the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2010) **61**, 928–940 doi: 10.1111/j.1365-313X.2010.04119.x

ARABIDOPSIS: A RICH HARVEST 10 YEARS AFTER COMPLETION OF THE GENOME SEQUENCE

Linking genotype to phenotype using the Arabidopsis unimutant collection

Ronan C. O'Malley and Joseph R. Ecker*

Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92307, USA

Received 22 September 2009; revised 30 November 2009; accepted 10 December 2009.
*For correspondence (fax +1 858 558 6379; e-mail ecker@salk.edu).

SUMMARY

The large collections of *Arabidopsis thaliana* sequence-indexed T-DNA insertion mutants are among the most important resources to emerge from the sequencing of the genome. Several laboratories around the world have used the Arabidopsis reference genome sequence to map T-DNA flanking sequence tags (FST) for over 325 000 T-DNA insertion lines. Over the past decade, phenotypes identified with T-DNA-induced mutants have played a critical role in advancing both basic and applied plant research. These widely used mutants are an invaluable tool for direct interrogation of gene function. However, most lines are hemizygous for the insertion, necessitating a genotyping step to identify homozygous plants for the quantification of phenotypes. This situation has limited the application of these collections for genome-wide screens. Isolating multiple homozygous insert lines for every gene in the genome would make it possible to systematically test the phenotypic consequence of gene loss under a wide variety of conditions. One major obstacle to achieving this goal is that 12% of genes have no insertion and 8% are only represented by a single allele. Generation of additional mutations to achieve full genome coverage has been slow and expensive since each insertion is sequenced one at a time. Recent advances in high-throughput sequencing technology open up a potentially faster and cost-effective means to create new, very large insertion mutant populations for plants or animals. With the combination of new tools for genome-wide studies and emerging phenotyping platforms, these sequence-indexed mutant collections are poised to have a larger impact on our understanding of gene function.

Keywords: T-DNA, mutagenesis, insertion mutant, salk, WiscDSLox, high-throughput genotyping, high-throughput phenotyping, high-throughput sequencing, homozygous lines.

T-DNA INSERTIONAL MUTAGENESIS

Agrobacterium, a ubiquitous genus of soil bacterium, infects a wide variety of plants, causing the diseases: crown gall, cane gall, and hairy root (Gelvin, 2009). The causative agent in the crown gall tumor is a small region of transferred DNA (T-DNA), originating from the Ti-plasmid, which is inserted into the plant's genome (Chilton *et al.*, 1977). The T-DNA encodes genes for the synthesis of auxin and cytokinin which are responsible for driving the neoplastic growth characteristic of this disease (Akiyoshi *et al.*, 1984; Schroder *et al.*, 1984). It also encodes genes for the synthesis of opines which the bacterium consumes (Akiyoshi *et al.*, 1984; Schroder *et al.*, 1984; Zambryski *et al.*, 1989). T-DNA transformation requires a second set of virulence genes found on the Ti-plasmid but not in the T-DNA region (Veluthambi *et al.*, 1989; Zambryski *et al.*, 1989; Gelvin, 2009). The development of a binary vector system in which the virulence genes and the T-DNA are located on two separate replicons greatly simplified the process of engineering the T-DNA, making plant transformation much more straightforward (de Framond *et al.*, 1983; Hoekema *et al.*, 1983; Lee and Gelvin, 2008). Two imperfect direct 25-bp repeats on the left and right borders of the T-DNA are required for transfer but the rest of the sequence can be replaced with foreign genetic material which can then be transferred into a plant's genome (Wang *et al.*, 1984). Being able to easily dictate the content of the transferred DNA without having to insert significant amounts of additional vector sequence, such as is the case with transposons or

viruses, makes T-DNA an excellent vehicle for genetic engineering in plants (Bevan *et al.*, 1983; Fraley *et al.*, 1983; Herrera-Estrella *et al.*, 1983).

While T-DNA transformation is widely used to shuttle genes into plants, it was also recognized that the insertion of the large T-DNA sequence into the open-reading-frame or promoter of a gene can directly disrupt function (Feldmann *et al.*, 1989; Koncz *et al.*, 1989; Marks and Feldmann, 1989; Koncz-Kalman *et al.*, 1990). Although the mechanism of site selection of T-DNA insertion is not fully understood and sequencing biases are known (Feldmann and Marks, 1987; Howden *et al.*, 1998; Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003; Li *et al.*, 2006), for all practical purposes these effect are minimal and the integration process can be considered largely random. Still, in order to identify an insertion in any particular gene, a very large number of independently transformed seed lines is required to reach gene-space saturation (Krysan *et al.*, 1999). The initial *Agrobacterium* transformation technique was laborious since it required regeneration of plants from tissue culture (Lloyd *et al.*, 1986). A number of improvements including a seed transformation protocol developed by Feldmann and Marks, transformation of intact plants by Bechtold *et al.*, and the 'floral dip' method introduced by Clough and Bent greatly simplified the transformation process (Feldmann and Marks, 1987; Bechtold *et al.*, 1993; Clough and Bent, 1998). These advances allowed several laboratories in the 1990s to produce collections of T-DNA insertion mutants in Arabidopsis. Initially, these T-DNA collections were designed as pooled sets of lines that could be probed by PCR with a gene- and T-DNA-specific primer pair to search for inserts in the gene of interest (McKinney *et al.*, 1995; Krysan *et al.*, 1996). Amplification of the same flanking sequence tag in a set of intersecting pools allowed researchers to locate a seed line with an insert in or near their gene of interest. To avoid the time, expense, and duplication of effort inherent in screening for genes one at a time, by the late 1990s several laboratories around the world took an alternative approach, creating indexed insertion line libraries by capture and direct sequencing of the flanking genome DNA in many thousands of individual lines (Parinov *et al.*, 1999; Tissier *et al.*, 1999; Galbiati *et al.*, 2000; Samson *et al.*, 2002; Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003; Woody *et al.*, 2007). The GABI-KAT, SAIL, Salk, WISC, FLAG, and more recently SK lines were created and sequence-indexed by various labs around the world resulting in a total of over 325 000 publicly available Arabidopsis T-DNA insertion lines.

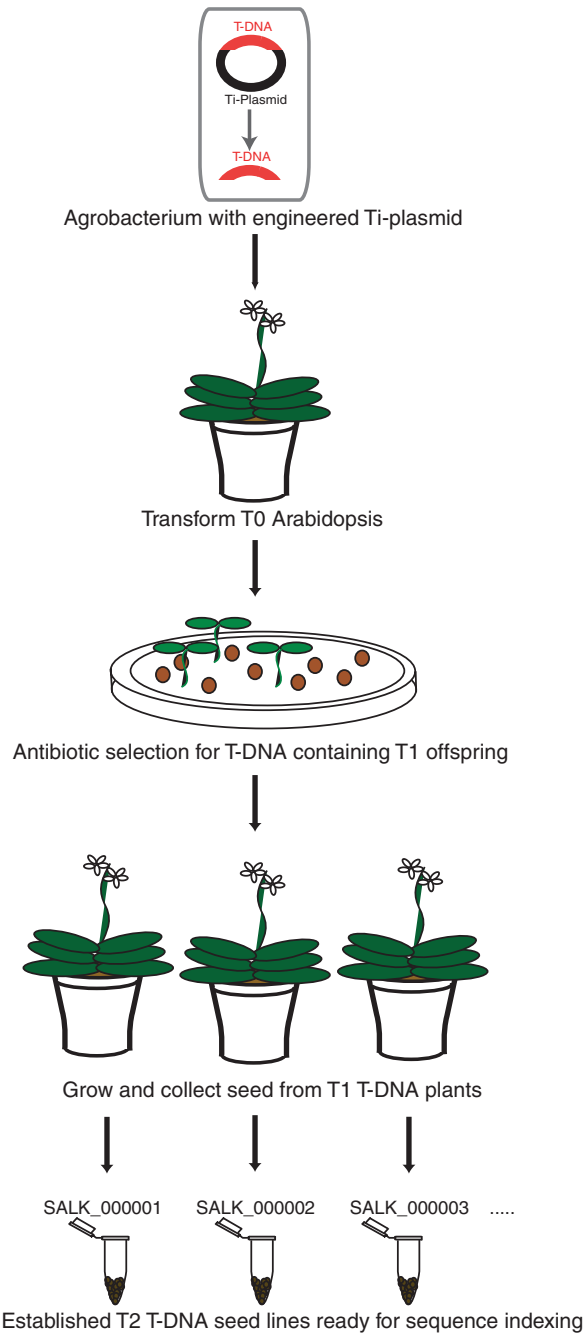

Figure 1. Creating a T-DNA insertion mutant collection.
Agrobacterium containing a Ti-plasmid transfers a T-DNA (red arc) with an antibiotic selection gene into the germline of a T0 Arabidopsis plant. The resulting T1 progeny are grown on antibiotic containing media to select for T-DNA containing individuals. Surviving plants are grown and the T2 seed is collected to establish the insertion line.

CREATING AND ACCESSING THE T-DNA MUTANT COLLECTION

The same basic approach was used to create all existing indexed T-DNA collections (Figure 1). The first step is to create a binary vector containing the desired T-DNA sequence. Minimally, the T-DNA contains an antibiotic resistance gene driven by a promoter capable of expression in Arabidopsis. Using the floral dip method, the T-DNA is transferred from the *Agrobacterium* to the genomic DNA of developing Arabidopsis ovaries. After the transformed ovaries are

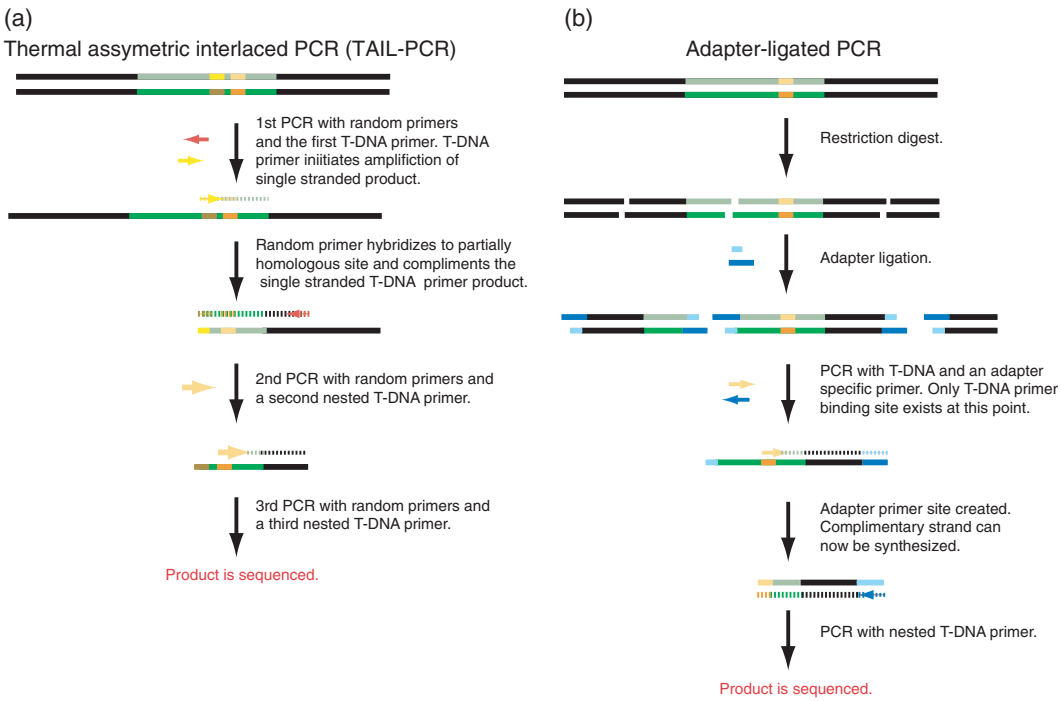

Figure 2. Methods for capturing T-DNA flanking sequence tags (FSTs).
(a) Thermal asymmetric interlaced PCR (TAIL-PCR) (Liu *et al.*, 1995) uses a series of nested T-DNA specific primers paired with a random primer set. The first T-DNA primer (yellow arrow) hybridizes to a specific site (brown line) in the T-DNA region (green line) synthesizing a product containing the T-DNA left border and the flanking genomic sequence (black line). The random primers (red arrow) anneal to any partially homologous regions in the genomic DNA. Over many cycles the T-DNA specific primer product is reverse complemented by one of the random primers, enriching the FST. Second and third rounds of PCRs are then used to further enrich the FST over non-specific amplification. The second round primer (pink arrow) and target site (orange line) are depicted in the figure. The enriched FST can now be sequenced.
(b) Adapter-ligated PCR (O'Malley *et al.*, 2007) is initiated by restriction digest of the genomic DNA. A double-stranded adapter with a long (dark blue line) and short arm (light blue line) is ligated onto the cut sites. The 3′ of the short arm is blocked by an amino group to prevent polymerase extension. A PCR reaction is run with a T-DNA specific primer (yellow arrow) and an adapter specific primer (blue arrow). If the short arm of the adapter complemented the full length of the long arm, an adapter binding site would be present, but due to the truncation, no adapter binding site is present. The T-DNA specific primer does have an annealing site (orange line) and a complementary strand is synthesized creating the binding site for the adapter specific primer. With the two requisite primer annealing sites, this molecule is amplified by PCR, and the resulting FST can be sequenced.

self-pollinated, the resulting T1 seeds are grown with the appropriate antibiotic to select for individuals containing a T-DNA. The seed from the T-DNA containing T1 plants are individually collected, each establishing a T-DNA line.

To identify the locations of T-DNA insertion sites in the Arabidopsis genome, the genomic DNA flanking each insertion site is captured and its sequence determined. A critical step in accurately mapping the genomic locations of the T-DNA is the protocol employed for the recovery of the flanking sequence tag (FST). Two different PCR-based approaches have been utilized for FST capture. The SAIL collection FSTs were produced using thermal asymmetric interlace PCR (TAIL-PCR). Here multiple rounds of PCR with a series of nested T-DNA specific primers, combined with random degenerate primers for synthesis of the reverse strand, are used to selectively amplify the FST (Liu *et al.*, 1995; Sessions *et al.*, 2002) (Figure 2a). A modified TAIL sequencing approach, initially developed for the SAIL collection (Sessions *et al.*, 2002), was also used for recovery of FSTs from the WISC collection (Woody *et al.*, 2007). For the Salk (Alonso *et al.*, 2003) and GABI-KAT (Rosso *et al.*, 2003) collections, adapter ligation mediated PCR was used to capture the FSTs. In this method, plant DNA is digested with a restriction enzyme that does not cut within the T-DNA border sequence (O'Malley *et al.*, 2007) (Figure 2b). After digestion, an asymmetric double stranded adapter is ligated to the overhanging ends. The long arm of the adapter contains a 'cryptic' primer binding site which will only be revealed by synthesis of its complementary strand, while the short arm is blocked at the 3′ end to prevent premature polymerase extension. Once the adapter is attached, a T-DNA specific primer is used to prime synthesis through the T-DNA into the flanking genomic sequence and ultimately copy the long arm of the adapter revealing a new adapter primer binding site. As only molecules containing a T-DNA binding site flanked by an adapter will amplify, this method will selectively enrich only the T-DNA-genomic junction region which can then be sequenced.

When a researcher uses a T-DNA insertion line to investigate a phenotype, they will typically want to work with a

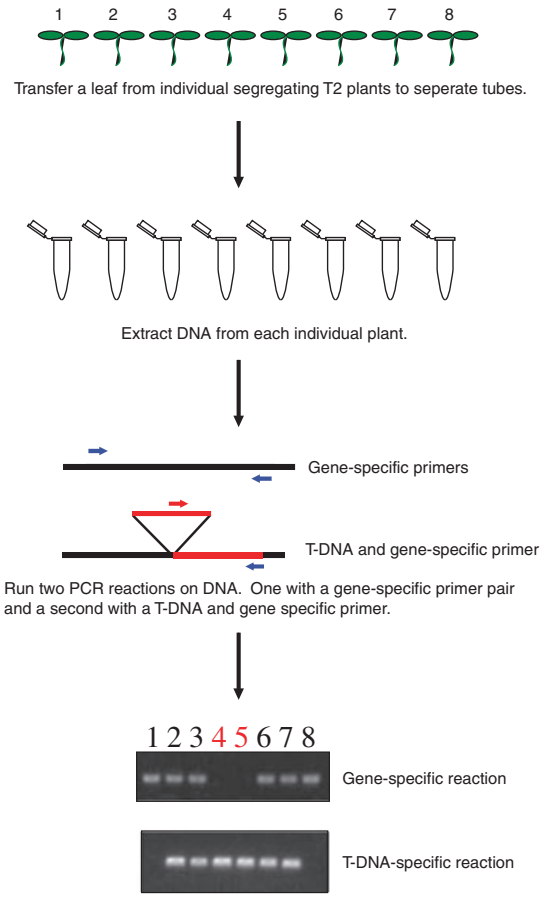

Figure 3. Genotyping of segregating T-DNA insertion lines to identify homozygous individuals.
A single leaf is cut from multiple progeny seedlings of a T-DNA line and transferred to individual tubes. DNA is extracted from these leaves. PCR and gel electrophoresis is used to genotype the individual's seedlings DNA samples. A primer pair specific for regions flanking the insertion site are used to check for the presence of a wild type, undisrupted allele of the gene. A separate PCR reaction using a T-DNA-specific primer and a gene-specific primer are used to test for the presence of a T-DNA insertion in the gene of interest. A homozygous plant will produce a T-DNA insertion product, but no wild-type product as can be seen for individual 4 and 5 in the electrophoresis gel image.

plant which is homozygous for the insertion so that both copies of the gene are disrupted. As the T1 lines are hemizygous for the insertion, multiple individual T2 plants are genotyped to identify homozygous individuals. To do this DNA is extracted from a single leaf of T-DNA segregating T2 progeny and PCR amplified with two primer pairs (Figure 3). The first pair of primers contains two genome specific sequences that hybridize to regions flanking the known insertion site. After PCR, if no product is amplified this result indicates, but does not prove, that both wild type copies contain a T-DNA, suggesting that the plant is homozygous for the T-DNA insertion (Figure 3). Successful amplification of a PCR product indicates that one or both copies of the wild-type gene are present, indicative of a hemizygote or wild-type plant (Figure 3). A second primer pair consisting of one T-DNA specific primer and one of the genome specific primers is then used to confirm the presence of the T-DNA insertion in the homozygous individuals (Figure 3). One difference between the GABI-KAT and all the other collections is that individual GABI-KAT segregating T2 plants were collected separately so a seed line will either be wild type, homozygous, or segregating the T-DNA depending on the parent. Therefore, 'a set' (10–12) of T2 seed packs is typically obtained for any particular GABI-KAT insertion line from the stock centers and DNA can be extracted from multiple seeds/line to directly determine if the parent plant was homozygous, heterozygous or contained no T-DNA (wild type).

PUBLICLY AVAILABLE T-DNA MUTANT COLLECTIONS

To date, over 325 000 T-DNA insertion lines have been isolated and sequenced. Theses lines may be browsed at a variety of web portals including: TAIR (http://www.arabidopsis.org/), FLAG (http://urgv.evry.inra.fr/projects/FLAGdb++/HTML/index.shtml), GABI-KAT (http://www.gabi-kat.de/db/), and Salk T-DNA Express (http://signal.salk.edu/cgi-bin/tdnaexpress). Two thirds of these lines, primarily from the Salk, SAIL, WiscDSlox (WISC) and GABI collections, are available for order from the Arabidopsis Biological Resource Center (ABRC) (http://www.biosci.ohio-state.edu/~plantbio/Facilities/abrc/abrchome.htm) or the Nottingham Arabidopsis Stock Center (NASC) (http://arabidopsis.info/). The recently produced Saskatoon (SK) lines will soon be available from ABRC (Samson *et al.*, 2002). The FLAG collection is available directly from l'institut National de la Recherche Agronomique (INRA).

The remaining insertion mutant populations found at the Salk Institute Genomic Analysis Laboratory (SIGnAL) T-DNA Express website are not exclusively T-DNA lines, but include the genomic locations of transposons launched from a T-DNA donor site. These are included in the T-DNA Express database as they can provide dependable loss-of-function mutants to supplement the larger T-DNA collections. The CSHL, RIKEN, IMA Ds collections were created using a Ds transposition and are available directly from the laboratories that generated them (Sundaresan *et al.*, 1995; Ito *et al.*, 2002). The SLAT and JIC-SM collection, also based on transposon insertion, use the Enhancer/Suppressor Mutator element from Maize and are available from NASC (http://arabidopsis.info/CollectionInfo?id=33) (Tissier *et al.*, 1999).

These populations represent a potential loss-of-function mutant for most Arabidopsis genes, defined here as the region from 500 bp upstream of the annotated transcriptional start site until the stop codon. However, despite the large size of the T-DNA mutant collections, 18.5% of TAIR9 annotated genes lack an insert and an additional 11% are

Table 1 Gene coverage of the insertional mutant collections

	Annotation class	Location of T-DNA insert	Number		Number (%)	
			TAIR9 genes	T-DNA hits	No T-DNA hit	One T-DNA hit
Line available from ABRC and NASC	All genes	500 bp UPS	33 239	28 052	5187 (18.5)	3875 (11)
		Exon or intron	33 239	25 579	7642 (23)	5978 (21)
	Protein- coding genes	500 bp UPS	27 344	24 044	3300 (12)	2709 (10)
		Exon or intron	27 344	20 672	6672 (24)	4216 (15.4)
All insertion lines (T- DNA and transposon)	All genes	500 bp UPS	33 239	29 154	4085 (12.2)	2735 (8.2)
		Exon or intron	33 239	25 887	7352 (22)	4766 (14.3)
	Protein- coding genes	500 bp UPS	27 344	24 896	2448 (9)	1669 (6)
		Exon or intron	27 344	22 552	4792 (17.5)	3621 (13.2)

UPS, upstream of transcriptional start site to stop codon.

only represented by one allele in those lines available from ABRC and NASC (Table 1). Coverage increases when considering all 385 000 existing T-DNA and transposon insertion mutants. Yet even with these additional lines there remains 12.2% of genes with no insert and 8.2% with only one (Table 1). While original estimates posited that greater than 99% of genes would be 'hit' with a collection of 325 000, the actual gene space coverage is much lower. This probably is due to a combination of factors. Firstly, T-DNA insertions favor intergenic regions (Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003). An analysis of all Salk, GABI, and FLAG FSTs shows insertion frequency peaks at the sites of transcriptional initiation and polyadenylation (Li *et al.*, 2006). This suggests a possible role for host transcriptional machinery or an open chromatin structure in the integration process (Li *et al.*, 2006). Also, genes lacking the restriction enzyme sites used in the FST capture protocols are overrepresented in the category of 'genes with no insertion' (Li *et al.*, 2006). Mutations effecting fertility or gametophytic development can reduce or abolish transmission, which has been estimated to occur for approximately 1% of T-DNA integration events (Feldmann and Marks, 1987; Howden *et al.*, 1998). Finally, smaller genes are less likely to have hits simply based on their smaller target size (Alonso *et al.*, 2003). Additional insertion lines will be needed to fill in the gaps in the collection and provide confirming alleles for phenotype-gene verification.

While the same basic approach was used to generate all of the publicly available collections, each one utilized a unique T-DNA sequence in their creation. In the case of the Salk and the WISC collection, only one plasmid was used in their respective collection. The SAIL and GABI-KAT collection utilized two and four distinct, but related plasmids in the creation of their respective collections (Table 2). Importantly, both the SAIL and the GABI-KAT collections preserved the left border region in all their plasmid versions so a single GABI-KAT or SAIL specific left border primer can be used to screen each collection. All of these T-DNAs contain a different promoter and antibiotic selection combination (Table 2). The Salk and SAIL collections were designed specifically for gene disruption and are the simplest in terms of the genetic information content of the T-DNA, primarily

Table 2 T-DNA collection specific information

Collection	Ecotype	Total lines	*Agrobacterium* strain	Plasmid	T-DNA size, bp	Selection	Genetic elements (LB-RB)
Salk	Col-0	137 181	C58	pROK2	4501	Kanamycin	35CaMV NTPII
SAIL	Col-0(52%); Col-3(48%)	57 671	GV3101	pCSA110	7541	BASTA	BASTA resistance cassette; pBluescript SKII+; Lat52 promoter-GUSB
			GV3101	pDAP101	4763	BASTA	BASTA resistance cassette; pBluescript SKII+
GABI-KAT	Col-0	64 058 (6676 T1 lines)	EHA2260	pAC106	5875	Sulfadiazine	35SCaMV SUL, 35SCaMv
			GV3101	pAC161	5773	Sulfadiazine	1′–2′ promoter SUL ORF, 35SCaMv
			GV3102	pGABI1	2507	Sulfadiazine	UBI 4–2, 1′–2′ promoter, Pea; RUBISCO Targeting Signal/SUL1 ORF
			GV3103	pADIS1	2279	Sulfadiazine	UBI 4–2, Pea RUBISCO; Targeting Signal/SUL1 ORF
WISC-DS Lox	Col-0	17 190	n.a.	pDS-Lox	8929	BASTA	35S Camv, Loxp; DS,MAS prom:BAR:PolyA; LoxP,Luc,DS,HYG
FLAG	WS	41 369	n.a.	pGKB5	7420	BASTA	GUS, nos 3′,ocs 3′,Kan,nos, 35SCamV, BASTA
SK	Col-4	15 507	GV3101	pSKI015	3443	BASTA	35SCamV,pBstKS+; BASTA

n.a., not available.

consisting of an antibiotic selection gene and promoter (Table 2). The WISC and GABI-KAT T-DNAs contain extra genetic features that allow applications in addition to insertional mutagenesis (Sessions *et al.*, 2002; Woody *et al.*, 2007). WISC T-DNA insertions can serve as a vehicle to create large deletions of genomic regions. This latter function relies on the presence of a loxP site and a Dissociation (Ds) element in the WISC T-DNA (Woody *et al.*, 2007). Though the process of creating a deletion is fairly involved, the capability of removing tandemly arrayed genes could be very valuable considering that 17% of all gene family members are in this configuration (Woody *et al.*, 2007). The GABI-KAT T-DNA contains a 35SCaMV adjacent to the T-DNA right-border, enabling the possibility of 'activation-tagging' of genes. In an 'activation-tagged' line if the T-DNA's 35SCaMV promoter has, by chance, inserted upstream of or in a gene it may cause over-expression of the gene product altering normal cellular processes resulting in a mutant phenotype (Weigel *et al.*, 2000).

CAVEAT EMPTOR

For the purpose of gene function studies, a researcher will generally want to prioritize the choice of mutation to maximize the chance that the gene's function will be knocked out, or at least severely reduced. This is typically done based on the locations of the insertion sites within a gene; in the preferred order of exon, intron, 5′UTR and promoter. A recent review provides an excellent summary of the effect of the location of the T-DNA insertion on the likelihood of a disruption of transcript and translated product (Wang, 2008). This review covers a decade of published literature of over 1084 insertion lines described in 648 publications. For the majority of published inserts no full length transcript was detected by RT-PCR in 99, 89 and 83% of insertions in exons/introns, upstream of 5′UTR, and downstream of the stop codon. In 136 cases where the translated product was also measured, no protein was detected for 85% of exon or intron insertions. Yet only 41% of inserts upstream resulted in no detectable protein. However, an additional 53% of these upstream insertions reduced protein expression compared to the wild type. In general, the further upstream from the 5′UTR the less likely it is that an insert will effect the transcript level. Still, 78% of insertion lines in the region 500–1000 bp upstream of 5′UTR resulted in reduced or abolished transcript levels. Knock ups, or up-regulation of expression, have been observed in 3% of inserts upstream of the start codon and 8% after the stop codon. There is at least one published example of a promoter insertion of a SAIL line causing upregulation which is notable as these lines were not designed for activation tagging (Lou *et al.*, 2007). While a complete knockout is often desired hypomorphic or constitutive alleles can be valuable, particularly where the knockout is lethal (Xiong *et al.*, 2009).

A common feature of all T-DNA collections is the presence of multiple insertion events occurring in a single plant. Such events can occur at distinct genomic locations or occur as tandem events at a single site of integration. The number of unlinked insertions in the Salk, SAIL and GABI-KAT collections were tested by segregation of antibiotic resistance, and all three projects observed around 1.5 inserts per line (Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003). One important implication of a second, unlinked insertion is the possibility that the non-indexed T-DNA is responsible for an observed phenotype. This situation can be clearly identified when the phenotype does not segregate with the known insertion. Multiple insertions at a single site occur frequently (Zambryski *et al.*, 1982; Jorgensen *et al.*, 1987; Koncz *et al.*, 1992; Azpiroz-Leehan and Feldmann, 1997). To test the frequency of tandem insertions in the SAIL collection, 96 lines were sequenced out of both borders. It was found that 25% of the left border and 62% of the right border sequenced into another T-DNA (Sessions *et al.*, 2002). A higher frequency of left border flanking the genomic sequence is consistent with the general observation by the groups developing the Salk, SAIL, GABI-KAT and WISC collections, that sequencing from the left border more frequently recovered genomic information. As a result, all these collections almost exclusively used the left-border primer for flanking sequence tag (FST) indexing (Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003; Woody *et al.*, 2007). An analysis of the complete set of SAIL left-border sequences found that 11.4% of left border sequence was T-DNA-only. Yet, this is likely to be an underestimation, as the inverted repeat at the borders of tandemly inserted T-DNA can be difficult to sequence (De Buck *et al.*, 1999; Sessions *et al.*, 2002). In a DNA gel blot analysis of 300 GABI-KAT single locus inserts, 60% showed banding patterns that were consistent with a tandem array of T-DNAs at a single site of insertion (Rosso *et al.*, 2003).

One limitation of all of the methods for capture of T-DNA flanking genomic regions is that they can dependably recover only a single FST even if there are multiple insertions in a single line. Since amplification is initiated from a T-DNA specific primer and the recovered products were not individually isolated (by cloning or gel electrophoresis), all FSTs produced by capillary sequencing begin at the same position of the T-DNA and produce the same, overlapping sequence trace (Figure 4, red box). Since the flanking chromosomal regions will be unique to each insert, when the sequencing reaction reaches the plant genomic DNA, distinct, overlapping signals are generated from each of the independent FST creating an incoherent sequence trace (Figure 4, green box). Only the longest piece of recovered genomic DNA in which the adapter or degenerate primer is most distant from the T-DNA primer binding site will be readable as it will have 'emerged' with a unique set of peaks in the chromatogram (Figure 4, blue

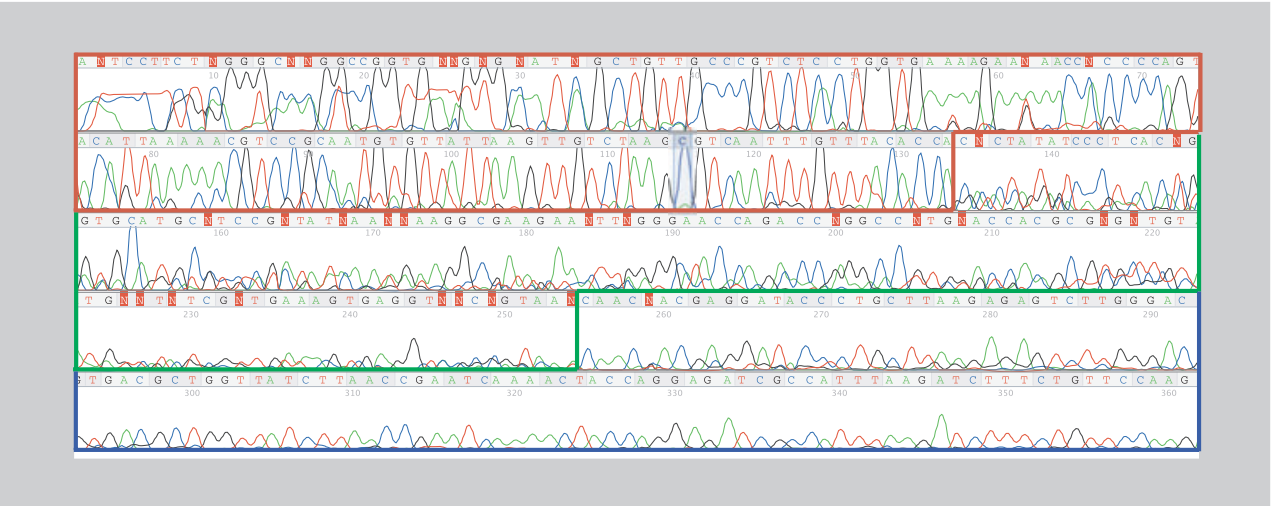

Figure 4. Overlapping traces from two T-DNA inserts in the same line.
The sequenced region enclosed in the red box matches the T-DNA left border with a blast *E*-value of 1e−125. Because the two independent insertions share the same sequence in the T-DNA region the resulting traces reinforce each other and the sequence can be read. In the green box, the sequence does not have a blast match (*E*-value cutoff 0.1) to the T-DNA sequence or the Arabidopsis genome. In this region, the sequence products are coming from the flanking genomic sequence but because two FSTs were recovered and each has a distinct sequence, an incoherent trace file is produced. Note how the signal in the green box and blue box drop by approximately half from the red box region, as each trace is now generated by distinct sequence. In the blue box, the sequence matches the Arabidopsis genome with a blast *E*-value of 2e−157. One of the two T-DNA sequences has been completed and is no longer contributing to the signal. The remaining T-DNA FST is now interpretable and will be the only one associated with this specific T-DNA line.

box). As a result, only one insert point will be recovered in any independent attempt to rescue a T-DNA sequence despite the presence of additional genomic DNA fragments. In the case of the SAIL collection, an effort was made to deconvolute the sequence traces based on the observation that the shorter T-DNA FST would, on average, have a higher representation in the mixture due to preferential amplification (Sessions *et al.*, 2002). Despite this attempt, the near absence of second FSTs /line in the SAIL collection is not representative of the 1.5 insertions per line suggested by genetic segregation of antibiotic resistance, indicating that identification of more than one insertion in any given line is still problematic. One possible solution to this problem may be the recently developed 'Trace Recalling' algorithm designed to deconvolute double traces (Tenney *et al.*, 2007). The algorithm extracts a degenerate sequence representing the double trace, searches for a unique match in the reference genome to find a primary match. When a primary match is found it will be subtracted out of the degenerate sequence to expose the second intertwined sequence. When tested on 38 000 traces from the Salk collection, this algorithm was able to recover two distinct FSTs 8.7% of the time, though the predictions have yet to be tested.

Not all indexed lines are equally likely to actually contain the T-DNA designated by the FST. One large class of insertion lines that may not contain the annotated insertion is the 'contaminated set'. These plant lines arise from DNA cross-contamination between sample wells during the FST sequencing process and are characterized by two or more T-DNA inserts sharing the same or nearly the same start site in the database. However, these can easily be identified by their 'coincidental' insertion site in the genome and by line numbers which indicates they were sequenced on the same plate (e.g. Salk_128569 and Salk_128571 or SAIL_872_H07 and SAIL_872_D06). The chances of identical, independent insertion lines being sequenced in the same 384 well plate is extremely low, so it can be assumed when this phenomenon is observed that it is most likely the result of cross-contamination. All of the described T-DNA collections contain some level of cross-contamination. Because there are examples where a specific gene is only represented by 'contaminated lines', this material is still quite valuable and so these events have not been removed from the collections. In such cases, a researcher can order all suspected contaminated lines and using PCR, confirm which of these lines contains an insertion in their gene of interest.

The original publications of the SAIL and GABI-KAT collections both reported that not all T-DNA predicted insertions were confirmed by a quality control PCR (Sessions *et al.*, 2002; Rosso *et al.*, 2003). For both collections, several hundred lines were tested with a left-border and genome-specific primer set. Both projects reported a 76% re-confirmation rate, a surprisingly identical result, considering that the projects were completely independent and used different approaches for FST sequencing. A project to identify a set of homozygous T-DNA insertions in every

gene in the genome was initiated several years ago (http://methylome.salk.edu/cgi-bin/homozygotes.cgi). To date, 59 601 Salk and 7495 SAIL lines have been genotyped by PCR and gel electrophoresis. Insertion lines which lack amplification of any T-DNA product were found in each of these indexed T-DNA populations: 12.6% of Salk and 14.5% of SAIL lines. In these cases, the DNA samples from the non-confirming lines did produce products of the predicted size with primer pairs designed to the wild-type allele but failed to produce a T-DNA product, indicating that the locus did not contain the predicted insert. A full list of all lines with confirmed (http://methylome.salk.edu/tdna_tpj/T-DNA.Confirmed) and non-confirmed (http://methylome.salk.edu/tdna_tpj/T-DNA.Unconfirmed) T-DNA insertions is available (http://methylome.salk.edu/tdna_tpj).

An added complication to gene function studies based on T-DNA insertion mutants is that the process of integration may have unwanted effects on genomic sequences surrounding the point of insertion. Small changes such as deletions, insertions, and the addition of unrecognizable 'filler sequence' at the T-DNA border are well documented (Gheysen *et al.*, 1987, 1991, Gorbunova and Levy 1997, Mayerhofer *et al.*, 1991, Ohba *et al.*, 1995). Larger genomic rearrangements associated with T-DNA insertions have also been characterized including: duplications, megabase chromosomal inversions and interchromosomal translocations (Nacry *et al.*, 1998; Laufs *et al.*, 1999; Tax and Vernon, 2001; Curtis *et al.*, 2009). For example, one study found that in a set of 36 T-DNA insertions, 20% resulted in translocation or inversion events (Castle *et al.*, 1993). Larger rearrangements that do not alter the gene content will not always effect the phenotype, but if gene content is altered it can (Tax and Vernon, 2001). T-DNA induced inversions and translocations can affect recombination rates and gametophyte survival, respectively. This can pose a problem when trying to cross distinct T-DNA lines to create multiple mutants (Tax and Vernon, 2001; Curtis *et al.*, 2009). There are, however, numerous examples of successfully created multigenic T-DNA mutants, which indicate that it is straightforward to generate multiple insertion mutants (To *et al.*, 2004; Okushima *et al.*, 2005; Prigge *et al.*, 2005). Aborted insertions are another class of genomic alterations which can result in mutations where the T-DNA fails to integrate at that same site. Such an event can be a problem, as it may not be detected through Southern blot or T-DNA FST sequencing (Tax and Vernon, 2001). Because of these additional potential genomic alterations, backcrossing of each T-DNA insertion line to a wild-type plant is strongly advised to remove potential unlinked, additional T-DNA or genomic disruptions (Krysan *et al.*, 1996; Alonso *et al.*, 2003). In addition, the identification of a second independent T-DNA allele in the same gene recapitulating the phenotype can serve to confirm that a phenotype is caused by the specific gene disruption as an independently derived T-DNA insertion line is very unlikely to share the same off-site genomic defects and/or secondary T-DNA inserts. In the case where a second allele is unavailable, complementation of the T-DNA mutant with a wild type copy of the gene provides an alternative way to confirm the phenotype.

When multiple inserts are present, cases of epigenetic silencing induced by T-DNA inserts are very common. One study in petunia found that a T-DNA containing a chalcone synthase gene caused silencing of the genomic copy of the gene (Stam *et al.*, 1997, 1998). Detailed analysis of the T-DNA insertion sites revealed that all the silenced plants had the T-DNA inserted in a head-to-head orientation (left borders pointing into the plant genome) while non-silenced plants had single copy T-DNA insertion events. The authors proposed that the inverted repeat resulting from the head-to-head insertion was responsible for activating post-transcriptional gene silencing of the native chalcone synthase gene. This conclusion was supported by the observation that a somatic reversion which resulted in loss of silencing was associated with T-DNA truncation which lead to the loss of an inverted repeat structure (Stam *et al.*, 1997). Trans-silencing was also observed in Arabidopsis, where a silenced CaMV 35S promoter in Salk and GABI-KAT lines caused silencing of a second non-T-DNA based CaMV 35S promoter driving expression of a reporter gene (Daxinger *et al.*, 2008). 50% of the Salk and GABI-KAT lines crossed to a plant expressing a CaMV 35S::GUS resulted in silencing (Daxinger *et al.*, 2008). The SAIL T-DNA lines, which do not contain a CaMV 35S promoter, did not cause silencing. This may make SAIL lines a better choice when experiments require a cross to a CaMV 35S driven transgene (Daxinger *et al.*, 2008). A highly significant example of silencing is of the T-DNA antibiotic selection genes, a persistent feature common to all the collections in post T1 generations (Sessions *et al.*, 2002; Alonso *et al.*, 2003; Rosso *et al.*, 2003; Woody *et al.*, 2007). While the exact mechanism of this silencing has not been established, it is probably related to the high frequency of tandem duplications (both inverted and direct) of T-DNA insertions at most integration sites. Because of this silencing, antibiotic selection should not be used to confirm the presence or absence of a T-DNA insert in a plant and instead it should be done by PCR genotyping or southern blotting.

HOMOZYGOUS 'UNI-MUTANT' INSERTION COLLECTIONS

T-DNA insertion lines have been widely used to test gene function (Esch *et al.*, 2003; Cheng *et al.*, 2004; Alonso and Ecker, 2006). However, these experiments have been done primarily on a small scale despite the availability of inserts in the majority of genes. One hindrance to large scale deployment of these studies is the isolation of individual plants homozygous for the insert, as this requires an extra geno-

typing step. One of the promises of the T-DNA collections is that it will allow large-scale screens with homozygous mutants, moving from individual gene function studies to understanding biological phenomena on a genome-wide scale. Until recently, the lack of a sufficiently large collection of homozygous T-DNA insertional mutants made a genome-wide reverse genetics approach untenable.

Identifying two independent homozygous T-DNA insertion alleles for every gene in *Arabidopsis thaliana* would enable the possibility of full genome wide forward genetic screens. This resource would allow for the unbiased identification of genes important to understanding basic plant biology and genes whose functions are important to further improvement of food, biomass and energy production. The current TAIR9 release of Arabidopsis gene annotation contains 27 379 protein coding genes. As of today, the Salk Unimutant collection consists of 31 033 total homozygous lines which have been identified and made public. These homozygous lines represent 18 506 individual genes, with 9276 genes covered by two alleles. The availability of two independent alleles for each gene is critical for this purpose, as it allows an observed phenotype to be immediately confirmed, thus, preventing false positive gene function assignments.

A remaining hurdle for the completion of the Salk Unimutant collection is the significant number of genes that have no inserts or just one currently available. Fortunately, with the advent of next-generation sequencing, in which hundreds of millions (and, soon, billions) of individual sequence reads can be produced in a single experiment (reviewed in Mardris 2008), the time and cost associated with creating new T-DNA mutant collections can be greatly reduced. Additionally, it should be possible to map all inserts in any given line, thus providing researchers with critical information regarding all possible segregating T-DNA in the plant under study. Using native transposons in petunia and Roche's 454 platform, a recent proof of principle of this approach has been demonstrated (Vandenbussche *et al.*, 2008). High throughput sequencing technologies also allow for rapid sequence and assembly of new plant genomes. Coupled with an inexpensive method to create indexed T-DNA libraries in any transformable species, the possibility exists in the near future for T-DNA collections in many different plant species.

PHENOTYPING

The stable monogenic mutant alleles provided by T-DNA insertions are the fastest, most accessible means to permanently eliminate or reduce gene activity. Researchers have taken advantage of T-DNA insertion lines to demonstrate and confirm gene function in an enormous variety of biological processes including: hormone biosynthesis (Liu and Zhang, 2004; Staswick *et al.*, 2005), hormone signaling (Alonso *et al.*, 2003; Guo and Ecker, 2003; Lorenzo *et al.*, 2004; Tyler *et al.*, 2004; Mallory *et al.*, 2005; Okushima *et al.*, 2005; Chini *et al.*, 2007) hormone cross-talk (Anderson *et al.*, 2004), disease resistance (Zipfel *et al.*, 2004; Katiyar-Agarwal *et al.*, 2006; Knoth *et al.*, 2007), smRNA regulation (Gasciolli *et al.*, 2005; Mallory *et al.*, 2005), floral development (He *et al.*, 2003; Wellmer *et al.*, 2004; Prigge *et al.*, 2005) analysis of cell wall biosynthesis and structure (Motose *et al.*, 2004; Somerville *et al.*, 2004; Brown *et al.*, 2005; Persson *et al.*, 2005; Yong *et al.*, 2005), metabolic regulation (Hussain *et al.*, 2004), and embryonic development (McCormick, 2004; Meinke *et al.*, 2008). As there have been over 2000 cumulative citations of the original Salk, WISC, GABI-KAT, and SAIL collections, the above list represents only a tiny subset of studies in which T-DNA mutants have been successfully used to probe gene function in Arabidopsis (Wang, 2008).

One common application of T-DNA insertion lines is as a follow up to a forward genetics approach. After a phenotype has been assigned to a particular gene, T-DNA mutant alleles can be used to confirm that a particular gene is responsible for the phenotype (Esch *et al.*, 2003; Cheng *et al.*, 2004). In cases where multiple candidate genes are found in an interval, testing T-DNA mutants for each gene can expedite the discovery of the correct causative mutation. This approach can save significant time and effort. Another area in which T-DNA lines have proven invaluable is in the assignation of individual gene function for gene families. The majority of genes in the Arabidopsis genome are members of gene families. Individual genes frequently have strong sequence similarity to their closest relatives. This results in significant functional overlap (AGI, 2000; Shiu and Bleecker, 2001; Gagne *et al.*, 2002; Shiu *et al.*, 2004). Due to their dependence on random mutagenesis, forward genetic approaches will generally fail when a loss-of-function in more than one gene is required in order to expose a phenotype. A reverse genetics approach utilizing gene-indexed T-DNA insertions allows stable multigenic mutants of closely related genes to be systematically generated by genetic crosses to relieve redundant gene functions. The study of To *et al.* (2004) with the type A response regulators (ARRs) provides an excellent example of this approach. The ARR family members were originally identified as potential components of the cytokinin pathway because they are strongly upregulated by cytokinin (To *et al.*, 2004). Using loss-of-function mutants in six of the 10 ARRs, it was demonstrated that, although none of the single mutants showed a phenotype, combinations up to and including the hextuple mutant resulted in a progressive increase in cytokinin sensitivity (To *et al.*, 2004). This established a role for the ARRs as repressors of cytokinin signaling. Additionally, complex regulatory roles and gene-specific functions among the members may be a feature common to multigene clades. A reverse genetic analysis of higher order mutants of the five members of the Class III homeodomain leucine zipper (HD-Zip III) family members provides another

example of how multigenic mutants can reveal distinct and overlapping roles for each gene in the family. Although the HD-Zip III genes already had well-established individual pleiotropic effects upon development, combinations of loss-of-function T-DNA alleles resulted in complex patterns of overlapping, distinct, antagonistic and tissue specific roles for these genes (To *et al.*, 2004; Prigge *et al.*, 2005). This could not have been predicted from phylogeny or expression pattern analysis.

A distinct and complementary strategy disregards familial relationships and focuses solely upon co-expression with genes known to be involved in a specific biological process. Two different groups have used co-expression of genes with specific cellulose synthase genes as a criterion to identify possible players in secondary cell wall synthesis. Subsequent T-DNA mutant analysis implicated seven new genes in that process (Brown *et al.*, 2005; Persson *et al.*, 2005). As insertion mutations in all gene family members are not always available, researchers have used artificial microRNA (amiRNA) as a complementary method to reduce the activity of the remaining genes in the family (Schwab *et al.*, 2006). A recent study of the nine functional members of the 1-aminocyclopropane-1-carboxylate synthase (ACS) family combined the seven available T-DNA mutants with a amiRNA targeting the remaining two genes to create a large set of higher order mutants (Tsuchisaka *et al.*, 2009). The resulting mutants reveal a combinatorial code of homo- and heterodimeric ACSs that is the central regulator of ethylene production during plant development (Tsuchisaka *et al.*, 2009).

Just as microarray and phylogenetic analyses have proven extremely valuable in reverse genetics studies, newly emerging genomic strategies and tools promise to greatly enhance the application of the T-DNA homozygous collection in the future. The protein–protein interaction methods of yeast two hybrid and protein microarrays are becoming much higher throughput providing genome-wide interaction networks (Rual *et al.*, 2005). These protein interaction maps should prove to be a gold mine for reverse genetics as it will provide valuable information regarding which nodes to eliminate by mutation, which pathways will be affected, and what types of phenotypes may result. Coupling this information to co-expression data from microarrays, next generation RNA-Seq, and protein mass spectroscopy can further refine targets by providing temporal and tissue specific expression of genes and proteins (Greenbaum *et al.*, 2003; Domon and Aebersold, 2006; Marioni *et al.*, 2008). Next generation sequencing can also produce high quality information regarding epigenetic marks, smRNA, transcription factor binding sites, and additional genomes for comparative phylogenies (Cokus *et al.*, 2008; Lister *et al.*, 2008, 2009; Ossowski *et al.*, 2008). As networking software is further developed to integrate these heterogenous data sets, high throughput reverse genetics with the homozygous mutant collections is poised to flourish (Cline *et al.*, 2007).

As the Arabidopsis homozygous mutant collection nears completion, moving beyond focused studies of a handful of genes and into full genomic-scale reverse genetics screens is becoming a possibility. Unlike a forward genetic screen, loss-of-function genes identified in a reverse genetic screen will be immediately identified, and quickly confirmed with a second T-DNA allele allowing for much greater throughput. Several groups have developed batteries of conditional challenges to expose phenotypes not visible under normal greenhouse growth conditions (Christensen and Feldmann, 2007). These pipelines can be established to allow a single pass through the collection to test a multitude of conditions. Continuous image capture from camera based systems are being developed and along with automated image analysis software, should help identify more subtle or temporal alterations in plant growth responses (Tsaftaris and Noutsos, 2004; Miller *et al.*, 2007; Wang *et al.*, 2009). Large scale phenotyping platforms are being established in several countries for high throughput assaying of various plant species (Finkel, 2009). These approaches will allow for very large numbers of individuals to be analyzed for a wide variety of important traits such as biofuel and food production (Finkel, 2009). While the ultimate goal of these efforts is to define traits in agricultural plants, Arabidopsis, as the reference plant, and T-DNA mutant collections as the most accessed resource to test plant gene function, will undoubtedly play an important role in this process. As we move ahead, these approaches may begin to close the gap between gene annotation and gene functional assignment, bridging the genotype to phenotype divide.

ACKNOWLEDGEMENTS

The authors thank Rhiannon Biddick for her thorough reading of the manuscript, Mary Galli for helpful discussions, and Andrew Kuruzar for early assistance with figure preparation. Research in our laboratory is supported by grants from the National Science Foundation, the Department of Energy, the National Institutes of Health and the Mark K. Chapman foundation.

REFERENCES

AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Akiyoshi, D.E., Klee, H., Amasino, R.M., Nester, E.W. and Gordon, M.P. (1984) T-DNA of *Agrobacterium tumefaciens* encodes an enzyme of cytokinin biosynthesis. *Proc. Natl Acad. Sci. USA*, **81**, 5994–5998.

Alonso, J.M. and Ecker, J.R. (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat. Rev. Genet.* **7**, 524–536.

Alonso, J.M., Stepanova, A.N., Leisse, T.J. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.

Anderson, J.P., Badruzsaufari, E., Schenk, P.M., Manners, J.M., Desmond, O.J., Ehlert, C., Maclean, D.J., Ebert, P.R. and Kazan, K. (2004) Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in *Arabidopsis*. *Plant Cell*, **16**, 3460–3479.

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

Bechtold, N., Ellis, J. and Pelletier, G. (1993) *In planta Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *Comptes Rendus de l'Academie des Sciences Serie 3 Sciences de la Vie*, **316**, 1194–1199.

Bevan, M.W., Flavell, R.B. and Chilton, M.D. (1983) A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. *Biotechnology*, **24**, 367–370.

Brown, D.M., Zeef, L.A.H., Ellis, J., Goodacre, R. and Turner, S.R. (2005) Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell*, **17**, 2281–2295.

Castle, L.A., Errampalli, D., Atherton, T.L., Franzmann, L.H., Yoon, E.S. and Meinke, D.W. (1993) Genetic and molecular characterization of embryonic mutants identified following seed transformation in *Arabidopsis*. *Mol. Gen. Genet.* **241**, 504–514.

Cheng, Y., Dai, X. and Zhao, Y. (2004) AtCAND1, a HEAT-repeat protein that participates in auxin signaling in *Arabidopsis*. *Plant Physiol.* **135**, 1020–1026.

Chilton, M.D., Drummond, M.H., Merio, D.J., Sciaky, D., Montoya, A.L., Gordon, M.P. and Nester, E.W. (1977) Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of crown gall tumorigenesis. *Cell*, **11**, 263–271.

Chini, A., Fonseca, S., Fernandez, G. *et al.* (2007) The JAZ family of repressors is the missing link in jasmonate signalling. *Nature*, **448**, 666–671.

Christensen, C.A. and Feldmann, K.A. (2007) Biotechnology approaches to engineering drought tolerant crops. In *Advances in Molecular Breeding Toward Drought and Salt Tolerant Crops* (Jenks, M.A., Hasegawa, P.M. and Jain, S.M., eds). The Netherlands: Springer, pp. 333–357.

Cline, M.S., Smoot, M., Cerami, E. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocols*, **2**, 2366–2382.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Curtis, M.J., Belcram, K., Bollmann, S.R., Tominey, C.M., Hoffman, P.D., Mercier, R. and Hays, J.B. (2009) Reciprocal chromosome translocation associated with TDNA-insertion mutation in *Arabidopsis*: genetic and cytological analyses of consequences for gametophyte development and for construction of doubly mutant lines. *Planta*, **229**, 731–745.

Daxinger, L., Hunter, B., Sheikh, M., Jauvion, V., Gascioli, V., Vaucheret, H., Matzke, M. and Furner, I. (2008) Unexpected silencing effects from T-DNA tags in *Arabidopsis*. *Trends Plant Sci.* **13**, 4–6.

De Buck, S., Jacobs, A., Van Montagu, M. and 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**, 295–295.

Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.

Esch, J.J., Chen, M., Sanders, M., Hillestad, M., Ndkium, S., Idelkope, B., Neizer, J. and Marks, M.D. (2003) A contradictory GLABRA3 allele helps define gene interactions controlling trichome development in *Arabidopsis*. *Development*, **130**, 5885–5894.

Feldmann, K.A. and Marks, M.D. (1987) Agrobacterium-mediated transformation of germinating seeds of *Arabidopsis thaliana*: a non-tissue culture approach. *Mol. Gen. Genet.* **208**, 1–9.

Feldmann, K.A., Marks, M.D., Christianson, M.L. and Quatrano, R.S. (1989) A dwarf mutant of *Arabidopsis* generated by T-DNA insertion mutagenesis. *Science*, **243**, 1351–1354.

Finkel, E. (2009) With 'phenomics', plant scientists hope to shift breeding into overdrive. *Science*, **325**, 380–381.

Fraley, R.T., Rogers, S.G., Horsch, R.B. *et al.* (1983) Expression of bacterial genes in plant cells. *Proc. Natl Acad. Sci. USA*, **80**, 4803–4807.

de Framond, A.J., Barton, K.A. and Chilton, M.D. (1983) Mini–Ti: a new vector strategy for plant genetic engineering. *Nat. Biotechnol.* **1**, 262–269.

Gagne, J.M., Downes, B.P., Shiu, S.-H., Durski, A.M. and Vierstra, R.D. (2002) The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **99**, 11519–11524.

Galbiati, M., Moreno, M.A., Nadzan, G., Zourelidou, M. and Dellaporta, S.L. (2000) Large-scale T-DNA mutagenesis in *Arabidopsis* for functional genomic analysis. *Funct. Integr. Genomics*, **1**, 25–34.

Gasciolli, V., Mallory, A.C., Bartel, D.P. and Vaucheret, H. (2005) Partially redundant functions of *Arabidopsis* DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr. Biol.* **15**, 1494–1500.

Gelvin, S.B. (2009) *Agrobacterium* in the genomics age. *Plant Physiol.* **150**, 1665–1676.

Gheysen, G., Montagu, M.V. and Zambryski, P. (1987) Integration of Agrobacterium tumefaciens transfer DNA (T-DNA) involves rearrangements of target plant DNA sequences. *Proc. Natl. Acad. Sci. USA*, **17**, 6169–6173.

Gheysen, G., Villarroel, R. and Van Montagu, M. (1991) Illegitimate recombination in plants: a model for T DNA integration. *Genes & Dev.* **5**, 287–297.

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

Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117–125.

Guo, H. and Ecker, J.R. (2003) Plant responses to ethylene gas are mediated by SCFEBF1/EBF2-dependent proteolysis of EIN3 transcription factor. *Cell*, **115**, 667–677.

He, Y., Michaels, S.D. and Amasino, R.M. (2003) Regulation of flowering time by histone acetylation in *Arabidopsis*. *Science*, **302**, 1751–1754.

Herrera-Estrella, L., Block, M.D., Messens, E., Hernalsteens, J.P., Montagu, M.V. and Schell, J. (1983) Chimeric genes as dominant selectable markers in plant cells. *EMBO J.* **2**, 987–995.

Hoekema, A., Hirsch, P.R., Hooykaas, P.J.J. and Schilperoort, R.A. (1983) A binary plant vector strategy based on separation of vir-and T-region of the *Agrobacterium tumefaciens* Ti-plasmid. *Nature*, **303**, 179–180.

Howden, R., Park, S.K., Moore, J.M., Orme, J., Grossniklaus, U. and Twell, D. (1998) Selection of T-DNA-tagged male and female gametophytic mutants by segregation distortion in *Arabidopsis*. *Genetics*, **149**, 621–631.

Hussain, D., Haydon, M.J., Wang, Y., Wong, E., Sherson, S.M., Young, J., Camakaris, J., Harper, J.F. and Cobbett, C.S. (2004) P-Type ATPase heavy metal transporters with roles in essential zinc homeostasis in *Arabidopsis*. *Plant Cell*, **16**, 1327–1339.

Ito, T., Motohashi, R., Kuromori, T., Mizukado, S., Sakurai, T., Kanahara, H., Seki, M. and 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.

Jorgensen, R., Snyder, C. and Jones, J.D.G. (1987) T-DNA is organized predominantly in inverted repeat structures in plants transformed with *Agrobacterium tumefaciens* C58 derivatives. *Mol. Gen. Genet.* **207**, 471–477.

Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Zhu, J.K., Staskawicz, B.J. and Jin, H. (2006) A pathogen-inducible endogenous siRNA in plant immunity. *Proc. Natl Acad. Sci. USA*, **103**, 18002–18007.

Knoth, C., Ringler, J., Dangl, J.L. and Eulgem, T. (2007) *Arabidopsis* WRKY70 is required for full RPP4-mediated disease resistance and basal defense against *Hyaloperonospora parasitica*. *Mol. Plant Microbe Interact.* **20**, 120–128.

Koncz, C., Martini, N., Mayerhofer, R., Koncz-Kalman, Z., Korber, H., Redei, G.P. and Schell, J. (1989) High-frequency T-DNA-mediated gene tagging in plants. *Proc. Natl Acad. Sci. USA*, **86**, 8467–8471.

Koncz, C., Németh, K., Rédei, G.P. and Schell, J. (1992) T-DNA insertional mutagenesis in *Arabidopsis*. *Plant Mol. Biol.* **20**, 963–976.

Koncz-Kalman, Z., Christiane Nawrath, B.R. and Redei, P. (1990) Isolation of encoding novel chloroplast protein by T-DNA tagging in *Arabidopsis thaliana*. *EMBO J.* **9**, 1337–1346.

Krysan, P.J., Young, J.C., Tax, F. and Sussman, M.R. (1996) Identification of transferred DNA insertions within *Arabidopsis* genes involved in signal transduction and ion transport. *Proc. Natl Acad. Sci. USA*, **93**, 8145–8150.

Krysan, P.J., Young, J.C. and Sussman, M.R. (1999) T-DNA as an insertional mutagen in *Arabidopsis*. *Plant Cell*, **11**, 2283–2290.

Laufs, P., Autran, D. and Traas, J. (1999) A chromosomal paracentric inversion associated with T-DNA integration in *Arabidopsis*. *Plant J.* **18**, 131–139.

Lee, L.-Y. and Gelvin, S.B. (2008) T-DNA binary vectors and systems. *Plant Physiol.* **146**, 325–332.

Li, Y., Rosso, M.G., Ülker, B. and Weisshaar, B. (2006) Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. *Genomics*, **87**, 645–652.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523–536.

Lister, R., Gregory, B.D. and Ecker, J.R. (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.* **12**, 107–118.

Liu, Y. and Zhang, S. (2004) Phosphorylation of 1-aminocyclopropane-1-carboxylic acid synthase by MPK6, a stress-responsive mitogen-activated protein kinase, induces ethylene biosynthesis in *Arabidopsis*. *Plant Cell*, **16**, 3386–3399.

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

Lloyd, A.M., Barnason, A.R., Rogers, S.G., Byrne, M.C., Fraley, R.T. and Horsch, R.B. (1986) Transformation of *Arabidopsis thaliana* with *Agrobacterium tumefaciens*. *Science*, **234**, 464–466.

Lorenzo, O., Chico, J.M., Sanchez-Serrano, J.J. and Solano, R. (2004) JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in *Arabidopsis*. *Plant Cell*, **16**, 1938–1950.

Lou, Y., Gou, J.Y. and Xue, H.W. (2007) PIP5K9, an *Arabidopsis* phosphatidylinositol monophosphate kinase, interacts with a cytosolic invertase to negatively regulate sugar-mediated root growth. *Plant Cell*, **19**, 163–181.

Mallory, A.C., Bartel, D.P. and Bartel, B. (2005) MicroRNA-directed regulation of *Arabidopsis* AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. *Plant Cell*, **17**, 1360–1375.

Marioni, J., Mason, C., Mane, S., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.

Marks, M.D. and Feldmann, K.A. (1989) Trichome development in *Arabidopsis thaliana*. I. T-DNA tagging of the GLABROUS1 gene. *Plant Cell*, **1**, 1043–1050.

Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402.

Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C. *et al.* (1991) T-DNA integration: a mode of illegitimate recombination in plants. *EMBO J.* **10**, 697.

McCormick, S. (2004) Control of male gametophyte development. *Plant Cell*, **16**, 142–153.

McKinney, E.C., Ali, N., Traut, A., Feldmann, K.A., Belostotsky, D.A., McDowell, J.M. and Meagher, R.B. (1995) Sequence-based identification of T-DNA insertion mutations in *Arabidopsis*: actin mutants act2–1 and act4–1. *Plant J.* **8**, 613–622.

Meinke, D., Muralla, R., Sweeney, C. and Dickerman, A. (2008) Identifying essential genes in *Arabidopsis thaliana*. *Trends Plant Sci.* **13**, 483–491.

Miller, N.D., Parks, B.M. and Spalding, E.P. (2007) Computer-vision analysis of seedling responses to light and gravity. *Plant J.* **52**, 374–381.

Motose, H., Sugiyama, M. and Fukuda, H. (2004) A proteoglycan mediates inductive interaction during plant vascular development. *Nature*, **429**, 873–878.

Nacry, P., Camilleri, C., Courtial, B., Caboche, M. and Bouchez, D. (1998) Major chromosomal rearrangements induced by T-DNA transformation in *Arabidopsis*. *Genetics*, **149**, 641–650.

Ohba, T., Yoshioka, Y., Machida, C. and Machida, Y. (1995) DNA rearrangement associated with the integration of T-DNA in tobacco: an example for multiple duplications of DNA around the integration target. *Plant J.* **7**, 157–164.

Okushima, Y., Overvoorde, P.J., Arima, K. *et al.* (2005) Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in *Arabidopsis thaliana*: unique and overlapping functions of ARF7 and ARF19. *Plant Cell*, **17**, 444–463.

O'Malley, R.C., Alonso, J.M., Kim, C.J., Leisse, T.J. and Ecker, J.R. (2007) An adapter ligation-mediated PCR method for high-throughput mapping of T-DNA inserts in the *Arabidopsis* genome. *Nat. Protocols*, **2**, 2910–2917.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033.

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

Persson, S., Wei, H., Milne, J., Page, G.P. and Somerville, C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl Acad. Sci. USA*, **102**, 8633–8638.

Prigge, M.J., Otsuga, D., Alonso, J.M., Ecker, J.R., Drews, G.N. and Clark, S.E. (2005) Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *Plant Cell*, **17**, 61–76.

Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K. and Weisshaar, B. (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol. Biol.* **53**, 247–259.

Rual, J.-F., Venkatesan, K., Hao, T. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.

Samson, F., Brunaud, V., Balzergue, S., Dubreucq, B., Lepiniec, L., Pelletier, G., Caboche, M. and Lecharny, A. (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucleic Acids Res.* **30**, 94–97.

Schroder, G., Waffenschmidt, S., Weiler, E.W. and Schroder, J. (1984) The T-region of Ti plasmids codes for an enzyme synthesizing indole-3-acetic acid. *Eur. J. Biochem.* **138**, 387–391.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N. and Weigel, D. (2006) Highly specific gene silencing by artificial micrornas in *Arabidopsis*. *Plant Cell*, **18**, 1121–1133.

Sessions, A., Burke, E., Presting, G. *et al.* (2002) A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell*, **14**, 2985–2985.

Shiu, S.H. and Bleecker, A.B. (2001) Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proc. Natl Acad. Sci. USA*, **98**, 10763–10768.

Shiu, S.-H., Karlowski, W.M., Pan, R., Tzeng, Y.-H., Mayer, K.F.X. and Li, W.-H. (2004) Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell*, **16**, 1220–1234.

Somerville, C., Bauer, S., Brininstool, G. *et al.* (2004) Toward a systems approach to understanding plant cell walls. *Science*, **306**, 2206–2211.

Stam, M., de Bruin, R., Kenter, S., van der Hoorn, R.A.L., van Blokland, R., Mol, J.N.M. and Kooter, J.M. (1997) Post-transcriptional silencing of chalcone synthase in *Petunia* by inverted transgene repeats. *Plant J.* **12**, 63–82.

Stam, M., Viterbo, A., Mol, J.N.M. and Kooter, J.M. (1998) Position-dependent methylation and transcriptional silencing of transgenes in inverted T-DNA repeats: implications for posttranscriptional silencing of homologous host genes in plants. *Mol. Cell. Biol.* **18**, 6165–6177.

Staswick, P.E., Serban, B., Rowe, M., Tiryaki, I., Maldonado, M.T., Maldonado, M.C. and Suza, W. (2005) Characterization of an *Arabidopsis* enzyme family that conjugates amino acids to indole-3-acetic acid. *Plant Cell*, **17**, 616–627.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H. and Martienssen, R. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9**, 1797–1810.

Tax, F.E. and Vernon, D.M. (2001) T-DNA-Associated Duplication/translocations in *Arabidopsis*. Implications for mutant analysis and functional genomics. *Plant Physiol.* **126**, 1527–1538.

Tenney, A.E., Wu, J.Q., Langton, L., Klueh, P., Quatrano, R. and Brent, M.R. (2007) A tale of two templates: automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices. *Genome Res.* **17**, 212–218.

Tissier, A.F., Marillonnet, S., Klimyuk, V., Patel, K., Torres, M.A., Murphy, G. and Jones, J.D.G. (1999) Multiple independent defective suppressor-mutator transposon insertions in *Arabidopsis*: a tool for functional genomics. *Plant Cell*, **11**, 1841–1852.

To, J.P.C., Haberer, G., Ferreira, F.J., Deruere, J., Mason, M.G., Schaller, G.E., Alonso, J.M., Ecker, J.R. and Kieber, J.J. (2004) Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin. *Plant Cell*, **16**, 658–671.

Tsaftaris, S.A. and Noutsos, C. (2004) Plant phenotyping with low cost digital cameras and image analytics. In *Information Technologies in Environmental Engineering* (Marx-Gomez, J., Sonnenshein, M., Muller, M., Welsch, H. and Rautenstrauch, C., eds). Berlin: Springer, pp. 238–251.

Tsuchisaka, A., Yu, G., Jin, H., Alonso, J., Ecker, J., Zhang, X., Gao, S. and Theologis, A. (2009) A combinatorial interplay among the 1-amino-1-carboxylate isoforms regulates ethylene biosynthesis in *Arabidopsis thaliana*. *Genetics*, **183**, 979–1003.

Tyler, L., Thomas, S.G., Hu, J., Dill, A., Alonso, J.M., Ecker, J.R. and Sun, T. (2004) DELLA proteins and gibberellin-regulated seed germination and floral development in *Arabidopsis*. *Plant Physiol.* **135**, 1008–1019.

Vandenbussche, M., Janssen, A., Zethof, J. *et al.* (2008) Generation of a 3D indexed *Petunia* insertion database for reverse genetics. *Plant J.* **54**, 1105–1114.

Veluthambi, K., Krishnan, M., Gould, J.H., Smith, R.H. and Gelvin, S.B. (1989) Opines stimulate induction of the vir genes of the *Agrobacterium tumefaciens* Ti plasmid. *J. Bacteriol.* **171**, 3696–3703.

Wang, Y.H. (2008) How effective is T-DNA insertional mutagenesis in *Arabidopsis*? *J. Biochem. Technol.* **1**, 11–20.

Wang, K., Herrera-Estrella, L., Van Montagu, M. and Zambryski, P. (1984) Right 25 bp terminus sequence of the nopaline T-DNA is essential for and determines direction of DNA transfer from *Agrobacterium* to the plant genome. *Cell*, **38**, 455–462.

Wang, L., Uilecan, I.V., Assadi, A.H., Kozmik, C.A. and Spalding, E.P. (2009) HYPOTrace: image analysis software for measuring hypocotyl growth and shape demonstrated on *Arabidopsis* seedlings undergoing photomorphogenesis. *Plant Physiol.* **149**, 1632–1637.

Weigel, D., Ahn, J.H., Blazquez, M.A. *et al.* (2000) Activation tagging in *Arabidopsis*. *Plant Physiol.* **122**, 1003–1013.

Wellmer, F., Riechmann, J.L., Alves-Ferreira, M. and Meyerowitz, E.M. (2004) Genome-wide analysis of spatial gene expression in *Arabidopsis* flowers. *Plant Cell*, **16**, 1314–1326.

Woody, S.T., Austin-Phillips, S., Amasino, R.M. and Krysan, P.J. (2007) The WiscDsLox T-DNA collection: an *Arabidopsis* community resource generated by using an improved high-throughput T-DNA sequencing pipeline. *J. Plant. Res.* **120**, 157–165.

Xiong, Y., DeFraia, C., Williams, D., Zhang, X. and Mou, Z. (2009) Characterization of *Arabidopsis* 6-phosphogluconolactonase T-DNA insertion mutants reveals an essential role for the oxidative section of the plastidic pentose phosphate pathway in plant growth and development. *Plant Cell Physiol.* **50**, 1277–1291.

Yong, W., Link, B., O'Malley, R. *et al.* (2005) Genomics of plant cell wall biogenesis. *Planta*, **221**, 747–751.

Zambryski, P., Depicker, A., Kruger, K. and Goodman, H.M. (1982) Tumor induction by *Agrobacterium tumefaciens*: analysis of the boundaries of T-DNA. *J. Mol. Appl. Genet.* **1**, 361–370.

Zambryski, P., Tempe, J. and Schell, J. (1989) Transfer and function of T-DNA genes from *Agrobacterium* Ti and Ri plasmids in plants. *Cell*, **56**, 193–201.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E.J., Jones, J.D.G., Felix, G. and Boller, T. (2004) Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature*, **428**, 764–767.

