FRANCIS, KIRK ERIC. The Effects of T-DNA Integration Sites on Transgene Expression in *Arabidopsis*. (Under the direction of Dr. Steven Spiker and Dr. William Thompson)

Several recent investigations of T-DNA integration sites in *Arabidopsis thaliana* have reported "cold spots" of integration, especially near centromeric regions. These observations have contributed to the ongoing debate over whether T-DNA integration is random, or if integration preferentially occurs in transcriptionally active regions. When transgenic plants are identified by selecting or screening for transgene activity, transformants with integrations into genomic regions that suppress transcription, such as heterochromatin, may not be identified. This phenomenon, which we call selection bias, may explain the perceived non-random distribution of T-DNA integration in previous studies. In order to investigate this possibility, I have characterized the sites of T-DNA integration in the genomes of transgenic plants identified by pooled PCR, a procedure that does not required expression of the transgene and is therefore free of selection bias.

Over 100 transgenic *Arabidopsis* plants were identified by PCR and compared to kanamycin-selected transformants from the same T1 seed pool. A higher perceived transformation efficiency and a higher frequency of transgene silencing were observed in the PCR-identified lines. Together, the data suggest approximately 30% of transformation events may result in non-expressing transgenes that would preclude identification by selection. Genomic integration sites in PCR-identified lines were compared to those in existing T-DNA integration databases. In PCR-identified lines with silenced transgenes, the integration sites mapped to regions significantly underrepresented by T-DNA integrations in studies where transformants were identified

by selection. The data presented here suggest that selection bias can account for at least some of the observed non-random integration of T-DNA into the *Arabidopsis* genome.

THE EFFECTS OF T-DNA INTEGRATION SITES ON TRANSGENE EXPRESSION IN ARABIDOPSIS

by

KIRK ERIC FRANCIS

A dissertation submitted to the Graduate Faculty of
North Carolina State University in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy

GENETICS

Raleigh, North Carolina

2004

APPROVED BY:

_____	_____
STEVEN SPIKER	WILLIAM THOMPSON
Co-chair of Advisory Committee	Co-chair of Advisory Committee
_____	_____
MARK CONKLING	ARTHUR WEISSINGER

Personal Biography

Kirk Eric Francis was born June 17, 1971 in Creston, Iowa. He grew up on a small family farm and was exposed to many aspects of agriculture including raising cattle and swine and growing corn and soybeans. He attended the Creston, Iowa public school system and graduated from Creston High School in 1989.

Following high school graduation, he worked on the family farm for one year while attending Southwestern Community College in Creston. In 1990 he transferred to Iowa State University in Ames, Iowa and went on to receive a Bachelor of Science in Genetics with a minor in Botany in 1993. While attending Iowa State University, He worked in several plant molecular biology laboratories and received two competitive internships to conduct research at Oak Ridge National Laboratory in Oakridge, Tennessee.

In 1994 He was accepted into the Plant Breeding and Plant Genetics program at the University of Wisconsin in Madison, Wisconsin. While there, he conducted his thesis research on Genetic Transformation and Transgene Analysis of Hybrid Poplar NM6 (*Populus nigra* X *Populus maximowiczii*), and received a Master of Science in Plant Breeding and Plant Genetics in 1996.

From Wisconsin, he accepted a senior research specialist position with Westvaco, a major paper and wood products company. While at Westvaco, he managed the Hardwood Transformation Facility of the Forest Science Laboratory in Summerville, South Carolina and conducted research leading to the development and improvement of transformation systems for several economically important hardwood tree species.

He was admitted to the Genetics Department at North Carolina State University in 1998, as a candidate for the degree of Doctor of Philosophy. He has undertaken his research under the guidance of Dr. Steven Spiker and Dr. William F. Thompson.

Acknowledgements

I would like to acknowledge the guidance, assistance, and friendship of Dr. Steven Spiker. I also wish to thank my fellow graduate students Chris Halweg and Rick Hall for their advice and assistance throughout the past several years. I wish to thank Mara Massel for keeping the lab and equipment in working order. I would like to acknowledge and thank Jose Alonso and Anna Stepanova for providing protocols and information necessary to perform much of the T-DNA mapping and sequence analysis reported in Chapter 2, and I would like to acknowledge and thank Anton Calloway for advice and protocols used throughout this dissertation. I wish to thank my committee, Steven Spiker, Bill Thompson, Arthur Weissinger, and Mark Conkling for providing advice and guidance, and also for providing funding (either directly or indirectly) for various aspects of the research reported in this dissertation. Finally, I wish to thank my wife, Laura Schenkman, for her support and scientific advice throughout my tenure as a graduate student.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2,4-D	2,4-dichlorophenoxyacetic acid
ABRC	*Arabidopsis* Biological Resource Center
BAC	bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
cDNA	complementary DNA
CM medium	callus maintenance medium
DNA	deoxyribonucleic acid
DSB	double strand break
EDTA	ethylene diamine tetraacetic acid
ELISA	enzyme-linked immunosorbent assay
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridization
GUS	β-glucuronidase
IAA	indole-3-acetic acid
IPAR	6-(γ,γ-dimethylallylamino)-purine riboside
LB	left border
MAR(s)	matrix attachment region(s)
MS medium	Murashige and Skoog medium
MUG	4-methyl umbelliferyl glucuronide
MU	methylumbelliferone
NCBI	National Center for Biotechnology Information
NOR	nucleolar organizer region
PCR	polymerase chain reaction
PTGS	post-transcriptional gene silencing
RB	right border
RdDM	RNA-dependent DNA methylation
RFU(s)	relative fluorescence unit(s)
RLU(s)	relative luminescence unit(s)
RNA	ribonucleic acid
rRNA	ribosomal DNA
SIGnAL	Salk Institute Genomic Analysis Laboratory
SG medium	seed germination medium
SGE medium	seed germination enhanced medium
T-DNA	transferred-DNA
TAIL PCR	thermal asymmetric interlaced PCR
Ti plasmid	tumor inducing plasmid
TGS	transcriptional gene silencing
UTR	untranslated region
wt	wild type
X-gluc	5-bromo-4-chloro-3-indoyl glucuronide

Chapter 1

Genomic Integration Preferences of T-DNA

Introduction

T-DNA integration has been described as being random and has alternatively been described as having integration preferences in regions of high gene density. T-DNA integration sites have been reported to be more common in gene-containing genomic regions, due to an apparent correlation between active transcription and T-DNA integration, and have also been reported to be less common in heterochromatin, due to the physical blocking of integration by a condensed chromatin state. Many attempts have been made to address the issue, but all previous studies are either too small to adequately assess randomness or have been subjected to a "selection bias" that may result in the inability to recover integrations in regions incapable of supporting transgene expression. The concept of selection bias is based on the idea that if transgenic DNA integrates into a region that suppresses transgene expression, selectable or screenable markers carried on that transgenic DNA will not be expressed, and transformed plants containing such integrations will not be identified under selective conditions.

A brief review of the research that has led to conflicting views of T-DNA integration preference is presented here. The focus will be on the specific experiments and how past results should be viewed in light of current knowledge. The idea of selection bias is presented, and a detailed discussion of how it should change the way integration data is interpreted will be made. An examination of T-DNA and direct transgene integration in plants and other organisms will also be addressed in light of selection bias. Finally, experiments will be outlined that may help to clarify the issue of T-DNA integration preference.

Early studies suggest T-DNA integration is random

Since it was first reported that DNA contained on the Ti plasmid of *Agrobacterium* can be incorporated into plant genomes (Chilton et al., 1977), the question of where in the genome the T-DNA integrates has been a frequently addressed issue. Subsequent research confirmed that T-DNA integrated into the nuclear genome and that independent integration events appeared to occur at different genomic loci (Chilton et al., 1980; Willmitzer et al., 1980). In efforts to identify the genomic integration sites of T-DNA, selectable markers in six transgenic petunia plants (Wallroth et al., 1986) and ten transgenic hybrid tomato plants (Chyi et al., 1986) were mapped using limited molecular and phenotypic markers. No chromosomal or locus preference was detected. Similar results were observed in five transgenic *Crepis capillaris* plants, where T-DNA integration sites were identified by *in situ* hybridization of T-DNA probes to metaphase chromosome spreads (Ambros et al., 1986). In a slightly larger experiment, the T-DNA integration sites of 37 transgenic tomato plants were fine-mapped using a large collection of molecular markers (Thomas et al., 1994). These authors all concluded that T-DNA integration appeared to be random (Ambros et al., 1986; Chyi et al., 1986; Wallroth et al., 1986; Thomas et al., 1994), although the size and scopes of the experiments limited the conclusions to randomness on a chromosomal level. Surprisingly, these reports are often cited as definitive evidence that T-DNA integration is random (e.g. Osborne et al., 1995; Tinland, 1996; Bundock et al., 2002). The first thorough attempt to assess the randomness of T-DNA integration was reported by Feldman (1991). More than 8000 *Arabidopsis* transformants were generated as part of an insertional mutagenesis project. Not all integrations were mapped, but a sample of T-DNA integration sites that resulted in mutant phenotypes were identified. It was

concluded that integration did not appear to be limited to a small number of loci (Feldmann, 1991). Feldman went on to make the assumption of random integration in estimating the number of transformants necessary to saturate the *Arabidopsis* genome with at least one T-DNA insertion into each gene.

Promoter-trapping experiments suggest T-DNA targets active genes

The view that T-DNA preferentially integrates into actively expressed genes has primarily arisen from promoter-tagging studies (reviewed in Walden, 2002). The idea of using T-DNA integration as a tool to identify and study plant gene regulatory sequences was first proposed and demonstrated by Koncz et al. (1989). In promoter-tagging experiments plants are transformed with a construct containing a selectable marker and a promoterless reporter gene at either the left or right T-DNA border. After selecting for T-DNA integration, the transformed plants are then screened for reporter gene activity. It is believed that without an upstream promoter in the T-DNA, the reporter gene will be incapable of expression unless it integrates downstream of an endogenous promoter as a transcriptional fusion. If such an event should occur, it would be expected that the reporter gene would assume the expression patterns of the endogenous promoter. This tagging is referred to as promoter-trapping and has been widely used in tobacco (Koncz et al., 1989; Herman et al., 1990; Topping and Lindsey, 1995) and *Arabidopsis* (Lindsey et al., 1993; Campisi et al., 1999; Karimi et al., 1999; Plesch et al., 2000). Early experiments involving promoter-trapping revealed that transformants displayed reporter gene activity at surprisingly high frequencies. Koncz et al. (1989) observed that over 25% of transformed tobacco plants had detectable reporter gene (*aph*II) activity. Even higher frequencies of promoterless reporter gene expression were subsequently reported by Herman et al. (1990) and Topping et al. (1991) of 50% and 78% respectively. There

are two reasons why these frequencies are surprisingly high. First, genic regions are believed to encompass only a relatively small portion of the tobacco genome (Fobert et al., 1991), and second, integration must occur in a specific orientation and distance from an endogenous promoter to drive reporter gene expression. If T-DNA integration were a completely random event with an equal chance of integration at any position in the tobacco genome, the events necessary to drive the expression of a promoterless reporter gene would be expected to occur at detectable, but low frequencies. The authors of these reports all concluded that the only way to explain their observations was that T-DNA integration specifically targeted genes that were either actively expressing or in a transcriptionally poised state (Koncz et al., 1989; Herman et al., 1990; Topping et al., 1991). This is partially supported by previous research that found integrated T-DNA to be in an open chromatin state (Coates et al., 1987). However this research did not address whether the T-DNA targeted a region with an existing open chromatin state (such as an actively transcribed gene) or whether the integration of the T-DNA itself resulted in an open chromatin state.

Further experiments supported these views. Even though different plant species possess genomes of vastly different sizes (Bennett and Leitch, 1995), it is thought that the proportion of those genomes that contain genes (the "gene space") may be fairly similar in size (Barakat et al., 1997). If different plant species have similar amounts of actively transcribed "gene space," and if T-DNA integration targets actively expressed genic regions, then it would be expected that different species should have similar frequencies of promoterless reporter gene expression, regardless of their genome size. That is indeed what was observed in some cases. Lindsey et al. (1993) conducted promoter trap experiments in tobacco, *Arabidopsis*, and potato. Very similar frequencies of promoterless *gusA* activity in leaf tissues were reported for all three species (22%, 28%,

and 23% respectively) even though their genome sizes vary by over 200-fold (Bennett and Leitch, 1995). These support previous observations of similar promoter-trap efficiencies in *Arabidopsis* and tobacco (Koncz et al., 1989). Although the frequencies of promoterless *gusA* expression in leaf tissues were similar for transformed tobacco, *Arabidopsis*, and potato, there were major differences in root and floral tissues. In tobacco, over 75% of transformants had promoterless *gusA* activity in roots, while only 30% of *Arabidopsis* and 9% of potato transformants had detectable activity in root tissues. In tobacco, over 90% of transformants had promoterless *gusA* activity in flowers, while the number of *Arabidopsis* transformants with GUS activity in floral tissues was estimated to be around 22% (Lindsey et al., 1993). These differences and similarities may reflect underlying mechanisms of T-DNA integration or may simply demonstrate the inherent variability of promoterless reporter gene experiments.

It has been suggested that expression patterns at the time of transformation may influence which genes receive T-DNA inserts (Martirani et al., 1999). This is what might be expected if T-DNA preferentially targets active genes. The data of Lindsey et al. (1993), however, do not support this model. Tobacco transformants were generated from leaf explants, but leaf tissues in transgenic tobacco had the lowest promoterless *gusA* expression frequency of all tissues tested (22% compared to 75% in roots and 92% in flowers). In an experiment designed to find root and nodule-specific promoters, Martirani et al. (1999) transformed root and hypocotyl explants with *Agrobacterium rhizogenes* containing a T-DNA with a promoterless *gusA* reporter gene. Twenty-three percent of root-transformed explants displayed GUS activity in roots, while only 13% of hypocotyl-transformed explants had detectable GUS activity in roots. This suggests that the expression profile of the tissues subjected to transformation may influence the sites of integration and subsequent expression patterns of transgenes. In keeping with this model,

it has been suggested that some forms of silencing may be related to the integration of T-DNA into regions active only at the time of transformation (Topping and Lindsey, 1995). In an experiment to test this, Topping and Lindsey (1995) cultured leaf explants from silenced transgenic plants on the same transformation medium initially used to transform those plants. Eleven silenced 35S:: *gusA* plants and eight non-expressing promoterless *gusA* plants were cultured in this way. All of the 35S::*gusA* tissues and five of the promoterless *gusA* tissues regained expression in callus generated on the transformation medium (Topping and Lindsey, 1995). The authors concluded that the restoration of expression was likely a result of a reactivation of the endogenous genes into which the T-DNAs had integrated. While this conclusion supports the underlying model, new research suggesting that post-transcriptional gene silencing (PTGS) may be deactivated in callus (Mitsuhara et al., 2002) may explain the results differently. This alternative explanation for the reactivation of *gusA* in callus does not necessarily invalidate the idea that T-DNA targets genes that are active in the plant tissue at the time of transformation.

In other experiments both transcriptional and translational fusions have been used to drive the expression of a promoterless reporter gene. Unlike transcriptional fusion cassettes, translational fusion cassettes lack a start codon (ATG) in the reporter gene and, therefore, must integrate into an exon and in frame in order to be expressed. Kertbundit et al. (1991) observed that 54% of *Arabidopsis* root transformants displayed GUS activity from a promoterless transgene designed to function as a transcriptional fusion. Only 1.6% of transformants expressed GUS from a similar construct designed to function as a translational fusion. It was concluded that T-DNA integration preference was much stronger for promoters and upstream untranslated regions than it was for exons (Kertbundit et al., 1991), although there are many reasons why a translational fusion may fail to express, including integrating out of translational frame. Koncz et al. (1989),

however, observed no difference between the frequencies of expression of promoterless *aphII* transcriptional and translational fusions. Differences in the results of these two similar studies are difficult to resolve and may again reflect the highly variable nature of promoterless reporter gene studies.

Not all evidence from gene-tagging experiments supports the model of integration preference. Fobert et al. (1991) conducted a very large project consisting of over 1000 transformed tobacco plants. Contrary to the reports of others, only 5% of the transgenic plants had detectable GUS activity. The authors felt this was consistent with what would be expected if integration were completely random and 95% of the transformants contained T-DNAs that had integrated into regions incapable of driving the expression of the promoterless *gusA* gene (Fobert et al., 1991).

Results of promoter-trapping experiments are often highly variable. Although most experiments detect promoterless reporter gene activity in 20% to 50% of transgenic plants, in some experiments frequencies as high as 92% have been reported (Lindsey et al., 1993). It is these surprisingly high results that frequently lead authors to conclude that T-DNA integration targets active genes. There are reasons, however, to question these observations. Even if T-DNA were exclusively targeted to integrate into promoters (or into 5' untranslated regions), the random orientation of the T-DNA integration would still limit the frequency of successful transcriptional fusions to 50%. It is possible that multiple T-DNA integrations could cumulatively increase the observed frequency, but most promoter-trapping reports claim to have single or low copy numbers (Fobert et al., 1991; Lindsey et al., 1993; Topping and Lindsey, 1995; Walden, 2002). The only way higher frequencies could be achieved would be if the integrating T-DNA had the ability to distinguish and target its integration to the sense strand. This is highly unlikely given the abundance of evidence suggesting that no strand preference exists (Alonso et al.,

2003) and considering that high frequencies of promoter-trapping have been achieved from both left border (LB) and right border (RB) fusions. Recent T-DNA mutagenesis projects have repeatedly demonstrated that T-DNA integration occurs throughout the *Arabidopsis* and rice genomes in promoters, introns, exons, terminators, and intergenic regions (Krysan et al., 1999; Barakat et al., 2000; Sessions et al., 2002; Szabados et al., 2002; Alonso et al., 2003; Chen et al., 2003; Ichikawa et al., 2003). However, there may be a slight preference for 5' and 3' regulatory region insertions (Krysan et al., 2002; Sessions et al., 2002; Szabados et al., 2002; Alonso et al., 2003).

It still remains unexplained how in some experiments such a high frequency of promoterless reporter gene expression was observed. One possible explanation is that promoterless reporter genes do not necessarily need to integrate near an endogenous promoter to be transcribed. Some non-promoter genomic regions might allow a very low level of transcription to occur. This level may be affected by the active translation of the adjacent selectable marker gene typically driven by a strong promoter. It takes only a low level of transcription to accumulate enough GUS product to be detected by fluorescence (Jefferson, 1987), and because of the high stability of GUS (Gallagher, 1992), low levels of transcription could eventually result in the accumulation of levels detectable by histochemical staining. It is also possible that due to complex T-DNA integrations, the reporter gene could be positioned downstream of the selectable marker. Read-through transcription from such an arrangement could result in low levels of *gusA* expression. In the few reports where promoterless *gusA* activities have been quantitatively reported, the levels for the majority of transformants are only slightly above background (Topping et al., 1991; Lindsey et al., 1993), indicating the majority of transformants with GUS activity are, indeed, expressing *gusA* at a very low level.

One important factor to consider is the possibility that selecting for transformants by requiring regenerating tissues to grow in the presence of a selective agent (typically kanamycin or hygromycin) may bias the population of transformants. T-DNAs that integrate into regions of the genome incapable of supporting expression of the selectable marker will not be identified. Consequently, all transformants identified by selection will be biased for integrations into regions supportive of transgene expression. This form of bias may be even more severe if the promoter driving the selectable marker is located at a T-DNA border so that surrounding plant chromatin is directly adjacent to the promoter. It is interesting that the promoter-trapping experiments with the highest frequencies of promoterless reporter gene expression have used selectable markers arranged in this manner (Kertbundit et al., 1991; Topping et al., 1991; Lindsey et al., 1993; Martirani et al., 1999) (Note: Figure 1 in Lindsey et al. (1993) incorrectly represents the direction of *nptII* transcription). Likewise, the promoter-trapping experiment with the lowest frequency (5% promoterless *gusA* expression), and the only promoter-trapping experiment to conclude that T-DNA integration is random, used a selectable marker with an internal promoter driving expression toward the left border (Fobert et al., 1991).

In order to completely avoid complications associated with the possibility of selection biased data resulting from the use of a selectable marker, an experiment would have to be conducted where transformants were identified without selection. Only one promoter-trapping experiment has taken this approach. Herman et al. (1990) transformed tobacco protoplasts with promoterless *nptII* located at either the left or right border. No other selectable marker or promoter was included in the T-DNA. Following cocultivation, all of the resulting calli were allowed to regenerate in the absence of selection. When the calli were screened for *nptII* activity, 26% were found to be kanamycin resistant. This frequency is very similar to what was observed when

transgenic calli were initially identified by hygromycin selection using a T-DNA that also contained *hpt* driven by the *nos* promoter. These results strongly suggest that T-DNA integration targets actively transcribed regions. However, the individual lines assayed by Herman et al. (1990) were not characterized to the point where it could be concluded that they were independent transformants. Furthermore, the individual calli were not screened in a manner that would identify the possibility of chimeric calli. Transgenic tobacco cell cultures are subject to chimerism following periods of culturing without selection (Chris Halweg, personal communication). What may appear to be a mini-callus derived from a single cell may actually be a mix of transgenic and non-transgenic cells growing in a sectored or interspersed group. Since transformants were identified by callus growth and not by the culturing of individual protoplasts, the possibility of chimerism cannot be ruled out. It is possible that many "independent" mini-calli contained a few cells that were all derived from a limited number of transformation events. When the calli were transferred to medium containing hygromycin, the untransformed and non-expressing cells died. The non-independently derived hygromycin resistant cells would have then grown and given the appearance of multiple promoterless *hpt*-expressing events. Without additional information, this possibility cannot be excluded.

Although many authors have suggested that T-DNA integration may target actively transcribed regions, relatively few models have been presented as to how this may occur. Mutants deficient in the ability to repair single-strand nicks or double-strand breaks have been found to also be deficient in T-DNA integration (Sonti et al., 1995). It is therefore reasonable to conclude that DNA repair machinery may play a leading or supportive role in T-DNA integration. It is also well established that some forms of DNA repair are coupled with transcription (Svejstrup, 2002). This potential relationship

between T-DNA integration and sites of active transcription has been proposed as a possible reason why T-DNA integration may preferentially occur in actively transcribed genes (Martirani et al., 1999). Indeed, the limited entourage of proteins associated with the T-DNA contributed by the *Agrobacterium* (only *vir*D2 and *vir*E2) suggests that integration must take advantage of the plant's own repair machinery (for review see Tzfira and Citovsky, 2002). Single-strand nicks and double-strand breaks have also been associated with DNA replication (Sutton and Walker, 2001). DNA damage associated with replication has been shown to play an important role in facilitating mitotic recombination (for review see Petes et al., 1991). Thus, just as it can be argued that there may be a relationship between transcription and T-DNA integration, it can also be argued that there may be a relationship between replication and T-DNA integration. A dependence on replication-associated repair systems would support the idea that T-DNA integration is random, since the entire genome undergoes replication regardless of transcriptional activity or chromatin state.

Upon an initial examination of promoter-trapping experiments, it is tempting to conclude that T-DNA integration must be somehow targeted to actively transcribed genomic regions. Indeed, many others have come to this conclusion (Barakat et al., 2000; Walden, 2002). However, a more complete examination suggests that alternate explanations may exist for the observed high frequencies of promoterless reporter gene expression.

Post-genomics era research reveals non-random T-DNA integration patterns

Experiments addressing the issue of the randomness of T-DNA integration can be divided into pre-genomic and post-genomic categories. Prior to the *Arabidopsis* genome project (*Arabidopsis* Genome Initiative, 2000), very little was known about T-DNA

integration preferences (or lack thereof). Experiments supporting the hypothesis that T-DNA integration was a random event were often insufficient in size to allow strong conclusions to be made. Furthermore, without established genomic maps, the precise sites of integration could not be determined. Promoter-tagging experiments that frequently led to the conclusion that T-DNA was targeted to active regions were subjected to selection bias and also suffered from the inability to identify precise sites of integration. Two breakthroughs that allowed for major advances in addressing the issue were the completion of the *Arabidopsis* (*Arabidopsis* Genome Initiative, 2000) and rice (Yu et al., 2002) genome projects and the widespread development of high-throughput techniques for identifying genomic sequences flanking T-DNA insertions (Liu et al., 1995; Siebert et al., 1995; Mathur et al., 1998). Combined with a rapid and highly efficient transformation system, such as floral dip in *Arabidopsis* (Clough and Bent, 1998), researchers can generate large numbers of transformants, identify genomic sequences flanking the T-DNA, and identify precisely where those T-DNAs have integrated.

Large scale T-DNA insertion mutagenesis in *Arabidopsis* was first proposed by Feldman (1991). Such projects were established with the goal of creating T-DNA insertions across the entire genome, thus saturating all known and predicted genes with mutations. One tremendous benefit of using T-DNA integration as a mutagen was that the mutations would be tagged with the integrating T-DNA, which would facilitate the rapid cloning of the mutated gene. The widespread cataloging of T-DNA insertions, regardless of whether they possessed an observed mutant phenotype, led to the development of reverse genetics. Unlike traditional or forward genetics, in which researchers start with a mutant phenotype and attempt to identify the genotype, reverse genetics allows researchers to start with a mutated genotype and look for a resulting

change in phenotype (Krysan et al., 1999). In order to facilitate reverse genetics, different projects were initiated to identify the genomic loci of thousands of T-DNA integration sites. Within a few years, an emerging picture of T-DNA integration was beginning to appear. Contrary to the predictions based on many previous studies involving promoterless reporter genes, integrations were observed not only in promoters, but also in exons, introns, downstream, and intergenic regions (Azpiroz-Leehan and Feldmann, 1997). Furthermore, integrations were found in all categories of genes, not just those genes thought to be active at the time of transformation. These observations contributed to the reemerging view that T-DNA integration was random (Azpiroz-Leehan and Feldmann, 1997). Further experiments supported these views (Krysan et al., 1999) and led to an accepted model of random integration that was frequently used to predict the number of transformants needed to uniformly saturate the genome with T-DNA inserts (Krysan et al., 2002).

In two studies, the distribution of approximately 1000 *Arabidopsis* T-DNA integration sites was examined (Szabados et al., 2002; Ichikawa et al., 2003). Szabadoz et al. (2002) found insertions distributed across all five chromosomes with a slight preference for integrations in regions with "high gene density." Within genic regions, there was a slight preference for integrations to be found in 5' and 3' regulatory regions (defined by Szabadoz et al., 2002, as 300 bp upstream of the start ATG and 300 bp downstream of the stop codon, respectively). These observations are somewhat similar to claims made based on promoter-tagging studies. However, the slight increase in promoter insertions reported by Szabados et al. (2002) cannot sufficiently explain the extremely high frequency of promoterless reporter gene activity in promoter-tagging studies. Ichikawa et al. (2003) do not report a high number of 5' or 3' regulatory region insertions but do state that the general pattern of integrations observed in their activation-

tagged collection is similar to that of Szabados et al. (2002). These two studies also report a clear decrease in the number of integrations that occurred near centromeric, telomeric, and rDNA repeat regions.

The first very large scale approach to identify T-DNA integrations for a high-throughput reverse genetics system was reported by Sessions et al. (2002). In this study, approximately 100,000 transformed lines were analyzed, and integration sites of 52,964 T-DNAs were identified. As in other reports, a slight bias for integrations in promoter regions was observed. Several loci appeared to be "hot spots" of integration with as many as 40-fold more integrations than would be expected if integrations were uniformly distributed. However, upon further examinations, specific integrations in these "hot spots" could not be confirmed. In fact, many other identified integrations (as many as 25%) were found to be artifactual and could not be verified experimentally (Sessions et al., 2002). Despite these problems, the collection is a valuable resource and appears to confirm observations of others, such as a clear depression in the number of integrations in centromeric regions.

The largest and most comprehensive analysis of T-DNA integration to date was recently reported by Alonso et al. (2003). In this monumental study 88,122 T-DNA junctions were sequenced and mapped to the *Arabidopsis* genome (currently the project has exceeded 140,000 mapped integrations (http://signal.salk.edu). As reported by others, integrations were observed across the entire genome. A statistical analysis found that the distribution did not follow a Poisson distribution and was not random. Regions were found that had both higher and lower than expected numbers of T-DNA integrations, leading to the idea of "hot spots" and "cold spots" of integration. An examination of the distribution of T-DNA integrations (available in Alonso et al., 2003, supplemental material) reveals numerous 5 to 10kb regions that lack T-DNA integrations.

A Poisson analysis for 88,122 integrations in the *Arabidopsis* genome indicates that any 10kb region with fewer than 4 integrations has significantly fewer integrations than would be expected if the distribution were random ($p = 0.02620$). Many of these cold spots contain both predicted and confirmed genes associated with ESTs or cDNAs. Additionally, a clear and pronounced decrease in the number of integrations was observed around the centromeres. This decrease appears to correspond closely with the decrease in predicted genes found near the centromeres. Although *Arabidopsis* is almost entirely composed of gene rich regions (Fransz et al., 1998; Barakat et al., 2000), Alonso et al. (2003) observed a close correlation between gene density and T-DNA integration. A bias for integrations in 5' UTRs, 3' UTRs, and promoters and against integrations in introns and exons was also observed. While this bias was significant, it was not severe and cannot wholly explain previous promoter-trapping results. In an effort to test previous claims that T-DNA integrates preferentially into genes that are transcriptionally active at the time of transformation, the distribution of integrations was compared to genomic tiling microarray data that measured genome-wide gene expression levels. No correlation was detected for genes preferentially expressed in floral tissues (the target tissues for floral dip transformation) or in any other tissues (Alonso et al., 2003; Yamada et al., 2003).

The success of these approaches in *Arabidopsis* has led to similar studies in rice. The composition and organization of the rice genome is very different from that of *Arabidopsis*. Unlike *Arabidopsis*, where the gene rich regions compose over 85% of the genome, the rice genome is predicted to contain only about 10 to 20% gene space (Barakat et al., 1997). In a comparison of T-DNA integration in rice and *Arabidopsis* Barakat et al.(1997) found that while integration in *Arabidopsis* occurred throughout the entire genome, integration in rice was limited to the gene space. The authors concluded

that T-DNA targets genic regions. A more comprehensive analysis found similar results. Chen et al. (2003) conducted an analysis of over 1000 T-DNA integration events in rice. As in previous *Arabidopsis* studies, a slight preference for integration in regions 5' and 3' of predicted genes was observed, but unlike *Arabidopsis*, T-DNA integration was primarily limited to only a few genomic regions. These regions corresponded to the gene-containing regions, and the authors concluded that T-DNA preferentially integrated into the gene space of the rice genome (Chen et al., 2003). However, it is important to consider that since such a large portion of the rice genome is considered to be non-gene space, transformants identified by selection may be especially prone to selection bias. If T-DNA integration were random, and only 10 to 20% of the rice genome were capable of supporting transgene expression, then 80 to 90% of transformants would not be recovered in rice.

Based on these reports of T-DNA integration in *Arabidopsis* and rice, three central conclusions can be made: (1) more than the expected number of integrations are observed in regions 5' and 3' of predicted open reading frames, (2) fewer than the expected number of integrations are observed near *Arabidopsis* and rice centromeres, and (3) the distribution of the observed sites of T-DNA integration is not random and appears to correlate with gene density.

It is conceivable that the preference for 5' and 3' regions could be a result of integration being dependent on transcriptional machinery (as previously described), but it is difficult to explain how such a preference would not also result in an increased frequency of coding region integrations (introns and exons). Some genes may be lethal in a heterozygous state (Bonhomme et al., 1998). T-DNA integrations that disrupted the coding regions of these genes would be selected against and would not be observed. However, integrations into the promoters and 5' UTRs of these genes would also be

selected against, so it is unlikely that this argument could explain the observed preference for 5' and 3' regions over exons and introns. It has recently been reported that T-DNA integration most likely occurs by a form of illegitimate recombination that favors integration at the end of a stretch of five nucleotides that are predominantly A-T rich (Brunaud et al., 2002). This is not a required motif, but an apparent preference does exist for this short pre-insertion site. Such a motif could exist anywhere in the genome, but if certain promoters were higher in A-T content, this may account for the subtle but significant preference for T-DNA integration in 5' UTRs. However, Alonso et al. (2003) failed to observe a correlation between A-T content and T-DNA integration (Alonso et al. 2003, supplemental material).

Many authors have attempted to explain the observation of fewer than expected T-DNA integrations in centromeric regions in terms of accessibility (Krysan et al., 2002; Sessions et al., 2002; Szabados et al., 2002; Alonso et al., 2003; Ichikawa et al., 2003). Centromeres are composed largely of heterochromatin that is in a highly condensed state (Copenhaver et al., 1999). Condensed chromatin is believed to physically block interactions with enzymes such as DNAse I (van Holde, 1989; Wolffe, 1998), and it has been suggested that heterochromatin may physically block T-DNA integration as well (see above references). An effect on integration could occur if centromeric heterochromatin prevented the invading T-DNA from interacting with centromeric DNA or blocked access of the molecular machinery used in T-DNA integration. The observation that recombination is severely repressed in *Arabidopsis* centromeres (Copenhaver et al., 1999) has led to the suggestion that the infrequency of centromeric integration may be related to the absence of required recombination or repair machinery in plant centromeres (Krysan et al., 2002). Whatever process or condition that is responsible for the reduced integration observed in centromeric regions, it is obviously

not sufficient to block all centromeric integrations. Although centromeric integrations are uncommon, they have been reported (Alonso et al., 2003; Chen et al., 2003).

The non-random distribution of T-DNA integration is difficult to explain. If T-DNA integration is observed more often in genic regions than non-genic regions, it could be caused by a preference for genic regions, an avoidance for non-genic regions, or a combination of both.

Other research relating to integration site selection and selection bias

T-DNA integration in non-plant species

In the past few years, researchers have demonstrated the ability to use *Agrobacterium* in the transformation of several non-plant species. The most success has been achieved in fungal species (Bundock et al., 1995; de Groot et al., 1998; Combier et al., 2003; Rolland et al., 2003). In many fungal species, transformation (by direct methods) can be achieved by homologous recombination (Petes et al., 1991). Similarly, homologous recombination can occur in *Agrobacterium*-mediated transformation in fungi and has been successfully used to target T-DNA integration to specific loci (Gouka et al., 1999). However, in the absence of homologous sequences in the T-DNA, integration occurs by illegitimate recombination in a manner believed to be similar to T-DNA integration in plants (Bundock and Hooykaas, 1996). The application of T-DNA integration as an insertional mutagen has been applied to several fungal species (Mullins et al., 2001; Bundock et al., 2002; Combier et al., 2003). Of these, only *Saccharomyces cerevisiae* has a completed genome project, which allows integration events to be mapped. In a pilot study of the application of T-DNA insertional mutagenesis in *S. cerevisiae*, Bundock et al. (2002) identified the genomic positions of 54 T-DNA integration events. Integrations were observed in all but one of the yeast chromosomes,

and the distributions along the chromosomes were fairly consistent with the exception of a potential "hot spot" on chromosome VII associated with a region that may be prone to double strand breaks (DSBs) and repairs. DSBs have previously been associated with transgene integrations in plants (Salomon and Puchta, 1998). While this distribution is intriguing, there are far too few integrations to draw strong conclusions about integration preferences. Heterochromatin associated with *S. cerevisiae* telomeres and mating-type loci has been extremely well characterized (Herskowitz et al., 1991; Gasser and Cockell, 2001). It is interesting that no T-DNA integrations were reported in these regions, although analysis of more events will need to be assessed before strong conclusions can be drawn. A single integration was reported in rDNA repeats. Although the highly variable rDNA region in *S. cerevisiae* is often considered to be in a condensed state, the heterochromatin composition is considered to be different from that of telomeres and mating-type loci (Kasulke et al., 2002). In fact, integration by homologous recombination is readily achieved in the yeast rDNA repeats, and the integrated genes typically express, often at very high levels (Olson, 1991). Regardless of T-DNA integration patterns in *S. cerevisiae*, it is likely that a better understanding of the requirements of T-DNA integration will be learned from studies in yeast (van Attikum et al., 2003).

The only other organism with a sequenced genome that has been reported to be receptive to *Agrobacterium*-mediated transformation is human. *Agrobacterium*-mediated transformation of several types of human cells has been reported (Kunik et al., 2001). However, transformation efficiencies are extremely low, and only a few T-DNA integration sites have been identified.

Direct and transposon-mediated integration in plants

Transgene integration by non-*Agrobacterium*-mediated approaches in plants has also been studied. Integration sites are difficult to obtain from transformants generated by direct gene transfer methods. Because circularized plasmid DNA is typically used in these transformations, precise transgene junctions cannot be predicted. Without a precise, known transgene end, methods such as ligation-mediated PCR (Siebert et al., 1995), TAIL PCR (Liu et al., 1995), and some forms of inverse PCR (Mathur et al., 1998) cannot be used to identify flanking genomic sequences. Very few analyses of direct transfer integration sites exist, but a common observation is that integrations appear to occur in close proximity to matrix attachment regions (MARs) (Sawasaki et al., 1998) and in close proximity to genomic regions with similarities to known MARs (Makarevitch et al., 2003). Although apparently not as common, T-DNA integration has also been reported in close proximity to MARs (Dietz et al., 1994). Even if this phenomenon is real, it is still uncertain weather it is the proximity of a MAR, or simply the proximity of any A+T rich region (a characteristic common to nearly all MARs) is important. The observation of transgenes in close proximity to MARs is interesting in the context of selection bias. MARs included in transformation vectors appear to increase and stabilize transgene expression, possibly through the reduction of transcriptional gene silencing (Allen et al., 2000; Bode et al., 2000). If endogenous MARs have the same effect on transgenes, and if selection requires that transgenes are active in order to be identified, then it may be a consequence of selection bias that transgenes are occasionally found near MARs. Although integrations near MARs are often observed, the expression of transgenes in close proximity to endogenous MARs is not necessarily affected by those MARs (Qin et al., 2003).

The transposition patterns of transposable elements designed for insertional mutagenesis have been extensively studied in several plant species. While the specific mechanisms of integration are still unknown, many studies suggest that some transposons preferentially integrate into transcriptionally active or transcriptionally poised genomic regions (Chin et al., 1999; Ito et al., 2002; Kolesnik et al., 2004). The apparent preference of some transposable elements for genic regions has been utilized to increase the rate of sequencing non-repeat-containing regions in maize by enriching for bacterial artificial chromosomes (BACs) containing an engineered *Mutator* (*Mu*) transposable element (Raizada et al., 2001; Palmer et al., 2003). This target site preference has been repeatedly observed, but it is important to remember that lines containing transposed inserts are identified by selecting or screening for the expression of transgenes contained in the transposon (Ito and Shinozaki, 2002). Again, selection bias would prevent the identification of transposition events that involved integration into heterochromatin capable of silencing those transgenes. Furthermore, transposable elements engineered for easy identification and recovery are initially introduced to the target plant by standard transformation methods (typically *Agrobacterium*-mediated approaches) and are therefore subject to selection bias at the initial integration sites. Since the majority of transposition events of some transposable elements occur in sites that are genetically linked to the initial locus (Parinov et al., 1999; Ito et al., 2002; Kolesnik et al., 2004), any initial bias of the starting integration sites may be reflected in the transposition sites. A transposition bias could also be explained in terms of nuclear address. It is thought that genomic regions with similar chromatin states may be localized to similar interphase nuclear positions (Brown et al., 1997). Similarly, the locations of entire chromosomes are believed to assume specific positions in higher eukaryotes (Cockell and Gasser, 1999; Zink et al., 1999). The majority of transposon integration analyses have been conducted

in maize and rice. It has been demonstrated that the interphase nuclei in several species closely related to maize and rice assume a Rabl configuration, where the telomeres and centromeres associate with opposite poles of the nucleus (Heslop-Harrison et al., 1993; Abranches et al., 1998). Transposition may preferentially occur into chromosomes or genomic regions that are physically close to the initial locus within the nucleus at the time of transposition. Such transposition to nearby regions may explain why some hot-spots of transposition have been identified that are not genetically linked to the initial locus and why some chromosomes receive relatively few transposition events from some loci (Parinov et al., 1999; Kolesnik et al., 2004). These views are supported by the observation that transposons at different starting loci have different transposition patterns (Ito et al., 2002).

Direct and targeted integration in other organisms

Although *Agrobacterium*-mediated transformation of mouse has not been reported, limited analyses of the integration sites of direct gene transfer have been conducted. Direct transgene delivery into the mouse genome is primarily considered to be random (Smith, 2001). Even though the inclusion of several kilobases of homologous sequences in the transformation vector can result in targeted integration (Lubahn et al., 1993), the frequency of integration into a non-targeted locus is thought to be 1000 to 10,000-fold higher (Smith, 2001). Like T-DNA integration in plants and fungi, the integration of linear transgenes in mice appears to occur by illegitimate recombination at sites of microhomology (one to three nucleotides) (Hamada et al., 1993). Non-targeted integration sites have been analyzed for the purpose of generating insertional muatagenesis collections. Studies that have identified integration sites based on fluorescent *in situ* hybridization (FISH) analysis of chromosomal spreads have led to the

conclusion that integration does not appear to be completely random (Nakanishi et al., 2002). Some genomic regions appear to be hot spots of integration, while others appear to exclude integration events. It was recognized by the authors that due to the selective nature of transformant identification, integrations into heterochromatin would likely not be identified.

Targeted integration in mouse embryo-derived stem cells occurs by homologous recombination. Since integration is believed to be concurrent with DNA replication, it is thought that chromatin structure should not affect the ability to target a specific locus (Thomas and Capecchi, 1987; Smith, 2001). The chromatin states of particular loci are, however, thought to contribute to the activities of transgenes targeted to those sites (Hatada et al., 1999). Although targeted integration may be possible at all loci, if transformants are identified by selection, selectable markers may not be expressed sufficiently at some target sites to allow identification. This appears to be the case, given that integration efficiencies are much lower at loci believed to be in a heterochromatic state in embryo-derived stem cells. However if transformants are identified without selection, differences are not observed (Oliver Smithies, personal communication).

Plant transformation studies that have avoided selection bias

Selection bias can be avoided only by screening for the presence of the transgenic DNA. Any identification methods that are dependent on expression of any transgene contained in the transferred DNA, such as negative and positive selectable markers as well as any screenable marker, are subject to selection bias.

Relatively few reports exist in the literature where transformants have been identified without selection. The earliest report was a promoter-trapping study described above (Herman et al., 1990). In this study, it was estimated that approximately 25% of

calli regenerated without selection expressed transgenes. A similar study was conducted by the same group several years later. These researchers were unable to identify transgenic *Arabidopsis* without selection but estimated that approximately 18% of tobacco transformants regenerated without selection were able to express the contained transgenes (De Buck et al., 1998). A small sample of non-expressing tobacco regenerants were screened, but only about 4% appeared to contain non-truncated transgenes that did not express. The authors concluded that the immediate or rapid silencing of transgenes following integration was an uncommon event.

The desire to obtain marker-free transformants for commercial purposes has driven the development of several approaches. Proposed methods include transposition excision (Goldsbrough et al., 1993; Ebinuma et al., 1997), intrachromosomal recombination (Zubko et al., 2000), and segregation of a selectable marker following cotransformation (Permingeat et al., 2003). While these approaches offer some potential for commercial applications, they are all dependent on initial selectable marker expression and therefore subject to selection bias. There have recently been reports of obtaining transformants in wheat (via particle bombardment) (Permingeat et al., 2003) and in potato and cassava (via *Agrobacterium*) (de Vetten et al., 2003) without the use of selection. These transformant populations were identified by PCR-based screening for the presence of transgenes, and should be free of selection bias. In potato the transformation efficiency was approximately 4.5% with about 45 to 60% of transformants displaying the desired transgenic phenotype (de Vetten et al., 2003). Since the transgenic phenotype was based on cosuppression of a gene required for amylose synthesis, it cannot be directly determined if the 40 to 55% of transgenic plants that were not amylose-free were unable to transcribe the transgene or unable to cosuppress the endogenous gene. In wheat the transformation efficiency was 2%, but no expression

data were provided (Permingeat et al., 2003). The most comprehensive analysis of transformation without selection and subsequent transgene expression was conducted in Mexican lime (Dominguez et al., 2002). Using a PCR-screening strategy, transgenic lime plants were identified and compared to transformants identified by a stringent selection regime requiring continued *nptII* and *gusA* expression for over three months. Whereas all plants identified by selection expressed *nptII* and *gusA*, only about 70% of PCR-identified transformants with full-length T-DNAs demonstrated *nptII* and *gusA* expression at detectable levels. Transformants that were silenced for *nptII* were also silenced for *gusA* and a viral coat protein gene also contained in the T-DNA. Although some silenced lines appeared to have been silenced by PTGS, methylation analysis and nuclear run-on assays suggest that some silencing was transcriptional (Dominguez et al., 2002). It is important to consider, however, that TGS may be triggered by PTGS (Wassenegger, et al., 1994), therefore PTGS can not be completely ruled out as an initiator of silencing. These results clearly suggest that some transgenes may rapidly silence upon integration and that this silencing may be a form of transcriptional gene silencing (TGS). The data support the idea that transgenes that integrate into certain genomic regions (such as heterochromatin) may not be identified under selective conditions.

Summary of previous research addressing the randomness of T-DNA integration

The three predominant schools of thought mentioned briefly in the introduction will be more thoroughly addressed here. Expected observations for the ideas that 1) T-DNA integration preferentially occurs in actively transcribed regions, 2) T-DNA integration is suppressed in heterochromatin, and 3) T-DNA integration is random, will

be discussed in light of previous experimental outcomes. The effects of selection bias on these predicted models will also be addressed.

1) Is T-DNA integration preferentially targeted to actively transcribed genes?

If T-DNA integration were preferentially targeted to actively transcribed genes, a bias would be observed for integrations in genes active at the time of transformation. Ovule tissues are the target of *Agrobacterium* in floral dip transformations (Ye et al., 1999). However, no preference for integrations in genes active in flowers or any other plant tissues was observed by Alonso et al. (2003). It is possible that a gene may only need to be transcriptionally poised and not necessarily in an actively transcribing state to be considered a target by T-DNA (Herman et al., 1990). However, most models for active-gene-targeting by T-DNA involve a dependency on repair enzymes associated with the transcriptional machinery, which would require any integration site to have been actively transcribed at the time of integration. The lack of preferential targeting to genes active at the time of transformation and the inability to explain how transcriptionally poised genomic regions could be preferential targets weaken the overall argument for the preferential targeting of T-DNA to actively transcribed genes.

If T-DNA integration were preferentially targeted to actively transcribed genes, no or very few integrations would be expected in inactive genes and in non-genic regions. An analysis of the distribution of T-DNA integrations cataloged by the Salk Institute Genomic Analysis Laboratory (SIGnAL) at http://signal.salk.edu reveals that many integrations have been identified in or near predicted genes that are not supported by ESTs or cDNAs. Considering the breadth of tissues that have been sampled, it is unlikely (although possible) that these genes are highly expressed under normal conditions. The presence of T-DNA integrations in these genes does not support the idea of targeted

integration in active genes. Integrations can also be readily found in intergenic regions and may even occur more frequently in intergenic regions than in open reading frames (Alonso et al., 2003, supplemental material). Furthermore, a recent genome-wide analysis of transcriptional activities in various *Arabidopsis* tissues (Yamada et al., 2003) reveals no correlation with SIGnAL project T-DNA integrations. Analyses of T-DNA integrations in centromeric regions reveal that integrations do occur in these regions but at a lower frequency than in non-centromeric regions. Since centromeric regions and pericentromeric regions are not frequently transcribed (Yamada et al., 2003), it is unlikely that integrations would be observed in these regions if transcription were a major influence.

If T-DNA integration were preferentially targeted to actively transcribed genes, the frequency of promoterless reporter gene activity would be high. High frequencies of promoterless reporter gene activities have been repeatedly observed in promoter-trapping experiments. On first examination, the data from these experiments appear to provide very strong evidence that T-DNA integration is preferentially targeted to actively transcribed genes. However, these experiments are subject to selection bias, which may result in the reduced recovery of transformants with integrations in non-genic regions. Furthermore, the results of some promoter-trapping experiments are questionable in that the frequencies of promoterless reporter gene activity are higher than would be expected even if integration were targeted to transcriptionally active regions (see above). It is also possible that the detection of a promoterless reporter gene's product is not dependent on transgene integration into an actively transcribed genomic region. If this is the case, all instances of reporter gene activity may not be indicative of integration in or near endogenous promoters. Collectively, these points suggest that data from promoter-

trapping experiments do not necessarily indicate a preferential targeting to transcriptionally active regions.

2) Is T-DNA integration suppressed in heterochromatin?

If T-DNA integration were suppressed in heterochromatin, few integrations would be found in centromeric or rDNA regions. Many large scale investigations into the distributions of T-DNA integrations in *Arabidopsis* and rice genomes have reported fewer integration events in centromeric regions than would be expected if integration were completely random. Such a distribution could result from a true integration avoidance, or it could be a consequence of selection bias. Centromeric and pericentromeric regions are largely composed of repeated sequences, are typically in a condensed chromatin state (Fransz et al., 1998; Fransz et al., 2002), and are reduced in recombination rates (Copenhaver et al., 1999). Although small regions capable of transcription have been reported, the majority of centromeric and pericentromeric regions appear to be untranscribed (Yamada et al., 2003). When centromeric or pericentromeric integrations are further examined, it is often discovered that the reporter genes on those integrations are capable of expression (Forsbach et al., 2003). It is possible that many of the identified centromeric integrations have occurred in "transcriptional islands." This, of course, could be a result of an avoidance of heterochromatin, a preference for transcription, or selection bias.

T-DNA integrations in rDNA regions appear to follow similar patterns. Typically, rDNA regions comprise only a small fraction of all observed integrations (Szabados et al., 2002). The majority of rDNA sequences in plants are not transcriptionally active and are typically in a heavily condensed chromatin state (Fransz et al., 2002). In *Arabidopsis*, the rDNA regions are located in nucleolar organizing

regions (NORs) at the telomeres of chromosomes 2 and 4 (Copenhaver and Pikaard, 1996) and in pericentromeric regions of chromosomes 3, 4, and 5 (Copenhaver et al., 1999). The majority of sequences in these regions are typically found in a heavily condensed chromatin state (Fransz et al., 1998; Fransz et al., 2002) and are composed almost entirely of repeats. For this reason it is nearly impossible to specifically identify sites of T-DNA integration in rDNA regions, and many T-DNA mapping projects have chosen to not report or analyze these integrations (e.g. Alonso et al., 2003). Integrations in rDNA genes do occur (Mathur et al., 1998), but at a very low frequency. Again, this could be a result of heterochromatin avoidance or selection bias.

If T-DNA integration were suppressed in heterochromatin, integrations in yeast mating-type loci and telomeres would be uncommon or absent. Few heterochromatic regions have been studied as thoroughly as the mating-type loci and telomeres in the yeast *S. cerevisiae* (for review see Herskowitz et al., 1991). A recent effort to identify T-DNA integration sites in *S. cerevisiae* found integrations distributed throughout the genome, but no integrations were observed in either the mating-type loci or the telomeres (Bundock et al., 2002). While this evidence is intriguing, the number of integration events examined was insufficient to rule out the possibility that such integrations could occur.

3) Is T-DNA integration random?

If T-DNA integration were random, integrations would be expected to occur throughout the genome. Larger insertional mutagenesis projects in *Arabidopsis* have generated populations containing integrations throughout the genome, however the observed distribution of integrations is not random (Alonso et al., 2003). As previously mentioned, the observed non-random distribution could be a consequence of selection

bias. Integration does occur throughout the *Arabidopsis* genome (with the exceptions of small 5 to 10 kb interspersed regions), although the densities of integration in some regions are low. The presence of regions with reduced numbers of integration events can be explained by the possibility that multiple T-DNA integrations may allow for some integrations to occur free from selection bias (see Chapter 2).

If T-DNA integration were random, some integrations would be expected in regions incapable of supporting transcription and would be silent. It is well established that transgenes that integrate into different sites can vary in expression (Matzke and Matzke, 1995). Unfortunately, selection bias would preclude the identification of transgenes completely incapable of expression or whose expression is below detection limits. Experiments where transformants have been identified without selection have resulted in the identification of some transformants that do not express their transgenes (De Buck et al., 1998; Dominguez et al., 2002). This evidence strongly supports the idea that integration could be random.

If T-DNA integration were random, different target tissues would be expected to have similar patterns of integration. Sessions et al. (2002), Ichikawa et al. (2003), and Alonso et al. (2003) all used floral dip transformation where ovule tissues are the target of *Agrobacterium* (Ye et al., 1999). However, Szabados et al. (2002) based their analysis primarily on transformants generated via root transformation. The patterns observed by all authors were similar, suggesting that different target tissues do not lead to different integration patterns.

If T-DNA integration were random, the frequency of promoterless reporter gene expression should reflect the density of promoters and genes in the target genome. The frequency of promoterless reporter gene activity is usually much higher than would be expected if integration were completely random, however, the high frequencies reported

by some researchers should probably be questioned. Furthermore, nearly all studies have been subject to selection bias. Frequencies that may reflect the densities of promoters and genes in tobacco have been reported (Fobert et al., 1991), but these results are not typical.

Proposed experimental approaches to study T-DNA integration

Selection bias is the largest obstacle that must be overcome before a more accurate view of T-DNA integration can be obtained. Unfortunately the identification of transformants without the use of selection can be laborious, and attempts in some transformation systems have been unsuccessful (De Buck et al., 1998). Pooled PCR strategies have been successful for the identification of transformants with considerably reduced efforts (Dominguez et al., 2002). However, the space and labor requirements for the growth and maintenance of most plant species would still be large. A low transformation efficiency would further complicate efforts. *Arabidopsis* provides an obvious solution to many of these problems. Transformation efficiencies by floral dip methods are reported to be as high as 3% (Clough and Bent, 1998), many plants can be grown in a very small space, and the available genome sequence facilitates rapid mapping of integration sites. Furthermore, large databases exist containing integration information for hundreds of thousands of integration events (Sessions et al., 2002; Alonso et al., 2003; Pan et al., 2003). Any integration events identified without selection bias may be compared to these databases to determine if those integrations occurred in regions previously missed by projects involving selection.

In the following chapter, experiments are described that take this approach to addressing the issue of selection bias. Over 100 transgenic *Arabidopsis* plants were identified by PCR and compared to kanamycin-selected transformants from the same T1

seed pool. A higher observed transformation efficiency and a higher frequency of transgene silencing were observed in the PCR-identified lines. Together, the data suggest approximately 30% of integration events may result in non-expressing transgenes that would preclude identification by selection. Genomic integration sites were identified and compared to existing T-DNA integration databases. The integration sites of non-expressing PCR-identified lines mapped to regions with significantly fewer integrations than expressing lines. The data presented in Chapter 2 suggest that T-DNA integration sites may be more evenly distributed across the *Arabidopsis* genome than previously reported.

References

Abranches R, Beven AF, Aragon-Alcaide L, Shaw PJ (1998) Transcription sites are not correlated with chromosome territories in wheat nuclei. J Cell Biol **143:** 5-12

Allen GC, Spiker S, Thompson WF (2000) Use of matrix attachment regions (MARs) to minimize transgene silencing. Plant Mol Biol **43:** 361-376

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D, Carter DE, Marchand T, Risseeuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301:** 653-657

Ambros PF, Matzke AJ, Matzke MA (1986) Localization of *Agrobacterium rizogenes* T-DNA in plant chromosomes by *in situ* hybridization. EMBO J **5:** 2073-2077

***Arabidopsis* Genome Initiative T** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815

Azpiroz-Leehan R, Feldmann KA (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. Trends Genet **13:** 152-156

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci U S A **94:** 6857-6861

Barakat A, Gallois P, Raynal M, Mestre-Ortega D, Sallaud C, Guiderdoni E, Delseny M, Bernardi G (2000) The distribution of T-DNA in the genomes of transgenic *Arabidopsis* and rice. FEBS Lett **471:** 161-164

Bennett MD, Leitch IJ (1995) Nuclear-DNA amounts in angiosperms. Annals of Botany **76:** 113-176

Bode J, Benham C, Knopp A, Mielke C (2000) Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). Crit Rev Eukaryot Gene Expr **10:** 73-90

Bonhomme S, Horlow C, Vezon D, de Laissardiere S, Guyon A, Ferault M, Marchand M, Bechtold N, Pelletier G (1998) T-DNA mediated disruption of essential gametophytic genes in *Arabidopsis* is unexpectedly rare and cannot be inferred from segregation distortion alone. Mol Gen Genet **260:** 444-452

Brown KE, Guest SS, Smale ST, Hahm K, Merkenschlager M, Fisher AG (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. Cell **91:** 845-854

Brunaud V, Balzergue S, Dubreucq B, Aubourg S, Samson F, Chauvin S, Bechtold N, Cruaud C, DeRose R, Pelletier G, Lepiniec L, Caboche M, Lecharny A (2002) T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. EMBO Rep **3:** 1152-1157

Bundock P, den Dulk-Ras A, Beijersbergen A, Hooykaas PJ (1995) Trans-kingdom T-DNA transfer from *Agrobacterium tumefaciens* to *Saccharomyces cerevisiae*. EMBO J **14:** 3206-3214

Bundock P, Hooykaas PJ (1996) Integration of *Agrobacterium tumefaciens* T-DNA in the *Saccharomyces cerevisiae* genome by illegitimate recombination. Proc Natl Acad Sci U S A **93:** 15272-15275

Bundock P, van Attikum H, den Dulk-Ras A, Hooykaas PJ (2002) Insertional mutagenesis in yeasts using T-DNA from *Agrobacterium tumefaciens*. Yeast **19:** 529-536

Campisi L, Yang Y, Yi Y, Heilig E, Herman B, Cassista AJ, Allen DW, Xiang H, Jack T (1999) Generation of enhancer trap lines in *Arabidopsis* and characterization of expression patterns in the inflorescence. Plant J **17:** 699-707

Chen S, Jin W, Wang M, Zhang F, Zhou J, Jia Q, Wu Y, Liu F, Wu P (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. Plant J **36:** 105-113

Chilton MD, Drummond MH, Merlo DJ, Sciaky D, Montoya AL, Gordon MP, Nester EW (1977) Stable incorporation of plasmid DNA into higher plant-cells - molecular basis of crown gall tumorigenesis. Cell 11: 263-271

Chilton MD, Saiki RK, Yadav N, Gordon MP, Quetier F (1980) T-DNA from Agrobacterium Ti Plasmid Is in the Nuclear-DNA Fraction of Crown Gall Tumor-Cells. Proc Natl Acad Sci U S A **77:** 4060-4064

Chin HG, Choe MS, Lee SH, Park SH, Koo JC, Kim NY, Lee JJ, Oh BG, Yi GH, Kim SC, Choi HC, Cho MJ, Han CD (1999) Molecular analysis of rice plants harboring an Ac/Ds transposable element-mediated gene trapping system. Plant J **19:** 615-623

Chyi YS, Jorgensen RA, Goldstein D, Tanksley SD, Loaizafigueroa F (1986) Locations and stability of *Agrobacterium*-mediated transfer DNA insertions in the *Lycopersicon* genome. Mol Gen Genet **204:** 64-69

Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis* thaliana. Plant J **16:** 735-743

Coates D, Taliercio EW, Gelvin SB (1987) Chromatin structure of integrated T-DNA in crown gall tumors. Plant Mol Biol **8:** 159-168

Cockell M, Gasser SM (1999) Nuclear compartments and gene regulation. Curr Opin Genet Dev **9:** 199-205

Combier JP, Melayah D, Raffier C, Gay G, Marmeisse R (2003) *Agrobacterium tumefaciens*-mediated transformation as a tool for insertional mutagenesis in the symbiotic ectomycorrhizal fungus *Hebeloma cylindrosporum*. FEMS Microbiol Lett **220:** 141-148

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, McCombie WR, Martienssen RA, Marra M, Preuss D (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science **286:** 2468-2474

Copenhaver GP, Pikaard CS (1996) RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. Plant J **9:** 259-272

De Buck S, Jacobs A, Van Montagu M, Depicker A (1998) *Agrobacterium tumefaciens* transformation and cotransformation frequencies of *Arabidopsis thaliana* root explants and tobacco protoplasts. Mol Plant-Microbe Int **11:** 449-457

de Groot MJ, Bundock P, Hooykaas PJ, Beijersbergen AG (1998) *Agrobacterium tumefaciens*-mediated transformation of filamentous fungi. Nat Biotechnol **16:** 839-842

de Vetten N, Wolters AM, Raemakers K, van der Meer I, ter Stege R, Heeres E, Heeres P, Visser R (2003) A transformation method for obtaining marker-free plants of a cross-pollinating and vegetatively propagated crop. Nat Biotechnol **21:** 439-442

Dietz A, Kay V, Schlake T, Landsmann J, Bode J (1994) A plant scaffold attached region detected close to a T-DNA integration site is active in mammalian cells. Nucleic Acids Res **22:** 2744-2751

Dominguez A, Fagoaga C, Navarro L, Moreno P, Pena L (2002) Regeneration of transgenic citrus plants under non selective conditions results in high-frequency recovery of plants with silenced transgenes. Mol Genet Genomics **267:** 544-556

Ebinuma H, Sugita K, Matsunaga E, Yamakado M (1997) Selection of marker-free transgenic plants using the isopentenyl transferase gene. Proc Natl Acad Sci U S A **94:** 2117-2121

Feldmann KA (1991) T-DNA insertion mutagenesis in *Arabidopsis* - mutational spectrum. Plant J **1:** 71-82

Fobert PR, Miki BL, Iyer VN (1991) Detection of gene regulatory signals in plants revealed by T-DNA-mediated fusions. Plant Mol Biol **17:** 837-851

Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R (2003) A comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis thaliana* genome. Plant Mol Biol **52:** 161-176

Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G (1998) Cytogenetics for the model system *Arabidopsis thaliana*. Plant J **13:** 867-876

Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I (2002) Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. Proc Natl Acad Sci U S A **99:** 14584-14589

Gallagher SR (1992) Quantitation of GUS activity by fluorometry. *In* SR Gallagher, ed, GUS Protocols: Using the GUS Gene as a Reporter of Gene Expression. Academic Press, Inc., San Diego, pp 47-59

Gasser SM, Cockell MM (2001) The molecular biology of the SIR proteins. Gene **279:** 1-16

Goldsbrough AP, Lastrella CN, Yoder JI (1993) Transposition mediated repositioning and subsequent elimination of marker genes from transgenic tomato. Bio-Technology **11:** 1286-1292

Gouka RJ, Gerk C, Hooykaas PJ, Bundock P, Musters W, Verrips CT, de Groot MJ (1999) Transformation of *Aspergillus awamori* by *Agrobacterium tumefaciens*-mediated homologous recombination. Nat Biotechnol **17:** 598-601

Hamada T, Sasaki H, Seki R, Sakaki Y (1993) Mechanism of chromosomal integration of transgenes in microinjected mouse eggs: sequence analysis of genome-transgene and transgene-transgene junctions at two loci. Gene **128:** 197-202

Hatada S, Kuziel W, Smithies O, Maeda N (1999) The influence of chromosomal location on the expression of two transgenes in mice. J Biol Chem **274:** 948-955

Herman L, Jacobs A, Van Montagu M, Depicker A (1990) Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. Mol Gen Genet **224:** 248-256

Herskowitz I, Rine J, Strathern JN (1991) Mating-type determination and mating-type interconversion in *Saccharomyces cerevisiae*. *In* EW Jones, JR Pringle, JR Broach, eds, The Molecular and Cellular Biology of the Yeast *Saccharomyces*, Vol 2. Cold Spring Harbor Laboratory Press, Plainview, New York

Heslop-Harrison JS, Lietch AR, Schwarzacher T (1993) The physical organization of interphase nuclei. *In* JS Heslop-Harrison, RB Flavell, eds, The Chromosome. Bios Scientific Publishers Ltd., Oxford, pp 221-232

Ichikawa T, Nakazawa M, Kawashima M, Muto S, Gohda K, Suzuki K, Ishikawa A, Kobayashi H, Yoshizumi T, Tsumoto Y, Tsuhara Y, Iizumi H, Goto Y, Matsui M (2003) Sequence database of 1172 T-DNA insertion sites in *Arabidopsis* activation-tagging lines that showed phenotypes in T1 generation. Plant J **36:** 421-429

Ito T, Motohashi R, Kuromori T, Mizukado S, Sakurai T, Kanahara H, Seki M, Shinozaki K (2002) A new resource of locally transposed Dissociation elements for screening gene-knockout lines in silico on the *Arabidopsis* genome. Plant Physiol **129:** 1695-1699

Ito T, Shinozaki K (2002) Random insertional mutagenesis in *Arabidopsis*. *In* JS M., ed, Molecular Techniques in Crop Improvement. Kluwer Academic Publishers, Netherlands, pp 409-425

Jefferson RA (1987) Assaying chimeric genes in plants: the GUS gene fusion system. Plant Mol Biol Rep **5:** 387-405

Karimi M, Van Montagu M, Gheysen G (1999) Hairy root production in *Arabidopsis thaliana*: cotransformation with a promoter-trap vector results in complex T-DNA integration patterns. Plant Cell Rep **19:** 133-142

Kasulke D, Seitz S, Ehrenhofer-Murray AE (2002) A role for the *Saccharomyces cerevisiae* RENT complex protein Net1 in HMR silencing. Genetics **161:** 1411-1423

Kertbundit S, De Greve H, Deboeck F, Van Montagu M, Hernalsteens JP (1991) In vivo random beta-glucuronidase gene fusions in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **88:** 5212-5216

Kolesnik T, Szeverenyi I, Bachmann D, Kumar CS, Jiang S, Ramamoorthy R, Cai M, Ma ZG, Sundaresan V, Ramachandran S (2004) Establishing an efficient Ac/Ds tagging system in rice: large-scale analysis of Ds flanking sequences. Plant J **37:** 301-314

Koncz C, Martini N, Mayerhofer R, Koncz-Kalman Z, Korber H, Redei GP, Schell J (1989) High-frequency T-DNA-mediated gene tagging in plants. Proc Natl Acad Sci U S A **86:** 8467-8471

Krysan PJ, Young JC, Jester PJ, Monson S, Copenhaver G, Preuss D, Sussman MR (2002) Characterization of T-DNA insertion sites in *Arabidopsis thaliana* and the implications for saturation mutagenesis. Omics **6:** 163-174

Krysan PJ, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell **11:** 2283-2290

Kunik T, Tzfira T, Kapulnik Y, Gafni Y, Dingwall C, Citovsky V (2001) Genetic transformation of HeLa cells by *Agrobacterium*. Proc Natl Acad Sci U S A **98:** 1871-1876

Lindsey K, Wei W, Clarke MC, McArdle HF, Rooke LM, Topping JF (1993) Tagging genomic sequences that direct transgene expression by activation of a promoter trap in plants. Transgenic Res **2:** 33-47

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

Lubahn DB, Moyer JS, Golding TS, Couse JF, Korach KS, Smithies O (1993) Alteration of reproductive function but not prenatal sexual development after insertional disruption of the mouse estrogen receptor gene. Proc Natl Acad Sci U S A **90:** 11162-11166

Makarevitch I, Svitashev SK, Somers DA (2003) Complete sequence analysis of transgene loci from plants transformed via microprojectile bombardment. Plant Mol Biol **52:** 421-432

Martirani L, Stiller J, Mirabella R, Alfano F, Lamberti A, Radutoiu SE, Iaccarino M, Gresshoff PM, Chiurazzi M (1999) T-DNA tagging of nodulation- and root-related genes in *Lotus japonicus*: Expression patterns and potential for promoter trapping and insertional mutagenesis. Mol Plant-Microbe Int **12:** 275-284

Mathur J, Szabados L, Schaefer S, Grunenberg B, Lossow A, Jonas-Straube E, Schell J, Koncz C, Koncz-Kalman Z (1998) Gene identification with sequenced T-DNA tags generated by transformation of *Arabidopsis* cell suspension. Plant J **13:** 707-716

Matzke MA, Matzke A (1995) How and why do plants inactivate homologous (trans)genes? Plant Physiol **107:** 679-685

Mitsuhara I, Shirasawa-Seo N, Iwai T, Nakamura S, Honkura R, Ohashi Y (2002) Release from post-transcriptional gene silencing by cell proliferation in transgenic tobacco plants: possible mechanism for noninheritance of the silencing. Genetics **160:** 343-352

Mullins ED, Chen X, Romaine P, Raina R, Geiser DM, Kang S (2001) *Agrobacterium-mediated* transformation of *Fusarium oxysporum*: an efficient tool for insertional mutagenesis and gene transfer. Phytopathology **91:** 717 - 726

Nakanishi T, Kuroiwa A, Yamada S, Isotani A, Yamashita A, Tairaka A, Hayashi T, Takagi T, Ikawa M, Matsuda Y, Okabe M (2002) FISH analysis of 142 EGFP transgene integration sites into the mouse genome. Genomics **80:** 564-574

Olson MV (1991) Genome structure and organization in *Saccharomyces cerevisiae*. *In* JR Broach, JR Pringle, EW Jones, eds, The Molecular and Cellular Biology of the Yeast *Saccharomyces*, Vol 1. Cold Spring Harbor Laboratory Press, Plainview, New York, pp 1 - 40

Osborne BI, Wirtz U, Baker B (1995) A system for insertional mutagenesis and chromosomal rearrangement using the Ds transposon and Cre-Lox. Plant J **7:** 687-701

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science **302:** 2115-2117

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. Nucleic Acids Res **31:** 1245-1251

Parinov S, Sevugan M, Ye D, Yang WC, Kumaran M, Sundaresan V (1999) Analysis of flanking sequences from dissociation insertion lines: a database for reverse genetics in *Arabidopsis*. Plant Cell **11:** 2263-2270

Permingeat HR, Alvarez ML, Cervigni GD, Ravizzini RA, Vallejos RH (2003) Stable wheat transformation obtained without selectable markers. Plant Mol Biol **52:** 415-419

Petes TD, Malone RE, Symingtone LS (1991) DNA recombination. *In* JR Broach, JR Pringle, EW Jones, eds, The Molecular and Cellular Biology of the Yeast *Saccharomyces*. Cold Spring Harbor Laboratory Press, Plainview, New York, pp 407-521

Plesch G, Kamann E, Mueller-Roeber B (2000) Cloning of regulatory sequences mediating guard-cell-specific gene expression. Gene **249:** 83-89

Qin H, Dong Y, von Arnim AG (2003) Epigenetic interactions between *Arabidopsis* transgenes: characterization in light of transgene integration sites. Plant Mol Biol **52:** 217-231

Raizada MN, Nan GL, Walbot V (2001) Somatic and germinal mobility of the RescueMu transposon in transgenic maize. Plant Cell **13:** 1587-1608

Rolland S, Jobic C, Fevre M, Bruel C (2003) *Agrobacterium*-mediated transformation of *Botrytis cinerea*, simple purification of monokaryotic transformants and rapid conidia-based identification of the transfer-DNA host genomic DNA flanking sequences. Curr Genet **44:** 164-171

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

Sawasaki T, Takahashi M, Goshima N, Morikawa H (1998) Structures of transgene loci in transgenic *Arabidopsis* plants obtained by particle bombardment: junction regions can bind to nuclear matrices. Gene **218:** 27-35

Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, Clarke JD, Cotton D, Bullis D, Snell J, Miguel T, Hutchison D, Kimmerly B, Mitzel T, Katagiri F, Glazebrook J, Law M, Goff SA (2002) A high-throughput *Arabidopsis* reverse genetics system. Plant Cell **14:** 2985-2994

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res **23:** 1087-1088

Smith K (2001) Theoretical mechanisms in targeted and random integration of transgene DNA. Reprod Nutr Dev **41:** 465-485

Sonti RV, Chiurazzi M, Wong D, Davies CS, Harlow GR, Mount DW, Signer ER (1995) *Arabidopsis* mutants deficient in T-DNA integration. Proc Natl Acad Sci U S A **92:** 11786-11790

Sutton MD, Walker GC (2001) Managing DNA polymerases: coordinating DNA replication, DNA repair, and DNA recombination. Proc Natl Acad Sci U S A **98:** 8342-8349

Svejstrup JQ (2002) Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol **3:** 21-29

Szabados L, Kovacs I, Oberschall A, Abraham E, Kerekes I, Zsigmond L, Nagy R, Alvarado M, Krasovskaja I, Gal M, Berente A, Redei GP, Haim AB, Koncz C (2002) Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. Plant J **32:** 233-242

Thomas CM, Jones DA, English JJ, Carroll BJ, Bennetzen JL, Harrison K, Burbidge A, Bishop GJ, Jones JD (1994) Analysis of the chromosomal distribution of transposon-carrying T-DNAs in tomato using the inverse polymerase chain reaction. Mol Gen Genet **242:** 573-585

Thomas KR, Capecchi MR (1987) Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell **51:** 503-512

Tinland B (1996) The integration of T-DNA into plant genomes. Trends Plant Sci **1:** 178-184

Topping JF, Lindsey K (1995) Insertional mutagenesis and promoter trapping in plants for the isolation of genes and the study of development. Transgenic Res **4:** 291-305

Topping JF, Wei W, Lindsey K (1991) Functional tagging of regulatory elements in the plant genome. Development **112:** 1009-1019

Tzfira T, Citovsky V (2002) Partners-in-infection: host proteins involved in the transformation of plant cells by *Agrobacterium*. Trends Cell Biol **12:** 121-129

van Attikum H, Bundock P, Overmeer RM, Lee LY, Gelvin SB, Hooykaas PJ (2003) The *Arabidopsis* AtLIG4 gene is required for the repair of DNA damage, but not for the integration of *Agrobacterium* T-DNA. Nucleic Acids Res **31:** 4247-4255

van Holde K (1989) Chromatin. Springer-Verlag, New York

Walden R (2002) T-DNA tagging in a genomics era. Crit Rev Plant Sci **21:** 143-165

Wallroth M, Gerats AGM, Rogers SG, Fraley RT, Horsch RB (1986) Chromosomal localization of foreign genes in *Petunia hybrida*. Mol Gen Genet **202:** 6-15

Willmitzer L, Debeuckeleer M, Lemmers M, Vanmontagu M, Schell J (1980) DNA from Ti plasmid present in nucleus and absent from plastids of crown gall plant-cells. Nature **287:** 359-361

Wolffe AP (1998) Chromatin: Structure and Function, Ed 3rd. Academic Press, San Diego

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MM, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, Banh J, Banno F, Bowser L, Brooks S, Carninci P, Chao Q, Choy N, Enju A, Goldsmith AD, Gurjal M, Hansen NF, Hayashizaki Y, Johnson-Hopson C, Hsuan VW, Iida K, Karnes M, Khan S, Koesema E, Ishida J, Jiang PX, Jones T, Kawai J, Kamiya A, Meyers C, Nakajima M, Narusaka M, Seki M, Sakurai T, Satou M, Tamse R, Vaysberg M, Wallender EK, Wong C, Yamamura Y, Yuan S, Shinozaki K, Davis RW, Theologis A, Ecker JR (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science **302:** 842-846

Ye GN, Stone D, Pang SZ, Creely W, Gonzalez K, Hinchee M (1999) *Arabidopsis* ovule is the target for *Agrobacterium* in planta vacuum infiltration transformation. Plant J **19:** 249-257

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science **296:** 79-92

Zink D, Bornfleth H, Visser A, Cremer C, Cremer T (1999) Organization of early and late replicating DNA in human chromosome territories. Exp Cell Res **247:** 176-188

Zubko E, Scutt C, Meyer P (2000) Intrachromosomal recombination between attP regions as a tool to remove selectable marker genes from tobacco transgenes. Nat Biotechnol **18:** 442-445

Chapter 2

The Identification of *Arabidopsis thaliana* Transformants Without Selection Reveals a High Occurrence of Silenced T-DNA Integrations

Introduction

Agrobacterium tumefaciens-mediated integration of T-DNA sequences into plant genomes has recently been exploited as an insertional mutagen in several plant species (Krysan et al., 1999; Chen et al., 2003). Insertional mutagenesis has proven to be very effective in *Arabidopsis thaliana* and has been responsible (in one case) for insertions in or near ~74% of the predicted *Arabidopsis* genes (Alonso et al., 2003). Projects to map the locations of T-DNA integrations by sequencing flanking genomic regions have provided a genome-wide view of T-DNA integration (Sessions et al., 2002; Szabados et al., 2002; Alonso et al., 2003; Ichikawa et al., 2003; Pan et al., 2003). These and other efforts have repeatedly demonstrated that identified sites of transgene integration via *Agrobacterium* transformation do not appear to be randomly distributed across the *Arabidopsis* genome. Specifically, fewer T-DNA integrations have been observed in centromeric, non-genic, and other regions than would be expected if integration were random (Brunaud et al., 2002; Krysan et al., 2002; Szabados et al., 2002; Alonso et al., 2003). This non-random distribution is frequently suggested to be a consequence of a preference for T-DNA integration into genic regions or open chromatin (Barakat et al., 2000). These suggestions are consistent with long held beliefs that T-DNA preferentially targets actively transcribed genes (Koncz et al., 1989; Topping et al., 1991; Lindsey et al., 1993) and that non-genic regions of heterochromatin may physically prevent T-DNA integration (Herman et al., 1990; Topping and Lindsey, 1995). These views are in

contrast to reports that T-DNA integration is random (Fobert et al., 1991; Thomas et al., 1994; Azpiroz-Leehan and Feldmann, 1997; Forsbach et al., 2003).

There are, in fact, two possible explanations for the observed non-randomness of mapped T-DNA integration sites. It is possible that genic regions are preferred sites for T-DNA integration (either through a preference for genic regions or an avoidance of non-genic or heterochromatic regions). However, it is also possible that T-DNA integration occurs randomly throughout the entire genome, but integrations into some regions are not recognized due to the repressive nature of the surrounding chromatin or nearby transcriptional regulators. If the selectable or screenable markers contained in the T-DNA are not expressed, those transgenic plants will not be identified. The result would be a pattern of identified integration sites with a bias for genic or active chromatin regions. We refer to this phenomenon as selection bias, and believe that selection bias is at least partially (if not entirely) responsible for the perceived non-randomness of T-DNA integration.

Recently, several transformation systems have been developed that avoid the use of selectable markers (Dominguez et al., 2002; de Vetten et al., 2003; Permingeat et al., 2003). These systems use PCR to identify transformants regenerating in the absence of selection and are therefore free from selection bias. It is very interesting that some of these selection-free systems have reported a high incidence of transcriptional gene silencing (Dominguez et al., 2002).

We have conducted a screen to identify transgenic *Arabidopsis* seedlings based on the presence of integrated transgenes, not their activities. Our screen revealed that more transgenic seedlings are recovered when selectable marker expression is not a prerequisite for identification. We also demonstrate that PCR-identified lines are more likely to contain inactive or silenced transgenes than kanamycin-selected lines. Many

identified transgenic plants contain inactive transgenes that appear to have integrated into regions characterized by a low density of previously mapped T-DNA integrations (Alonso et al., 2003). The data presented here suggest that many successful transformation events are not identified by conventional methods and that T-DNA integration may be more nearly random than previously thought.

Materials and Methods

Agrobacterium transformation

All *Agrobacterium tumefaciens* transformations were conducted in strain GV3101:pMP90 (Koncz and Schell, 1986). All transformations were generated by electroporation. Electrocompetent *Agrobacterium* was prepared by growing a 100 ml culture of GV3101:pMP90 in YEP medium (10g/l yeast extract, 10g/l peptone, and 5g/l sodium chloride) at 28°C until the OD at 600nm was between 0.500 and 0.700. Cells were chilled on ice and centrifuged at 4°C for 10min at 1200g (2500rpm in Beckman AccuSpinFR). Pellets were resuspended in 10ml ice cold 10% glycerol by vortexing vigorously and centrifuged as above. This wash was repeated four additional times to ensure complete removal of salts. After the fifth wash, the cells were resuspended in 1ml ice cold 10% glycerol, aliquoted to 100µl volumes, rapidly frozen in liquid nitrogen, and stored at -80°C.

Electrocompetent GV3101:pMP90 was transformed with binary vector pBI121 (Jefferson et al., 1987). Electrocompetent cells were thawed on ice, and 40µl were added to an ice cold, sterile Gene Pulser cuvette with a 0.2cm gap (BioRad). Two µl of 100ng/µl pBI121 were added directly to the *Agrobacterium* in the cuvette. The cuvette was then electroporated at 2.50kV, 25µF, and 600Ω. One ml of ice cold YEP was

immediately added to the cuvette, and transformed cells were allowed to recover for 2hrs at 28°C with gentle shaking. After recovery, 50µl were plated on YEP with 50mg/l kanamycin. Plates were kept at 28°C for 48hrs. Kanamycin-resistant colonies were selected and grown as individual cultures. Binary vector plasmid preparations were carried out using a QIA-spin miniprep kit (Qiagen). The manufacturer's instructions were followed except for the modification of adding 15mg lysozyme to each ml of P1 buffer, and a 5min incubation step was added after the resuspension of the bacterial pellet in the modified P1 buffer. The presence of the pBI121 binary vector was confirmed by restriction digestion of plasmid preparations.

Arabidopsis transformation

Arabidopsis transformation was performed by the floral dip method, essentially as described in Clough and Bent (1998). Twenty *Arabidopsis thaliana* (var. Columbia) plants were grown at 22°C in long-day (16hr light, 8hr dark) conditions for 6 weeks. Initial floral bolts were cut back to promote secondary bolting. All siliques and opened flowers were removed.

A 5ml culture of *Agrobacterium* strain GV3101:pMP90 (Koncz and Schell, 1986) containing the binary vector pBI121 was started from a frozen stock and grown in YEP medium overnight at 28°C. This culture was then used to start a 50ml culture in YEP medium grown under similar conditions. The 50ml culture was grown for 16hrs or until the OD at 600nm was approximately 1.000. The culture was then centrifuged at 1200g (2500rpm in Beckman AccuSpinFR) for 10min at room temperature, and the pellet was resuspended in an equal volume of a 5% sucrose solution. Silwet L-77 (Lehle Seeds, Round Rock, TX) was then added to a final concentration of 0.05%. Flowering plants were inverted and the unopened and developing flowers were immersed in the

Agrobacterium solution. A pipette was used to apply the *Agrobacterium* solution directly to any flowers that were not immersed. Plants were then placed in a sealed transparent container to maintain high humidity for 48hrs and returned to the long-day growth chamber. After 48hrs the container was opened. Seven days after the initial floral dip, the procedure was repeated to target new floral buds. No siliques or flowers were removed for the second dip.

Plants were allowed to grow and produce seed for four to five weeks after the second dip, then water was withheld from the plants to promote seed drying. Seed was collected 2 to 3 weeks later. An estimated 10,000 seeds (the total from all twenty dipped plants) were combined into a single 1.5ml microcentrifuge tube. The seeds were then repeatedly and thoroughly mixed in the tube to ensure that subsequent sampling resulted in a random sample of seeds from the original twenty T0 plants. Repeated sampling consistently resulted in transformation frequencies via kanamycin selection of around 2.5%.

Seed sterilization

All seed was sterilized by one of two methods. If a large number of seeds (>500) from relatively few lines were needed, seeds were sterilized via a wet sterilization procedure. About 500 seeds were transferred to a clean microcentrifuge tube, and 1ml of sterilization solution (20% commercial bleach, 0.01% Triton X-100) was added. Tubes were then placed on a shaker at 100rpm for 15 minutes. The sterilization solution was then removed and replaced with 1ml of fresh sterilization solution, and the tubes were returned to the shaker for 5 minutes. Following the second wash, the sterilization solution was removed, and the seeds were rinsed six times with sterile water. This

method was used to sterilize T1 seed used in kanamycin selection and PCR-identification screens.

If only a few seeds (<100) were needed from many lines, seeds were sterilized by a chlorine gas vapor-phase procedure (Clough and Bent, 1998). About 100 seeds from each line were transferred to clean microcentrifuge tubes. The tubes were then placed in racks with their caps open inside a vacuum bell inside a fume hood. A beaker containing 100ml of commercial bleach was placed in the vacuum bell. Three ml of concentrated HCl was added directly to the bleach, and the lid was immediately placed on the vacuum bell. After 4hrs, the beaker was removed from the vacuum bell in the fume hood, and the open tubes were moved to a laminar flow hood to briefly ventilate in a clean environment. After 15min, the tubes were capped and either immediately used or stored for up to 4 weeks. This method was used to sterilize T2 seed used in T2 kanamycin resistance assays.

Identification of transgenic T1 seedlings by PCR

A schematic of the pooled sample PCR-screening approach used is presented in Figure 2-2. Samples of seed from the pool of T1 seed described above were plated on Seed Germination (SG) medium [MS salts (Gibco) at 0.5X concentration, 5g/l sucrose, pH 5.7, solidified with 6g/l phytagar (Gibco)] containing 200mg/l Timentin. A low sucrose concentration was used to slow the growth rate of any contaminants that may result from repeated sampling and transfers expected in the PCR screening. Either 49 seeds were plated in a 7 X 7 grid (Screen 1), or 36 seeds were plated in a 6 X 6 grid (Screens 2, 3, 4, and 7). Plates were wrapped with micropore tape (3M) and placed in a dark 4°C chamber for 4 days. After cold treatment, plates were transferred to a 22°C short-day (12hr light, 12hr dark) chamber, where seedlings germinated within 24hrs.

When seedlings had reached a stage where the first two true leaves were approximately 3mm across and the 3rd and 4th true leaves were beginning to emerge (10 to 14 days after germination), tissues were sampled. From each seedling, the two 3mm leaves were excised, pooled with the other leaves from the same row and column to form row pools and column pools, placed in a microcentrifuge tube containing ~0.2ml of 1mm glass beads, and frozen in liquid nitrogen.

Pooled tissues were removed from liquid nitrogen and immediately homogenized for 5sec using a dental amalgamator. After homogenization, the fine powder was suspended in 0.5ml DNA Extraction Buffer [1.4M NaCl, 20mM EDTA, 100mM Tris-HCl (pH 8.0), 3% CTAB (Calbiochem)] by an additional 5 second treatment in the dental amalgamator. After a 30 minute incubation at 68°C, an equal volume of chloroform was added, and the phases were separated by centrifugation. DNA was precipitated from the aqueous phase with an equal volume of isopropanol and resuspended in 100µl sterile water.

PCR was performed on each DNA sample of pooled tissues. For each reaction, 1µl of isolated DNA template was used in a 25µl reaction with 200µM of each dNTP, 0.5 units of taq polymerase (Fisher), and Reaction Buffer A (Fisher) to a concentration of 1X resulting in a final Mg concentration of 1mM. Primers were added to final concentration of 400nM. Each reaction consisted of two primer sets, which resulted in two distinct product sizes. To verify the presence of T-DNA, primers A (5'-atgacgcacaatcccactat-3') and B (5'-gtggtgtagagcattacgct-3') were used to generate a 648bp product corresponding to a region in the center of the T-DNA (see Figure 2-1). As a control to verify that *Arabidopsis* genomic DNA is present and of amplifiable quality, primers DDM1 (5'-caccttttcttttgcgtccac-3') and DDM2 (5'-tggggtgttctgtaaatgggtc-3') were used to generate a 490bp product corresponding to the single-copy *Arabidopsis* gene *ddm1* (AY333120)

(Jeddeloh et al., 1998). Alternatively, an internal control corresponding to a 200bp fragment of *HMGA* (X99116) was amplified with HMGup (5'-gctaatcatggtgaagaac-3') and HMGdown (5'-ccggtaatgaccttctcctgg-3'). The amplifications were carried out as follows: 94°C for 5min, followed by 35 cycles of 94°C for 30sec, 60°C for 30sec, and 72°C for 60sec. This was followed by 5min at 72°C and a hold at 4°C. PCR products were separated by gel electrophoresis in 1% agarose gels containing ethidium bromide (20ng/ml) run at 100 volts. Following separation, gels were imaged using Gel Doc 2000 imaging system (BioRad). Seedlings that corresponded to positive T-DNA amplification for both row pools and column pools were resampled and amplified by PCR to confirm T-DNA amplification. A single leaf from each candidate seedling was harvested, and DNA was isolated as described above. For situations where initial amplification was detected only in row pools but not column pools (or *vice versa*), all seedlings in the positive row or column were individually sampled and amplified by PCR to confirm T-DNA amplification. Seedlings that individually tested positive for T-DNA amplification were transferred to Seed Germination Enhanced (SGE) medium [MS salts (Gibco) at 1X concentration, 10g/l sucrose, pH 5.7, solidified with 6g/l phytagar (Gibco)] containing 200mg/l Timentin in 20mm deep plates. Amplification of *VirG* (AE007924) for the detection of potentially contaminating *Agrobacterium* was achieved with primers *VirG*-up (5'-gcaatgatttctctcaactgctcg-3') and *VirG*-dn (5'-gatttcagacgatagccctggtaac-3').

Identification of transgenic T1 seedlings by kanamycin selection

Samples of sterilized seed from the same T1seed pool used for PCR-screening were plated on SG medium containing 200mg/l Timentin and 50mg/l kanamycin. Between 100 and 200 seeds were plated on each plate. Plates were wrapped with micropore tape (3M) and placed in a dark 4°C chamber for 4 days. After cold treatment,

plates were transferred to a 22°C short-day (12hr light, 12hr dark) chamber, where they germinated within 24hrs. Kanamycin tolerance was scored at two weeks. Any seedling that showed any degree of kanamycin tolerance above that of untransformed wt seedlings was considered to be kanamycin resistant and was transferred to SGE medium containing 200mg/l Timentin and no kanamycin in 20mm deep plates.

Culture of transgenic seedlings

PCR-identified and kanamycin-selected seedlings were cultured in SGE medium containing 200mg/l Timentin in 20mm deep plates. At five to six weeks (post-germination) seedlings were transferred to 1.5inch plug trays containing Metro-Mix (Scotts). Seedlings were grown under clear plastic domes for one week and then acclimated to chamber humidity for one week by gradually sliding back the clear plastic domes. At seven weeks post-germination, trays were moved to long-day (16hr light, 8hr dark) growth chamber to promote floral bolting. Plants were allowed to self-fertilize and seed was collected for each line.

Characterization of T-DNA by PCR

The presence of non-truncated reporter genes was verified by PCR with primers C (5'-aggctattcggctatgactggc-3') and D (5'-tttcatagatggcggcggtg-3') for *nptII* verification and primers E (5'-gatagtggaaaaggaaggtggctc-3') and F (5'-ttgttgattcattgtttgcctcc-3') for *gusA* verification. The structure and complexity of T-DNA integration was also partially assessed by PCR. Tandem repeats of T-DNA units have been previously reported as a common artifact of *Agrobacterium* transformation. Tandem repeats can be identified by PCR amplification using a forward primer near the right border of the T-DNA and a reverse primer near the left border. All transgenic plants were assayed for the presence of

direct tandem repeats using primers G (5'-tgatagtgaccttaggcgac-3') and H (5'-gaaaaccctggcgttacc-3') (see Figure 2-1). A product resulted only if T-DNAs integrated as direct tandem repeats. Inverted T-DNA repeats will not produce a PCR product, probably because an inverted repeat hairpin structure will out-compete the primers for binding (De Buck et al., 1999).

The presence of non-T-DNA binary vector sequence was identified by PCR amplification with primers I (5'-tgtttatcggcagttcgtagagc-3') and J (5'-tgtggcagcaggtgttggag-3'), which amplify a 672bp fragment in the pBI121 vector approximately 2kb outside of the left border. The amplification of an I/J fragment indicated that vector sequence was present in the plant, but did not necessarily indicate that vector sequence was adjacent to the T-DNA.

Screening for kanamycin tolerance in T1 seedling tissue

To determine if T1 seedlings were kanamycin tolerant, samples of T1 tissues were cultured *in vitro* in the presence of kanamycin using two different approaches. For the first approach, a previously described method (Mathur and Koncz, 1998) of generating and maintaining *Arabidopsis* callus was modified for this purpose. One or two leaves from each four to six-week-old *in vitro* grown seedlings (grown in the absence of kanamycin) were harvested. The base of each leaf was excised with a scalpel, removing the entire petiole, but leaving the remainder of the leaf intact. The surface of each leaf was then lightly scarred with a scalpel tip, and the leaf was placed adaxial side down on CM medium [MS salts (Gibco) at 1X concentration, 30 g/l sucrose, 0.5 mg/l 6-(γ,γ-dimethylallylamino)-purine riboside (IPAR)(Sigma), 0.5 mg/l 2,4-dichlorophenoxyacetic acid (2,4-D)(Sigma), 2 mg/l indole-3-acetic acid (IAA)(Sigma), solidified with 6g/l phytagar (Gibco)] containing 50 mg/l kanamycin. Cultures were then placed in short-day

growth chambers in conditions previously described and assayed for callus production in three to four weeks. In the presence of kanamycin, leaves from transgenic *Arabidopsis* expressing *nptII* were able to produce callus on this medium, while wild-type *Arabidopsis* leaves were not.

A second in vitro assay for *nptII* activity involved the ability of excised leaves to regenerate roots in the presence of kanamycin. It was observed that leaves from four to six-week-old wild-type *Arabidopsis* seedlings could produce roots from the bases of petioles when placed on MS medium (Murashige and Skoog, 1962). One or two leaves from each seedling were harvested so that the petioles remained in tact. The leaves with attached petioles were then placed adaxial side down in MS medium [MS salts at 1X concentration (Gibco), 30 g/l sucrose, solidified with 6 g/l phytagar (Gibco)] with 50 mg/l kanamycin. Cultures were then placed in short-day growth chambers in conditions previously described and assayed for petiolar root growth in three to four weeks. In the presence of kanamycin, leaves from transgenic *Arabidopsis* expressing *nptII* were able to produce roots on this medium, while leaves from wild type *Arabidopsis* were not.

Segregation of kanamycin tolerance in T2 seedlings

Kanamycin tolerance in T2 seedlings was measured by plating between 18 and 72 T2 seed from PCR-identified and kanamycin-selected lines on SG medium containing 50mg/l kanamycin. Seeds were plated as described above, and kanamycin tolerance was scored two weeks after germination. Seedlings displaying tolerance at any level above that of untransformed *Arabidopsis* were considered to be demonstrating some degree of kanamycin tolerance. Kanamycin tolerance frequencies for each line were calculated based on the number of germinated seedlings showing kanamycin tolerance. For each line with a kanamycin tolerance frequency lower than 0.75, a chi square goodness of fit

test was conducted against the predicted 3:1 segregation at the $\alpha = 0.01$ level. Lines with significantly fewer kanamycin tolerant seedlings (lines that failed the chi square test) were considered to be silenced for *nptII*. Lines with frequencies of kanamycin tolerance significantly higher than the predicted 3:1 segregation ratio were likely the result of multiple unlinked integration events. Because a very liberal definition of kanamycin tolerance was applied, we consider this a very conservative estimate of silencing. All other lines were considered to be *nptII*-expressing.

Qualitative Measurement of GUS activity

GUS activity in leaf tips of PCR-identified and kanamycin-selected T1 seedlings was analyzed by X-gluc histochemical staining as described by Jefferson et al. (1987). When seedlings were approximately 3-weeks-old, the distal 2mm of the largest leaf was excised from each seedling. Leaf samples were covered with X-gluc buffer [100mM sodium phosphate (pH 7.0), 10mM EDTA, 500µM potassium ferricyanide, 500µM potassium ferrocyanide, 0.1% Triton-X100, 0.5mg/ml 5-bromo-4-chloro-3-indoyl glucuronide] and incubated for 24hrs at 37°C. X-gluc buffer was then removed by pipetting, and 70% ethanol was added to cover each sample. After 4 hrs, 70% ethanol was removed by pipetting, and fresh 70% ethanol was added to cover each sample. Tissues were then analyzed by visual assessment. The presence of blue pigment in the leaf tissue was considered to be a positive indicator of GUS activity.

Quantitative Measurement of GUS activity

GUS activity was measured by monitoring cleavage of the β-glucuronidase substrate 4-methylumbelliferyl β-D-glucuronide (MUG) (Jefferson et al., 1987; Gallagher, 1992). The assay was adapted so that large numbers of samples could be

assayed and measured in a 96-well plate format. Tissues were collected into microcentrifuge tubes containing ~200µl of 1mm glass beads and frozen in liquid nitrogen. Tissues were homogenized using a dental amalgamator. Samples were removed from liquid nitrogen, immediately placed in the amalgamator, and homogenized for 5sec. After homogenization, 1ml GUS Extraction Buffer [150mM Sodium Phosphate pH 7.0, 10mM EDTA, 10mM β-mercaptoethanol, 0.1% Triton X-100, 0.1% sarcosyl, 140µM PMSF] was added, and samples were mixed by an additional 5sec in the amalgamator. Debris was pelleted by centrifugation at 13000rpm in microcentrifuge for 15min at room temperature. For each sample, 200µl of supernatant was collected and kept on ice for use in MUG assays and for protein quantification.

The assay reaction consisted of 10µl sample extract and 130µl Assay Buffer [GUS Extraction Buffer containing 1.2mM 4-methylumbelliferyl β-D-glucuronide (MUG)(Sigma)]. The reaction was carried out in a dark 37°C water bath. After 20 minutes, 10µl of the reaction was transferred to 190µl Stop Buffer [200mM sodium carbonate] in an opaque 96-well plate. Fluorescence was measured on a FLUO-star plate reader (BMG Labtechnologies Inc., Durham, NC) at 460nm when excited at 355nm. A standard curve corresponding to 50, 25, 5, 2.5, 0.5, 0.25, and 0µM 4-Methylumbelliferone (MU) was included with every plate,and used to calculate the amount of liberated MU produced by each sample.

Values from the fluorescence assay were converted to moles of MU/minute, and then standardized by protein concentration to accommodate differences in tissue sampling and extraction. Protein concentrations were determined by methods described by Bradford (1976) and analyzed on a FLUO-star plate reader (BMG Labtechnologies Inc., Durham, NC) set to measure absorbance at 600nm. Final GUS activity values were recorded as pMoles MU/min/mg protein.

Detection of neomycin phosphotransferase using ELISA

Quantitative measurements of neomycin phosphotransferase were determined for a subset of PCR-identified lines. Pools of 10 *in vitro* grown T2 seedlings were collected from plates containing SG medium without kanamycin at seven days after germination. Samples were frozen in liquid nitrogen and homogenized using a dental amalgamator as described for GUS quantification above. Concentrations of neomycin phosphotransferase were determined using an NPTII ELISA kit (Agdia Inc., Elkhart, IN). The procedure included with the kit was followed with minor modifications. Because the neomycin phosphotransferase concentration was very high in some sample extracts, all samples were diluted to be within the sensitivity range of the assay. Samples were extracted into 1ml PEB1 buffer (provided with the NptII ELISA kit) and then diluted 1:1 with PEB1 resulting in a 0.5X sample concentration. All samples were measured in duplicate, and empty wells were positioned between samples to prevent cross-contamination during washes. Plates were read on a FLUO-star plate reader (BMG Labtechnologies Inc., Durham, NC) set to measure absorbance at 450nm. Samples were standardized by protein concentration to accommodate differences in tissue sampling and extraction. Because of buffer incompatibilities, protein concentrations could not be determined by the methods of Bradford (1976). Instead, a DC Protein Quantification kit (BioRad) using reagents and methods described by Lowrey et al. (1951) was used. After protein standardization, values from the neomycin phosphotransferase assay were recorded as pg of NptII/mg protein.

Identification of flanking genomic sequences

Flanking genomic sequences adjacent to T-DNA left borders were identified by ligation-mediated PCR (Siebert et al., 1995) as described in Alonso et al. (2003).

Adapters for ligation to EcoRI-cut ends were generated by annealing oligos ADAPS-E1 (5'-aattcacctgcccgg/3AmMc7/-3') with a 3' amino-terminal end and ADAPL-E1 (5'-ctaatacgactcactatagggctcgagcggccgcccgggcaggtg-3'). Primary products were generated by amplification with primers AP1 (5'-ggatcctaatacgactcactatagggc-3') and PBI121LB-WP1 (5'-ctgttgcccgtctcactggt-3'). Nested amplification was achieved with primers AP2 (5'-tatagggctcgagcggccg-3') and PBI121LB-WP2 (5'-aagaaaaaaccaccccagtac-3'). Nested PCR products were verified by agarose gel electrophoresis and directly sequenced by the University of North Carolina Lineberger Comprehensive Cancer Center DNA sequencing facility using primer PBI121LB-WP2. Genomic integration sites were identified by BLAST (Altschul et al., 1990) at the SIGnAL (signal.salk.edu) and NCBI (www.ncbi.nlm.nih.gov/BLAST/) websites.

Analysis of Salk SIGnAL Project data

All comparisons to the Salk Institute's SIGnAL project T-DNA integrations are based on integration data available in the supplemental data of Alonso et al. (2003). The data available in the SIGnAL database (signal.salk.edu) contains artifactual data resulting from errors in sequencing, sample contamination, and difficulties in analyzing integrations into repeat-containing genomic regions. The SIGnAL project integration data available in the supplemental material of Alonso et al. (2003) have been filtered to remove integrations that are likely artifacts and are therefore a more appropriate resource to use.

Results

Pooled PCR is an effective method for the identification of transgenic *Arabidopsis*

In order to ensure that a leaf from a single transgenic seedling could be detected in the presence of a pool of non-transgenic leaves, a preliminary study was conducted. Leaves from two week old seedlings of a known transgenic *Arabidopsis* line and wild type *Arabidopsis* were harvested. Genomic DNA was prepared as described above from the following samples: a single transgenic leaf, a single transgenic leaf with six wild type leaves, a single transgenic leaf with nine wild type leaves, two transgenic leaves with five wild type leaves, two transgenic leaves with eight wild type leaves, and seven transgenic leaves. As controls, genomic DNA preparations from a single wild type leaf and from seven wild type leaves were made. The samples were amplified by PCR using primers A and B for the pBI121 T-DNA region and primers HMGup and HMGdown for the endogenous *Arabidopsis HMGA*. The products were analyzed by gel electrophoresis (Figure 2-3). Amplification of the T-DNA fragment (648bp) occurred in all samples where at least one transgenic leaf was sampled. No amplification of the T-DNA fragment occurred in samples containing only wild type samples. The genomic *HMGA* fragment (200bp) was amplified in all samples prepared from leaf tissues. These results indicate that the pooled sampling procedure described above should be capable of detecting a single transformant in a pool of at least up to nine untransformed seedlings. These results also demonstrate that the genomic DNA isolation methods used are efficient enough to isolate DNA from a single two week old leaf.

Five separate screens were conducted on samples taken from the pooled seed described above. Each screen resulted in the identification of numerous transgenic plants (Table 2-1). In most cases only a single transgenic seedling was identified in each

matrix. In some cases multiple transgenic seedlings were identified, with the largest number of transgenic seedlings identified in a single matrix being four. A typical example of a screen is presented in Figure 2-4. Some plates revealed positive amplification in one, but not both pools. In these cases, each individual seedling within that pool was analyzed by PCR. Transgenic seedlings were occasionally identified by this approach, however sometimes no transgenic seedlings were found when individual seedlings from a PCR-positive pool were screened, suggesting that both false-positive and false-negative PCR results from the pooled samples occurred. False-positive results were uncommon (only two cases in the entire study) and quickly identified in subsequent analysis. Efforts were taken to reduce the occurrence of false-negative results, however, it is likely that some transgenic seedlings were not identified, and therefore our PCR-based estimations of transformation are conservative.

Significantly more transgenic seedlings are identified by PCR-screening than by kanamycin-selection

In addition to PCR-screening, kanamycin selection was used to identify transformants from the same pool of T1 seed. Estimated transformation efficiencies for PCR-identified and kanamycin-selected approaches are presented in Table 2-1. Five screens were conducted in order to obtain at least 100 PCR-identified lines. In total, 2959 seedlings were screened by PCR for the presence of T-DNA sequences, and 104 transgenic seedlings were identified (3.5% transformation efficiency). It should be noted that screens 5 and 6 were conducted on seed pools from plants transformed with different constructs and are not appropriate for comparison with screens 1, 2, 3, 4, and 7. The results of screens 5 and 6 will not be discussed further in this dissertation. Three screens included a subset of seedlings that were plated on medium containing kanamycin at

50mg/l. In total, 1895 seeds were plated on kanamycin, and 47 kanamycin resistant seedlings were identified (2.5% transformation efficiency). Since all screens were conducted on samples from the same homogenized pool of seed, any differences in the estimates of transformation efficiency should be due to differences in the methods of detection. Within the screens identifying transgenic seedlings by PCR, all estimated efficiencies are similar with the exception of screen 4, which is considerably lower (see Table 2-1). It should be noted that this particular screen was largely carried out by an undergraduate assistant and is also markedly different in that over one third of the plated seedlings were lost to either contamination or physical injury during harvest. The seedlings identified in this screen are, however, clearly transgenic and the decision was made to include them in the final analysis. No discrepancies were noted for the kanamycin-selection component of screen 4. The net effect of including screen 4 in the total analysis is a reduction in the estimation of the total PCR-identified efficiency from 3.9% to 3.5%. A statistical comparison of the methods used to identify transgenic seedlings using Fisher's exact test indicated that PCR-screening identified significantly more transgenic seedlings than kanamycin selection ($p = 0.0250$).

In *Agrobacterium*-mediated transformation, non-T-DNA vector sequences have been reported to integrate concurrently with T-DNA sequences (for review see Smith et al., 2001). Binary vector sequences can integrate in tandem with T-DNAs or independently (Kononov et al., 1997; Wenck et al., 1997). Although it has been suggested (Kononov et al., 1997), it has not been demonstrated that binary vector sequences can integrate in the complete absence of T-DNA. Such events would be unidentified in systems where selection was a prerequisite for transgenic plant identification. Using the same pooled DNA samples isolated for screen 7 T-DNA screening, seedlings were screened for the presence of binary vector sequence using

primers KEF8 and KEF9. Twenty-one transgenic seedlings were identified by screening for binary vector sequences, however each of these seedlings was also identified by screening for T-DNA sequences. Although no transgenic seedlings containing exclusively binary vector sequences were identified, this approach did demonstrate the effectiveness of the PCR-screening methods used. The absence of screen 7 plants with exclusively binary vector integrations does not imply that such events are not possible.

Verification and analysis of T-DNA integration

In both PCR-identified and kanamycin-selected lines, full-length open reading frames for *nptII* and *gusA* were verified by PCR. T-DNA integration into each seedling was verified by PCR for full length *gusA* with primers E and F (see Figure 2-1). Amplification was achieved in the majority of lines tested. In all cases where amplification failed, histochemical staining with X-gluc confirmed the presence of at least one full length *gusA* gene copy (see Appendix B). Full length *nptII* was verified by PCR with primers C and D (see Figure 2-1). Amplification was achieved in every PCR-identified line except S2-ZC6, S7-BB4, S7-FD2, and S7-PA4. However, the presence of a functional *nptII* in these lines was confirmed by the presence of kanamycin resistant seedlings in the T2 generation (see Appendix B). While line S2-ZC6 failed the chi square test for 3:1 segregation, it did demonstrate kanamycin resistance in 15 of 35 T2 seedlings. Therefore, it can be concluded that each line contained at least one copy of *gusA* and one copy of *nptII* in the T1 generation.

PCR was used to verify the presence of non-T-DNA sequences and to determine integration characteristics of some lines. Amplification with primers I and J was conducted to determine the frequency of pBI121 vector (backbone) integration. Non-T-DNA pBI121 vector sequences were present in 69 of 102 PCR-identified lines (68%) and

in 33 of 42 kanamycin-selected lines (79%). These frequencies are similar to those reported by others (Wenck et al., 1997). The presence of tandem T-DNA integrations (LB-RB) was detected by amplification with primers G and H. Tandem integrations were present in 50 of 71 PCR-identified lines (70%) and in 43 of 63 kanamycin-selected lines (68%). The results from these analyses are summarized in Table 2-2. Attempts to determine the presence of inverted T-DNA sequences around the left border or the right border were unsuccessful, probably due to self-competition (De Buck et al., 1999). A chi square test comparing the presence of tandem repeats and binary vector sequences in PCR-identified and kanamycin selected lines was conducted. No significant difference was observed ($p = 0.2578$), suggesting that differences between the two identification methods are reflected by neither the presence of non-T-DNA vector sequences nor tandem T-DNA repeats. An examination of lines containing both tandem T-DNA repeats and vector sequences reveals that a greater than expected number of lines contain either both features or neither feature (++ and -- in Table 2-2). An analysis of the combined PCR-identified and kanamycin-selected data reveals more than twice the expected number of lines containing neither feature if tandem T-DNA repeats and vector sequences are independent (27 lines observed compared to an expected 12 lines). Chi square analysis suggests that these differences are significant ($p = 8.1\times10^{-8}$ for the combined data sets). It can therefore be concluded that the presence of tandem T-DNA repeats and vector sequences are not independent of one another. Lines containing tandem repeats were more likely to also contain non-T-DNA vector sequences.

It is possible that *Agrobacterium* from the floral dip of the T0 plants could remain as a systemic contaminant in the T1 plants. This could potentially result in false positive amplification by PCR for sequences present in pBI121. To minimize this possibility, all T1 seedlings were grown in the presence of Timentin at 200mg/ml to eradicate

contaminants. It is still possible, however, that low levels of bacteria could remain in the plant tissues and result in false PCR findings (Cubero et al., 1999). To test for the presence of contaminating *Agrobacterium*, all DNA samples from T1 plants were subjected to PCR with primers for endogenous *Agrobacterium VirG*. No plant genomic DNA samples ever amplified the *VirG* fragment, while the positive control preparation of plant tissue dipped in an overnight culture of *Agrobacterium* strain GV3101:pMP90 always amplified the *VirG* fragment.

A higher frequency of *nptII* silencing occurs in PCR-identified lines

Kanamycin tolerance in T1 tissues

Tissues from some T1 seedlings were assayed for the ability to generate callus or roots in the presence of kanamycin. Select plants from screen 1, 2, and 3 were assayed (Table 2-3). Early results suggested that tissues from PCR-identified lines were more sensitive to kanamycin in rooting assays, but later results appeared to contradict these observations. A large number of samples from kanamycin-resistant seedlings failed to generate roots or callus in the assay suggesting a tissue culture-based assay may be too stringent to estimate the kanamycin sensitivity of young seedlings. The success of the assays appeared to be highly dependent on the quality of the starting leaf material, and the additional tissue sampling led to contamination and eventual loss of some transgenic T1 plants. It was eventually decided to discontinue the assay in subsequent screens.

Kanamycin tolerance in T2 seedlings

Because standard kanamycin selection is carried out on germinating seedlings, it is impossible to directly determine if PCR-identified transgenic seedlings (identified after two weeks) would have survived kanamycin selection. Thus, in order to gain an estimate of kanamycin resistance of the T1 plants, T2 seedlings from PCR-identified lines were

assayed for kanamycin tolerance and compared to those of kanamycin-selected lines. Between 18 and 72 sterilized T2 seed from all T1 plants that produced seed were plated on medium containing kanamycin at 50mg/l. At two weeks, seedlings were scored by comparing the number of kanamycin tolerant seedlings to the total number of germinated seedlings for that line. In general germination frequencies were highly consistent (average = 94.2%), however three lines (two PCR-identified lines and one kanamycin-selected line) had germination frequencies lower than 80%. Low T2 germination frequencies could be an indicator of integration into essential genes. Such lines would not be expected to conform to a 3:1 segregation ratio for kanamycin resistance, therefore they were not included in subsequent analysis. The segregation ratios for kanamycin tolerance of the remaining 89 PCR-identified lines and 42 kanamycin-selected lines were tested against the expected 3:1 ratio by chi square goodness of fit tests. Lines with significantly fewer kanamycin tolerant seedlings than expected were considered to be experiencing some degree of silencing of the *nptII* transgene and are referred to as "*nptII*-silencing" lines throughout this study. All other lines are considered to be "*nptII*-expressing." One-third of the PCR-identified lines (30 lines, 34%) displayed *nptII* silencing, while only around one-tenth of the kanamycin-selected lines (5 lines, 12%) displayed *nptII* silencing. The proportions of the PCR-identified and kanamycin-selected lines that failed the chi square test were compared to each other using Fisher's exact test. A significantly higher proportion of the PCR-identified lines were found to deviate from the expected ratio than the kanamycin-selected lines ($p = 0.0046$). As the *nptII* genes are unaltered, we attribute the absence of kanamycin tolerance to *nptII* silencing.

In addition to the difference in the frequency of silencing between PCR-identified and kanamycin-selected lines, the distribution of kanamycin tolerance frequencies within each group is strikingly different. Figure 2-5 shows the histograms of the frequencies of

kanamycin tolerance in PCR-identified and kanamycin-selected lines. As expected, the distribution of kanamycin tolerance frequencies within the kanamycin-selected lines indicates that the vast majority of the lines have frequencies of 0.75 or higher, a few lines are slightly lower than 0.75, and a single line (S4-KR19) has a very low kanamycin tolerance frequency. Frequencies higher than 0.75 could be accounted for by two or more unlinked copies of *nptII*. The distribution of kanamycin tolerance frequencies in the PCR-identified lines is very different. The distribution is clearly bimodal. The upper mode is largely composed of lines with frequencies over 0.75, while the lower mode is heavily skewed towards zero (Figure 2-5b). If the lines in the lower mode (frequencies below 0.40) of the PCR-identified distribution are excluded, only 10 of the remaining 69 lines (14%) have significantly fewer kanamycin tolerant seedlings than expected. This is very similar to the number of kanamycin-selected lines failing the chi square test (12%). A comparison of the upper mode of the PCR-identified lines and all of the kanamycin-selected lines using Fisher's exact test reveals no significant difference in the number of lines that fail the chi square test. Collectively, these data suggest that the PCR-identified lines contain a subset of kanamycin-sensitive lines, represented by the lower mode of Figure 2-5b, that are absent from the kanamycin-selected population.

In order to determine if kanamycin resistance correlated with levels of *nptII* expression, ELISAs were conducted on pools of two-week old T2 seedlings from PCR-identified lines to quantitatively measure the levels of neomycin phosphotransferase (NptII) present in the seedlings. Samples of 12 *nptII*-silencing lines and 12 *nptII*-expressing lines, ranging in T2 segregation ratios from 0 to 0.95, were assayed. The levels of NptII present in samples were found to weakly correlate with segregation ratios for kanamycin tolerance ($R^2 = 0.31$) (Figure 2-6a). The *nptII*-expressing lines had significantly higher levels of NptII than the *nptII*-silencing lines ($p = 0.0218$) (Figure 2-

6b). The *nptII*-silencing lines included values as low as 9pg NptII/mg (which was comparable to untransformed control levels), while the lowest value obtained from a *nptII*-expressing line was 1781pg NptII/mg. These results suggest that T2 kanamycin tolerance segregation frequencies reflect the levels of *nptII* expression.

In order to help determine if the *nptII*-silencing lines are undergoing post-transcriptional gene silencing (PTGS), seedlings were grown at 15°C to block or reduce the effects of PTGS (Szittya et al., 2003). Although a few lines demonstrated slight changes in kanamycin tolerance, the overwhelming majority of lines were not significantly different at 15°C. Additionally, the PCR-identified lines continued to follow a bimodal distribution (data not shown). This consistency between 22°C and 15°C suggests transcriptional gene silencing (TGS) may be the underlying cause of the increased kanamycin sensitivity in the PCR-identified lines.

Correlation of *nptII* and *gusA* expression

Histochemical staining of T1 tissues

Even though we are testing for selection bias and are therefore primarily interested in the expression of *nptII*, we have also assayed the expression of GUS in PCR-identified and kanamycin-selected lines. Evidence in the literature indicates that expression of linked transgenes may be disparate (Peach and Velten, 1991; Mlynarova et al., 2002). Nevertheless, GUS expression in T1 plants (as assayed by histochemical staining (Jefferson et al., 1987)), was found to be in general agreement with *nptII* expression in the T2 generation. At approximately 4 weeks post-germination, leaf tips from seedlings were sampled for GUS by histochemical X-gluc staining. The staining intensity of each seedling was determined by a qualitative assessment of strong, weak, or no staining (see Table 2-4). All kanamycin-selected transgenic seedlings demonstrated

strong histochemical staining, while only 73% of PCR-identified transgenic seedlings stained at this intensity (7% had weak staining and 20% had no staining). A comparison of the categorical X-gluc staining intensities of kanamycin-selected and PCR-identified lines was made. The absence of kanamycin-selected samples with weak or no staining is significant ($p = < 1 \times 10^{-10}$).

GUS activity in T1 plants

In order to quantitatively measure GUS activity, MUG assays were conducted on older PCR-identified and kanamycin-selected plants (see Figure 2-7). Each plant was sampled between one and three times between 40 and 80 days post-germination. Most plants were at similar developmental stages when sampled, but due to early flowering in some lines (possibly caused by stress from repeated sampling - see above) some developmental variation existed between plants resulting in differing availabilities of tissues for sampling. Each sample consisted of 3 rosette leaves between 1.5 and 3cm in length if available. Values from plants sampled multiple times were averaged. In general, repeated samplings resulted in consistent values. GUS activity values (pmole MU/min/mg) obtained for both PCR-identified and kanamycin-selected lines were not normally distributed. Further analysis of the data revealed heavily skewed distributions in both PCR-identified and kanamycin-selected data sets. Attempts to normalize the data by transformation were unsuccessful. Such data cannot be compared by T-tests or other statistical tests that assume normal distributions. The Mann-Whitney two-sample rank test for the comparison of groups lacking normal distributions was used to compare GUS activities in PCR-identified and kanamycin-selected T1 plants, and no significant difference was detected.

GUS activity in T2 seedlings

GUS activities for PCR-identified and kanamycin-selected T2 seedlings were also measured. Pools of 10 T2 seedlings germinated on SG medium without kanamycin were harvested at 14 days post-germination and used for quantitative measurements of GUS activity. The distribution was strongly skewed toward zero in both PCR-identified and kanamycin-selected lines. A natural log transformation was applied to the data to obtain a normal distribution so that the data could be compared by a T-test. There was not a significant difference in levels of GUS activity between PCR-identified and kanamycin-selected T2 lines (see Figure 2-8a). However, *nptII*-silencing PCR-identified lines had significantly lower levels of GUS activity than *nptII*-expressing PCR-identified lines ($p = 0.0073$), suggesting a correlation between *nptII* and GUS expression (see Figure 2-8b).

Analysis of genomic integration

Analysis of SIGnAL integration data

Included in the supplemental online material for Alonso et al. (2003) is information related to the SIGnAL project's T-DNA insertion lines, including sites of genomic integration. These integration data have been filtered to remove integration sites that are in genomic repeat regions (such as centromeric and ribosomal DNA regions) since a precise integration position cannot be determined for these lines based on flanking sequence homology alone. Additionally, integration data that are likely artifactual, primarily from cross-contamination, have been removed. The remaining data encompass 101,329 integrations over 117,260,472 base pairs. This represents a genome-wide average of 0.8641 integrations/kb. If the distribution of integrations is random, it should follow a Poisson distribution. The Poisson probability for the expected number of integrations within a particular window size can be calculated using the equation $P(X) =$

$\mu^X/e^\mu X!$ (Zar, 1996), where P(X) is the probability of X integrations in a particular window (i.e. 10kb), and μ is the number of integrations in that particular window based on the population mean (i.e. 8.641 integrations/10kb). If the distribution of T-DNA integrations follows a random Poisson distribution, any 10kb window with fewer than 3 or greater than 14 integrations would be considered to fall outside of the expected range ($\alpha = 0.05$). A cursory examination of the distribution of T-DNA integrations reveals many regions do indeed fall outside of this expected range. However, as a result of the methods used by Alonso et al. (2003) to identify the T-DNA flanking sequences (Siebert et al., 1995), known biases in the identification of integration sites were introduced. In other words, even though some lines may have contained T-DNA integrations in particular genomic regions, those integrations were unidentified. This is because the methods used to identify flanking sequences require that i) the integration be in close proximity to an EcoRI restriction site and ii) the region between that restriction site and the point of integration be capable of amplification by PCR (Siebert et al., 1995). Considering this bias, Alonso et al. (2003) determined that even though the absence of T-DNA integrations into some genomic regions could be explained by the absence of nearby EcoRI restriction sites, there were significantly more low-density and high-density regions than could be explained by chance alone.

Identification of sites of integration

Using a ligation-mediated PCR approach initially reported by Seibert et al. (1995), and modified by Alonso et al. (2003), the genomic sites of T-DNA integration in several of the transgenic lines were identified. Candidate lines for sequencing were initially determined based on kanamycin resistance frequencies in T2 seedlings. Approximately equal numbers of PCR-identified *nptII*-expressing and PCR-identified *nptII*-silencing lines were chosen for analysis. Some kanamycin-selected lines were also

analyzed, but because very large numbers would be needed to detect differences between all kanamycin-selected lines and all PCR-identified lines, we focused on detecting differences between PCR-identified lines with differences in kanamycin tolerance. Due to the nature of the ligation-mediated PCR approach used (Siebert et al., 1995), attempts to identify flanking genomic sequences from some lines failed. All procedures commonly used to identify genomic sequences flanking T-DNA integration sites have inherent biases which may exclude the identification of integrations in some lines due to the nature of the integration locus (Hui et al., 1998; Cottage et al., 2001). Possible causes of failure include the absence of appropriate restriction enzyme sites in close proximity to the integration site and the presence of flanking sequence not amenable to PCR amplification (see above). Because our intentions were to compare our integrations with those of the SIGnAL project (Alonso et al., 2003), it was critical that we used an identical approach to mapping our integrations. All resulting sequences were analyzed for homology by BLAST comparison against *Arabidopsis* sequences in the SIGnAL database (http://signal.salk.edu) and all existing sequences in Genebank (http://www.ncbi.nlm.nih.gov/blast/). Table 2-5 includes the genomic integration information of mapped PCR-identified and kanamycin-selected lines identified in this study.

The distribution of T-DNA inserts encompassed all five chromosomes of *Arabidopsis* for both *nptII*-silencing and *nptII*-expressing PCR-identified lines. Two *nptII*-silencing lines contain T-DNA integrations in regions frequently associated with heterochromatin. Line S3-GC3 contains an integration site in the *Arabidopsis* 25S rDNA repeat region. The 25S rDNA regions in *Arabidopsis* are found exclusively in large repeats contained in the nucleolar organizing regions (NOR2 and NOR4) adjacent to the telomeres of chromosomes II and IV (Copenhaver and Pikaard, 1996). It can then be

concluded that this line contains a T-DNA integration in a NOR, although further mapping analyses would be necessary to determine if the integration occurred on NOR2 or NOR4. Line S7-BB1 contains a T-DNA integration in the gene At4g05150 located near the centromere of chromosome IV. In total, 2 of 14 (14%) mapped *nptII*-silencing lines have inserts in known heterochromatic regions. No mapped integrations in *nptII*-expressing lines identified in this study appear to have occurred in similar regions.

All mapped integrations were analyzed for T-DNA left border (LB) truncation. It has been reported that large LB truncations can occur frequently during T-DNA integrations (Tinland, 1996), although several recent reports indicate that truncation may be less severe than previously thought (Brunaud et al., 2002; Forsbach et al., 2003). In general, truncation was common, but limited to fewer than 25bp in most lines (see Figure 2-9). No difference was observed in the frequency or degree of truncation between *nptII*-expressing and *nptII*-silencing lines.

Comparison to SIGnAL integration profiles

Using the T-DNA integration data reported in the supplemental material of Alonso et al. (2003), a sliding window analysis was conducted across the entire *Arabidopsis* genome for the number of SIGnAL inserts within 5kb windows at 1kb steps. It was then possible to graphically represent the SIGnAL integration densities around the sites of our PCR-identified T-DNA integrations. Integration profiles of two PCR-identified lines are represented in Figure 2-10. It is interesting that the majority of low or non-expressing PCR-identified lines (such as S3-BB1 and S3-EB4) integrated into regions underrepresented by T-DNA integrations in the SIGnAL project. In contrast, the majority of expressing PCR-identified lines and kanamycin-selected lines integrated into regions with either moderate or high numbers of SIGnAL integrations.

T-DNA integration profiles similar to those shown for two transgenes in Figure 2-10 have been constructed for all the mapped PCR-identified lines. For each line, the number of SIGnAL project T-DNA integrations in 1kb windows were determined for a 100kb region centered on the point of integration. The number of integrations in each window is averaged for 13 *nptII*-silencing lines in the lower mode in Figure 2-5b (lines that appear to represent a population of lines with severe silencing) and plotted in Figure 2-11a. In Figure 2-11b, the averages for 12 *nptII*-expressing lines are plotted. In comparison of these two profiles, the most striking observation is the presence of a clear depression in the average number of SIGnAL integrations within a few kb of the integration sites of *nptII*-silencing lines. No such depression is observed for the *nptII*-expressing lines. Furthermore, the average number of SIGnAL integrations is lower for the *nptII*-silencing lines than for the *nptII*-expressing lines (dotted lines in Figures 2-11a and 2-11b).

Fewer SIGnAL integrations are observed in regions where *nptII*-silenced PCR-identified lines have integrated. The number of SIGnAL integrations within 5kb, 10kb, 100kb and 200kb centered on the integration sites of each mapped PCR-identified line is presented in Table 2-6 along with the T2 segregation frequencies for kanamycin resistance. The average numbers of integrations for all of the *nptII*-silencing PCR-identified lines and for the subset of those lines that compose the lower mode in Figure 2-5b are presented in Figure 2-12. For comparison, the average numbers of integrations for *nptII*-expressing PCR-identified lines are also presented. The average number of integrations in 5, 10, 100, and 200kb windows centered on *nptII*-silencing lines is significantly lower than the number of integrations in *nptII*-expressing lines ($p = 0.0073$, 0.0206, 0.0365, and 0.0485 respectively for each of the windows). No significant difference is observed for 500kb windows (see Figure 2-12). If only the lines with the

most severe silencing (lower mode of Figure 2-5b) are compared to the *nptII*-expressing lines, the difference is even more pronounced ($p = 0.0014$, 0.0041, 0.0346, and 0.0378 respectively).

It is important to consider that the true difference in the average integration profiles and the number of SIGnAL integrations within specific windows between *nptII*-silenced and *nptII*-expressing lines is likely to be even greater than reported here. Because the *nptII*-silenced line S3-GC3 integrated into an rDNA repeat-containing region, the precise point of integration for this line cannot be determined. Similarly, the SIGnAL database does not include integrations that have occurred in repeated regions because of the difficulties in determining the precise points of integration. If the density of SIGnAL integrations in the rDNA repeats is lower than the densities in the remaining genome, as has been reported in other T-DNA insertion collections (Szabados et al., 2002), it is likely that values calculated here for the average number of SIGnAL integrations near integration sites of kanamycin-sensitive lines is higher than it would be if the data for integrations in the rDNA repeats were included.

Analysis of predicted Matrix Attachment Regions

It has been reported that transgenes frequently integrate in close proximity to Matrix Attachment Regions (MARs) (Dietz et al., 1994; Sawasaki et al., 1998; Makarevitch et al., 2003). In some reports MARs were identified by *in vitro* binding assays (Dietz et al., 1994; Sawasaki et al., 1998; Shimizu et al., 2001), while in other reports "candidate" MARs were identified by using MAR-Wiz (http://www.futuresoft.org/MAR-Wiz/) (Makarevitch et al., 2003; Qin et al., 2003) or other approaches based on sequence motifs and patterns commonly found in known MARs (Takano et al., 1997). If MARs increase the level of expression of nearby transgenes as has often been suggested (Allen et al., 2000; Bode et al., 2000), it is

possible that selection bias may result in the preferential identification of transformants that have integrated in close proximity to MARs. The presence of MAR candidates within a 10kb region centered on the point of integration was estimated using MAR-Wiz (Table 2-6). Nearby MAR candidates were predicted for only about half of the sites analyzed, and the presence of MAR candidates was not correlated with the kanamycin tolerance of the lines examined. It has also been suggested that MARs may facilitate transgene integration because of their A+T-richness, DNA bending propensity, and presence of topoisomerase sites (Sawasaki et al., 1998). For this reason, the presence of MAR candidates was also examined in regions of high SIGnAL T-DNA integration density, however, no relationship was observed (data not shown). Collectively, these data suggest that the presence of predicted endogenous MAR candidates does not correlate with the integration sites of T-DNAs in transgenic plants identified by PCR-screening or by kanamycin selection.

Discussion

Differences in perceived transformation efficiencies

Significantly more transgenic *Arabidopsis* seedlings from the same pool of T1 seeds were identified by PCR screening than by selecting for kanamycin resistance. If PCR-identification represents a conservative estimate of the true integration efficiency, we can conclude that identification of transgenic seedlings by kanamycin selection fails to identify about 30% of all integration events *Arabidopsis*. It is likely that kanamycin selection failed to identify some transformants because the selectable marker (*nptII*) was not expressed or expressed at a low level at the time of selection. Many possibilities exist as to why a transgene may not express. These can be summarized in three general

categories: i) mutations in the transgene, ii) post-transcriptional gene silencing (PTGS), and iii) chromatin-related transcriptional gene silencing (TGS).

Of the various possible mutations that could result in transgene inactivity, truncation associated with T-DNA integration is the most likely. Many reports of T-DNA truncation exist in the literature (reviewed in Tinland, 1996). Truncations at the left border are more commonly reported, although right border truncations also occur (Gheysen et al., 1990). Recently, several high-throughput approaches to examining T-DNA integration have reported that left and right border truncations are perhaps less common than previously thought (Brunaud et al., 2002; Forsbach et al., 2003). No truncation that would be expected to preclude transgene expression was observed in any of the lines described here. Point mutations and internal deletions are also possible, although rarely reported for *Agrobacterium*-mediated integration and more frequently associated with particle bombardment-mediated integration (Sawasaki et al., 1998; Ülker et al., 2002). Because at least some T2 plants from nearly all of the PCR-identified lines showed kanamycin tolerance, it is very unlikely that our data can be explained by mutations in the transgenes.

Complex transgene arrangements and high expression levels have been reported to result in PTGS in some cases (Elmayan and Vaucheret, 1996; De Buck et al., 2001). We have used PCR to detect direct repeats and the presence of binary vector sequences in the transgenic lines. Examples of such complex integrations were found in several lines, but there was no difference in frequency between kanamycin-selected and PCR-identified lines. The presence of inverted repeats, more commonly thought to be associated with PTGS (Ma and Mitra, 2002), was determined for a sample of lines by DNA blot hybridization and found to occur at similar frequencies for PCR-identified and kanamycin-selected lines (data not shown). It has recently been shown that PTGS is

significantly less active in very young and actively dividing tissues (Mitsuhara et al., 2002). Since kanamycin selection of transgenic seedlings was performed within the first 10 to 14 days after germination, it is unlikely that PTGS played a major role in silencing *nptII* in the very young T1 seedlings.

The assays we have made under conditions that minimize PTGS and our studies on the sites of T-DNA integrations suggest that transcriptional gene silencing (TGS) accounts for the higher perceived transformation efficiency when PCR is used to identify transformants. That is, some transgenic lines may not be identified by kanamycin selection because the T-DNAs in those lines have integrated into genomic regions that repress transgene expression. Such repression of transgene expression by chromatin or nearby genomic elements has been characterized in many systems (for general reviews see van Holde, 1989; Wolffe, 1998) and likely accounts for the higher percentage of transformants identified by PCR.

Silencing in PCR-identified lines

PCR-identified lines are more likely than kanamycin-selected lines to demonstrate *nptII* silencing in the T2 generation. Lines with T2 kanamycin tolerance ratios significantly lower than the expected 3:1 ratio were considered to be *nptII*-silenced. Silencing of *nptII* was observed in 34% of the PCR-identified lines, while only 12% of kanamycin-selected lines had *nptII* silencing. The majority of PCR-identified lines with *nptII* silencing demonstrated severe silencing (lower mode in Figure 2-5b), while only one kanamycin-selected line demonstrated silencing to this degree. Furthermore, in 7 of 89 PCR-identified lines analyzed, none of the T2 seedlings showed kanamycin tolerance. No kanamycin-selected line demonstrated this degree of *nptII*-silencing. It is interesting that the 34% of PCR-identified lines with *nptII* silencing corresponds closely to the

approximately 30% increase in transformation efficiency when transformants were identified by PCR. Similar levels of silencing have been reported in Mexican lime plants identified without selection (Dominguez et al., 2002).

Again, it is possible that T2 silencing could be a result of PTGS, however kanamycin tolerance was assayed in very young seedlings that may not be capable of PTGS (Mitsuhara et al., 2002). Furthermore, kanamycin tolerance for PCR-identified and kanamycin-selected T2 seedlings was not significantly different at 15°C, a temperature reported to block PTGS (Szittya et al., 2003). PTGS can cause meiotically stable changes in transgene methylation that can result in a heritable loss of transgene activity (Wassenegger et al., 1994). This phenomenon, called RNA-directed DNA methylation (RdDM) (Aufsatz et al., 2002), may explain why some kanamycin-selected lines showed a partial loss of kanamycin resistance in the T2 generation but cannot explain why more silencing was observed in PCR-identified lines. Combined, these data suggest that PTGS may not have been a major contributor to silencing of *nptII* in PCR-identified T2 seedlings. It is therefore likely that the silencing observed in the T2 generation was largely a result of TGS. Although it seems unlikely, we can not rule out the possibility that the small number PCR-identified lines with no sign of kanamycin tolerance in the T2 generation could be the result of point mutations, rearrangements, or truncations not detected by the approaches used.

Differences in active and silent T-DNA integration loci

Genomic integration sites were identified in a sample of PCR-identified lines. Two lines with significant *nptII* silencing appeared to have integrated into genomic regions usually considered to have chromatin structures that would preclude transcription. Line S7-BB1 integrated into the pericentromeric region of chromosome 4,

and line S3-GC3 integrated into rDNA repeats in either NOR2 or NOR4. For lines with integrations in non-repeat-containing regions, by comparing the identified integration sites with the integration sites of over 100,000 transgenic lines reported in the supplemental material of Alonso et al. (2003), it could be determined if any integrations occurred in integration "cold-spots" (i.e. regions with few or no SIGnAL T-DNA integrations). A clear difference in the density of SIGnAL integrations at the sites of integration between PCR-identified expressing and silencing lines was seen. PCR-identified lines with *nptII* silencing integrated into regions with roughly half the number of SIGnAL integrations when compared to *nptII*-expressing lines (Table 2-6). Furthermore, several of the lines with complete or nearly complete *nptII* silencing had T-DNA integrations in regions either completely devoid of SIGnAL integrations or containing only a single integration within 5kb and 10kb windows around the integration locus. None of the *nptII*-expressing lines (either PCR-identified or kanamycin-selected) appeared to have integrated into one of these "cold spots."

It is possible that these regions are identified as "cold spots" because they are either incapable of supporting transgene expression or are capable of supporting expression only at a level too low to facilitate identification by kanamycin selection. This may explain why PCR-identified lines that have integrated into these regions have little or no detectable transgene expression. We have clearly demonstrated that integrations into some "cold spots" are possible. It is therefore unlikely that some genomic regions exist that contain very few or no T-DNA integrations because the physical structure of the surrounding chromatin physically blocks T-DNA integration. A more plausible explanation is that integrations into these regions can occur but are rarely reported because the selectable markers contained within those T-DNAs are not expressed. It is important to emphasize, though, that simply because integration is

possible in some "cold spots," there still may be some genomic regions physically incapable of receiving T-DNA integrations.

The "Silent Second T-DNA" model

An examination of the distribution of SIGnAL T-DNA integrations reveals numerous 5 to 10kb regions that completely lack integrations. There are also many considerably larger regions that contain only a very few integrations. A complete absence of integrations could be explained by either the complete inability of T-DNAs to integrate into those regions or by the complete inability of those regions to support transgene expression. Regions where identified T-DNA integrations exist, but with low frequency, could be explained by the occurrence multiple T-DNA integrations. *Agrobacterium*-mediated transformation often results in the insertion of multiple T-DNAs during a single transformation event (De Buck et al., 1998). These multiple integrations can occur at a single locus or at multiple loci. Indeed, Alonso et al. (2003) found the average number of unlinked loci per line to be ~1.5. This number is very consistent with other studies (Feldmann, 1991; Rios et al., 2002; Ichikawa et al., 2003). In the data presented here, over 25% of the kanamycin-selected lines demonstrated T2 segregation ratios that suggest the presence of multiple active loci. The presence of multiple independent integrations can play a role in the effect of selection bias. Under selective conditions, only one T-DNA integration is required to have occurred in a transcriptionally competent region. Since the expression of all other independently integrated T-DNAs in that line are not required, they may be considered as being free of selection bias. In order to identify sites of integration in a high-throughput manner, the methods used by Alonso et al. (and used here as well), involved the identification of a single integration site per line, even if that line contained multiple integrations. The

mapping methods used do not distinguish between active and inactive T-DNAs. It is therefore possible that some of the integration sites identified by the SIGnAL project (and here as well) are integration sites of inactive T-DNAs and are thus free of selection bias. We refer to these integrations as "second silent T-DNAs" and suggest that integrations into silent chromatin may only be possible through these means when transformants are identified under selective conditions. Furthermore, some SIGnAL integrations may be located in regions that have a transcriptionally repressive chromatin structure at the time of selection, but are activated through changes in chromatin structure at a later developmental time (e.g. flowering). If selection bias precludes identification of transformants with T-DNA insertions in or near such genes, it may be impossible to obtain such insertion mutants except by "silent second T-DNAs."

Because they were identified by selection, all of the SIGnAL lines contain at least one integration site that was active at the time of selection. If we consider that approximately half of the SIGnAL lines contain multiple unlinked integrations, and we assume that the mapped integration site in these lines was chosen at random, we can conservatively conclude that approximately 1/4 of the mapped SIGnAL sites were identified without selection bias. Based on the data presented here, about 29% of transgenic *Arabidopsis* would not have been identified by conventional methods due to selection bias, so at least 29% of integration sites may not support expression. It can therefore be concluded that about 29% of 1/4 of the mapped SIGnAL sites (or ~7%) are mapped to genomic regions that may not support expression at the time of selection.

It is well known in the *Arabidopsis* community, that it is not uncommon for some SIGnAL lines obtained from the *Arabidopsis* Biological Resource Center (ABRC) to have "lost" kanamycin resistance (for comments see signal.salk.edu). While silencing of some previously active T-DNAs may have occurred, it is possible that some T-DNAs

were never active because they integrated into genomic regions that repress expression. If these regions segregate from actively expressed T-DNA insertions, it would appear as though the transgenic line had retained the T-DNA insert but that the transgene had been silenced subsequent to selection.

Consequences of selection bias

It has been suggested that only a minor proportion of the small and simple *Arabidopsis* genome is comprised of heterochromatin (Barakat et al., 2000; Fransz et al., 2002). This may explain why T-DNA integrations can be identified throughout the majority of the genome with only a few regions lacking or low in observed integrations. Organisms with more complex genomes and greater proportions of heterochromatin, such as rice (*Oryza sativa*) (Barakat et al., 1997), might be expected to display a different pattern of identified integrations. That is, indeed, what is observed. In rice, identified sites of T-DNA integration occur predominately in genic regions, with fewer than 3% occurring in repetitive regions (Barakat et al., 2000; Chen et al., 2003). Such repetitive regions would likely be in a heterochromatic state and may be incapable of supporting transgene expression. It would be interesting to see if an approach similar to the one taken here with *Arabidopsis* would result in the identification of significantly more integrations into repetitive genomic regions in rice.

Under most circumstances, selection bias should not have a significant effect on experimental outcome. Indeed, a transgenic plant that is incapable of expressing its transgenes is of limited experimental and practical use. There are, however, particular circumstances where selection bias must be considered. Selection bias could result in the inability to completely saturate any genome by insertional mutagenesis. It may be particularly difficult to achieve a large number of insertional mutants in regions inactive

at the time of selection (i.e. developmental or stress-regulated regions), even though these regions may be active later in the plant lifecycle. Integrations in these regions may be achievable, however, through "second silent T-DNA" integrations. Studies concerning the effects of particular elements such as MARs or insulators could be heavily impacted by selection bias. If such elements act by allowing transgene expression in genomic regions normally repressed by surrounding chromatin, they may functionally act to decrease the effects of TGS-related selection bias in transformants containing those elements. On the other hand, organisms transformed with control constructs lacking these elements would be subject to selection bias, and transformants lacking gene expression would be unidentified. The elimination of non-expressing transformants from the control populations would result in observing an artificially smaller effect on transgene expression levels than what may be truly occurring. One other possible consequence of selection bias relates to transformation-recalcitrant species. Many species considered to be difficult to transform contain very large and complex genomes (Wang et al., 2001; Emani et al., 2002; Janakiraman et al., 2002; Gelvin, 2003). It is possible that these organisms are actually transformed at a relatively high rate, but the fraction of the genome capable of supporting transgene expression is so small, that under selective conditions, transformants are only rarely identified. It is also possible that some species are so effective at identifying and silencing foreign DNA insertions that the majority of integrations are immediately silenced and subsequently unidentified. If either of these possibilities were the cause of transformation-recalcitrance, it would be prudent to take a different approach in improving transformation efficiencies, such as including elements that may increase the likelihood of transgene expression. It has been previously demonstrated that the inclusion of MARs or insulators can improve transformation efficiencies in both plants (Han et al., 1997) and *Drosophila* (Roseman et al., 1995). It is

quite possible that these elements function by allowing transgene expression in normally repressed genomic regions, therefore reducing selection bias, and allowing for the identification of transformants under selective conditions.

Conclusions

We have shown that using PCR to detect transformants results in the identification of T-DNA integration sites not likely to be found by selection. However, it is not our intention to suggest that T-DNA integration is completely random. Rather, we simply wish to raise the issue that selection bias does exist and has possibly clouded previous assessments of T-DNA integration patterns. Our results suggest that true T-DNA integration efficiencies in *Arabidopsis* are significantly higher than the efficiencies observed when expression of selectable or screenable markers is a prerequisite for identification. It is likely that these observations will also apply to other plant species. We have also demonstrated that plants containing silenced transgenes may contain integrations in genomic regions that are largely under-represented in T-DNA insertional mutagenesis collections. These findings may lead to advancements in the general understanding of T-DNA integration and in the goal of generating T-DNA insertions in every gene in *Arabidopsis* and in other important plant species.

Table 2-1. Transformation efficiencies based on method of identification

	PCR-Identified			Kanamycin-Selected		
	Screened[a]	PCR Positive[b]	Efficiency	Screened[a]	Kanamycin Resistant[c]	Efficiency
Screen 1	411	17	4.1%			
Screen 2	783	27	3.4%	843	19	2.3%
Screen 3	412	15	3.6%	284	9	3.2%
Screen 4	612	12	2.0%	768	19	2.5%
Screen 7[d]	741	33	4.5%			
Total	2959	104	3.5%	1895	47	2.5%

[a] Seeds that failed to germinate and aborted seedlings were not considered.
[b] Based on A/B amplification from individual seedlings.
[c] Based on kanamycin tolerance at 2 weeks on kanamycin at 50mg/l.
[d] Screens 5 and 6 were performed with pools of seed transformed with different binary vectors, and therefore, are not included in this comparison.

Table 2-2. Presence of tandem repeats and non-T-DNA vector sequences in PCR-identified and kanamycin-selected transformants.

	Tandem Repeats [a]	Vector Sequences [b]	Number of Plants
PCR-identified	+	+	41
	+	-	9
	-	+	7
	-	-	13
Kanamycin-selected	+	+	37
	+	-	5
	-	+	6
	-	-	14
Combined [a]	+	+	78
	+	-	14
	-	+	13
	-	-	27

[a] The presence of tandem repeats was determined by PCR amplification with primers G and H.
[b] The presence of non-T-DNA binary vector sequences was determined by PCR amplification with primers I and J.
[c] Data from PCR-identified and kanamycin-selected lines were combined for analysis of the independence of the presence of tandem repeats and binary vector sequences.

Table 2-3. Kanamycin tolerance in T1 tissues

	PCR-Identified				Kanamycin-Selected			
	Callus Assay[a]		Rooting Assay[b]		Callus Assay[a]		Rooting Assay[b]	
	+	-	+	-	+	-	+	-
Screen 1			7	7				
Screen 2	24	3	8	19	8	1	6	6
Screen 3[c]	5	0	4	0	6	0	7	1
Total	29	3	19	26	14	1	13	7

[a] Plus signs (+) indicate lines were able to generate callus from leaf explants in the presence of kanamycin at 50mg/l. Minus signs (-) indicate lines were not able to generate callus from leaf explants in the presence of kanamycin at 50mg/l.
[b] Plus signs (+) indicate lines were able to generate roots from leaf explants in the presence of kanamycin at 50mg/l. Minus signs (-) indicate lines were not able to generate roots from leaf explants in the presence of kanamycin at 50mg/l.
[c] Assays were discontinued after Screen 3.

Table 2-4. GUS activity in seedling leaf tips[a].

	PCR-Identified				Kanamycin-Selected			
	Strong	Weak	None	Total	Strong	Weak	None	Total
Screen 1	9	3	4	16				
Screen 3	14	0	0	14	9	0	0	9
Screen 4	8	2	3	13	16	0	0	16
Screen 7	25	0	8	33				
Total	56	5	15	76	25	0	0	25

[a] Histochemical X-gluc staining to detect GUS activity was conducted on excised leaf tips of 3-week-old PCR-identified and kanamycin-selected lines.

Table 2-5. Integration sites of a sample of PCR-identified and kanamycin-selected lines.

PCR-Identified *nptII*-silencing[a]

	Homology	Chromosome	Position[b]	Gene	Feature[c]	
S1-CG6	*A. thaliana*	IV	+2587	At4g00890	IG	
S3-EB4	*A. thaliana*	V	-147	At5g49550	5'	
S3-GC3	*A. thaliana*	II/IV				25S rDNA
S4-QB5	*A. thaliana*	I	-504	At1g54095	5'	
S4-SA3	*A. thaliana*	I	-559	At1g11380	5'	
S4-WE2	*A. thaliana*	I	+6568	At1g05570	I	
S7-BB1	*A. thaliana*	IV	+1185	At4g05150	I	Centromere
S7-BD5	*A. thaliana*	I	+3942	At1g36035	E	
S7-HB5	*A. thaliana*	I	+74	At1g01490	I	
S7-JD6	*A. thaliana*	III	+211	At3g44150	E	
S7-KB2	*A. thaliana*	V	-26	At5g58270	5'	
S7-LA6	*A. thaliana*	III	-1438	At3g29642	IG	
S7-LD1	*A. thaliana*	III	-45	At3g14050	5'	
S7-OE1	*A. thaliana*	II	-635	At2g44350	5'	

PCR-Identified *nptII*-expressing[a]

	Homology	Chromosome	Position	Gene	Feature
S3-AD1	*A. thaliana*	I	-711	At1g61820	5'
S3-CC5	*A. thaliana*	II	-1179	At2g02010	IG
S4-CE2	*A. thaliana*	III	-430	At3g10330	5'
S4-GB6	*A. thaliana*	IV	-1262	At4g23690	IG
S4-YC1	*A. thaliana*	V	+187	At5g08100	I
S7-AC5	*A. thaliana*	V	+4364	At5g37830	3'
S7-AC6	*A. thaliana*	I	+732	At1g64310	E
S7-FB5	*A. thaliana*	V	-2094	At5g66880	IG
S7-GA3	*A. thaliana*	III	+2570	At3g15410	I
S7-IC1	*A. thaliana*	V	+2515	At5g67610	3'
S7-NA5	*A. thaliana*	I	-1899	At1g29660	IG
S7-TE3	*A. thaliana*	V	-236	At5g53920	5'

Kanamycin-Selected *nptII*-expressing[a]

	Homology	Chromosome	Position	Gene	Feature
S3-KR6	*A. thaliana*	V	+428	At5g41490	E
S3-KR3	*A. thaliana*	I	-170	At1g77760	5'

[a] *nptII*-silencing and *nptII*-expressing lines determined by T2 segregation on kanamycin.
[b] Position relative to predicted translational start site of nearest open reading frame.
[c] Integration occurring within 1kb upstream of gene (5'), within 1kb downstream of gene (3'), in intron (I), or in exon (E). All other integrations were considered to be in intergenic regions (IG).

Table 2-6. Summary of SIGnAL integrations and MAR potentials centered on T-DNA integration sites for *nptII*-silencing and *nptII*-expressing lines.

	T2	SIGnAL integrations within window[h]				Nearest MAR[d]	
Line	Kan. Res.	5kb	10kb	100kb	200kb	kb	score
S3-EB4	0.00 [a,b]	0	0	79	163	none	-
S7-OE1	0.00 [a,b]	1	7	107	205	none	-
S7-BB1	0.00 [a,b]	1	3	68	142	1.8	305
S7-LA6	0.00 [a,b]	2	4	43	94	1.0	223
S4-WE2	0.05 [a,b]	1	1	77	153	none	-
S7-LD1	0.08 [a,b]	2	4	84	172	1.1	209
S7-JD6	0.13 [a,b]	3	9	57	96	3.4	308
S7-HB5	0.14 [a,b]	6	9	110	216	1.1	298
S7-KB2	0.25 [a,b]	1	1	63	139	none	-
S7-BD5	0.30 [a,b]	1	2	55	79	none	-
Average (integrations/kb)[e]		1.8[c](0.36)	4.0[c](0.40)	74[c](0.74)	146[c](0.73)		
S4-QB5	0.44 [a]	7	16	92	166	4.4	154
S4-SA3	0.48 [a]	5	10	99	211	none	-
S1-CG6	0.49 [a]	7	10	86	185	1.7	486
S3-GC3	0.53 [a]	-	-	-	-	na	na
Average (integrations/kb)[f]		2.9[c](0.57)	5.8[c](0.58)	78[c](0.78)	155[c](0.78)		
S3-CC5	0.64	4	5	93	168	3.9	402
S4-YC1	0.69	6	6	108	200	3.2	128
S7-FB5	0.69	13	16	115	192	none	-
S7-TE3	0.75	7	10	87	197	1.7	245
S3-AD1	0.75	2	3	105	218	none	-
S4-CE2	0.76	4	10	82	171	4.4	231
S7-IC1	0.76	7	20	na	na	none	-
S7-NA5	0.79	9	12	82	175	0.6	142
S4-GB6	0.94	5	16	87	159	none	-
S7-GA3	0.94	3	7	106	199	none	-
S7-AC5	0.95	5	5	58	128	2.2	240
S7-AC6	0.95	4	12	99	193	none	-
Average (integrations/kb)[g]		5.8 (1.15)	10.2 (1.02)	93 (0.93)	181 (0.91)		

[a]T2 kanamycin resistance frequencies are significantly lower than expected 3:1 ratio ($\alpha = 0.01$).
[b]*nptII*-silencing lines represented in the lower mode of Figure 2-5b.
[c]Number of integrations is significantly lower than *nptII*-expressing lines ($\alpha = 0.05$).
[d]MAR candidates based on MAR-Wiz (www.futuresoft.com).
[e]Average integrations/kb of *nptII*-silencing lines in lower mode of Figure 2-5b.
[f]Average integrations/kb of all *nptII*-silencing lines.
[g]Average integrations/kb of *nptII*-expressing lines.
[h]Number of SIGnAL integrations reported within windows of varying sizes (kb) centered on T-DNA integration sites of PCR-identified lines.

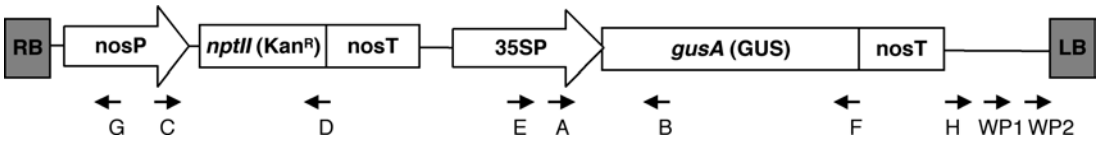

Figure 2-1. Map of pBI-121 T-DNA.
nptII is driven by the *nos* promoter (nosP), which is located at the right border (RB). *gusA* is driven by the *35S* promoter (35SP) and is located internally. PCR primer sites and orientations are indicated by solid black arrows. Walking primers (WP1 and WP2) are located just inside the left border (LB). Regions and primer sites are not drawn to scale.

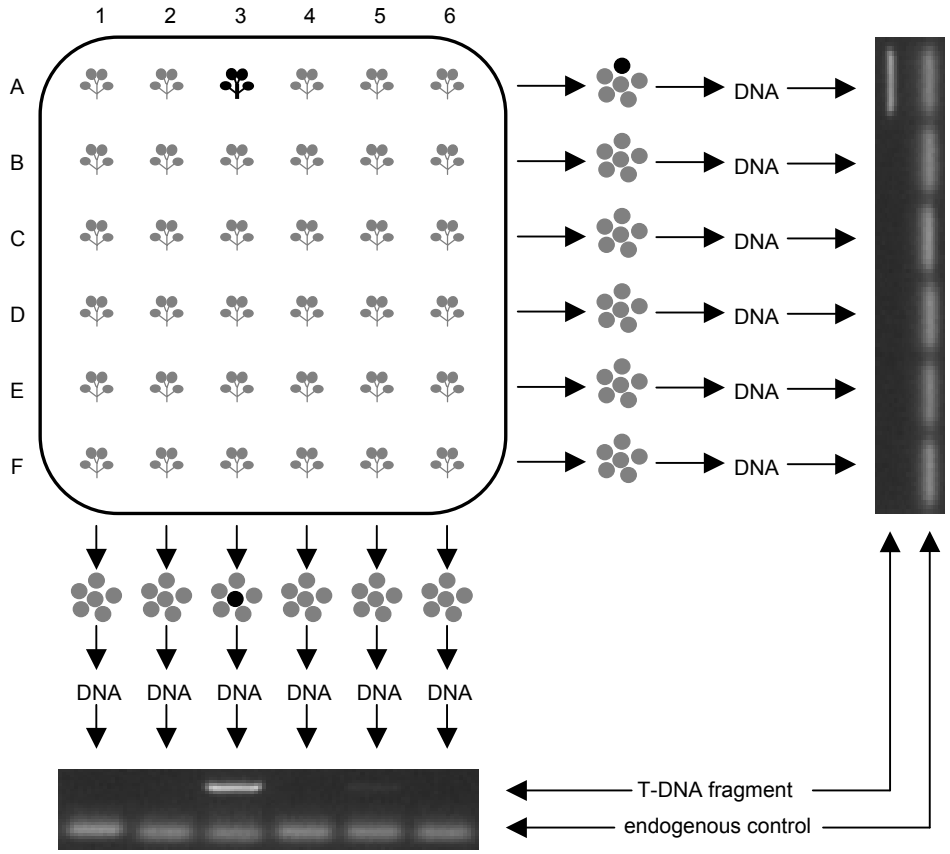

Figure 2-2. Schematic of pooled-sample PCR-screening approach.
T1 seed from a central pool of seed was plated in grids on medium without selection. As each seedling germinated and developed, its two cotyledons were harvested and pooled in either row-pools or column-pools with other cotyledons from the same rows and columns. DNA was prepared from these pools and amplified with primers for T-DNA regions and endogenous controls. Transgenic seedlings were identified by positive T-DNA amplification from both row and column pools.

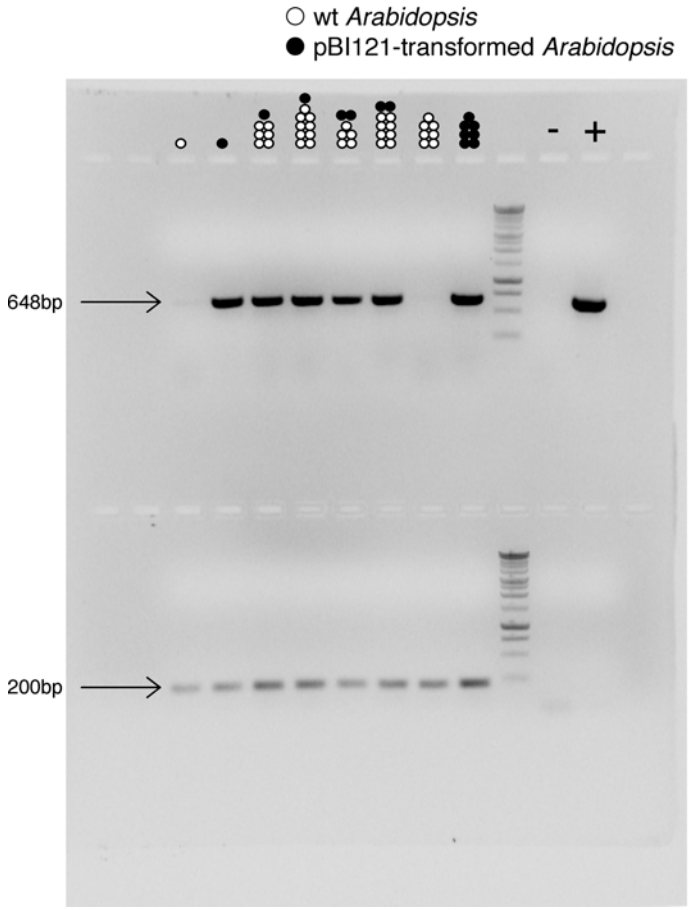

Figure 2-3. Sensitivity of PCR of pooled samples.

The ability of PCR to amplify a fragment corresponding to primers A and B (see Figure 2-1) from DNA prepared from a pool of transgenic and untransformed cotyledons was tested. Sources of template DNA used in PCR reactions are indicated above the top lanes. DNA was prepared from pooled samples of combinations of untransformed (○) and pBI121-transformed (●) *Arabidopsis* cotyledons. Control lanes include no template (-) and pBI121 plasmid (+). Amplification of the 648bp A/B product (top set of lanes) was seen from all samples containing at least one pBI121-transformed cotyledon, while no amplification was seen in samples prepared from only untransformed cotyledons. In separate reactions, amplification of the 200bp control *HMGA* fragment (lower set of lanes) was seen in all samples.

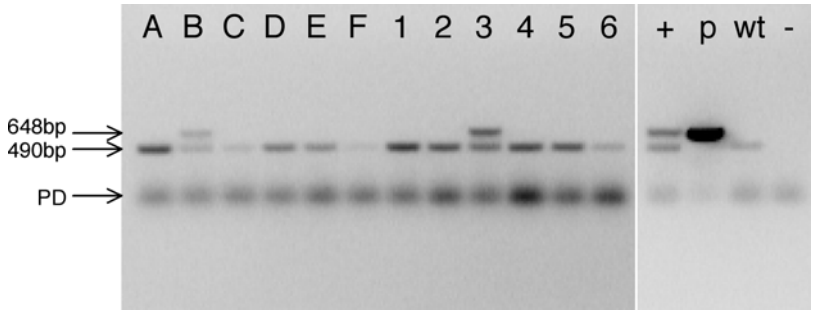

Figure 2-4. A typical PCR screen with controls.

The PCR-screen corresponding to plate M of Screen 2 is presented. PCR-products from row-pools are indicated with letters (A-F), and PCR-products from column-pools are indicated with numbers (1-6). A fragment corresponding to the predicted 648bp fragment resulting from amplification with primers A and B (see Figure 2-1) is seen in lanes B and 3. Amplification of the 490bp *ddm1* control fragment is seen in all sample lanes. Control lanes include a known pBI121-transformed *Arabidopsis* line (+), pBI121 plasmid (p), untransformed wildtype *Arabidopsis* (wt), and no template (-). The amplification conditions used typically resulted in the generation of a ~150bp fragment believed to be primer-dimers (PD).

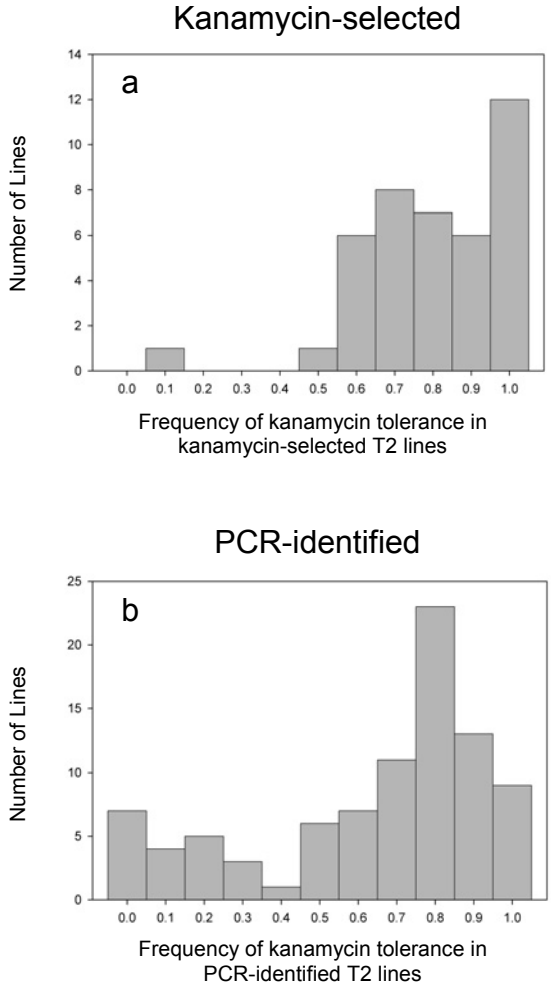

Figure 2-5. Kanamycin tolerance in kanamycin-selected and PCR-identified lines. Histograms for kanamycin tolerance in 2-week-old T2 seedlings are presented for 42 kanamycin-selected (a) and 89 PCR-identified (b) lines. In the absence of silencing, the expected frequency for a single-locus line is 0.75. Low frequencies are indicative of silencing, and high frequencies are indicative of multiple loci. PCR-identified lines are bimodally distributed with frequencies lower than 0.40 belonging to the lower mode.

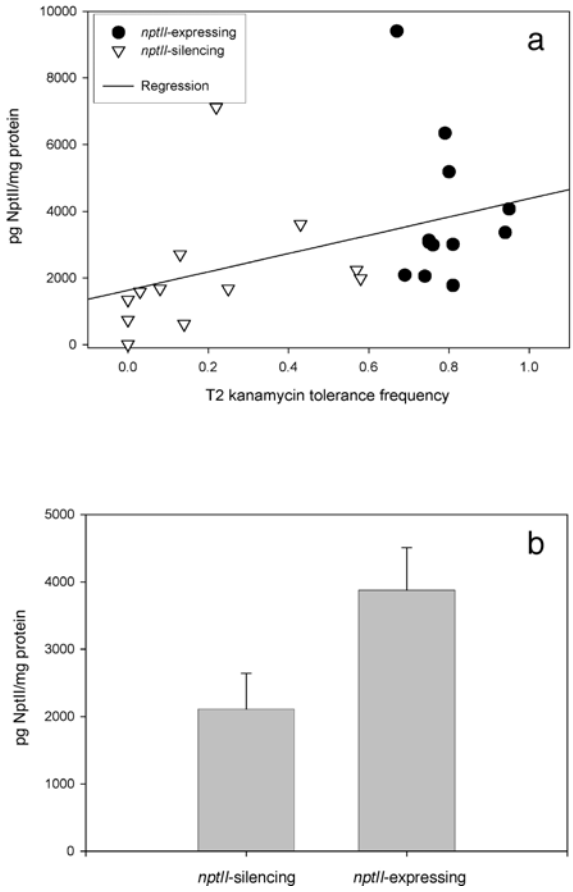

Figure 2-6. Correlation between NptII levels and kanamycin tolerance.
The relationship between the frequency of kanamycin tolerance in T2 seedlings and the level of NptII expression (as determined by ELISA) is illustrated in panel a. Open triangles (▽) represent 12 *nptII*-silencing PCR-identified lines (lines with segregation frequencies significantly lower than 0.75), and solid circles (●) represent 12 *nptII*-expressing PCR-identified lines. The significant difference ($p = 0.0218$) in levels of NptII expression between *nptII*-silencing and *nptII*-expressing lines is illustrated in panel b. Error bars indicate standard errors. Levels of NptII were standardized by the amount of total soluble protein in each sample extract.

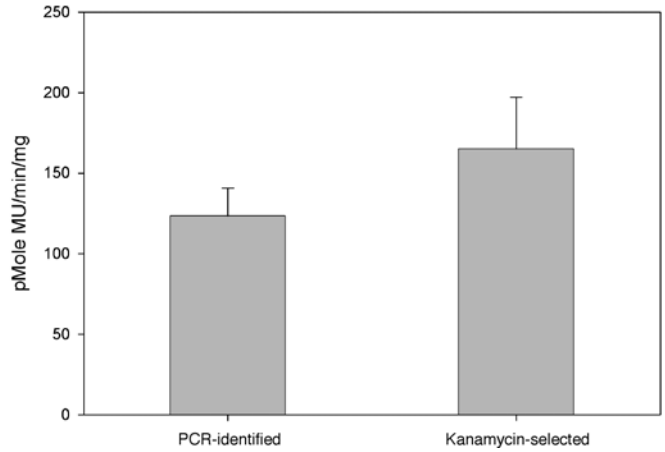

Figure 2-7. GUS activities in mature PCR-identified and kanamycin-selected T1 plants. GUS activities from 99 PCR-identified lines and 54 kanamycin-selected lines were determined based on MUG assays of pools of three mature rosette leaves. The difference in average GUS activities between PCR-identified and kanamycin-selected lines is not significant. Error bars indicate standard errors.

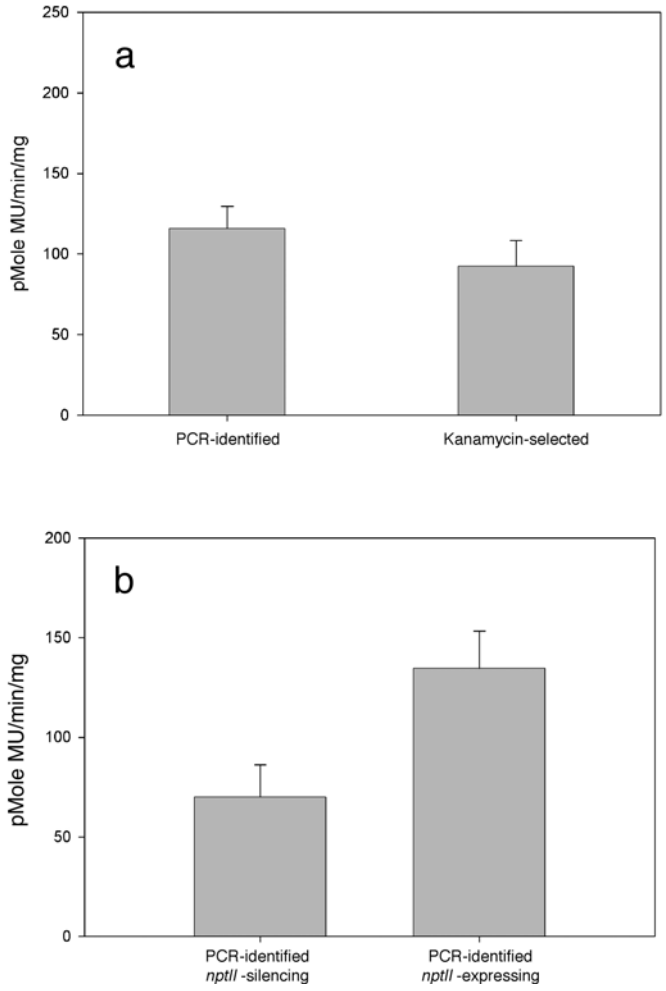

Figure 2-8. GUS activities in T2 seedlings.

GUS activities based on MUG assays of pools of ten T2 seedlings at 14 days post-germination from 89 PCR-identified lines and 41 kanamycin-selected lines are presented in plot a. The difference in average GUS activities between PCR-identified and kanamycin-selected lines is not significant. GUS activities of PCR-identified lines in plot a are presented as either *nptII*-silencing or *nptII*-expressing lines in plot b. Lines with *nptII* silencing have significantly less GUS activity than *nptII*-expressing lines ($p = 0.0073$). Error bars indicate standard errors.

```
AAAACGTCCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATATCCTGCCAccagccagccaacagctccc

AAAACGTNCCGCAATGTGTTA-
AAAACGTACCGCACTGTGTTATT-
AAAACGTACCGCCATGTGTTATTAA-
AAAACGTNCCGCAATGTGTTATTAAG-
AAAACGTNCCGCAATGTGTTATTAAGTTG-
AAAACGTNCCGCAATGTGTTATTAAGTTGTC-
AAAACCGTCCGCAATGTGTTATTAAGTTGTCTA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAG-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGC-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCG-
AAAACGGTCCGCAATGTGTTATTAAGTTGTCTAAGCGTC-
AAAACGTNCCGCANTGTGTTATTAAGTTGTCTAAGCGTC-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCA-
AAAACGTNCCGCACTGTGTTATTAAGTTGTCTAAGCGTCA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAA-
AAAACGTCCCGCAATGTGNTATTAAGNTGTCTAAGCGTCAAT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAAT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTG-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGNCAATTTGT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGT-
AAAACGTGCCGCACTGTGTTATTAAGTTGTCTAAGGG.CAATTTGTTT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACAC-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATAT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATAT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATAT-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATA-
AAAACACGTCCGCCTGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATA-
AAAACGT.CCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATA-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTNACNCCACAATATAT-
AAAACGTGGCGANNCGTGTTATTAAGTTGTCTAAGCGCCAATTTGTTTACACCACAATATATCC-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATATCCTG-
AAAACGTNCCGCNNTGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATATCCTG-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACAATATATCCTG-
CCTGCCCGGNCCGGCCGCTCGAGGCCCTATATAAGCGTCAATTTGTTTACACCACAATATATCCTG-
AAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTNTACACCACAATATATCCTGNCAccagccagccaacagc-
```

Figure 2-9. T-DNA splice junctions for 39 transgenic lines.

Precise T-DNA splice junctions are provided for 39 PCR-identified and kanamycin-selected lines. The top line represents the established sequence of pBI121 for the region examined. Sequences corresponding to the T-DNA region are in capital letters, left border sequences are underlined, and non-T-DNA binary vector sequences adjacent to the left border are in lower case letters. Plant genomic sequences, non-adjacent pBI121 vector sequences, or sequences with no known homology were found to the right of each sequence. These sequences are not shown and are indicated by a dash. Truncation past the left border sequences occurred in 18 of 39 lines (46%).

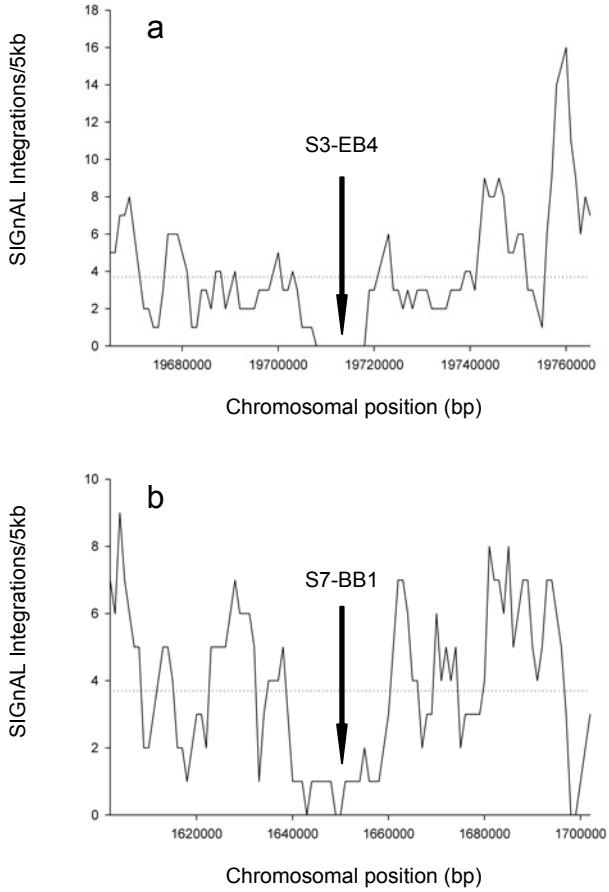

Figure 2-10. Profiles of SIGnAL project integrations surrounding two PCR-identified integrations.

Bold arrows indicate the position of integration of PCR-identified lines S3-EB4 in chromosome V (panel a) and S7-BB1 in chromosome IV (panel b). Surrounding these sites, the profiles of integrations are shown as the number of SIGnAL project (Alonso et al., 2003) integrations within a 5kb window with incremental steps of 1kb. The dotted lines indicate the average number of SIGnAL integrations per 5kb across the 100kb regions.

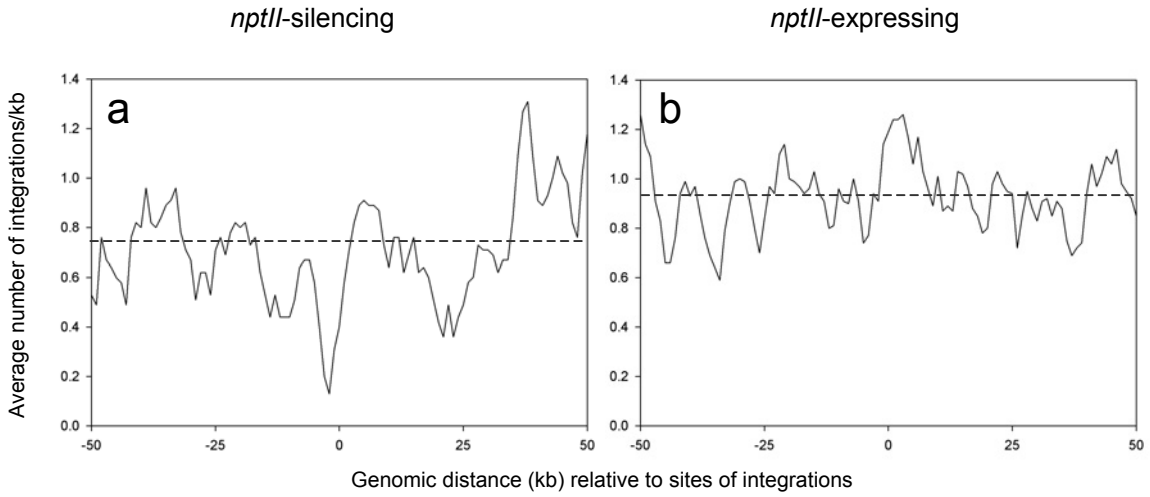

Figure 2-11. Average profiles of SIGnAL project integrations surrounding integrations of *nptII*- silencing and expressing lines identified by PCR.

Profiles of SIGnAL project integrations (Alonso et al.,2003) surrounding 13 *nptII*-silencing PCR-identified lines (panel a) and 12 *nptII*-expressing PCR-identified lines (panel b) are presented as the average number of integrations in 1kb windows. These profiles are plotted at positions relative to the sites of integration (0 on X-axis) in PCR-identified transformants. Dashed lines indicate the average number of integrations across the entire100 kb window.

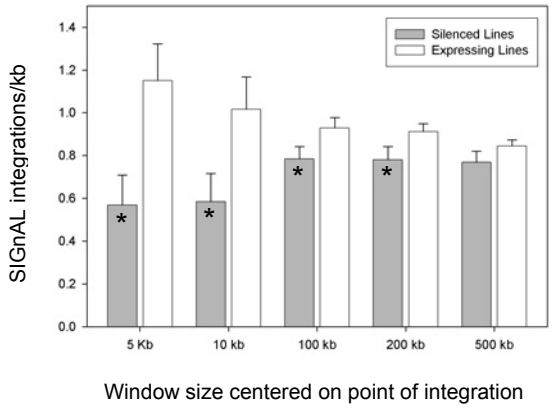

Figure 2-12. Comparison of average numbers of SIGnAL project integrations in regions surrounding *nptII*-silencing and *nptII*-expressing PCR-identified lines.
Bars represent the average number of SIGnAL project (Alonso et al., 2003) integrations in 5, 10, 100, 200, and 500kb windows centered on integration sites of PCR-identified lines. Lines with signs of *nptII* silencing are represented with gray bars, and lines with expected patterns and levels of *nptII* expression are represented with white bars. Error bars represent the standard error of the mean. Asterisks indicate significant differences at the $\alpha = 0.05$ level.

References

Allen GC, Spiker S, Thompson WF (2000) Use of matrix attachment regions (MARs) to minimize transgene silencing. Plant Mol Biol **43:** 361-376

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D, Carter DE, Marchand T, Risseeuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301:** 653-657

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403-410

Aufsatz W, Mette MF, van der Winden J, Matzke AJ, Matzke M (2002) RNA-directed DNA methylation in *Arabidopsis*. Proc Natl Acad Sci U S A **99 Suppl 4:** 16499-16506

Azpiroz-Leehan R, Feldmann KA (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. Trends Genet **13:** 152-156

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci U S A **94:** 6857-6861

Barakat A, Gallois P, Raynal M, Mestre-Ortega D, Sallaud C, Guiderdoni E, Delseny M, Bernardi G (2000) The distribution of T-DNA in the genomes of transgenic *Arabidopsis* and rice. FEBS Lett **471:** 161-164

Bode J, Benham C, Knopp A, Mielke C (2000) Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). Crit Rev Eukaryot Gene Expr **10:** 73-90

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem **72:** 248-254

Brunaud V, Balzergue S, Dubreucq B, Aubourg S, Samson F, Chauvin S, Bechtold N, Cruaud C, DeRose R, Pelletier G, Lepiniec L, Caboche M, Lecharny A (2002) T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. EMBO Rep **3:** 1152-1157

Chen S, Jin W, Wang M, Zhang F, Zhou J, Jia Q, Wu Y, Liu F, Wu P (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. Plant J **36:** 105-113

Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. Plant J **16:** 735-743

Copenhaver GP, Pikaard CS (1996) RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. Plant J **9:** 259-272

Cottage A, Yang AP, Maunders H, de Lacy RC, Ramsay NA (2001) Identification of DNA sequences flanking T-DNA insertions by PCR-walking. Plant Mol Biol Rep **19:** 321-327

Cubero J, Martinez MC, Llop P, Lopez MM (1999) A simple and efficient PCR method for the detection of *Agrobacterium tumefaciens* in plant tumours. J Appl Microbiol **86:** 591-602

De Buck S, Jacobs A, Van Montagu M, Depicker A (1998) *Agrobacterium tumefaciens* transformation and cotransformation frequencies of *Arabidopsis thaliana* root explants and tobacco protoplasts. Mol Plant-Microbe Int **11:** 449-457

De Buck S, Jacobs A, Van Montagu M, Depicker A (1999) The DNA sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. Plant J **20:** 295-304

De Buck S, Van Montagu M, Depicker A (2001) Transgene silencing of invertedly repeated transgenes is released upon deletion of one of the transgenes involved. Plant Mol Biol **46:** 433-445

de Vetten N, Wolters AM, Raemakers K, van der Meer I, ter Stege R, Heeres E, Heeres P, Visser R (2003) A transformation method for obtaining marker-free plants of a cross-pollinating and vegetatively propagated crop. Nat Biotechnol **21:** 439-442

Dietz A, Kay V, Schlake T, Landsmann J, Bode J (1994) A plant scaffold attached region detected close to a T-DNA integration site is active in mammalian cells. Nucleic Acids Res **22:** 2744-2751

Dominguez A, Fagoaga C, Navarro L, Moreno P, Pena L (2002) Regeneration of transgenic citrus plants under non selective conditions results in high-frequency recovery of plants with silenced transgenes. Mol Genet Genomics **267:** 544-556

Elmayan T, Vaucheret H (1996) Expression of single copies of a strongly expressed 35S transgene can be silenced post-transcriptionally. Plant J **9:** 787-797

Emani C, Sunilkumar G, Rathore KS (2002) Transgene silencing and reactivation in sorghum. Plant Sci **162:** 181-192

Feldmann KA (1991) T-DNA Insertion Mutagenesis in *Arabidopsis* - Mutational Spectrum. Plant J **1:** 71-82

Fobert PR, Miki BL, Iyer VN (1991) Detection of gene regulatory signals in plants revealed by T-DNA-mediated fusions. Plant Mol Biol **17:** 837-851

Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R (2003) A comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis thaliana* genome. Plant Mol Biol **52:** 161-176

Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I (2002) Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. Proc Natl Acad Sci U S A **99:** 14584-14589

Gallagher SR (1992) Quantitation of GUS activity by fluorometry. *In* SR Gallagher, ed, GUS Protocols: Using the GUS Gene as a Reporter of Gene Expression. Academic Press, Inc., San Diego, pp 47-59

Gelvin SB (2003) Improving plant genetic engineering by manipulating the host. Trends Biotechnol **21:** 95-98

Gheysen G, Herman L, Breyne P, Gielen J, Vanmontagu M, Depicker A (1990) Cloning and sequence-analysis of truncated T-DNA inserts from *Nicotiana-tabacum*. Gene **94:** 155-163

Han KH, Ma CP, Strauss SH (1997) Matrix attachment regions (MARs) enhance transformation frequency and transgene expression in poplar. Transgenic Res **6:** 415-420

Herman L, Jacobs A, Van Montagu M, Depicker A (1990) Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. Mol Gen Genet **224:** 248-256

Hui EK, Wang PC, Lo SJ (1998) Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. Cell Mol Life Sci **54:** 1403-1411

Ichikawa T, Nakazawa M, Kawashima M, Muto S, Gohda K, Suzuki K, Ishikawa A, Kobayashi H, Yoshizumi T, Tsumoto Y, Tsuhara Y, Iizumi H, Goto Y, Matsui M (2003) Sequence database of 1172 T-DNA insertion sites in *Arabidopsis* activation-tagging lines that showed phenotypes in T1 generation. Plant J **36:** 421-429

Janakiraman V, Steinau M, McCoy SB, Trick HN (2002) Recent advances in wheat transformation. In Vitro Cellular & Developmental Biology-Plant **38:** 404-414

Jeddeloh JA, Bender J, Richards EJ (1998) The DNA methylation locus DDM1 is required for maintenance of gene silencing in *Arabidopsis*. Genes Dev **12:** 1714-1725

Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO J **6:** 3901-3907

Kohli A, Griffiths S, Palacios N, Twyman RM, Vain P, Laurie DA, Christou P (1999) Molecular characterization of transforming plasmid rearrangements in transgenic rice reveals a recombination hotspot in the CaMV 35S promoter and confirms the predominance of microhomology mediated recombination. Plant J **17**: 591-601

Koncz C, Martini N, Mayerhofer R, Koncz-Kalman Z, Korber H, Redei GP, Schell J (1989) High-frequency T-DNA-mediated gene tagging in plants. Proc Natl Acad Sci U S A **86:** 8467-8471

Koncz C, Schell J (1986) The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimeric genes carried by a novel type of *Agrobacterium* binary vector. Mol Gen Genet **204:** 383-396

Kononov ME, Bassuner B, Gelvin SB (1997) Integration of T-DNA binary vector 'backbone' sequences into the tobacco genome: evidence for multiple complex patterns of integration. Plant J **11:** 945-957

Krysan PJ, Young JC, Jester PJ, Monson S, Copenhaver G, Preuss D, Sussman MR (2002) Characterization of T-DNA insertion sites in *Arabidopsis thaliana* and the implications for saturation mutagenesis. Omics **6:** 163-174

Krysan PJ, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell **11:** 2283-2290

Lindsey K, Wei W, Clarke MC, McArdle HF, Rooke LM, Topping JF (1993) Tagging genomic sequences that direct transgene expression by activation of a promoter trap in plants. Transgenic Res **2:** 33-47

Lowry OH, Rosebrough NJ, Farr AL, Randall RJ (1951) Protein measurement with the Folin phenol reagent. J Biol Chem **193:** 265-275

Ma C, Mitra A (2002) Intrinsic direct repeats generate consistent post-transcriptional gene silencing in tobacco. Plant J **31:** 37-49

Makarevitch I, Svitashev SK, Somers DA (2003) Complete sequence analysis of transgene loci from plants transformed via microprojectile bombardment. Plant Mol Biol **52:** 421-432

Mathur J, Koncz C (1998) Callus culture and regeneration. *In* J Martinez-Zapater, J Salinas, eds, Methods in Molecular Biology, Vol. 82: *Arabidopsis* Protocols, Vol 82. Humana Press Inc., Totowa, NJ, pp 31-34

Mitsuhara I, Shirasawa-Seo N, Iwai T, Nakamura S, Honkura R, Ohashi Y (2002) Release from post-transcriptional gene silencing by cell proliferation in transgenic tobacco plants: possible mechanism for noninheritance of the silencing. Genetics **160:** 343-352

Mlynarova L, Loonen A, Mietkiewska E, Jansen RC, Nap JP (2002) Assembly of two transgenes in an artificial chromatin domain gives highly coordinated expression in tobacco. Genetics **160:** 727-740

Murashige T, Skoog F (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiologia Plantarum **15:** 473-475

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. Nucleic Acids Res **31:** 1245-1251

Peach C, Velten J (1991) Transgene expression variability (position effect) of CAT and GUS reporter genes driven by linked divergent T-DNA promoters. Plant Mol Biol **17:** 49-60

Permingeat HR, Alvarez ML, Cervigni GD, Ravizzini RA, Vallejos RH (2003) Stable wheat transformation obtained without selectable markers. Plant Mol Biol **52:** 415-419

Qin H, Dong Y, von Arnim AG (2003) Epigenetic interactions between *Arabidopsis* transgenes: characterization in light of transgene integration sites. Plant Mol Biol **52:** 217-231

Rios G, Lossow A, Hertel B, Breuer F, Schaefer S, Broich M, Kleinow T, Jasik J, Winter J, Ferrando A, Farras R, Panicot M, Henriques R, Mariaux JB, Oberschall A, Molnar G, Berendzen K, Shukla V, Lafos M, Koncz Z, Redei GP, Schell J, Koncz C (2002) Rapid identification of *Arabidopsis* insertion mutants by non-radioactive detection of T-DNA tagged genes. Plant J **32:** 243-253

Roseman RR, Johnson EA, Rodesch CK, Bjerke M, Nagoshi RN, Geyer PK (1995) A P element containing suppressor of hairy-wing binding regions has novel properties for mutagenesis in *Drosophila melanogaster*. Genetics **141:** 1061-1074

Sawasaki T, Takahashi M, Goshima N, Morikawa H (1998) Structures of transgene loci in transgenic *Arabidopsis* plants obtained by particle bombardment: junction regions can bind to nuclear matrices. Gene **218:** 27-35

Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, Clarke JD, Cotton D, Bullis D, Snell J, Miguel T, Hutchison D, Kimmerly B, Mitzel T, Katagiri F, Glazebrook J, Law M, Goff SA (2002) A high-throughput *Arabidopsis* reverse genetics system. Plant Cell **14:** 2985-2994

Shimizu K, Takahashi M, Goshima N, Kawakami S, Irifune K, Morikawa H (2001) Presence of an SAR-like sequence in junction regions between introduced transgene and genomic DNA of cultured tobacco cells: its effect on transformation frequency. Plant J **26**: 375-384

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res **23:** 1087-1088

Smith N, Kilpatrick JB, Whitelam GC (2001) Superfluous transgene integration in plants. Crit Rev Plant Sci **20:** 215-249

Szabados L, Kovacs I, Oberschall A, Abraham E, Kerekes I, Zsigmond L, Nagy R, Alvarado M, Krasovskaja I, Gal M, Berente A, Redei GP, Haim AB, Koncz C (2002) Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. Plant J **32:** 233-242

Szittya G, Silhavy D, Molnar A, Havelda Z, Lovas A, Lakatos L, Banfalvi Z, Burgyan J (2003) Low temperature inhibits RNA silencing-mediated defence by the control of siRNA generation. EMBO J **22:** 633-640

Takano M, Egawa H, Ikeda JE, Wakasa K (1997) The structures of integration sites in transgenic rice. Plant J **11**: 353-361

Thomas CM, Jones DA, English JJ, Carroll BJ, Bennetzen JL, Harrison K, Burbidge A, Bishop GJ, Jones JD (1994) Analysis of the chromosomal distribution of transposon-carrying T-DNAs in tomato using the inverse polymerase chain reaction. Mol Gen Genet **242:** 573-585

Tinland B (1996) The integration of T-DNA into plant genomes. Trends Plant Sci **1:** 178-184

Topping JF, Lindsey K (1995) Insertional mutagenesis and promoter trapping in plants for the isolation of genes and the study of development. Transgenic Res **4:** 291-305

Topping JF, Wei W, Lindsey K (1991) Functional tagging of regulatory elements in the plant genome. Development **112:** 1009-1019

Ülker B, Weissinger AK, Spiker S (2002) *E. coli* chromosomal DNA in a transgene locus created by microprojectile bombardment in tobacco. Transgenic Res **11:** 311-313

van Holde K (1989) Chromatin. Springer-Verlag, New York

Wang ZY, Hopkins A, Mian R (2001) Forage and turf grass biotechnology. Crit Rev Plant Sci **20:** 573-619

Wassenegger M, Heimes S, Riedel L, Sanger HL (1994) RNA-directed *de novo* methylation of genomic sequences in plants. Cell **76:** 567-576

Wenck A, Czako M, Kanevski I, Marton L (1997) Frequent collinear long transfer of DNA inclusive of the whole binary vector during *Agrobacterium*-mediated transformation. Plant Mol Biol **34:** 913-922

Wolffe AP (1998) Chromatin: Structure and Function, Ed 3rd. Academic Press, San Diego

Zar JH (1996) Biostatistical Analysis, Ed Third. Pretence Hall, Upper Saddle River, New Jersey

Appendices

Appendix A

Modification and Testing of a MUG Assay for Use in 96-Well Plates

Introduction

The use of the *E. coli gusA* (β-glucuronidase (GUS)) gene as a reporter gene has revolutionized plant biotechnology. Since its first reported use in plants by Jefferson et al. (1987), *gusA* has been used as an effective reporter gene to test the strength and specificity of promoters and other DNA elements, it has been used as a fusion protein to determine the localization of endogenous plant proteins, and it has even been used as a selectable marker (Joersbo and Okkels, 1996). One of the key attributes of GUS is that it can be detected by either a histochemical staining assay or highly quantitative activity assays (Jefferson, 1987). GUS detection assays function by monitoring the cleavage of a substrate which releases a colormetric, fluorescent, or luminecent compound that can be readily detected and in some cases quantitatively measured. For the histochemical staining assay, 5-bromo-4-chloro-3-indoyl glucuronide (X-gluc) is used as a substrate. X-gluc staining can provide highly detailed information about the localization of GUS in various plant tissues, however, it can not be used as an accurate measure of the quantity of GUS present in those tissues. Uniform substrate penetration into various tissues and rapid saturation present major problems in attempting to quantify GUS activity with X-gluc (Stomp, 1992).

Fortunately, quantitative alternatives exist. The most commonly used quantitative GUS assay utilizes the substrate 4-methyl umbelliferyl glucuronide (MUG) and is appropriately referred to as a MUG assay (Jefferson, 1987; Gallagher, 1992). MUG assays are extremely quantitative but are highly time consuming and logistically cumbersome. This is especially true if large numbers of samples need to be analyzed. MUG assays involve the extraction of the highly stable GUS enzyme from the sampled plant tissues. The extract is then used in a controlled reaction to cleave the MUG substrate. When MUG is cleaved, the fluorescent compound methylumbelliferone (MU)

is liberated. As the reaction progresses, aliquots are removed from the reaction vessel and transferred to a solution of sodium carbonate. The sodium carbonate acts to stop the cleavage of MUG and greatly increases the fluorescence of the liberated MU (Gallagher, 1992). Multiple aliquots are typically sampled at specific time points. By comparing the fluorescence of each sample to an internal control curve based on the fluorescence of known concentrations of MU, the amount of liberated MU in each sample can be determined. The rate of the reaction for each sample is then determined based on the molar amount of MU present in aliquots from each time point.

For the experiments described in this dissertation, large numbers of samples needed to be quantitatively assessed for GUS activity at multiple time points. The logistics of conducting MUG assays as described by Jefferson (1987) and Gallagher(1992) essentially limits the total number of samples that can be measured in one reaction to around twenty. A rapid and efficient method to quantify GUS activity from multiple samples was needed.

A kit for the quantification of GUS activity in a 96-well format is commercially available. The GUS-Light System (Applied Biosystems) is based on the decomposition of a proprietary chemiluminescent substrate by the GUS enzyme. The use of this GUS detection system has previously been used in our lab for analysis of GUS activity in tobacco plants (Ülker et al., 1999) and appeared to offer a convenient solution to the need for a rapid and efficient method to quantify GUS activity in *Arabidopsis*. However, further investigation revealed that extracts from *Arabidopsis* plants differentially inhibited the GUS-Light reaction and resulted in variable results. Furthermore, the GUS-Light System kits are expensive and a cheaper alternative was desirable.

Results and Discussion

Arabidopsis and Tobacco extracts inhibit GUS-Light assays

To determine if plant extracts can inhibit the GUS-Light reaction or the measurement of the GUS-Light luminescent product, a series of standards were tested in

the GUS-Light assay. Manufacturer's instructions were followed, and standard curves encompassing a range from 10^{-6} to 10^{-2} GUS units/µl were prepared in extracts from *Arabidopsis* rosette leaves, tobacco leaves, and NT1 cells. Extracts were prepared as previously described (Ülker et al., 1999). For each tissue type, 50mg of tissue was homogenized in 300µl GUS-Light lysis buffer (provided with kit). A control without plant extract was also prepared in lysis buffer. Figure A-1 clearly demonstrates that standard curves prepared in extracts from *Arabidopsis* rosette leaves, tobacco leaves, or NT1 cells result in lower luminescence than curves prepared in lysis buffer without plant extracts for GUS concentrations greater than 10^{-5} units/µl. Furthermore, the standard curves prepared in plant extracts are flat at the lower ends of the curves.

While an effect of the plant extracts was clearly seen, it can not be determined by the experiments presented here what caused the lower luminescence values. It is possible that a compound (or series of compounds) present in plant extracts interfered with the GUS-Light assay resulting in reduced cleavage of the chemiluminescent substrate. It is also possible that light absorbing pigments present in the plant extracts are absorbing some luminescence from the cleaved substrate. This may explain why less luminescence was observed in *Arabidopsis* and tobacco leaf extracts than in NT1 cell extracts.

Even though plant extracts affect the GUS-Light luminescence values, it may be possible to successfully use the system if the samples can be diluted to a level where inhibition is no longer occurs. Figure A-2 illustrates that concentrations as low as 1% *Arabidopsis* leaf extract (prepared as described above and diluted in lysis buffer) can result in lower GUS-Light luminescence. In order to determine if extracts from similar plants can differentially inhibit GUS-Light luminescence, leaves were harvested from four different untransformed *Arabidopsis* full-sib plants from an inbred line grown together. From each plant three young leaves (1 to 1.5cm) and three old leaves (4 to 6cm) were harvested and used to prepare extracts (three young leaf extracts and three old leaf extracts from each plant). GUS to a final concentration of 10^{-2} units/µl was added to each sample. The results of GUS-Light assays performed on the spiked extracts are reported in Figure A-3. Extracts from similar plants resulted in very different GUS-Light

luminescence values. The samples were not standardized by protein concentration, however the variations between leaves from the same plant are very low (see error bars), suggesting extractions were uniform. No consistent trend between young and old leaves from the same plant was observed. A control containing no plant extract was also included in the assay. The luminescence of the control was 804600 RLUs.

Arabidopsis leaf extracts do not affect MUG assays

Because the degree of inhibition of GUS-Light luminescence by *Arabidopsis* extracts was highly variable, an alternative approach to measuring GUS activity in a rapid and efficient manner was desired. A MUG assay procedure (Jefferson et al., 1987) was modified for use in a 96-well plate. In order to determine if extracts from similar plants differentially affected MUG assay results, as was seen with GUS-Light results, a similar approach was taken. Leaves were harvested from three different untransformed *Arabidopsis* full-sib plants from an inbred line grown together. From each plant three young leaves (1 to 1.5cm) and three old leaves (4 to 6cm) were harvested and used to prepare extracts (three young leaf extracts and three old leaf extracts from each plant). Tissue samples were homogenized and suspended in 1ml GUS Extraction Buffer. A control with no plant extract was also assayed. GUS to a final concentration of 10^{-2} units/µl was added to each sample and control. The results of MUG assays performed on the spiked extracts are reported in Figure A-4. None of the extract samples were significantly different from the control.

Refinement of MUG Assay

Previously described protocols for MUG assays (Jefferson et al., 1987; Gallagher, 1992) typically involve measuring the reaction at several time points to determine the rate of MU liberation from MUG by the GUS enzyme. It is important to measure the reaction at several time points to determine if the rate is slowing down (i.e. if substrate is limiting). Repeated MUG assays of *Arabidopsis* leaf and seedling tissues using the conditions, concentrations, and volumes described above consistently resulted in a highly

linear rate for up to one hour after the reaction was started based on 5 measurements taken at 10 to 20 minute intervals (data not shown). If GUS activities could be reliably determined by taking a single MUG assay measurement, the number of samples that could be processed in a single plate and the number of plates that could be processed in a single day could each be increased 5-fold. In order to determine if GUS activity based on a single MUG assay time point taken at 20 minutes was a valid indicator of GUS activity based on a MUG assay involving five time points over one hour, a comparison of the two approaches was made. An extremely high correlation ($R^2 = 0.9946$) between the two approaches was observed (see Figure A-5), and it was concluded that the single-point approach was an appropriate assessment of GUS activity. It should be noted, however, that this conclusion applies only to the tissues analyzed and the methodologies used in this circumstance. A single-point approach to MUG assays may not be appropriate for other conditions. If samples with particularly high GUS activity were assayed, it would be necessary to ensure that RFUs were not exceeding the detection limits of the plate reader or that the MUG substrate was not rate limiting resulting in a plateau.

Real-time MUG Assays offer less sensitivity than stopped assays

The FLUO-star fluorescent plate reader used in these analyses has the ability to monitor fluorescence of a particular MUG reaction as that reaction progresses. Because of this ability to monitor fluorescence in real-time, it may be possible to monitor MU liberation at multiple time points without the inconvenience of generating multiple stopped samples for each reaction. However, since the fluorescence of MU is enhanced in basic conditions, the fluorescence in the assay reaction may be less than the fluorescence of a sample of that reaction stopped with the basic Stop Buffer solution. To test this, samples were prepared to emulate conditions similar to either a stopped sample of a reaction with a known concentration of liberated MU, or a real-time measurement of a reaction with a known concentration of liberated MU. Figure A-6 illustrates the difference in fluorescence by these two approaches. Fluorescence of a small sample of a reaction is greater than the fluorescence of the entire reaction if that small sample is

combined with Stop Buffer. Nevertheless, the curve of the real-time measurement was linear for MU concentrations greater than 0.4μM MU in the reaction. Because the real-time curve has a lower slope than the stopped curve, it would be less sensitive over the range of MU concentrations analyzed in Figure A-6. It could, however, cover a much larger range of MU concentrations than the stopped curve, since the FLUO-star's highest relative fluorescent unit (RFU) value is 65000. However, unlike the stopped curve, the real-time curve is flat at concentrations of MU below ~0.3μM MU. This is probably a result of the low level of fluorescence of uncleaved MUG. In the real-time measurement fluorescence is monitored in the assay buffer, which contains MUG. The stopped reaction contains very little MUG (only 1/13 of the uncleaved MUG in the assay buffer), and is probably only minimally affected by its fluorescence. As a result, the real-time measurement of MUG activity may lack the ability to distinguish samples with very low GUS activity from samples with no GUS activity. This appears to be the case. Figure A-7 provides GUS activities (pMole MU/min/mg) for three transgenic samples and an untransformed wild type control. GUS activities were determined by MUG assays using either fluorescence of stopped reactions or real-time fluorescence of the reaction as it progressed. Both approaches seem appropriate for determining relative differences between samples JA4 or HD2, however, only the stopped reaction approach (reactions run in assay tubes) was able to distinguish sample CC2 as having GUS activity above background levels.

In conclusion, MUG assays where a single measurement of fluorescence of the stopped reaction is made at 20 minutes appears to be the most efficient and convenient method analyzed for the measurement of GUS activities.

MUG Assay Protocol

Duplicate reps of 40 samples can be measured in each assay, one of which should be a wild type control. The last 8 cells must be reserved for a set of standards.

GUS Extraction:

Arabidopsis leaf tissue, whole leaves, and young whole seedlings (2-4 weeks) have been used. Total weight should not exceed 50mg.

1) Place a small scoop (~50 to 100µl) of 1mm glass beads in a microcentrifuge tube.
2) Place tissue in tube and freeze in liquid nitrogen (samples can be stored at –80C).
3) Remove sample tube from liquid nitrogen and immediately mix for 5sec in a dental amalgamator.
4) Add 1ml of GUS Extraction Buffer.
5) Mix again for 5sec in a dental amalgamator.

   Note: Sample tubes may leak at this point if glass beads interfere with a tight closure. I have found that Safelock tubes (Eppendorf) work well. Tapping town the glass beads before opening the tube to add buffer also helps.
6) Spin at max for 15min in a microcentrifuge at room temperature.
7) Collect 200µl supernatant and transfer to labeled microdilution tubes (USA Scientific) in the same pattern as will be used in the 96-well plate.
8) Keep extracts on ice.

Assay:

This assay makes heavy use of 12-chanel pipets. It is not practical to use conventional single-channel pipets for this assay. A timer set to go off at 20 minutes is very helpful.

1) Prepare MUG Assay Buffer.

   Note: 20ml is more than is needed, however making a smaller volume involves weighing < 5mg of MUG which is difficult to consistently do.
2) Label two sets of reaction tubes (microdilution tubes).
3) Add 10µl of each extract to each set of tubes.

4) Start timer set to go off at 20 minutes.

5) Add 130ul MUG Assay Buffer to each set of reaction tubes.

Note: Thorough mixing is important. By adding the 130µl Assay Buffer to the 10µl sample, this is achieved without the need to mix by pipeting.

6) Immediately place reaction tubes in 37°C water bath and cover.

7) While reaction is running, load 190ul Stop Buffer to every well in a 96-well plate.

8) At 20 minutes, transfer 10ul of each reaction to appropriate well in 96-well plate containing the Stop Buffer.

10) Cover 96-well plate with foil and store in dark place at room temperature.

Standards:

Make dilutions in extraction buffer with 7.1% wt extract to maintain an equivalent concentration of plant extract in each well (520µl GUS Extraction Buffer, 40µl wt extract).

1) make the following serial dilutions:

90µl	50µM MU standard	→	50	µM MU
90µl	50µM MU standard + 90µl 7.1% Ext.	→	25	µM MU
10µl	50µM MU standard + 90µl 7.1% Ext.	→	5	µM MU
10µl	25µM MU standard + 90µl 7.1% Ext.	→	2.5	µM MU
10µl	5µM MU standard + 90µl 7.1% Ext.	→	0.5	µM MU
10µl	2.5µM MU standard + 90µl 7.1% Ext.	→	0.25	µM MU
	90µl 7.1% Ext.	→	0.00	µM MU

2) add 10µl of each standard to the appropriate wells containing Stop Buffer in the 96-well plate

Note: The MU standards are not included in the 20 minute reaction. They are added directly to the Stop Buffer, and measured in the plate reader. Standards can be added to appropriate wells containing Stop Buffer while the 20 minute reaction is proceeding, or can be added after the sample reactions have been stopped. MU standards should always be protected from light.

Fluorescence Measurement:

The plates should be scanned on a fluorescent plate reader, such as a FLUO-star (BMG Scientific Inc., Durham, NC), with an excitation filter for 355nm and an emission filter for 460nm.

> Note: The FLUO-star can not read higher than 65000 relative fluorescent units (RFUs). If any values are greater than 65000, you should either run the reactions again with diluted extract or make a 10-fold dilution of the entire plate, and re-scan it. Inconsistent results have previously been associated with making dilutions, therefore repeating the assay with diluted extracts is recommended.

Protein Determinations:

The GUS Extraction Buffer in this assay is compatible with the Bio-Rad Protein Assay kit which utilizes the methods of Bradford (Bradford, 1976). Follow directions included with the kit.

Analysis:

Data should be converted to pMole of MU liberated per minute per mg of protein (pMole MU/min/mg). It is most convenient to set up a spreadsheet that automatically performs all of the calculations.

1) Using the MU standard curve, determine the moles of MU in each sample well.
2) Calculate the rate of each reaction by dividing by 20 minutes.
3) Standardize each sample by dividing by the total mg of protein added to each reaction.

Other Comments:

The volumes and concentrations in this protocol are optimized for measurement of GUS driven by the 35S promoter in *Arabidopsis* tissues. This optimization typically brings the activities within a range that is easily measured by the FLUO-star without extensive dilutions. It is very likely that other species, tissues,

promoters, or extraction methods will result in changes that could result in activities that are outside of the FLUO-star's range. If this occurs, the following could be employed to avoid the need for extensive dilutions:

-Change tissue/extraction buffer ratios to make the extract more or less concentrated.

-Change the amount of extract that is added to the assay reaction.

Do not attempt to load less than 10ul with 12-chanel pipets

-Change the time point that the reaction is stopped.

-Change the amount of reaction that is added to stop buffer.

Do not attempt to load less than 10ul with 12-chanel pipets

Solutions

GUS Extraction Buffer

150mM Sodium Phosphate pH 7.0
10mM EDTA
10mM β-mercaptoethanol
0.1% Triton X-100
0.1% Sarcosyl
140µM PMSF

MUG Assay Buffer (1mM MUG)

8.3mg MUG
20ml GUS Extraction Buffer

Stop Buffer (200mM Na_2CO_3)

21.1g Sodium Carbonate in 1L water

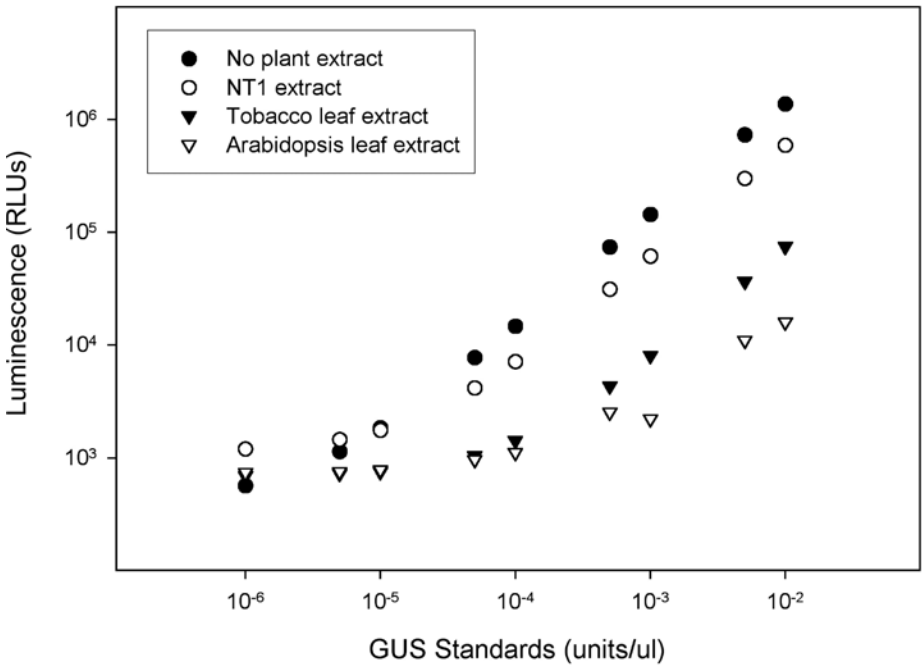

Figure A-1. Inhibition of GUS-Light luminescence by plant extracts. Untransformed plant extracts from NT1 tobacco cells (○), tobacco leaves (▼), and *Arabidopsis* rosette leaves (▽) were spiked with nine different concentrations of GUS ranging from 10^{-6} to 10^{-2} units/µl. Plant extracts consisted of 50 mg of tissue extracted into 300 µl GUS-Light lysis buffer. A set of controls containing no plant extract (●) were also assayed. For all samples containing plant extracts, luminescence was lower than controls at GUS concentrations greater than 10^{-5} units/µl and luminescence was higher than the control at the lowest GUS concentration (10^{-6} units/µl). Note log scales.

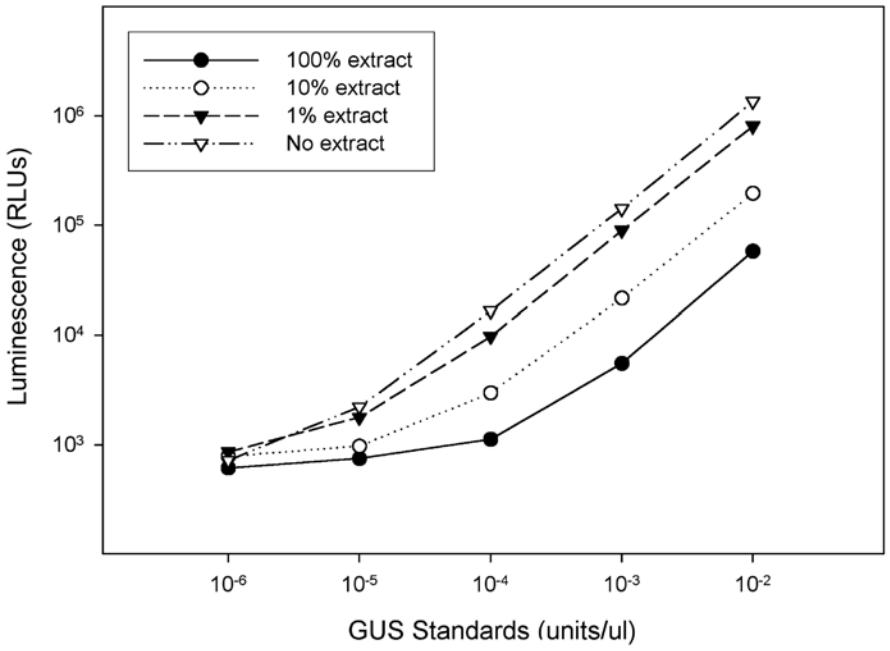

Figure A-2. Inhibition of GUS-Light luminescence by varying concentrations of *Arabidopsis* leaf extract.

Decreasing concentrations of untransformed *Arabidopsis* rosette leaf extract were compared for inhibition of GUS-Light luminescence. Concentrations of 100% (●), 10% (○), and 1% (▼) were compared against controls containing no plant extract (▽). Inhibition of luminescence in samples with GUS concentrations greater than 10^{-5} units/µl was detected for all *Arabidopsis* extract concentrations. Note log scales.

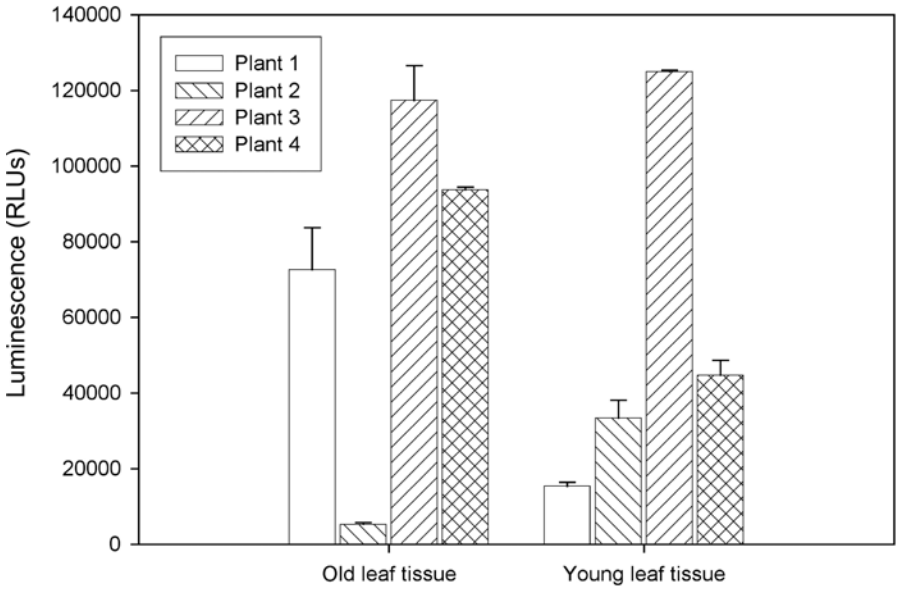

Figure A-3. Inconsistent inhibition of GUS-Light luminescence by *Arabidopsis* extracts from similar sources.

Extracts from three young (1 to 1.5cm) and three old (4 to 6cm) leaves were prepared from four different 6-week-old untransformed *Arabidopsis* plants grown in identical conditions (same tray). Extracts were spiked with GUS to a final concentration of 10^{-2} units/µl. Each bar represents the average GUS-Light luminescence from extracts prepared from each of the three young or old leaves from each plant. Error bars indicate standard errors. A control sample included in the assay is not included in the figure due to its much higher average luminescence (804600 RLUs).

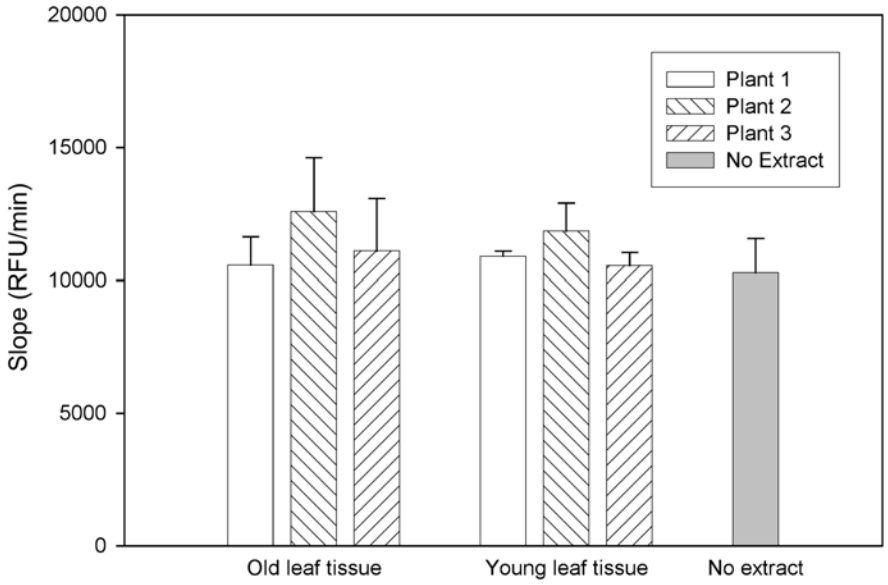

Figure A-4. No observed inhibition of MUG assay by *Arabidopsis* extracts. Extracts from three young (1 to 1.5cm) and three old (4 to 6cm) leaves were prepared from three different 6-week-old untransformed *Arabidopsis* plants grown in identical conditions (same tray). Extracts were spiked with GUS to a final concentration of 10^{-2} units/µl, and MUG assays were conducted. Each bar represents the average slope (RFU/min) from extracts prepared from each of the three young or old leaves from each plant. Error bars indicate standard errors. No significant difference exists between any *Arabidopsis* extract sample and the no extract control.

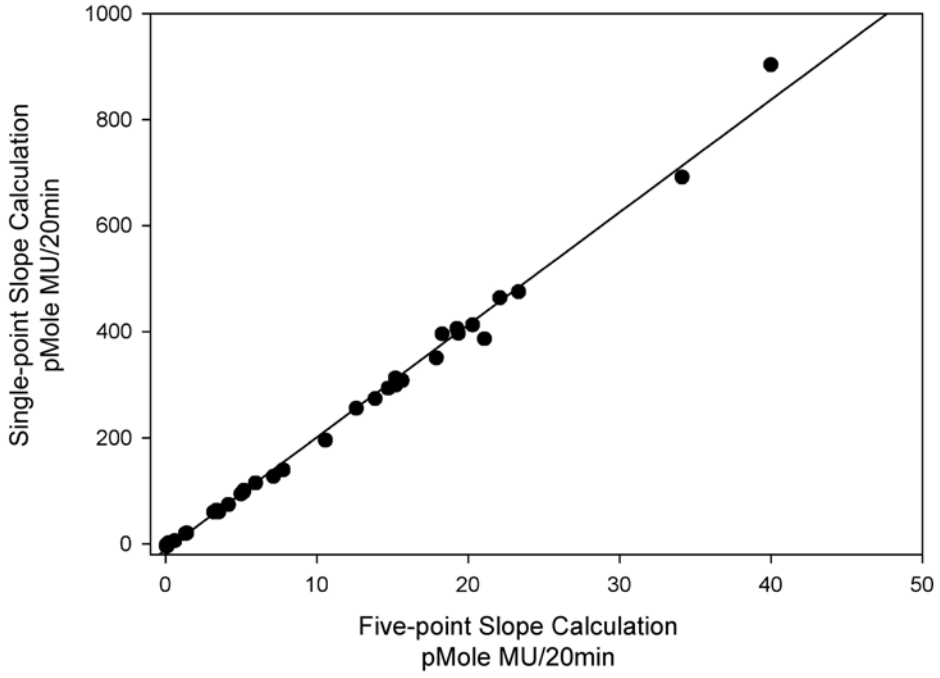

Figure A-5. Correlation between single-point and five-point calculations of MUG assay rate slopes.

The MU liberation rates (pMole MU/min) were determined for 41 samples using standard MUG assay approaches (Jefferson et al., 1987) incorporating five reaction stop points (1, 5, 10, 15, and 20 minutes). The MU liberation rates for the same samples were then calculated based solely on the 20 minute time point (pMole MU/20min). A very strong correlation between these two approaches was observed ($R^2 = 0.9946$)

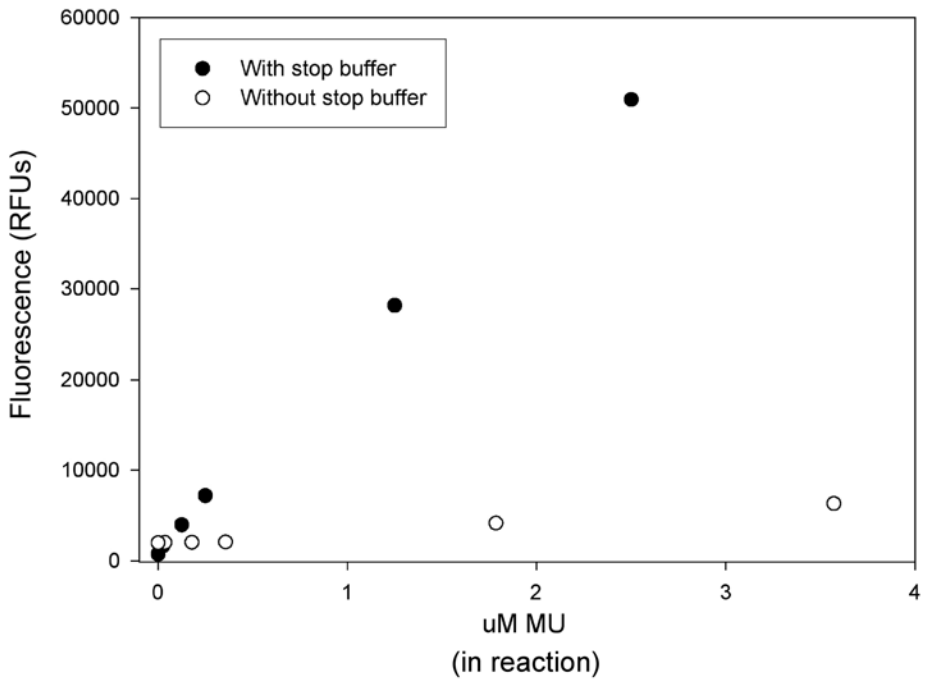

Figure A-6. Effects of MUG assay stop buffer on MU fluorescence. Fluorescence (Excitation = 355nm, Emission = 460nm) of different MU concentrations was measured by two approaches. The first approach (●) was designed to emulate conditions where MUG assays are run in reaction tubes and 10µl of the reaction is combined with 190µl of stop buffer (200mM Na_2CO_3). The second approach (○) was designed to emulate conditions where MUG assays are run in a 96-well plate and fluorescence of the 130µl reaction volume is measured in real-time as the reaction progresses. Both approaches produce fairly linear curves, however the fluorescence of the real-time measurements is greatly reduced. Furthermore, the reaction buffer has a low-level of fluorescence (probably from uncleaved MUG) that results in a flat response at the low end of the real-time curve.

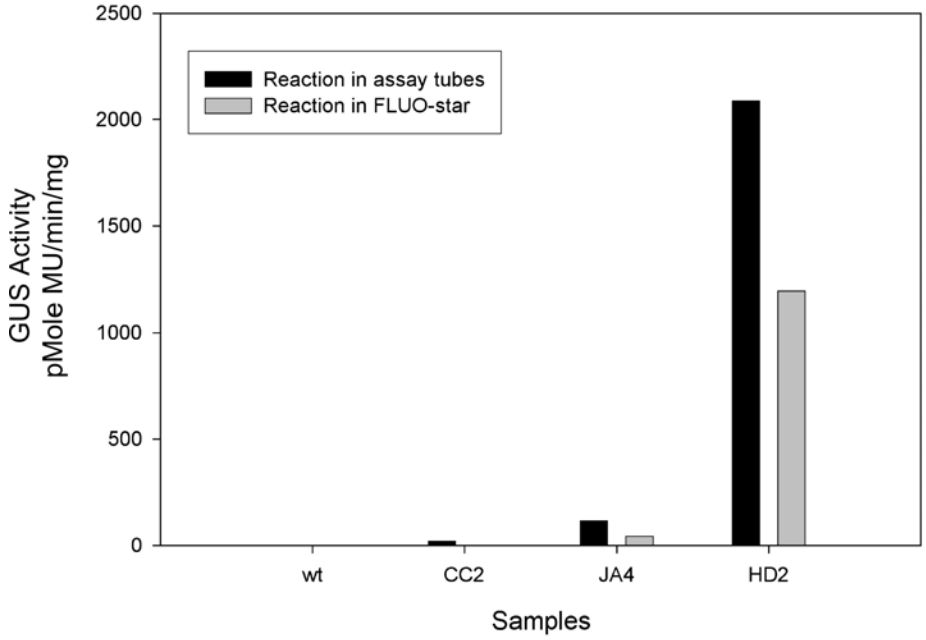

Figure A-7. Comparison of GUS activities using two different approaches.
GUS activities for three samples and one untransformed control were determined using two different MUG assay approaches. In the first approach (black bars) reactions were run in assay tubes, and a 10µl of the reaction was transferred to 190µl of stop buffer in a 96-well plate at 20 minutes. The stopped reactions were then measured for fluorescence. In the second approach (gray bars) reactions were run in a 96-well plate in the fluorescent plate reader. Fluorescence was measured every 60 seconds for 20 minutes as the reaction progressed. The two approaches provide different specific levels of GUS activity for samples JA4 and HD2, but generally yield similar results. However, for sample CC2, the GUS activity value from the real-time approach was similar to the value obtained from the untransformed control. The assay tube approach yielded a higher value for sample CC2, clearly distinguishing it from the untransformed control.

References

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem **72:** 248-254

Gallagher SR (1992) Quantitation of GUS activity by fluorometry. *In* SR Gallagher, ed, GUS Protocols: Using the GUS Gene as a Reporter of Gene Expression. Academic Press, Inc., San Diego, pp 47-59

Jefferson RA (1987) Assaying chimeric genes in plants: the GUS gene fusion system. Plant Mol Biol Rep **5:** 387-405

Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO **6:** 3901-3907

Joersbo M, Okkels FT (1996) A novel principle for selection of transgenic plant cells: Positive selection. Plant Cell Reports **16:** 219-221

Stomp A (1992) Histochemical localization of beta-glucuronidase. *In* SR Gallagher, ed, GUS Protocols: Using the GUS Gene as a Reporter of Gene Expression. Academic Press, Inc., San Diego, pp 103-114

Ülker B, Allen GC, Thompson WF, Spiker S, Weissinger AK (1999) A tobacco matrix attachment region reduces the loss of transgene expression in the progeny of transgenic tobacco plants. Plant J **18:** 253-263

Appendix B

Attributes of PCR-Identified and Kanamycin-Selected Lines

Table B-1. Attributes of Screen 1 PCR-identified and kanamycin-selected lines.

	Initial Screen (A-B)[a]	X-gluc staining[b]	*gusA* (E-F)[a]	*nptII* (C-D)[a]	Direct repeats (G-H)[a]	pBI121 vector (I-J)[a]	T2 kan tolerance[c]	GUS T1[d]	GUS T2[d]	Int. mapped[e]
S1-CA4	+	+		+	+	-	+	193.56	343.13	
S1-CF1	+	+	+	+	+	+		372.38		
S1-CG6	+	-	+	+	+	+	-	1.79	49.65	+
S1-FC1	+	+								
S1-FD7	+	-	+	+	+	+	-	9.48	337.07	
S1-FF3	+	-	+	+	+	+	-	0.66	178.40	
S1-GB6	+	~	+	+	+	-	+	0.45	20.90	
S1-GD5	+	-	+	+	+	+	-	213.03	10.23	
S1-GG5	+	+	+	+	+	+	+	114.74	75.93	
S1-HD2	+	+	+	+	-	-	+	200.56	259.72	
S1-IA1	+	+	+	+	+	+	+	122.37	103.02	
S1-JA4	+	+	+	+	+	+	+	252.56	57.53	
S1-JB6	+	+	+	+	+	-	-	165.66	38.24	
S1-JC7	+	+	+	+	+	+	+	188.15	358.35	
S1-JE1	+	~	+	+	+	+	+	1.45		
S1-JG2	+	~	+	+	+	+	-	48.48	102.91	
S1-KB7	+		+	+	-	-		48.46		
S1-KR1			-	+	-	-	+	3.85	9.59	
S1-KR2			-	+	-	-	+		5.41	
S1-KR3			+	+	+	+	+	2.65	107.15	
S1-KR4			-	+	+	-	+	28.22	20.58	
S1-KR5			-	+	-	-	+	5.32	32.25	
S1-KR6			+	+	+	+				
S1-KR7			-	+	+	-	+	608.55	329.53	
S1-KR8			+	+	+	+	+	3.65	31.42	
S1-KR9			+	+	+	+	+	2.76	14.14	
S1-KR10			+	+	+	+		753.59	107.15	
S1-KR11			-	+	-	-		778.58		
S1-KR12			-	+	-	-		813.28		
S1-KR13			+	+	+	+		705.40		
S1-KR14			-	+	-	-		68.18		
S1-KR15			+	+	+	+				
S1-KR16			+	+	+	+		392.60		
S1-KR17			+	+	+	+		3.14		
S1-KR18			+	+	-	-				
S1-KR19			+	+	+	+		4.55		
S1-KR20			-	+	+	-		3.43		
S1-KR21			+	+	+	+		2.25		
WT		-	-	-	-	-	-	-1.14	0.15	

*Footnotes for Table B-1 are located after Table B-5.

Table B-2. Attributes of Screen 2 PCR-identified and kanamycin-selected lines.

	Initial Screen (A-B)[a]	X-gluc staining[b]	*gusA* (E-F)[a]	*nptII* (C-D)[a]	Direct repeats (G-H)[a]	pBI121 vector (I-J)[a]	T2 kan tolerance[c]	GUS T1[d]	GUS T2[d]	Int. mapped[e]
S2-AE6	+		+	+	+	+	+	244.69	139.36	
S2-CB6	+		-	+	-	-	+	275.71	319.46	
S2-CD1	+		-	+	+	-				
S2-CD6	+		+	+	-	+		3.34		
S2-CF4	+		+	+	+	+	+	103.42	130.16	
S2-EB6	+		+	+	+	+	+	0.76	120.35	
S2-ED1	+		-	+	-	-				
S2-GA2	+		+	+	-	-		-0.62		
S2-GF2	+		+	+	+	+	+	334.17	18.09	
S2-JE3	+		-	+	+	-				
S2-KC1	+		+	+	+	+		10.22		
S2-KC6	+		+	+	+	+		4.98		
S2-LC4	+		-	+	+	+	+	15.58	427.83	
S2-MB3	+		-	+	-	-	-	0.77	25.37	
S2-NB2	+		-	+	-	-	+	0.92		
S2-QE5	+		+	+	+	+		256.82		
S2-SF4	+		+	+	+	+	+	3.16	93.74	
S2-TA3	+		+	+	+	+	+	-0.47	15.57	
S2-TC2	+		+	+	+	+	+	1.18	30.18	
S2-UB6	+		+	+	+	+	-	1.11	16.47	
S2-UE1	+		+	+	+	+	+	353.22	51.29	
S2-WF6	+		+	+	+	+	+	153.22	53.28	
S2-XB5	+		-	+	-	-	+	169.55	154.48	
S2-YB5	+		+	+	+	-	+	0.06	213.96	
S2-YE3	+		+	+	+	+				
S2-ZC6	+		-	-	-	-	-	2.58	152.23	
S2-ZE5	+		+	+	-	-	+	272.18	195.27	
S2-KR1			+	+	+	+			157.20	
S2-KR2			+	+	+	-	+	3.58	178.99	
S2-KR3			-	+	-	+	-	0.89	161.47	
S2-KR4			+	+	+	+	+	194.21	30.91	
S2-KR5			+	+	+	-			23.54	
S2-KR6			+	+	-	-	+	0.56	19.35	
S2-KR7			+	+	-	+				
S2-KR8			+	+	+	+			39.59	
S2-KR10			+	+	+	+				
S2-KR11			+	+	-	-	+	0.57		
S2-KR12			+	+	+	+				
S2-KR13			+	+	-	-	+	-0.15		
S2-KR14			+	+	+	+	-	476.85		
S2-KR15			-	+	+	+				
S2-KR16			+	+	+	+	+	148.48		
S2-KR17			+	+	+	+	+	258.44		
S2-KR18			+	+	+	+	+	400.15		
S2-KR19			+	+	+	+	+	415.61		
WT		-	-	-	-	-	-	-0.88	-0.13	

*Footnotes for Table B-2 are located after Table B-5.

Table B-3. Attributes of Screen 3 PCR-identified and kanamycin-selected lines.

	Initial Screen (A-B)[a]	X-gluc staining[b]	*gusA* (E-F)[a]	*nptII* (C-D)[a]	Direct repeats (G-H)[a]	pBI121 vector (I-J)[a]	T2 kan tolerance[c]	GUS T1[d]	GUS T2[d]	Int. mapped[e]
S3-AD1	+	+	-	+	+	-	+	0.83	7.04	+
S3-AE2	+	+	+	+	-	+	+	278.40	491.12	
S3-BE6	+	+	+	+	+	+	-	407.81	1.77	
S3-CC5	+	+	+	+	-	+	+	4.87	19.27	+
S3-CD1	+	+	+	+	-	+	+	220.48	203.45	
S3-EB4	+	+	+	+	-	+	-	557.48	95.04	+
S3-EF5	+	+	+	+	+	+	+	250.75	33.77	
S3-GC3	+	+	+	+	-	+	-	0.17	92.58	+
S3-GF6	+	+	+	+	+	+	+	288.66	5.35	
S3-HF2	+	+	+	+	+	-	+	0.08	14.56	
S3-JF5	+	+	+	+	+	-	+	429.56	312.06	
S3-KA4	+		+	+	+	+		332.37		
S3-KE5	+	+	+	+	+	+	-	300.94	7.00	
S3-LB4	+	+	+	+	+	+	+	124.13		
S3-OA4	+	+	+	+	+	+	+	180.56	6.22	
S3-KR1		+	+	+	+	+	+	4.57	157.20	
S3-KR2		+	+	+	-	-	+	226.99	178.99	
S3-KR3		+	+	+	+	+	+	4.86	161.47	+
S3-KR4		+	+	+	+	+	-	0.71	30.91	+
S3-KR5		+	+	+	+	+	+	-0.51	23.54	
S3-KR6		+	+	+	-	-	+	-0.55	19.35	
S3-KR7		+	+	+	+	+		2.28		
S3-KR8		+	+	+	+	+	+	112.77	39.59	
S3-KR9		+								
WT		-	-	-	-	-	-	-0.77	-0.11	

*Footnotes for Table B-3 are located after Table B-5.

Table B-4. Attributes of Screen 4 PCR-identified and kanamycin-selected lines.

	Initial Screen (A-B)[a]	X-gluc staining[b]	*gusA* (E-F)[a]	*nptII* (C-D)[a]	Direct repeats (G-H)[a]	pBI121 vector (I-J)[a]	T2 kan tolerance[c]	GUS T1[d]	GUS T2[d]	Int. mapped[e]
S4-AE3	+	+	+	+	+	+	+	247.08	165.06	
S4-CE2	+	~	+	+	-	-	+	-0.27		+
S4-EB6	+	+	+	+	+	+	+	228.67		
S4-GB6	+	~	-	+	-	+	+	-0.56	71.85	+
S4-HC1	+	-	+	+	+	+	+	231.27	134.97	
S4-HF6	+	-	+	+	+	-	+	0.08	11.21	
S4-PF6	+	+	+	+	+	+	+	51.55	76.91	
S4-QB5	+	+	+	+	+	+	-	354.08	62.38	+
S4-QE2	+	+	-	+	+	+	-	3.33	90.60	
S4-SA3	+	+	+	+	+	+	-	1.70	55.93	+
S4-WE2	+	+	+	+	+	+	-	23.36	2.50	+
S4-YC1	+	+	+	+	-	-	+	0.32	56.44	+
S4-KR1		+	+	+	+	+	+	3.87	54.62	
S4-KR2		+	+	+	+	+	+	0.20	278.53	
S4-KR3		+	+	+	-	+	+	0.31	42.55	
S4-KR4		+	+	+	+	+	+	148.00	28.12	
S4-KR5		+	+	+	+	+	-	358.18	51.17	
S4-KR6		+	+	+	-	+	+	270.14	413.82	
S4-KR7		+	+	+	+	+	+	353.86	64.38	
S4-KR8		+	-	+	-	-	+	206.21	177.31	
S4-KR9		+	+	+	+	+				
S4-KR10			+	+			+	-1.07		
S4-KR11		+	+	+	+	+	+	-0.32	15.02	
S4-KR12		+	+	+	+	+		1.69		
S4-KR13		+	+	+	+	+	+	386.46	307.08	
S4-KR14		+	+	+	+	+	+	217.22	104.42	
S4-KR15		+	+	+	-	+	+	62.20	37.82	
S4-KR16		+	+	+	-	+	+	-0.31	70.55	
S4-KR17		+	+	+	-	-		19.55		
S4-KR18			+	+			+	0.95	7.89	
S4-KR19			+	+			+	448.21		
WT		-	-	-	-	-	-	-0.63	0.07	

*Footnotes for Table B-4 are located after Table B-5.

Table B-5. Attributes of Screen 7 PCR-identified lines.

	Initial Screen (A-B)[a]	X-gluc staining[b]	*gusA* (E-F)[a]	*nptII* (C-D)[a]	Direct repeats (G-H)[a]	pBI121 vector (I-J)[a]	T2 kan tolerance[c]	GUS T1[d]	GUS T2[d]	Int. mapped[e]
S7-AC5	+	+	+	-		-	+	6.70	130.20	+
S7-AD5	+	+	+	+		-	+	6.76	85.57	
S7-BB1	+	+	+	+		+	-	182.56	-0.25	+
S7-BB4	+	-	+	-		+	+	5.28	16.61	
S7-BD5	+	+	+	+		-	-	3.38	71.82	+
S7-BE1	+	+	+	+		+	+	4.10	132.42	
S7-FB5	+	+	+	+		-	+	269.02	332.20	+
S7-FC1	+	+	+	+		-	+	150.52	117.51	
S7-FD2	+	-	+	-		-	+	2.00	9.78	
S7-GA3	+	+	+	+		-	+	5.31	73.58	+
S7-GE3	+	+	+	+		-				
S7-HB5	+	+	+	+		+	-	4.82	42.18	+
S7-HE6	+	-	+	+		+	-	1.94	10.18	
S7-IC1	+	-	-	+		+	+	2.97	193.69	+
S7-JB3	+	+	+	+		+	+	5.84	79.10	
S7-JD6	+	+	+	+		+	-	11.84	50.05	+
S7-KB2	+	+	+	+		-	-	6.38	70.52	+
S7-LA6	+	+	+	+		+	-	346.92	382.20	+
S7-LD1	+	+	+	+		+	-	6.91	48.56	+
S7-NA5	+	+	+	+		+	+	4.63	274.10	+
S7-ND4	+	+	+	+		+	-	4.92	120.80	
S7-OD2	+	-	+	+		+	+	2.91	27.58	
S7-OE1	+	+	+	+		+	-	305.81	37.50	+
S7-OE4	+	+	+	+		+	+	2.27	14.74	
S7-PA4	+	+	-	-		-	+	610.90	586.59	
S7-PB3	+	-	+	+		+	+	4.41		
S7-QA2	+	+	+	+		+	-	3.76	0.63	
S7-QC2	+	+	+	+		+	-	293.44		
S7-RE1	+	-	+	+		-	-	7.82	26.89	
S7-TB6	+	-	+	+		-	+	3.06	11.47	
S7-TE3	+	+	+	+		+	+	3.67	14.21	+
S7-UD2	+	+	+	+		+				
S7-UF3	+	+	+	+		+	-	3.33	72.20	
WT	-	-	-	-	-	-	-	0.49	-0.27	

[a] Presence of genes or structures determined by PCR amplification with primers indicated in parentheses.
[b] GUS activity based on histochemical staining of leaf tips with X-gluc. Staining was qualitatively assessed as strong (+), weak (~), or no staining (-).
[c] T2 kanamycin tolerance frequencies were used to determine if lines were nptII-expressing (+) or nptII-silencing (-).
[d] GUS activities in T1 and T2 generations based on MUG assays (pMole MU/min/mg).
[e] Lines with mapped T-DNA integration sites are indicated (+).

Summary of left border flanking genomic sequences of PCR-identified lines.

Key:

S3-AD1[a]

GTGTTATTAAGTTGTCTAAGCGTCAATTT[b]***GTTTACACCACAATATATCCTG***[c]NC
NTNTTNCCNNTCCNGGCNGGGNNAGGNGCNCACTTTTAACATATTGTTTCCGGTT
TANNTGTCTGNACCAATCAATTGATGGTNGAAGTTTAATTTCAAACCAAATTTAA
TTATTACAGTCATGATT[d]

SIGnAL: Chr1 22901523[e] (1e-18)[f]

Alonso et al: ~22443000[g]

[a] Name of line.

[b] Sequences in bold (**CAATGTGTTA**) correspond to pBI121 T-DNA sequences.

[c] Sequences in italic (***GTTTACACCA***) correspond to left border sequences.

[d] T-DNA flanking sequences (i.e. plant genomic, rearranged binary vector, or unknown sequences)

[e] SIGnAL: Genomic position as determined by BLAST against SIGnAL database (http://signsl.salk.edu).

[f] BLAST e-value.

[g] Alonso et al.: Genomic position (to nearest 1 kb) at the time of analysis for data included in supplemental material of Alonso et al. (2003).

S1-CG6

AACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGNCAATTT*GT*CTACNAATTCCCAGGTA
TGTGTACATATTTCACTTTTCAAAACTAAAAATAGTAGTAATAACATATTCCATGATGTTGTTT
TGGGTCAAACGTAGTAACATTTTAAATTGACGTAAGACAATCAAAACTTAAACCTACAACATG
TAAATTTTCAGCCTAAATAACAGAGGCTATATGTAAATAGAGATTATTCTTTTTGATAAAATT
GCAAACTAATAAATAAATAAAAATGATTCAAATACTAAACAAGAAAAATGTATTCAAAATTA
GAAGCTGGATAGCATCAGCTTGAAGAGCATAGAGACAAACATGCCTCAAAACTTGACAGCAT
GCTTTATTTGTATTTATTACTTCCCTAATTACAATGTCACCAAAACGTAGTCGTCGCCAGACTA
TCTGCATGCACATACACATCTGTATTTATTAGTTTTGAAATAGTAATATTTAATGATTTTTTCTA
TTTTAAAAATATAAAATTATGGATTATTTCAAATAACCAAAAAGTTTGGTTTAAGAAAAAGTA
TGGTGAAATATTGGGAGAACTATATAGTTAAAGAAATGAATATTGTGAAATATACTTAAACA
AAAAAACTACTATATATTGGGAACATCTATGGGGTTTAGACAGTGGCGTGAAGTTCCAAGTTA
AAAATACCCCATGCCATAAAGAAACAGGAGTAATGTATTCCCGTTGAACANCTTTTCCNCTAA
TACTTCCAATTAGCTTAGA
SIGnAL: Chr4 372763 (1e-103)
Alonso et al. (2003): 372000

S3-AD1

AACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAATATATCC TG*NCNTNTTNCCNNTCCNGGCNGGGNNAGGNGCNCACTTTTAACATATTGTTTCCGGTTTANN
TGTCTGNACCAATCAATTGATGGTNGAAGTTTAATTTCAAACCAAATTTAATTATTACAGTCAT
GATTAAGAAGATAAAGACTTACTAGGTAAGACTCTAGACCAGGAAATAGATAATCTGTAACT
ATTGACTCCAAGAAAATTCATTGATTGGATATCTTCCTGTAATCATATATAGTAGATAGATGTC
AATATAAAAAAGAATTAAAAACAACTAAAGTAAATAGTTTATTCAGGTAAAAGGAACAAAAT
ATTATAATAATGTATTTATTTTACCATATATCGATGATATTGGTCCGTAGCTATGTCTCCATTG
CTTCCATCAACTATTTTCCCTGAATTTTTTAATATAATATTGTCAATGGAATCAACATGGCATT
TATAAGATAAAAAAACTTTGTCGAGAACCAATCAAAGAAAGCTAAACTGTAATTTATTTTACC
TGCCCGGGCGGCCGCTCGGCCCCCTATAANNNNNNNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr1 22901523 (1e-30)
Alonso et al. (2003): 22443000

S3-CC5

TCCCGCAATGTGNTATTAAGNTGTCTAAGCGTCAATGTTGCTTACACCACAATTTGACNCG
TNNGCAACTCTAGAGAAGTAAACCGGNGGCGGTTACCCATAATTGCCGGNTCGGCGAGNGGN
GTCTCGGTCTCCATTTGCCGGTTCCCATTGACTAACGCGTTAGAGAGTGACTGATGGGTCGAC
TTTTGACTTTTACTATTTTTCTTCTTTTGTTTCTTACTGCCAGTTATGCCCCATTCCCTGCCTCTT
AATTACGTATGAGTGCCACTCACAAATATTTTTATTTCTGAAAATCATTTGGGGGAGACAAAA
AAAAACAAGATATAACAAACCTGTTGTTTTACTTTTGTTACTGACAAGTCCTTCAGTCAAAAA
ATAAAAACAAGATAGATAAAGAAAATGAAAACTGTATTTAATAATATCCTGAATAGTCAAAA
TAGAATTGTCTATAAAATCTGATTTTTTAAAAAATATAAACATATATGATGAAGTATTCTTGTT
TCATGTATCCATTCGAAAATATAAGTAAATCCTATGTCCAATTGAAAAAAAAAAAAAAGGGG
NTTTTAAAANNNNGGGNNNAAAACCNNGNGTTTTCNANGGNNTNACCCCNAGNCGGGGCNG
GGGGGTTTTNNTTTTANANAAAATTANCCCCAGNNGGGGTTTTTTAAAAGGTNNNTNAAACN
NGGGGGGGGNGGGAAAAACCNCCCACNTTTTTTTGNTNAATNGGGNCAACCNNGGGGNNGG
GGGNTTTTTNNAATTTTTATCCTTTNNGAGGGGGNANNGGGANNTNNCNNNAA
SIGnAL: Chr2 477673 (1e-54)
Alonso et al. (2003): 476735

S3-EB4

CGTCCGCAATGTGTTATTAAGTTGTCTACNCGTGGAGGCTTATTATTCGTTGATATCAACTA
TTGTCCATGTCTGTAAGCCTCTCCCTACGACAGAGGTTTTCAGAGCAGTAACAAACAACTTAC
ACATGAGAGTTAACATTGCACAGTCTGAGATTATACTCGCTAGATTCTTTTCAAGTCTTGAAG
AACAACATGTGCAGCGTTTCTTCCAGGTGCACCCATAACTCCGCCTCCAGGATGAGCTCCGCT
GCCACACAAGTAGAGACCTTTCAAAGGGCTTCTGTAATTTGACCTGTTTCACACAGTCCAACA
AAGTTTTGAATTTCATAATCATAAATCGTAAAGACTGTTTTGAGTACCTGCCCGGGCGGCCGC
TCGGCCCCTATAANGCNNNCATNNNNNGANNTANNNNNANNTNNNTNNNNNGNANNNNGNN
NGNNNCTNNNNCNNNCNNNGNGTNCNCCNNCNNNN
SIGnAL: Chr5 20124529 (1e-22)
Alonso et al. (2003): 19715000

S3-GC3

TAAAAACGTACCGCCATGTGTTATTAATTTGTCTAAGTCGTCNATTTGTTTACACCACAATA
TATCCTGAACTGGCGATGCGGGATGAACCGGANGCCGGNTTACGGTGCCCAACTGCGCGCTA
ACCTANAACCCACAAAGGGTGTTGGTCGATTAAGACAGCAGGACGGTGGTCATGGAAGTCGA
AATCCGCTAAGGAGTGTGTAACAACTCACCTGCCGAATCAACTAGCCCCGAAAATGGATGGC
GCTTAAGCGCGCGACCTATACCCGGCCGTCGGGGCAAGAGCCAGGCCTCGATGAGTAGGAGG
GCGCGGCGGTCGCTGCAAAACCTAGGGCGCGAGCCCGGGCGGAGCGGCCGTCGGTGCTGATC
TTGGTGGTAGTAGCAAATATTCAAATGAGAACTTTGAANGCCGAANAGGGGAAAGGTTCCAT
GTGAACGGCACTTGCACATGGGTTAGTCGATCCTAAGAGTCGGGGGAAACCCGTCTGATAGC
GCTTAANCGCGAACTTCGAAAGGGGATCCGGTTAAAATTCCGGAACCGGNACGTGGCGGTTG
ACGGCAACGTTAGGGAGTCCGGAGACGTCGGCGGGGGCCTCGGGAAGAGTTATCTTTTCTGTT
TAACAGCCTGCCCACCCTGGAAACGGCTCAGCCGGAGGTAGGGTCCAGCGGGCTGGAAGAGC
ACCGCACGTCGCGTGGTGTCCGGTGCGCCCCGGCGGCCCTTGAAAATCCGGAGGACCG

NCBI: homology to 25s ribosomal RNA (X52320) or 5.8s ribosomal RNA (X52322) which are located on chr 2 or chr 4

S3-KR3

CNNTGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAATATATCCTG*ACAAATG
TGAGTTTGTGTGGTCGCTGAGTAGAGGGACCCACCTACCTAGATCCTAAGGTTTGGCCTTTAA
GGCGGGAGATTTTTATTTCATAGGATATATATATTTGCTTTTAAATACTAATGCGATATGAAGT
TAAAAATAAAATGCGGGTCTAAGGATTCGATCCGTGGGCAATATCGTGGGAGGGTACTATGA
TTATGATACCTCTCTCACTCAACACTTAACTTTAATACAAACNCGANNNNNNCNCNNCNNNCN
NCNNCNNNNNNNNNNNNNNNNNNNCNCNNNNNNNNNNNNNNNNNNNNNNCNNNNNN
NTNNNNNNTNNCCNTNNNNNNNNNNNNNCNTNNNNANNNNCNNNNNNNNNNNNNCNNNNN
CNNCNNNCANCNNNNNTNNNNNNNCNNNANCNNNCTCNNCN

SIGnAL: Chr1 29306293 (1e-16)

Alonso et al. (2003): 28841000

S3-KR6

GCAATGTGTTATTAAGTTGTCTAAGCGTCAAGAGTTGTCCANAAATCAGTAGTCAANANCA
TAGATTNCATAGCAAACATTTTGGTTTTCTGGACTCTTCAGGTAAATAATCTAAGGAACCGGT
ACNATNNGGCNGCTNCNACAAGATCCTTAGANACTCGTATCCAAGGGCATATCTGAACCGAT
CCCATGNTGNTNGGACTNGGNGGGTGGNGANCTCNNCTNTNCNGNCCTGCCTGNGNGGGCCC
CGAGACAGANCCCNAACNTNAANCNNNTCCNCGNCCTGCCNNNNANGCTANNNNCNNNCNC
NNNNCNGNNCTNNTTATNCNNNCAGNNCTNCCTNNAGNACGANNNCGNTGGAANNCNNNGA
ATGNNTNTNTGNCGNNCTTCGCTNTTGNCTGNGGCNCNGCNTNCCNNNNNCTCCCGNCCTGTG
CGNNNGNCNAGTNGNNNACTCGTAACNCNTCNCGCCTNANNCTTNNNCGGNNNCTNNTNNTC
CNNNNTNCCTGNTCCCACNNNCGNGNCGACNNTNANNCCCGGNGCCTNNCANANNNCNTCGC
NCNGNN

SIGnAL: Chr5 16618101 (1e-17)

Alonso et al. (2003): 16208000

S4-CE2

ACGTGCCGCACTGTGTTATTAAGTTGTCTAAGGGCAATTT*GTTT*TTGTTTGTTCAGACTTCT
TCAGACCCATCTATCTATAATAGGGCTCTGAAACAAACCCGGCCCAACGTTTTGAAGCCCTAT
ATTGTTTTGGGTCTCTAAATGATTCTTTTTACTTTTATGNNCATGCACATCATCCATACGGCCA
TACCACATGAGATTATCACACATGGGTCATAAGTCATAACACACCAAACAGACACAGTATAT
GATACACAACTTTGGATAAAAATAAACATGGACATAACTAGGCAAATACCNANNAGCAATCN
CACATATTGNAAGACATATGNCNGCACATCATGCTNNCNAANCACGNANAGGAANGNNANN
NTTGNATNGNCNANNNTNNANGNNNNCGGNNNCGNCNNNNTNNAACNNNCANCNNNTNCNT
NNGCNAGGANCATCCNNGNNACCGNNTNNCANCNCCCNCCNNACGCNTCGNNCCGCNNCCC
GCCGNCNCCNGCCNNNNCNGNANNNNNGCCNNCAAANTNCNNAANNANCCATNGTTNGCNN
NNNNNGCNNNCACNNGTNCNTNATNNNCNATCNNNCGTCCNCCNNCNNCNCGCCNTCNNANT
NCNNCNCGCNANNNNNNNCAGNNCAGCANNCN
SIGnAL: Chr3 3199482 (2e-47)
Alonso et al: 3198000

S4-GB6

CGCACTGTGTTATTTGTTTGATTTTCAACTCTGTCTAACTTCTCCATTTTCAAAGGACTTTACA
GAGCTTTAGTTCAGATTTTTAGTATCTTTGGGGTTCATGGTTTCAAAGCAGTTTTTCTCTTTTCT
TTTTCTAATTTTCAAGATAGTTAGATCGTTGTGAAGTTTGTAAATAACATAATTGGTTTAGAAG
CTGTNNACTTTTCACATGGGTCTGCTCTGAAATCAAAGGANCAACCGNNNACNANCCNCCNN
AGGCCCCNAAAACCGCNTCCNANCNNNATAGCACGNNCNTNNGCCCNTCCCCGNNGCCCNGG
CCNCGNNCNANGGCACACTCCGCGNATGTTTCNNCNNCCNCNCTCNNNGACCCNNCCCNCCN
CCCCNNCCNNCGCCNCCNACCCNCCNCNACANNGGGGNCGCNCCNCCNCNCCCCCNTCACNN
NNNCNACCNCACCCNCCNCCTANCCNTANTCCGAGCCNCNTGCAAAACACNGCCNCGGCTCC
TNTCCNNCGNNNCNNNNNGCACNCNCTCNNGNCCNCNGNNGGCNCNCCCCNNNTGAGGACC
NCGNNGNNNNNCNNNGTNTGAAAANTCCCGTCNANCCNTCTTCNANCCCCNNCANGNNACNG
NNNNNAAGNGAGGATNNCCCCCANCGNCNNACNGNCNCTNNNACCNNCCCNATGCCATN
SIGnAL: Chr4 12340987 (2e-24)
Alonso et al. (2003): 11305000

S4-QB5

**ACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAATATATCCT
G***TTTTATTTTTACGGTCGACCCAAAAAAAAATGTTTTATTTACGGACGACACAAAGGATTTGTT
GAAGAGTCTCTTCAACAAAAATTCCTTTAAATTCCGACGATTCTCTTTCTTGATATCTGTCTCT
CCAAATCAAGAAATCTCTCCACTCGATCAAAAAGCTCTCTTGTCTAACAAAGCTTGCACTAAA
TTTCCAAAAAAAAANCAAAGCTNGCCCTTNGNTTNGATCAAANGTNGAGGCAAAAAAAANGN
CAAAAGGANCANCGNTGAANNGANGAATTCNCCNGCCNGGGNGGCCGNTNGGCCCCTNTAA
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr1 20258297 (3e-84)
Alonso et al. (2003): 19792000

S4-SA3

ACGTNCCGCACTGTGTTATTAAGTTGTCTAAGCGTCAGTATCTTAAAACGGAAAACACAGT
ATTAATGGAGGTAGATGCGACTATAAACTCATAGTAATTGCTAAATTAAAAACTTACCTATAA
ACTAAAGTTGCAATTNACCTGCCCGGGCGGCCGCTCGAGCCCTATAANNNNNNNNNNN
SIGnAL: Chr1 3832136 (3e-15)
Alonso et al. (2003): 3832000

S4-WE2

CGTNCCGCAATGTGTTATTAAGTTGTCTAAGNCGTCAATTTGTTTACACCACAATATATGC
CTGNAAATCGGTGANTNANNANCATCAATGTGAAGAAGTANGCNCCCTAGGTATAGCACCAT
ATNNGTATGNAGCCGTCNTCTAAGTGATCGATGCATTTAGGTTGGGTAGCACCCTGTTGNAGA
GNCTCCANCCCANAGCNTACTAGAGACTGCCCNCAANGGANNTNANGGANNGTGTGANNTCA
NGACCCCTCTTNAGCAACNGCNCAAATATNCCAGTCANCTTCGANTCCGAGATGATNCTCNNN
CGNANGCTNGGAAAGAAAAAGCNAGNNTCTTCCTCTTAAAGTTATTCTNCTCTCATTATTGTG
TTCCACCACAGTGGTGCANCTGATCTCTTTGGCATGAGGGATGTCTNAANGGNNTGTCTGCTT
GCCTCTGGNTTNGNCCTTNGGCAGATTAAGAGGCTNNATCNNTNGCNTNCTGNNGNGGAGTC
ANNTATGGNTGNTCCATCCACNTNGGAAGCTCGTAGGCNTNTTACTTANNNTNCAAANTCACT
ATNCNTGGACATGCCTCCTGCANCCGAGATCCANGACATGCTTTNCTTCTCNNAGGTTCCAGT
CNCGATGGNCATCTTCTCGAAGGCNTACTATATCTTAACTCTTCCATTTTGTATTTTNATANGG
TCCTTACGCCATACTTTNNANACNNTGTAANTGTGCTCCATATGGGGCCTAGNACCNNCNNNT
GAAACCAGAGTTTCTATNCNCATNTNCCTNGCCNANGAACCTTGCCAGGGNCCCCTTTTCCNN
CTGCCNTTAATTTNTNCNTTNGGANCCGCTAACTTGGNTTGGNCTATNAACNNAANNCATN
SIGnAL: Chr1 1652024 (1e-31)
Alonso et al. (2003): 1652000

S4-YC1

GTGGCGANNCGTGTTATTAAGTTGTCTAAGCGCCAATTT*GTTTACACCACAATATATCC*TGA
AATTGTGGGGTGGGCGATTGCGCTACACGGNGGAGCCGGTAGGACATTCCGATCGATCTCCCC
GACGAGCGACGTATGCCTCGTGAGAGCGCCCTCCGAANANTNNCTCGATCTTGGNATCTCCCT
CCTCAAATCCGGNANGNCTGCCTTGGACGTCGCCGAACTTGTCGTACGCNNTCNCTTTCTCGC
TGATTGCTCGCTTTTTNTTTGGGANATGCGCNCNCNNGCCGNNNTNCATNCNNNTNNGCNCAC
CANNGCGGNNNNGTCTAGGACTACTCTCGAATANNTNNGCAGCACTTCANAATGANTNTNAC
NACACATACACTNNNATAGTNCTNACGCTCCNTCCCNGCTNTCGNNNNTGNNCNGGTTATCAN
GNNTNANCTCNNNNANNGCGCNNNACNATACCTATNNNNCNGCAAGCGNNCAGNCACGATN
CACGNCNGNNANNNCNNCANNNGNNNNNCANACNANNNCNTCNACTAGNAANGTANCNNCC
NAGTCNNNTNGAACNGGGTCTCNNNTNGNNTCGTTNATGCTNCTTATACNTNNNTNTCCCNNN
AGCGTACCNANGGCCCTAAGCNANANNTNTGGNANNNGAGTNNNGCTNTGACGN
SIGnAL: Chr5 2594400 (1e-19)
Alonso et al. (2003): 2593000

S7-AC5

**ANTAAAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGNCGTCAATTT*GTTTACACCACA
ATAT***NTGNTGATACAATATNGNATCCACAATATTCGNTTAAATGAAATACCCTTTTGTTGCCTT
TTATTTTCTTTNGNATTCCTTTGTAAAAATTAATNAGAATTCACCTGCCCGGGCGGCCGCTCGA
GCCCTATAGTTGGTTGTGCTTTAGTGAACATGTACTCGAGATGCATGTGCATTGCTTCTGCTTG
TAGTGTGTTTAATAGCATCTCTGAGCCTGATTTAGTTGCGTGTTCTTCTTTGATAACTGGTTATT
CGAGATGCGGGAATCACAAGGAGGCTTTGCATCTGTTTGCTGAGCTGAGAATGAGTGGTAAG
AAGCCAGATTGTGTTCTTGTGGCGATTGTATTGGGGTCTTGTGCTGAGTTATCGGATTCAGTCT
CCGGAAAAGAAGTGCACAGTTATGTTATTCGACTAGGACTAGAACTTGATATAAAGGTTTGCT
CTGCTCTTATAGACATGTACTCAAAATGTGGGCTTCTAAAATGCGCGATGAGTCTTTTCGCAG
GAATTCACCTGCCCGGGCGGCCGCTCGAGCCCCTATAANNNNNNNNNNNNNNNN
SIGnAL: Chr5 15073391 (1e-38)
Alonso et al. (2003): ~14664000

S7-AC6

TNCCGCAATGTGTTATTAAGTTGTCTAAGNCGTCAATTT*GTTTACACCACAA*NNGNTGATA
CAATATNGNATCCACAATATTCGNTTAAATGAAATACCCTTTTGTTGCCTTTTATTTTCTTTNG
NATTCCTTTGTAAAATTTAATNAGAATTCACCTGCCCGGGCGGCCGCTCGAGCCCTATAGTTG
GTTGTGCTTTAGTGAACATGTACTCGAGATGCATGNGCATTGCTTCTGCTTGTAGTGTGTTTAA
TAGGATCTCTGAGCCTGATTTAGNTGCGTGTTCTTCTTTGATAACTGGTTATTCGAGATGCGGG
AAACACAAGGAGGCTTTGCATCTGTTTGCTGAGCTGAGAATGAGTGGTAAGAAGCCAGATTG
TGTTCTTGTGGCGATTGTATTGGGGTCTTGTGCTGAGTTATCGGATTCAGTCTCCGGAAAAGAA
GTGCACAGTTATGTTATTCGACTAGGACTAGAACTTGATATAAAGGTTTGCTCTGCTCTTATAN
NNNNNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr1 23932310 (1e-110)
Alonso et al. (2003): 23466000

S7-BB1

ACGTNCCGCAATGTGTTATTAAGTTGACCTNGNCNGGGCGNTGCGTTACNCGCNCNTNTNA
TCCTGTACANCTGNNANNCANCTGNTNCCCGTNNNNNTNTGTAANCACAGTAGNGNCCNCTC
GANACAAGCANNGNNTCTANNNGNNGCNGTGGCNNNCGCTGAGCCGTTACCCGGCGNNANA
CTNNGNTNATGTNACNTGNGNTTTTNTGGNGANCTGTGTCNACTNTGANTNATTGGNCTAAAA
AAANATTTNNGNNGNGGATNAAAGTATTTTGCAANCTTATNGNNGGNAANNGGNCNGANATN
NNGTTGGCCATGAGATCACCNGTCCGGGNGNTCANNCTTNNNATNTACCTGCCCGGGCGGCC
GCTCGAGCCCTATAATAATAATGATGATTGTTGCGTTGGAATATTNGGATAAGGNAAAAGCA
ACATANNCNNGNNNTACCNGCTNCCNNACGAANATNTNGNNGCNNNNNNTNNCNNNTNANC
NGNCNANGNNGNNNNNAACANGNNGGNNNNNNNCNNCNNNNNN
SIGnAL: Chr4 2661522 (3e-10)
Alonso et al. (2003): 2661000

S7-BD5

TAAAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCNGNACTGCANCCCANANCTCTGA
ATCAACCCTTGGGCATACATAGNCTGAGAGCATATGAACTCCTTCATCAGTCTGTTTAATCTGT
AGACCCAGGAAGTGGTTCATCTCTCCAACCATGCTCATCTCAAACTCTTTGGTCATGGATTCCA
CAAAGTCAGAGACAAGTTTCTGNTTTGTGCTTCAAAACACAATATCATCCACATAAATTTGCA
CAATGAGGATGTCCTTTTCATCAACAAGGATGAACAAAGTCTTGTCCACACTTCCACGTTTGA
AACTTTTTTCTATTAGGAACAGAGTGAGATGCTCATACCAAGCCATGGGTGCTTGCTTCAACC
CATAAAGAGCCTTTTTAAGCTTGTACACATGATCTGGAAGGTTTGAATCTTCAAAGCCTTTAA
GTTGAGACACAAACACTTCTTCTTGAATAACTCCATTCAGAAAAGCACTCTTGACATCCATCT
GAAATAGTTTGATTTTCAGCAAGCACGATATTCCAAGAAGAAGACGAATAGATTCCAGCCGA
GCTACTGGAGCAAAAGTCTCATCAAAATCAATTCCTTCAACCTGAGAGTATCCCTGAGCAACC
AGTCGAGCTCTGTTACGTGTCACATTGCCTTCTTCATCAGTTTTATTCTTAAAAATCCATTTAGT
ACCAACCACATTGACATGAACCTGTCTTGGAACTAAATCCCATACTTCATGACGACTGAAATG
AT
SIGnAL: Chr1 13445230 (e-value=0.0)
Alonso et al. (2003): 13447000

S7-FB5

TAAAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACAC*CNNCACTG
NNCNGNCNGNCGGCTCGGAGCNCTATAATCTGATCATGAGCGGAGCAATTAAGGGAGTCACG
TTATGACCCCCGCCGATGACGCGGGACAAGCCGTTTTACGTTTGGAACTGACAGAACCGCAAC
GTTGAAGGAGCCACTCAGCCGCGGGTTTCTGGAGTTTAATGAGCTAAGCACATACGTCAGAA
ACCATTATTGCGCGTTCAAAAGTCGCCTAAGGTCACTATCAGCTAGCAAATATTTCTTGTCNG
AAATGCTCCACTGACGTTCCATAATATTCCACCTGCCCGGGCGGCCGCTNGAGCCCTATAGNT
ATGTGATAAACCANTNGGGCANGTAACACTTCTNTTGACATNTAATTTACAAATAGGATTTGC
AACCAAAAACACTAATTTCTCNTTTNTTTNACTTGAACCAAAAGAAATGTCAAAAAAATATTA
ATAGGGTNCAAAGAAAAAAAAANGCNGGGGNTNTTTNGTNNNGGGNGGGNGGGGAANTTTT
TNGTNANANTNGTGCCNGGNNCTTTGNTTNNNACTTTNTCCGAAANNNGNGTGGNGCNNTCN
AAATNNNCCNNNNNNNNNNNNTTNTTNGNNNNNNNTNNNNNTTTNNNNGGGGGNGGGNNTT
TNNNNNNNGGNCNNCCNNNAAANNNNGNANAAANAAAANNNNNNTGGNNNCNNAANNCNT
NGNNNAANNNNNNNNTTNCNNCNGNA
SIGnAL: Chr5 26725829 (5e-25)
Alonso et al. (2003): 26317000

S7-GA3

AACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTACACCACACCATAAC
TACCGATAATAACTTCTAGTTTCTATTAAGAACATTGAATTCACCTGCCCGGGCGGCCGCTCG
AGCCCTATAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr3 5205957 (1e-9)
Alonso et al. (2003): 5207000

S7-HB5

ACGTNCCGCAATGTGTTAGTAAATTCTCTAATCCGAGGAATCAAATCTTGCATACGTGTAAA
GAAAGAGAGAGAGAGAGAGAGATCGACCTTCATTGCTCTCAGCTTTTTATGTAACTTGAATGA
ATTCACCTGCCCGGGCGGCCGCTCGAGCCCTATAAGNNNNNGNANCGNNCCCGCCTNNNCGN
CACCTNCNCCGCCCNGCCCGANCNGNNNTNACNGGANTCACNNGANCCNANNACNNNCNTTC
NNNNCCTNNGCTGCCCGCTCTANNNNCNCGTAANNCNNCNNTNNNANATNTGNNNTANCANN
TNNNNNTNNNTNNNGTTNCNTCNNCTCNTCGCGCCCNNTCNNTCNNNNNNNNNCNNNNNNN
SIGnAL: Chr1 181992 (1e-17)
Alonso et al. (2003): 182000

S7-IC1

**ATAAAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAAT
AT***TCATGATCATAGGTAAAGTTGCTTTCTTCTTTTATGACTTGTTCAGTGTTTTAGCTTTTTTTTT
NGGGGANGANNCCNCGATTNGTNGNANCGANCCTAATNGGGTTTTNAGGTNGCGNCANGTTN
GNGNGTANNCAANGCAANGANCATNCTTTTNCCTTTNANGANCTNGAAGGGGTNCANCTTTN
CANGGGGANGNGAANCTTTNAGTANGGATATAAANATNCNAATATTACAATATGATTANCTA
TGACANGGGAATANGGNCNANCANGNGGGNGGANAAAACNCANATTTTNCNCGAATNCNCC
NGCCCGGGCGGCCGCTCGGCCCCNATAANNNNNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr5 26978325 (2e-18)
Alonso et al. (2003): 26569000

S7-JD6

AACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GT*GTANTCCATNCNTNNTT
CTTCATAAACNAAACGGTTCTTCTNTNTNCCTGTGACCGCCGCCTTTCCCTCGCCGCCATGAAT
TCACCTGCCCGGGCGGCCGCTCGAGCCCTATAGCTGCGCAACTGTTGGGAAGGGCGATCGGTG
CGGGCCTCTTCNCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAANTTGG
GTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTCACCTGCCCG
GGCGGCCGCTCGAGCCCTATAAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAAAGGAGCGG
GCGCCATTCANGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTNTTCGCTATTACGC
CAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAANTTGGGTAACGCCAGGGTTTTCCCA
GTCACGACGTTGTAAAACGACGGCCAGTGAATTCACCTGCCCGGGCGGNCGNTCNAGCCCTA
TANNNNNNNNNNNNNNNNNN
SIGnAL: Chr3 15898815 (2e-14)
Alonso et al. (2003): 15895000

S7-KB2

ATTAAAAACGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTTGTTTCCCGTCAATT
TGTTGTATGAATTCACCTGCCCGAGCGGCCGCTCGAGCCCTATAANNCANANNNNCNCTNNNT
ACANGCAGNCNATCANTCCNGGACGGCNNNNNNGGNNGAGCNNTNCNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNANNNNNNNNNNNNNNNNNNNNNNNANN
SIGnAL: no homology found
NCBI: homology to mitochondrial half-ABC transporter; STA1 gene on Chr5 (3e-17)
Alonso et al. (2003): 23169000

S7-LA6

ACGTCCGCCTGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAATATA*AGCGTC
AATGAATATCTCAAGAACAATGGATGAATAGCTTATGTCTCTCCCAATAGTGTTTGCATAAAA
CTTGATAAAAGACTTGATAATGAGAACTAGGTCTACAAACAATATATATAGACCCTAAAAAA
GTCCAGGGACTAATAATGCAAATAGGAAAATCTTCTGGGGCAAATTTGATATTCTGTAAACTC
GAAAGCGTTTAGGGTTTTTGCTGGGCCGAAACTGGTGTCGATCGACACCAAGAGTGTGTCGAT
CGACACCAAGAGTGTGTCGATCGACACTCCTCGCTGATTCCTGAACCAAAATCAGCCTTAGCT
TTTGCTCCTAAATGTCTCCTTATCTTCATTGTTGTTTCATTGCATAGAATACCTGAAAAGACAC
TAAAAAGACTCGAGAAATAACATAAAGACTCAAAATCCTATACCTAAAAACATGGATAAAAT
CAGTAAAAATCGGGTTATATCAGCTACCAATGTTTTTTTAATCTTTTTGGCATGGATACAAATC
TTAGTGTTAAACTCTTTTTGTAAAGCCTCAAACTTTTATTTTATATGATATTCACATTCNTATGA
TAGCTTAATATACACAAAAGGATCTTTTCAAAAGATCAATTAAAAATAGGGGGAAATTAAAT
CTAAAGGAGTTCTACAAACAAAATTGTCCATATACACCTTTTTTTTGGAGAATGCCAATTTGG
SIGnAL: Chr3 11498929 (e-value=0.0)
Alonso et al. (2003): 11498000

S7-LD1

CGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAAGTTGTGTACGCCNCNTGNNNNCCT
GGACTCGNTNCTACGATTTTTCCCCTTCTTCAATTTCNTTCCAAACTCGTATAAATTGTACANA
CTTTAGGCTTTGTTCTGTTCTTCCTCTGTNTTAAAGTTCCTTGTNCAACTCAAAATCCATCTGGG
TTTCGTGTTAGAATTTGTGGGTTTTGTGTTAATCTGTTTGTTTCAATAGAAAAACGGTGACTTT
AAATTCCCGGGAAATTTGACGATTTTTGGGAGCTAATTCTGGGAATTGGATAATTTTAATTCTT
TGATTTCTGAGAGAGAAAAAACTGAACTTTAGGATTATAAAGCAGAACCAACGAGCTCCAAT
GGTGGTGGCGACGACTATAGCACTTTACGCGAGTCCACCAAGCAGCGTTTGTTCCACGCCACA
TCAGATCTCATGTGATCTCGATCTCACCTCCAGGAGTTCGTCAACATCCTCTTCCATGGCTTCC
TCGCCGCAGAAACCAATCGTTGGAGGTCTCTCTTCACTTTTCTCCAGCGCATCAGTCAAATCTT
CATCTTCTTCCTCTTGCTCATACTCAACCGGAGTAGATGAGTTCTCTTCACTACGCTATGATCG
AATTCACCTGCCCGGGCGGCCGCTCGGGCCCCNTATAANNNNNNN
SIGnAL: Chr3 4653566 (1e-25)
Alonso et al. (2003): 4654000

S7-NA5

TNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCATTGGATATACTATATTCCAAGGTACTAA
AAATGTTGGCAAAAAATATTCTATGTGTAAGTATCAATAAATAATACTGCGTAGGGTGTGCGG
ACACTAGCCGACATAACAGTGCATAGAGGGGTCCGATCCCTAAAGAGAGTGTTGTAGTATGC
TTCACTAATCCAACTTTTTGTGAAGCGAAAAAAAGAAAAGCAAAAATGGTGGCTTGAATTGA
ACCTTTCTCTTGTAAATTTGATTTGAATTCACCTGCCCGGGCGGCCGCTCGAGCCCCTATAANN
NNNNNNNNNNNNNNNNNNNN
SIGnAL: Chr1 10370042 (2e-17)
Alonso et al. (2003): 10370000

S7-OE1

CGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATTT*GTTTACACCACAATATA*GCCAA
AAATAAATGATGATTATCAACAATACCGTAAATAAGGCAAAGGAAATAAATGATGATTATCA
ACAATTGCCGAAAATAACCACACCGGTACACCGAAATTGTTTTTGAGATTCATAAAACCGGAA
TAATACGCCTAAAGGTTATTTTCTCCAATACCGGTTTAGTTTGGTTTGCCTTCTCAAAGTACTC
GAAGGTCAAAATTGGAATTTAGTAATAAACCTGCTCCAGCTATATACAAATGTCATTTGCTCA
CCTTTCGCCGATTGATTTTGTACATCACACGCAGACGCAGCTTTTAGTATTTTTTTCAACAATC
GCTTTTTTTTTTCTNGATCTCCGNGAAGCTCTCGAATCTCNGGTANGTTTCTNCGCTTCTTCATA
GGGTTTCTNATTTCCTTTGGGNCTNCCATAATTTTNGGAGAANCTAGNCCGGTTTTNCTCTTCT
TCNGGAANCNGAATTCACCNGCCCGGGNGGCCGCTCGGCCCCNTATAANNNNNNNNNNN
SIGnAL: Chr2 18323114 (1e-19)
Alonso et al. (2003): 18264000

S7-TE3

CGTNCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATGTGTGACGCACCCTAGAATTTTA
GGGTGTGNNTNAANTTTATTCATAGTGATANNATATTCANCGNGNNGGNCAGCCGGATNAAA
GCCCAATCTTTCAGGATTTTCATGGTTTTCAGATTAAATATTAAATATATACTTTATATTTAAG
GTTTCTTTTTTTTATTATGTATATCTTTGTATAATTAATTNACCCGCCCGGGCGGNCGCNCGNG
CCCTATCNAAAATATCTTTCTATCTACATTTTTGGAAATTTCCTCTTAATTTTTCTCTATTTGNG
ACTTGACAGCGTTTAAATTCGATAATTAGGTGAAAAAAAAAGAAGAAAANGTTAGCTTGGAA
TTCACCTGCCCGGGCGGCCGCTCGAGCCCTATAANNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNCNNNNNNNNNNNNNNNNCNCNNCCNNNNCCNNNNNNNNNCNNNGNNN
SIGnAL: Chr5 21909114 (1e-17)
Alonso et al. (2003): 21500000