nicotiana tabacum* and *Datura innoxia*." Plant Growth Regulation **3**: 293-307.

Villarejo, A., S. Buren, S. Larsson, A. Dejardin, M. Monne, C. Rudhe, J. Karlsson, S. Jansson, P. Lerouge, N. Rolland, G. von Heijne, M. Grebe, L. Bako and G. Samuelsson (2005). "Evidence for a protein transported through the secretory pathway *en route* to the higher plant chloroplast." Nature Cell Biology **7**(12): 1224-1231.

Vogel, J. P., D. F. Garvin, T. C. Mockler, J. Schmutz, D. Rokhsar, M. W. Bevan, K. Barry, S. Lucas, M. Harmon-Smith, K. Lail*, et al.* (2010). "Genome sequencing and analysis of the model grass *Brachypodium distachyon*." Nature **463**(7282): 763-768.

Vonheijne, G. (1986). "Why Mitochondria Need a Genome." Febs Letters **198**(1): 1-4.

Wada, S., H. Ishida, M. Izumi, K. Yoshimoto, Y. Ohsumi, T. Mae and A. Makino (2009). "Autophagy plays a role in chloroplast degradation during senescence in individually darkened leaves." Plant Physiol **149**(2): 885-893.

Wang, H. C., A. R. Perrault, Y. Takeda, W. Qin, H. Y. Wang and G. Iliakis (2003). "Biochemical evidence for Ku-independent backup pathways of NHEJ." Nucleic Acids Research **31**(18): 5377-5388.

Ward, B. L., R. S. Anderson and A. J. Bendich (1981). "The Mitochondrial Genome Is Large and Variable in a Family of Plants (Cucurbitaceae)." Cell **25**(3): 793-803.

Weterings, E. and D. J. Chen (2008). "The endless tale of non-homologous end-joining." Cell Research **18**(1): 114-124.

Wilkinson, J. E., D. Twell and K. Lindsey (1997). "Activities of CaMV 35S and *nos* promoters in pollen: implications for field release of transgenic plants." Journal of Experimental Botany **48**(307): 265-275.

Wilson, R. J. M., P. W. Denny, P. R. Preiser, K. Rangachari, K. Roberts, A. Roy, A. Whyte, M. Strath, D. J. Moore, P. W. Moore and D. H. Williamson (1996). "Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*." Journal of Molecular Biology **261**(2): 155-172.

Wilson, R. J. M. and D. H. Williamson (1997). "Extrachromosomal DNA in the Apicomplexa." Microbiology and Molecular Biology Reviews **61**(1): 1-&.

Windels, P., S. De Buck, E. Van Bockstaele, M. De Loose and A. Depicker (2003). "T-DNA integration in Arabidopsis chromosomes. Presence and origin of filler DNA sequences." Plant Physiol. **133**(4): 2061-2068.

Wolfe, K. H., W. H. Li and P. M. Sharp (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." Proceedings of the National Academy of Sciences of the United States of America **84**(24): 9054-9058.

Wolfe, K. H., C. W. Morden and J. D. Palmer (1992). "Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant." Proceedings of the National Academy of Sciences of the United States of America **89**(22): 10648-10652.

Xiao, Y. L. and T. Peterson (2000). "Intrachromosomal homologous recombination in Arabidopsis induced by a maize transposon." Molecular and General Genetics **263**(1): 22-29.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto and D. Bhattacharya (2004). "A molecular timeline for the origin of photosynthetic eukaryotes." Molecular Biology and Evolution **21**(5): 809-818.

Yu, H. S. and S. D. Russell (1994). "Populations of plastids and mitochondria during male reproductive cell maturation in *Nicotiana tabacum* L.: A cytological basis for occasional biparental inheritance." Planta **193**(1): 115-122.

Yukawa, M., T. Tsudzuki and M. Sugiura (2005). "The 2005 version of the chloroplast DNA sequence from tobacco (*Nicotiana tabacum*)." Plant Molecular Biology Reporter **23**(4): 359-365.

Zhu, Q. H., K. Ramm, A. L. Eamens, E. S. Dennis and N. M. Upadhyaya (2006). "Transgene structures suggest that multiple mechanisms are involved in T-DNA integration in plants." Plant Sci. **171**(3): 308-322.